No-Collapse Accurate Quantum Feedback Control via Conditional State Tomography

Sangkha Borah[1,2,3,*] and Bijita Sarma[2,3,†]

[1]*Max Planck Institute for the Science of Light, Staudtstraße 2, 91058 Erlangen, Germany*
[2]*Department of Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Staudtstraße 7, 91058 Erlangen, Germany*
[3]*Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan*

The effectiveness of measurement-based feedback control (MBFC) protocols is hindered by the presence of measurement noise, which impairs the ability to accurately infer the underlying dynamics of a quantum system from noisy continuous measurement records. To circumvent this limitation, a real-time stochastic state estimation approach is proposed in this work, that enables noise-free monitoring of the conditional dynamics, including the full density matrix of the quantum system, despite using noisy measurement data. This, in turn, enables the development of precise MBFC strategies that leads to effective control of quantum systems by essentially mitigating the constraints imposed by measurement noise, and has potential applications in various feedback quantum control scenarios. This approach is particularly important for machine learning-based control, where the AI controller can be trained with arbitrary conditional averages of observables, including the full density matrix, to quickly and accurately learn control strategies.

Future quantum technologies will rely on the ability to efficiently control quantum systems by manipulating the quantum states with reliable control protocols and feedback strategies [1–4]. In general, pure control strategies use open-loop pulse-based controls of quantum circuits, and such problems nowadays can be successfully solved using standard optimal control tools such as gradient-ascent pulse engineering and other open-loop methods [5–8]. These methods essentially rely on a differentiable model of the quantum dynamics, which can not be carried over to controls using feedback [5, 9] that require the choice of non-trivial control strategies to be discovered based on the conditional dynamics within the framework of continuous measurement. Such measurement-based feedback control (MBFC) techniques have been considered one of the most crucial and essential control strategies that can be used for real-time quantum control in laboratory experiments [10–18].

At a fundamental level, MBFC approaches suffer from limitations from two primary sources. First, such approaches often fail to control the dynamics beyond a specific limit set by the signal-to-noise ratio of the intrinsic and unavoidable measurement-induced noise to the measured quantity. The level of noise increases as $1/\sqrt{\kappa\delta t}$, where κ denotes the measurement rate, and δt is the measurement time interval, which given the fact that δt is related directly to the variance of the noise distribution (in the Wiener noise model) and $\delta t \ll 1$, the actual measured signal can be well hidden in the sea of random noise [19]. This makes it practically impossible for MBFC to find suitable control strategies for the system to achieve the desired dynamics. Second, the continuous measurement process naturally leads to the so-called measurement backaction, which makes the MBFC schemes highly non-intuitive and non-trivial in general [19–23]. It is, however, possible to control a system optimally if the precise signal is available by any means and the disruptive effect of measurement backaction can be exploited for one's advantage.

In this Letter, we research in this direction and propose an efficient MBFC protocol that can control the dynamics of a quantum system of interest precisely based on noisy continuous measurement records collected in real time. This becomes possible by building a measurement-based stochastic estimator that can extract the real-time state of the measured system noiselessly and without collapse (conditional tomography), using which the system dynamics can be controlled in any desirable manner. We show the efficiency of the scheme by applying it to control the dynamics of linear as well as nonlinear quantum systems, where the feedback applied is state-based or conditional. We also show the usefulness of the scheme for cases where control laws can be derived based on conditional moments (assuming perfect extraction of the measured signal out of the noisy data, which is typically not available in realistic experiments), which we illustrate with an example of the preparation of symmetric and anti-symmetric entangled states of two qubits. In addition to these, the scheme can also be used efficiently in real-time feedback with artificial controls.

Model-free reinforcement learning (RL) has recently been proven as a powerful new ansatz for control tasks, which, in the quantum domain, was first demonstrated for quantum error correction [22] and optimization of quantum phase transition in 2018 [24]. Following these initial studies, we have recently witnessed its applications in different sets of non-intuitive problems including applications in quantum control [19, 25–27], state transfer [28, 29], quantum state preparation and engineering [23, 30–32], and quantum error correction [33]. Very recently, the use of RL controls for real laboratory experiments of quantum system has become a reality [16, 17]. RL learns by exploring the system space (RL-environment) through trial-and-error interactions by applying specific controls (actions) to accumulate knowledge about the system (problem) over time, which makes the learning process relatively tricky compared to other forms of machine learning (ML). Indeed, such learning tasks become even more difficult and time-consuming when applied to stochastic dynamics, such as systems subjected to continuous measurement, wherein the main challenge for the controller is to identify the signals from the noisy measurement data. The optimal performance of the RL agent can be expected if it somehow becomes possible to use the exact measurement signal (without noise)

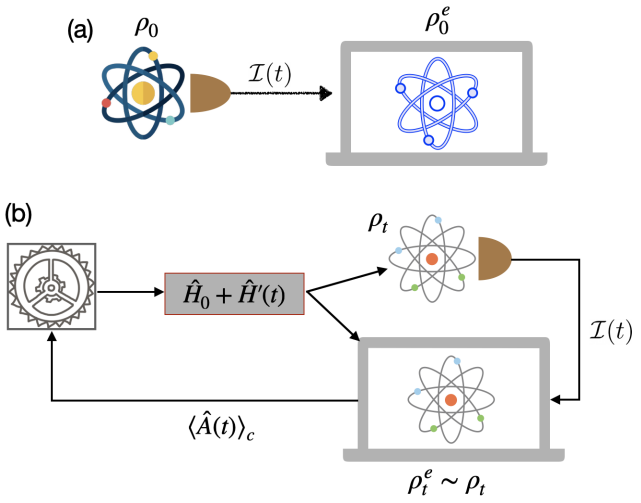

Figure 1. Schematic of the two operation stages involved - estimation and control. (a) In the estimation stage, the instantaneous measurement current $\mathcal{I}(t)$ from the physical quantum system (left) undergoing continuous measurement of the operator $\hat{A}$, is fed to the stochastic estimator (right) starting at a random state ρ_0^e. The estimator learns to follow the dynamical state of the measured quantum system accurately after several instances of the measurement record input. (b) In the control stage, a controller is used to apply accurate feedback onto the measured quantum system on the basis of the estimated noiseless signal obtained by the estimator.

or the conditional averages of one or more observables or conceivably the full density matrix of the system for training the RL-agent, which unfortunately is not available in real time with desired accuracy in realistic situations. [19, 23, 25]. We have shown how the stochastic estimator facilitates the use of RL as a controller, by making available the noiseless signals along with the full density matrix elements for training the neural network in real time for optimal and efficient training and control.

The protocol is shown schematically in Fig. 1. It consists of two operation steps - (a) the estimation stage and (b) the control stage. In the estimation stage, the to be controlled quantum system (shown on the left), with an unknown initial state (given by the density matrix ρ_0) is measured using a weak continuous measurement approach. The noisy measurement current streams are used to construct a stochastic estimator (shown on the right), which is a computational model of the measured quantum system, with the same Hamiltonian but with any random initial quantum state ρ_0^e. The estimator can track the dynamics of the measured quantum system in real-time after a while, as the conditional state of the estimator converges to that of the physical quantum system. In the control stage of operation, (b), a controller is developed to mediate between the real system and the estimator by applying feedback on the systems based on the conditional dynamics of the latter while continuing to control the systems through the real-time measured data of the physical quantum system.

We first describe the theory behind the measurement-based stochastic estimator and the feedback control method. Suppose the laboratory quantum system (shown in Fig. 1(a) on the left), is being measured continuously with a weak probe for the measurement operator (observable) $\hat{A}$ (suitably scaled to make it dimensionless). Such a continuous measurement process leads to conditional stochastic dynamics of the system density matrix in time $\rho_c(t)$ and is described by the so-called quantum stochastic master equation (SME),

$$d\dot{\rho}_c(t) = -i[\hat{H},\ \rho_c(t)] + \kappa \mathcal{D}[\hat{A}]\rho_c(t) + \sqrt{\kappa\eta}\mathcal{H}[\hat{A}]\rho_c(t)d\xi(t). \quad (1)$$

Here, κ is the measurement rate (the rate at which information is extracted from the detector), η is the measurement efficiency of the detector and $d\xi(t)$ represents an instantaneous random Wiener noise increment (white noise model with zero mean and variance $\sqrt{dt}$, where dt is the time interval between successive measurements). $\mathcal{D}[\hat{A}]$ and $\mathcal{H}[\hat{A}]$ are the superoperators describing respectively the backaction and diffusion terms in the SME [1, 3]. Probing the system with a weakly coupled meter that, in effect, has a broad probability distribution of the quantum state leads to noisy measurement records given by,

$$\mathcal{I}(t) = \langle \hat{A}(t) \rangle_c + \frac{1}{\sqrt{4\kappa\eta}} d\xi(t). \quad (2)$$

The first term on the right-hand side of the above equation denotes the conditional mean of the measurement operator (the signal) and the second term represents the contribution of the measurement noise, which depends on η and κ.

The estimator is a model quantum system with the same Hamiltonian $\hat{H}$ (as shown in Fig. 1(a), right) which is initialized in any arbitrary quantum state $\rho^e(0)$, and is driven by the noisy measurement current of the real laboratory quantum system, $\mathcal{I}(t)$ (Eq. 2). The dynamics of the estimator is described by the modified SME,

$$d\rho_c^e(t) = -i[\hat{H},\ \rho_c^e(t)]dt + \kappa \mathcal{D}[\hat{A}]\rho_c^e(t)dt + 2\kappa\eta\left[\mathcal{I}(t) - \langle A(t) \rangle_c^e\right]\mathcal{H}[\hat{A}]\rho_c^e(t)dt, \quad (3)$$

where $\rho_c^e(t)$ denotes the conditional density matrix of the estimator independent of the real system, and $\langle \hat{A}(t) \rangle_c^e = \mathrm{Tr}[\rho^e \hat{A}]$ is the conditional mean calculated for the estimator at time t. In essence, the estimator dynamics is driven by the noisy real-time measurement currents from the meter and the conditional means of the estimator itself. It can be shown that the overlap between the states $\rho(t)$ and $\rho^e(t)$ following Eqs. 1 and 3 monotonically increases until it reaches unity: $\delta \mathrm{Tr}[\rho\rho^e](t) \sim \mathrm{Tr}\left[\sqrt{\rho}(\hat{A}+\langle\hat{A}\rangle)\,\rho^e\left(\hat{A}+\langle\hat{A}\rangle\right)\times\sqrt{\rho}\right]\delta t$. Thus, provided the estimator gets sufficient amount of measurement data, the convergence of its dynamic state to that of the physical quantum system, i.e., $\rho_e(t) \sim \rho(t)$ can always be guaranteed for all the cases except for the trivial case, $[\hat{H}, \hat{A}] = 0$, a situation that is only of marginal importance since in a real problem the dynamics is always nontrivial and $[\hat{H}, \hat{A}] \neq 0$ [1, 21]. The convergence of the fidelity between the real and the estimator

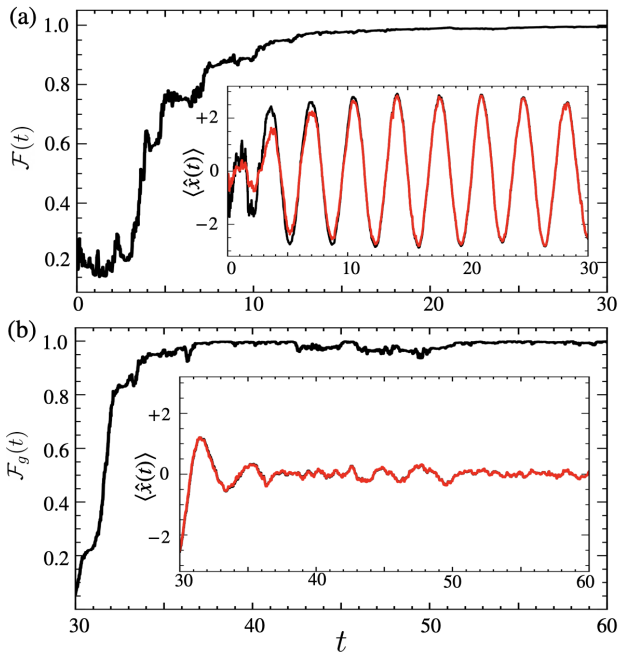

Figure 2. The control protocol is applied to the case of dynamical state control of a linear quantum harmonic oscillator for ground state preparation. In the estimation phase (a), the scheme's application leads to a gradual convergence of the estimator state to that of the measured quantum system. In the inset of (a), the conditional means of the physical system and the estimator are plotted in black and red colors, respectively. In the control phase (b), the state based controller leads to fast control of the particle's motion around the mean $\langle\hat{x}(t)\rangle_c = 0$ (see the inset) leading to accurate ground state preparation.

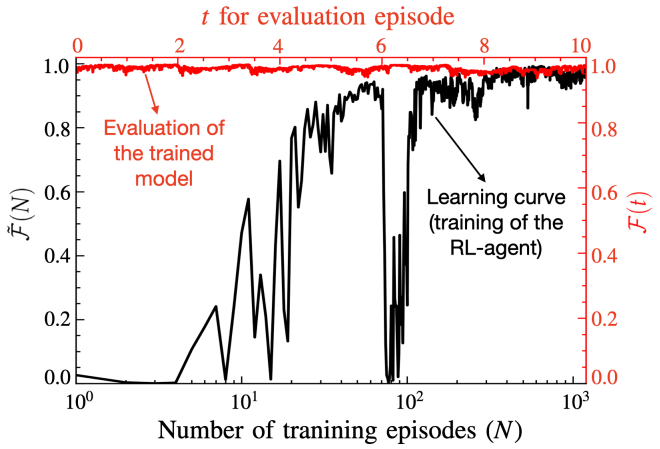

Figure 3. The protocol is applied to control a particle's motion in a nonlinear quartic potential to cool it to its dynamic ground state using RL-based control. The training process is shown in black colored lines as the average fidelity over each episode N with respect to the target state (ground state), $\bar{\mathcal{F}}(N)$, which is maximized through training. Note that the sudden drop at $N \sim 100$ is due to the exploration of the RL-agent. The performance of the trained agent is shown in red colored lines. See the main text for details.

states $\mathcal{F}(t)$ can always be guaranteed, irrespective of the values of η and κ, although the convergence time t_f is increased with decreasing values of η and κ (see Supplemental Material, where the protocol is demonstrated with the intuitive example of a qubit). Once this estimation stage is complete, the second stage of the MBFC scheme, namely the control stage, is initiated (see Fig. 1(b)).

We first apply the scheme for dynamic feedback cooling of a linear quantum harmonic oscillator and demonstrate how it becomes possible to employ accurate state-based feedback control to achieve this. The Hamiltonian of the linear quantum harmonic oscillator is given by $\hat{H}_0 = \hat{p}^2/2m + m\omega^2\hat{x}^2/2$, where $\hat{x}$ and $\hat{p}$ are the position and momentum operators respectively, m is the mass of the oscillator, and ω denotes the frequency of oscillation. Let us consider that we make a measurement of the position operator, so that $\hat{A} = \hat{x}$. We now use a state-based control strategy given by $\hat{H}(t) = \hat{H}_0 - \langle\hat{x}(t)\rangle_c\hat{p}$, where $\langle\hat{x}(t)\rangle_c$ denotes the conditional mean of $\hat{x}$ at time t. In Fig. 2(a), the instantaneous fidelity between the states of the real system and the estimator, $\mathcal{F}(t)$ is shown during the estimation stage of the control protocol. Shown in terms of the monotonically improved fidelity, the estimator starts mimicking the dynamics of the measured quantum system, also see in the inset of the Fig. 2(a), where the evolution of the conditional means of $\hat{x}$ for the measured system and estimator are compared. After the estimation stage is complete, which is typically smaller than κ^{-1}, the control stage is initialized. In this stage, conditional mean based feedback is applied on both the measured system and the estimator on the basis of the noiseless conditional mean of the position extracted by the estimator. The results are shown in Fig. 2(b), where it is found that the proposed control protocol leads to fast and accurate dynamic cooling of the quantum harmonic oscillator. The inset of Fig. 2(b) shows how the control protocol could keep the quantum state stick to a dynamical minima to any length of time, which is crucial.

Next we consider a non-linear quartic potential with the unperturbed Hamiltonian given by, $\hat{H}_0 = \hat{p}^2/2m + \lambda\hat{x}^4$, where, we have chosen $m = 1/\pi$ and $\lambda = \pi/25$ with proper dimensions. We apply artificial control *viz.* RL [34–36], to devise proper feedback strategies in this case. It is noteworthy that with the designed stochastic estimator, it is now possible to apply the full density matrix as well as the means and moments of the operators for choosing any accurate feedback scheme. In this particular case, we consider the state (observation) of the RL-agent as $s_t = \{\langle\hat{x}\rangle, \langle\hat{p}\rangle, \langle\hat{x}^2\rangle, \langle\hat{p}^2\rangle, \mathcal{F}\}$, $\mathcal{F}$ being the fidelity with the target state. Another advantage of the estimator control is that state fidelities are now realizable, which are usually pervasive in real experimental measurements. Therefore, given that we have access to the fidelity $\mathcal{F}(t)$ from the estimator, it can be used as a simplistic and efficient reward function that needs to be maximized by the RL-agent in the training process. The agent is first trained with a given initial state, which, due to the generalizability of the trained model, permits to be used for controlling the system started with other (random) initial states. The learning curve as the mean fidelity

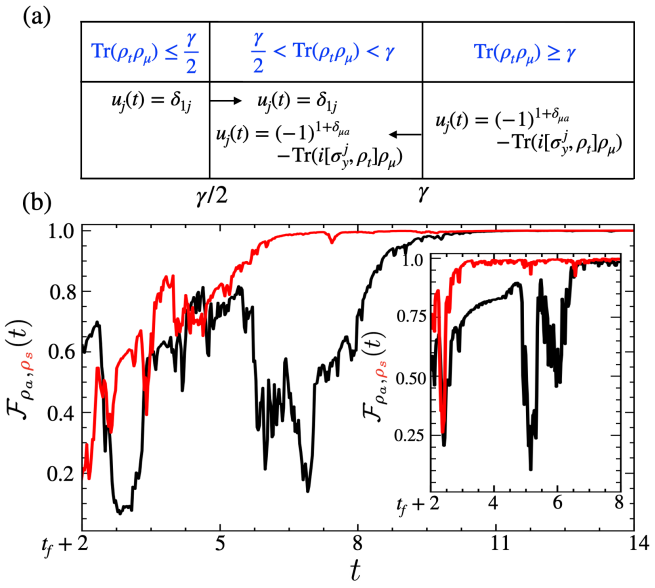

Figure 4. Demonstration of the proposed MBFC protocol for the preparation of symmetric, ρ_s, and antisymmetric, ρ_a, entangled states between two qubits as an example of use-case when it is possible to derive control laws based on conditional moments within stochastic dynamics. The control laws u_1 and u_2 are selected depending on the conditional value of $\rho_t\rho_\mu$, where $\mu \in \{s, a\}$ (symmetric and antisymmetric) being in the three regimes, conveniently demonstrated in (a), and the arrows represent direction of the entrance boundary of ρ_t to the middle section. γ is the damping parameter, the measurement rate κ is assumed to be 0.1, and the efficiency $\eta = 0.5$ for this simulation. After the estimation stage (not shown), these control laws are applied on conditional mean data (density matrices to compute instantaneous fidelity), which leads to convergence to the target states (ρ_a: black and ρ_s: red), shown in (b). In absence of such laws, RL can be used - the performance is shown in the inset of the figure (b) with similar color settings.

$\bar{\mathcal{F}}(N)$ over each training episode N is shown in black colored lines in Fig. 3. Using conditional means for training the RL-agent makes learning quicker and more accurate. The evaluated episodic fidelity variation $\mathcal{F}(t)$ is shown in red colored lines in Fig. 3 in the biaxial plot's second scale, demonstrating accurate feedback control by the trained RL-model.

Besides, it is often possible to derive control laws for systems undergoing continuous measurement based on the conditional means of observables (without the noise component). Although such control laws would not have much value in realistic situations due to the unavailability of accurate noiseless signal, we now, show in the following that in such context too, our proposed scheme would be useful. To illustrate it, we consider the preparation of symmetric (ρ_s) and antisymmetric (ρ_a) entangled states of two qubits, where

$$\rho_{(s/a)} = \frac{1}{2}(\psi_{\uparrow\downarrow} \pm \psi_{\downarrow\uparrow})(\psi_{\uparrow\downarrow} \pm \psi_{\downarrow\uparrow})^*. \tag{4}$$

Here, $\psi_{\uparrow\downarrow} = (\uparrow) \otimes (\downarrow)$ and $\psi_{\downarrow\uparrow} = (\downarrow) \otimes (\uparrow)$ are the tensor product states of the individual qubit states in the ground and excited states. The quantum filtering equation under feedback with control variables $u_1(t)$ and $u_2(t)$ is given by,

$$\begin{aligned} d\rho_t(t) =& -iu_1(t)[\sigma_y^{(1)}, \rho_t(t)]dt - iu_2(t)[\sigma_y^{(2)}, \rho_t(t)]dt \\ &-\frac{1}{2}[F_z, [F_z, \rho_t(t)]]dt + \sqrt{\eta}\big\{F_z\rho_t(t) + \rho_t(t)F_z \\ &-2\,\mathrm{Tr}[F_z\rho_t(t)]\rho_t(t)\big\}dW_t, \end{aligned} \tag{5}$$

where dW_t is the Winner noise increment at time t. σ_g^i, $g \in \{x, y, z\}$ and $i = \{1, 2\}$ are tensored Pauli operators for qubit i and $F_z = \sigma_z^1 + \sigma_z^2$ [37]. The control laws dictate non-intuitive choices of the control parameters $u_1(t)$ and $u_2(t)$ provided the real-time conditional fidelity between the current and the target states, ρ_s and ρ_a could be accurately extracted via conditional tomography of the quantum states, which is often a difficult task if not impossible. These are discussed in the Supplemental Material and conveniently represented in Fig. 4(a). Using these control laws with the MBFC scheme makes it possible to evaluate the controls $u_1(t)$ and $u_2(t)$ in real time that leads to a guaranteed preparation of the states ρ_a and ρ_s, shown in black and red colored lines respectively in Fig. 4(b). It becomes also possible to use RL for control similar to the case shown for quartic oscillator above, in which case, one can use the full density matrix for training along with conditional means, and the performance is shown in the inset of the figure. Compared to the control laws, the RL controller can help the system reach its target state in a shorter time-scale.

In typical closed-loop MBFC in laboratory experiments, one would continuously monitor a quantum system, extracting measurement currents at the output instead of the actual measurement signal, based on which certain feedbacks are applied to the input as functions of the measurement current so that a desired dynamics can be achieved. For such controls, various noise filtering methods such as LQR/LQG and Kalman filer are typically used. Also several pioneering experimental groups have attempted to use RL for such control tasks and have demonstrated it for real-time quantum control in a couple of milestone works [16, 17]. However, any feedback strategy that is operated as a function of the measurement current available from laboratory experiments would never be able to yield near-perfect control, which is only possible if one can perfectly isolate the actual signal from the noise. The proposed control protocol gives a way of avoiding the measurement-induced noise, and makes it feasible to develop accurate and optimal feedback quantum controls, by application of an external classical stochastic estimator. Essentially, it allows to perform conditional state tomography of a quantum system exposed to continuous measurement. This is expected to have revolutionizing implications in the fields of MBFC, ML, and other relevant areas of quantum research.

In conclusion, we have presented a scheme for accurately controlling quantum systems through the use of an external classical simulator and weak measurement, which can be used with state-based controls or RL for improved learning efficiency and control performance.

Supplemental Material

CONVERGENCE OF FIDELITY FOR A DRIVEN QUBIT

In the following, we use the intuitive example of a qubit to demonstrate the MBFC protocol, with the Hamiltonian given by,

$$\hat{H} = \frac{\varepsilon\hat{\sigma}_z}{2} + \frac{\Delta\hat{\sigma}_x}{2}, \tag{6}$$

where $\hat{\sigma}_i,\ i = (x, y, z)$ are Pauli operators, ε is the bare energy splitting, and Δ is the tunneling rate between the two states of the qubit system. We start the state of the physical qubit with excited state occupancy and see if the stochastic estimator started with a random state with an initial fidelity of ~ 0.6, can lead to a perfect estimate of the state in time. As shown in Fig. 5, the conditional state of the estimator gradually converges with time and perfectly reproduces the real system state. Note that this

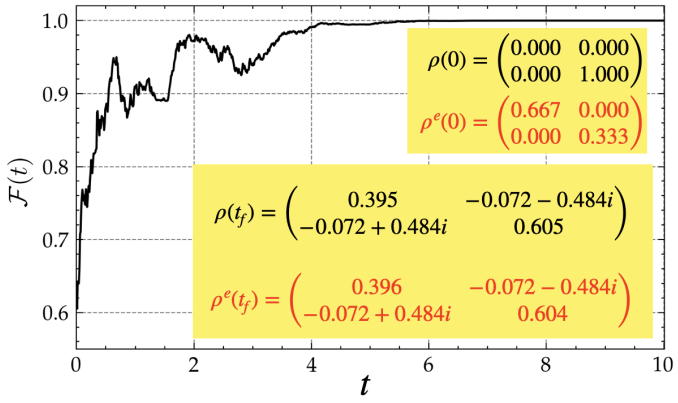

Figure 5. We demonstrate the convergence of the state of the estimator to the real quantum system state for the toy model of a qubit. In the insets, we show the initial and final states of the real and the estimator for this particular example. The initial fidelity of the real and estimator states is $\mathcal{F}(0) \sim 0.6$, which gradually improves until it reaches $\mathcal{F}(t_f) \approx 1$. This represents the estimation phase of the MBFC protocol shown in Fig. 1(a) in the main text. The parameters considered are $\varepsilon = 0.1,\ \delta = 1.0,\ \kappa = 1.0,\ \eta = 1.0$.

convergence can always be guaranteed regardless of whether the efficiency η is ideal or not. In the case of $\eta \neq 1$, the time t_f required to reach convergence becomes slightly longer. This is shown in Fig. 6(a) as a function of η. On the other hand, for detectors with larger measurement rate κ, t_f becomes smaller as shown in Fig. 6(b) for $\eta = 1$ as a function of κ. This behavior of the estimator is understandable, since intuitively the estimator would be able to learn the state faster if it had more accurate information (larger η) and less noisy measurement data (larger κ).

CONTROL LAWS FOR SYMMETRIC AND ANTISYMMETRIC ENTANGLED STATE PREPARATION

We consider the example of two qubits, which starting from random states can be prepared in symmetric and antisymmetric entangled states given by,

$$\rho_s = \frac{1}{2}(\psi_{\uparrow\downarrow} + \psi_{\downarrow\uparrow})(\psi_{\uparrow\downarrow} + \psi_{\downarrow\uparrow})^* \tag{7}$$

$$\rho_a = \frac{1}{2}(\psi_{\uparrow\downarrow} - \psi_{\downarrow\uparrow})(\psi_{\uparrow\downarrow} - \psi_{\downarrow\uparrow})^*, \tag{8}$$

where $\psi_{\uparrow\downarrow} = (\uparrow) \otimes (\downarrow)$ and $\psi_{\downarrow\uparrow} = (\downarrow) \otimes (\uparrow)$ are tensor product states of the individual qubit states in the ground and excited states. We consider the stochastic feedback controls given below [37].

The quantum filtering equation under feedback with control variables $u_1(t)$ and $u_2(t)$ is given by,

$$\begin{aligned} d\rho_t(t) = &- iu_1(t)[\sigma_y^{(1)}, \rho_t(t)]dt - iu_2(t)[\sigma_y^{(2)}, \rho_t(t)]dt - \frac{1}{2}[F_z, [F_z, \rho_t(t)]]dt + \sqrt{\eta}\big\{F_z\rho_t(t) + \rho_t(t)F_z \\ &-2\ \mathrm{Tr}[F_z\rho_t(t)]\rho_t(t)\big\}dW_t. \end{aligned} \tag{9}$$

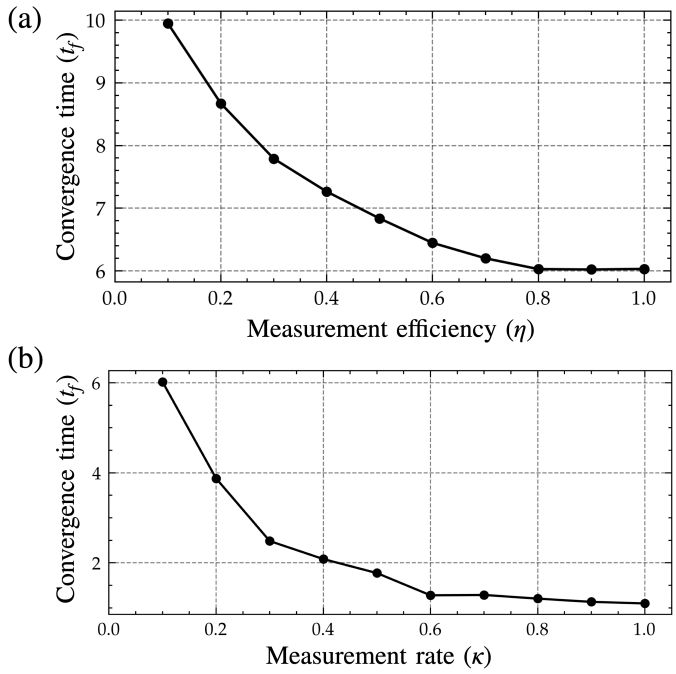

Figure 6. The convergence time t_f as a function of (a) measurement efficiency η and (b) measurement rate κ, showing the fact that t_f depends on the access to information obtained from noisy continuous measurements.

This can be written as,

$$d\rho_t(t) = -iu_1(t)[\sigma_y^{(1)}, \rho_t(t)]dt - iu_2(t)[\sigma_y^{(2)}, \rho_t(t)]dt + \mathcal{D}[F_z]\rho_t(t)dt + \sqrt{\eta}\mathcal{H}[F_z]\rho_t(t)dW_t,$$

where dW_t is the Winner noise increment at time t. σ_g^i, $g \in \{x, y, z\}$ and $i = \{1, 2\}$ are tensored Pauli operators for qubit i and $F_z = \sigma_z^1 + \sigma_z^2$. This can be rearranged to fit into the the general form of the SME as follows:

$$d\rho_t(t) = -i[H, \rho(t)]dt + \mathcal{D}[A]\rho(t)dt + \mathcal{H}[A]\rho(t)dW_t,$$

where,

$$H = u_1(t)\sigma_y^{(1)} + u_2(t)\sigma_y^{(2)}, \tag{10}$$

$$A = F_z. \tag{11}$$

For this, the control laws are given as follows. To stabilize ρ_a, the control laws are:

1. $u_1(t) = 1 - \mathrm{Tr}[i[\sigma_y^1, \rho_t]\rho_a],\ u_2(t) = 1 - \mathrm{Tr}[i[\sigma_y^2, \rho_t]\rho_a]$ if $\mathrm{Tr}[\rho\rho_a] \geq \gamma$;
2. $u_1(t) = 1,\ u_2(t) = 0$ if $\mathrm{Tr}[\rho\rho_a] \leq \gamma/2$;
3. If $\rho_t \in \mathcal{B}_a = \{\rho : \gamma/2 < \mathrm{Tr}[\rho\rho_a] < \gamma\}$, then $u_1(t) = 1 - \mathrm{Tr}[i[\sigma_y^1, \rho_t]\rho_a]$, $u_2(t) = 1 - \mathrm{Tr}[i[\sigma_y^2, \rho_t]\rho_a]$ if ρ_t last entered the set $\mathcal{B}_a$ through the boundary $\mathrm{Tr}[\rho\rho_a] = \gamma$; and $u_1(t) = 1,\ u_2(t) = 0$ otherwise.

Similarly, to stabilize ρ_s, the control laws are:

1. $u_1(t) = 1 - \mathrm{Tr}[i[\sigma_y^1, \rho_t]\rho_s]$,
   $u_2(t) = -1 - \mathrm{Tr}[i[\sigma_y^2, \rho_t]\rho_s]$ if $\mathrm{Tr}[\rho\rho_s] \geq \gamma$;
2. $u_1(t) = 1,\ u_2(t) = 0$ if $\mathrm{Tr}[\rho\rho_s] \leq \gamma/2$;
3. If $\rho_t \in \mathcal{B}_s = \{\rho : \gamma/2 < \mathrm{Tr}[\rho\rho_s] < \gamma\}$, then take $u_1(t) = 1 - \mathrm{Tr}[i[\sigma_y^1, \rho_t]\rho_s]$, $u_2(t) = -1 - \mathrm{Tr}[i[\sigma_y^2, \rho_t]\rho_s]$ if ρ_t last entered the set $\mathcal{B}_s$ through the boundary $\mathrm{Tr}[\rho\rho_s] = \gamma$; and $u_1(t) = 1,\ u_2(t) = 0$ otherwise.

We have demonstrated a convenient representation of the control laws in Fig. 4(a) of the main paper. Note that this feedback control law works only if it is possible to perform real time tomography of the qubits so that the instantaneous fidelities, $\mathrm{Tr}[\rho\rho_{s/a}]$ with the target symmetric, ρ_s and antisymmetric, ρ_a states can be computed based on which the feedback controls are decided.

REINFORCEMENT LEARNING CONTROLLER

We have used Proximal Policy Optimization (PPO) [35] which is a reinforcement learning algorithm that is used to optimize the policy of an agent in an environment that is designed to be both simple to implement and effective in practice. PPO is a variant of the popular actor-critic algorithm, which separates the policy (the actor) from the value function (the critic). The PPO algorithm can be broken down into the following steps:

1. Collect a batch of samples by interacting with the environment using the current policy. These samples consist of a sequence of state-action-reward tuples (s_t, a_t, r_t).

2. Estimate the value function $V(s_t)$ for each state in the batch using a neural network. The value function can be estimated using the Bellman equation, which gives the optimal value function $V^*(s)$ for each state,

$$V(s) = \max_a \mathbb{E}[R(s,a) + \gamma V(s')], \tag{12}$$

   where s' is the next state, and the expectation is taken over the distribution of next states given the current state and action. The value function can be estimated by iteratively updating the estimates using the Bellman equation, this process is known as dynamic programming [34]. One popular method to estimate the value function is using the temporal-difference learning algorithm, which is a type of online, model-free method for estimating the value function.

3. Estimate the advantage function $A(s_t, a_t)$ for each state-action pair in the batch. The advantage function is an estimate of the difference between the expected return and the value function,

$$A(s_t, a_t) = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k r_{t+k}] - V(s_t). \tag{13}$$

4. Use the samples to update the policy network, which is a neural network that maps states to a probability distribution over actions. PPO modifies the actor objective function by using a 'clip' function to ensure that the updated policy is not too far from the previous one. The PPO objective function is,

$$L_{\text{PPO}} = \min(r_t(\theta), \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon))A(s_t, a_t),$$

   where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the ratio of the new policy to the old policy, $\pi_\theta(a_t|s_t)$ is the probability of taking action a_t in state s_t under the current policy, $\pi_{\theta_{\text{old}}}(a_t|s_t)$ is the probability of taking action a_t in state s_t under the previous policy, $\text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)$ is a function that clips the ratio of the new policy to the old policy to the range $[1-\epsilon, 1+\epsilon]$.

The PPO algorithm is an example of trust region policy optimization algorithm [36] that ensures that the updated policy is not too far from the previous one. This makes the optimization process more stable and prevents the agent from overfitting to the current policy. For that it modifies the objective function of the actor to ensure that the updated policy is not too far from the previous one. The use of the clip function also helps to reduce the variance of the gradient estimates, which can lead to more stable and efficient training.

One of the key advantages of PPO is that it is relatively simple to implement compared to other state-of-the-art algorithms. PPO does not require the use of complex off-policy methods or value function approximations. Another advantage of PPO is that it is a sample-efficient algorithm. It allows the agent to learn from a relatively small number of samples, which makes it well-suited to applications where data collection is expensive or time-consuming. PPO also has a good performance on high-dimensional and continuous action spaces. It has proven to be a useful algorithm for problems involving quantum systems.

ACKNOWLEDGEMENTS

The authors thank Prof. Gerard Milburn, Prof. Jason Twamley and Dr. Michael Kewming for useful discussions.

* sangkha.borah@mpl.mpg.de
† bijita.sarma@fau.de

[1] H. M. Wiseman and G. J. Milburn, *Quantum Measurement and Control* (Cambridge University Press, Cambridge, 2009).

[2] J. Zhang, Y.-x. Liu, R.-B. Wu, K. Jacobs, and F. Nori, Quantum feedback: Theory, experiments, and applications, Phys. Rep. **679**, 1 (2017).

[3] K. Jacobs, *Quantum Measurement Theory and its Applications* (Cambridge University Press, Cambridge, England, UK, 2014).

[4] A. C. Doherty, S. Habib, K. Jacobs, H. Mabuchi, and S. M. Tan, Quantum feedback control and classical control theory, Phys. Rev. A **62**, 012105 (2000).

[5] P. de Fouquieres, S. G. Schirmer, S. J. Glaser, and I. Kuprov, Second order gradient ascent pulse engineering, J. Magn. Reson. **212**, 412 (2011).

[6] O. V. Morzhin and A. N. Pechen, Krotov method for optimal control of closed quantum systems, Russ. Math. Surv. **74**, 851 (2019).

[7] C. P. Koch, U. Boscain, T. Calarco, G. Dirr, S. Filipp, S. J. Glaser, R. Kosloff, S. Montangero, T. Schulte-Herbrüggen, D. Sugny, and F. K. Wilhelm, Quantum optimal control in quantum technologies. Strategic report on current status, visions and goals for research in Europe, EPJ Quantum Technol. **9**, 19 (2022).

[8] B. Sarma, S. Borah, A. Kani, and J. Twamley, Accelerated motional cooling with deep reinforcement learning, Phys. Rev. Res. **4**, L042038 (2022).

[9] T. Propson, B. E. Jackson, J. Koch, Z. Manchester, and D. I. Schuster, Robust Quantum Optimal Control with Trajectory Optimization, Phys. Rev. Appl. **17**, 014036 (2022).

[10] L. S. Martin, W. P. Livingston, S. Hacohen-Gourgy, H. M. Wiseman, and I. Siddiqi, Implementation of a canonical phase measurement with quantum feedback, Nat. Phys. **16**, 1046 (2020).

[11] S. Kuang, G. Li, Y. Liu, X. Sun, and S. Cong, Rapid Feedback Stabilization of Quantum Systems With Application to Preparation of Multiqubit Entangled States, IEEE Trans. Cybern. , 1 (2021).

[12] M. Rossi, D. Mason, J. Chen, Y. Tsaturyan, and A. Schliesser, Measurement-based quantum control of mechanical motion, Nature **563**, 53 (2018).

[13] R. Vijay, C. Macklin, D. H. Slichter, S. J. Weber, K. W. Murch, R. Naik, A. N. Korotkov, and I. Siddiqi, Stabilizing Rabi oscillations in a superconducting qubit using quantum feedback, Nature **490**, 77 (2012).

[14] 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).

[15] D. J. Wilson, V. Sudhir, N. Piro, R. Schilling, A. Ghadimi, and T. J. Kippenberg, Measurement-based control of a mechanical oscillator at its thermal decoherence rate, Nature **524**, 325 (2015).

[16] V. V. Sivak, A. Eickbusch, B. Royer, S. Singh, I. Tsioutsios, S. Ganjam, A. Miano, B. L. Brock, A. Z. Ding, L. Frunzio, S. M. Girvin, R. J. Schoelkopf, and M. H. Devoret, Real-time quantum error correction beyond break-even, arXiv 10.48550/arXiv.2211.09116 (2022), 2211.09116.

[17] K. Reuer, J. Landgraf, T. Fösel, J. O'Sullivan, L. Beltrán, A. Akin, G. J. Norris, A. Remm, M. Kerschbaum, J.-C. Besse, F. Marquardt, A. Wallraff, and C. Eichler, Realizing a deep reinforcement learning agent discovering real-time feedback control strategies for a quantum system, arXiv 10.48550/arXiv.2210.16715 (2022), 2210.16715.

[18] W. P. Livingston, M. S. Blok, E. Flurin, J. Dressel, A. N. Jordan, and I. Siddiqi, Experimental demonstration of continuous quantum error correction, Nat. Commun. **13**, 1 (2022).

[19] S. Borah, B. Sarma, M. Kewming, G. J. Milburn, and J. Twamley, Measurement-based feedback quantum control with deep reinforcement learning for a double-well nonlinear potential, Phys. Rev. Lett. **127**, 190403 (2021).

[20] S. Hacohen-Gourgy and L. S. Martin, Continuous measurements for control of superconducting quantum circuits, Advances in Physics: X **5**, 1813626 (2020).

[21] J. Zhang, Y.-x. Liu, R.-B. Wu, K. Jacobs, and F. Nori, Quantum feedback: Theory, experiments, and applications, Phys. Rep. **679**, 1 (2017).

[22] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Reinforcement Learning with Neural Networks for Quantum Feedback, Phys. Rev. X **8** (2018).

[23] R. Porotti, A. Essig, B. Huard, and F. Marquardt, Deep Reinforcement Learning for Quantum State Preparation with Weak Nonlinear Measurements, Quantum **6**, 747 (2022), 2107.08816v3.

[24] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Reinforcement Learning in Different Phases of Quantum Control, Phys. Rev. X **8**, 031086 (2018).

[25] Z. T. Wang, Y. Ashida, and M. Ueda, Deep reinforcement learning control of quantum cartpoles, Phys. Rev. Lett. **125**, 100401 (2020).

[26] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, Universal quantum control through deep reinforcement learning, npj Quantum Inf. **5**, 1 (2019).

[27] X.-M. Zhang, Z. Wei, R. Asad, X.-C. Yang, and X. Wang, When does reinforcement learning stand out in quantum control? A comparative study on state preparation, npj Quantum Inf. **5**, 1 (2019).

[28] R. Porotti, D. Tamascelli, M. Restelli, and E. Prati, Coherent transport of quantum states by deep reinforcement learning, Commun. Phys. **2** (2019).

[29] I. Paparelle, L. Moro, and E. Prati, Digitally stimulated Raman passage by deep reinforcement learning, Phys. Lett. A **384**, 126266 (2020).

[30] X.-M. Zhang, Z. Wei, R. Asad, X.-C. Yang, and X. Wang, When does reinforcement learning stand out in quantum control? A comparative study on state preparation, npj Quantum Inf. **5** (2019).

[31] J. Mackeprang, D. B. Rao Dasari, and J. Wrachtrup, A reinforcement learning approach for quantum state engineering, QuantumMachine Intell. **2**, 1 (2020).

[32] T. Haug, W.-K. Mok, J.-B. You, W. Zhang, C. Eng Png, and L.-C. Kwek, Classifying global state preparation via deep reinforcement learning, Mach. Learn. Sci. Technol. **2** (2020).

[33] H. P. Nautrup, N. Delfosse, V. Dunjko, H. J. Briegel, and N. Friis, Optimizing Quantum Error Correction Codes with Reinforcement Learning, Quantum **3**, 215 (2019).

[34] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. (The MIT Press, 2018).

[35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, Proximal policy optimization algorithms (2017), arXiv:1707.06347 [cs.LG].
[36] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, Trust region policy optimization (2017), arXiv:1502.05477 [cs.LG].
[37] M. Mirrahimi and R. Van Handel, Stabilizing Feedback Controls for Quantum Systems, SIAM J. Control Optim. (2007).