Preparation of cavity-Fock-state superpositions by reinforcement learning exploiting measurement backaction

Arthur Perret and Yves Bérubé-Lauzière ●*

Institut quantique and Département de génie électrique et de génie informatique, Faculté de génie, Université de Sherbrooke, Sherbrooke, Québec, J1K 2R1, Canada

Preparation of bosonic and general cavity quantum states usually relies on using open-loop control to reach a desired target state. In this work, a measurement-based feedback approach is used instead, exploiting the nonlinearity of weak measurements alongside a coherent drive to prepare these states. The extension of previous work on Lyapunov-based control is shown to fail for this task. This prompts for a different approach, and reinforcement learning (RL) is resorted to here for this purpose. With such an approach, cavity-eigenstate superpositions can be prepared with fidelities over 98% using only the measurements backaction as the nonlinearity, while naturally incorporating detection of cavity photon jumps. Two different RL frameworks are analyzed: an off-policy approach recently introduced called truncated quantile critic (TQC) and the on-policy method commonly used in quantum control, namely proximal policy optimization. It is shown that TQC performs better at reaching higher target state fidelity preparation.

I. INTRODUCTION

Quantum states in the Hilbert space of microwave cavities have been at the forefront of recent efforts towards a fault-tolerant quantum computer [1–3]. Improvements in the coherence times of cavities, along with the well-understood and highly biased noise hints at a scalable path using such devices. Indeed, a whole class of bosonic codes, that is, encoding information in the multiple energy levels of harmonic oscillators, is being developed [4,5].

Preparation of states with such encodings can, however, be challenging, using for instance dissipative or adiabatic approaches for the cat states [6–8]. More general bosonic states, such as binomial code states or arbitrary superpositions, rely on open-loop protocols [9–15]. While arbitrary states can be prepared this way, these techniques do not exploit the advantages made possible by feedback approaches, namely their added robustness. Indeed, even if full unitary control is available, there is no guarantee that one can recover from an error such as a cavity jump occurring during state preparation. By repeatedly monitoring the cavity state during state preparation, one can adjust the control action as necessary in a feedback loop, while using the same physical architecture as for an open-loop approach.

Deterministic state preparation using feedback has been demonstrated through the pioneering work of Haroche's group, whereby cavity eigenstates were prepared using the backaction from weak measurements as a decimation procedure [16,17]. Similar approaches have also been developed for both superconducting and semiconductor qubits, using either classically inspired feedback techniques [18] or

reinforcement learning approaches [19,20]. In these works, however, only eigenstates are being prepared. Recently, Porotti *et al.* [19] extended this idea using continuous-time multiplexed measurements [21] to prepare both single Fock states and superpositions. Preparing superpositions, however, proved to be more challenging with relatively low state preparation fidelities, and was restricted to a limited set of states.

The present work focuses on preparing cavity-state superpositions in a circuit quantum electrodynamic (cQED) architecture [22–24], using a generalized version of the measurements used by Haroche's group. It is shown that careful choice of these measurements, and therefore of their backaction on the cavity's state, allows reaching target states with high fidelity, but at the cost of a more elaborate optimization procedure. Indeed, Haroche's method crucially relied on having the cavity Fock states being the fixed points (eigenstates) of the measurement operators. The control problem was therefore simplified to bring the current cavity state near the targeted fixed point, with the measurement backaction playing the role of an attractor, hence allowing asymptotic convergence [25]. Extending Haroche's approach to prepare superpositions of cavity eigenstates is, however, not direct. As discussed in Refs. [26–29], degeneracies of the measurement operators eigenvalues, while necessary to stabilize superpositions, makes whole subspaces invariant under these operators. In such cases, relying on the measurements backaction alone will only ensure convergence towards the subspace containing the target state. To counteract this, one would need to add a second Hamiltonian control term to lift the degeneracies [28]. However, additional control introduces new error pathways, and makes the whole control protocol more complex. Such an approach will thus not be pursued here. Since previous approaches can only converge to subspaces, this raises the following question: Is it possible to find a controller that

---

*Yves.Berube-Lauziere@USherbrooke.ca

allows convergence to specific state superpositions with high fidelity without additional Hamiltonian control terms? This is the question addressed in the present work.

A further limitation in the present work is that only coherent displacements can be applied to drive the cavity. The limited control problem is interesting in its own right, as it puts forward the question of how to interact in an optimal way with the quantum backaction resulting from repeated weak measurements. A recent research direction addressing this question has been to make use of reinforcement learning (RL) [30] as an optimization method to find such optimal controls [31], through the underlying formalism of quantum-observable Markov decision processes [32], the quantum analog of the classical RL framework.

A significant body of work has thus studied RL methods to optimize the controls of quantum systems without feedback [33–45], focusing mostly on state transfer and gate synthesis. To note, Zhang *et al.* studied different optimization algorithms and found that deep Q-learning and policy gradient method to perform best in an open-loop state preparation task with discretized controls [36]. They, however, did not consider continuous controls to which off-policy actor critic architectures, which are known to have the best sample efficiency and performances in continuous actions settings [46–48], but are more fragile to train [30].

Of greater interest to the present work is another use of RL methods to design real-time controllers, thereby implementing quantum feedback. Along this line, Reuer *et al.* [20] showed the experimental viability with an RL agent doing real-time feedback during the readout of a superconducting qubit. RL has also been used to devise a feedback strategy for noise tailored error correction [49], for parameter estimation [50], to counteract the drift of a continuously monitored qubit [51], as well as to mitigate errors induced by nonadiabatic evolution [52]. Another line of work has been to employ RL to derive optimal policies for the cooling of potential wells, with added backaction induced by measurements in the quantum domain [53,54]. In addition, Refs. [55,56] used the measurement basis as a control variable for the RL agent to steer a given state to the target, while Porotti *et al.* [19] exploited instead the strength of continuous homodyne measurements to prepare superpositions of Fock states.

In the present work, the large state space along with the high variance of the dynamics coming from random measurement outcomes makes the preparation of state superpositions with RL challenging. Another objective of the present paper is thus also to pinpoint an RL method able to handle such dynamics and provide for convergence to specific desired cavity-state superpositions.

The rest of this paper is structured as follows. In Sec. II, the cQED model considered is defined with its allowed measurements and controls. We develop the theory showing that with the available measurements, only certain subspaces of superpositions can be stabilized. Considering this limitation, the control problem is then considered in the light of RL. Section III presents results for the ideal decoherence free setting as well as when decoherence is present. While decoherence reduces the state preparation fidelities that can be achieved, state-of-the-art devices operate in a regime where our approach allows the stabilization and recovery from

photon jumps in the cavity. In Sec. IV, the learned behavior of the RL agent is analyzed, comparing its optimal policy with other approaches, namely an on-policy RL agent along with a Lyapunov function-based controller.

II. SYSTEM DEFINITION AND CONTROL APPROACH

A. System considered

A standard cQED architecture is considered here, in which a qubit is dispersively coupled to one mode of a microwave cavity, with the following Hamiltonian:

$$H_{\text{system}} = \omega_c N + \frac{\omega_q}{2}\sigma_z + \chi N \sigma_z, \tag{1}$$

subject to a resonant control drive on the cavity given by

$$H_{\text{control}} = i(\varepsilon a^\dagger - \varepsilon^* a). \tag{2}$$

Here, $N = a^\dagger a$ is the number operator, with a and $a^\dagger$ being, respectively, the annihilation and creation operators of the cavity mode, and χ is the dispersive coupling strength between the qubit and the cavity mode. Associated with the control Hamiltonian is the displacement operator given by

$$D(\alpha) = e^{\alpha a^\dagger - \alpha^* a}, \tag{3}$$

with $\alpha = \varepsilon t$, the coherent resonant drive amplitude. Assuming the cavity to be in state $|\psi\rangle$, then after application of the resonant control drive it becomes $|\psi'\rangle = D(\alpha)|\psi\rangle$. In the case the density operator ρ is used to specify the state rather than the state vector $|\psi\rangle$, then the state ρ' after application of the drive is given by

$$\rho' = D(\alpha)\rho D(\alpha)^\dagger = D(\alpha)\rho D(-\alpha). \tag{4}$$

This transformation of ρ into ρ' can be written in superoperator form as [16]

$$\rho' = \mathbf{D}(\alpha)\rho. \tag{5}$$

In the dispersive regime, no energy exchange occurs between the qubit and the cavity mode. Instead, the cavity experiences a light shift depending on the state of the qubit, while the latter experiences a phase shift depending on the cavity's state. As detailed below, this interaction is at the core of the present proposal, with the phase information acquired by the qubit updating the knowledge about the cavity's state. It will be seen that the backaction from qubit measurements then acts as a decimation procedure to prepare a target quantum state in the cavity [17]. This backaction is thus a resource for controlling the cavity.

The measurement scheme's full sequence, as shown in Fig. 1, starts by preparing the qubit in state $|+\rangle = \frac{|e\rangle + |g\rangle}{\sqrt{2}}$ by applying a $\pi/2$ pulse on it, where $|g\rangle$ and $|e\rangle$ are, respectively, its ground and excited states, and then letting the qubit interact with the cavity for a given interaction time t_{int}. After the interaction, the qubit and cavity evolve to the state

$$|\Psi\rangle = \frac{1}{\sqrt{2}}\big(|g\rangle \otimes e^{-\frac{i}{2}(\phi_0 a^\dagger a + \delta\phi)}|\psi\rangle + |e\rangle \otimes e^{\frac{i}{2}(\phi_0 a^\dagger a + \delta\phi)}|\psi\rangle\big), \tag{6}$$

where $\phi_0 = t_{int} \cdot \chi$ is the phase shift per photon present in the cavity, which is an experimentally tunable parameter, $\delta\phi$ is a constant phase shift whose exact form is not important here,
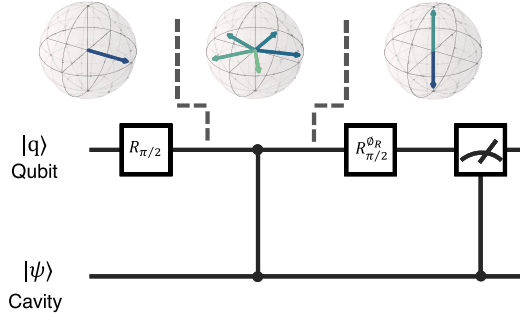
FIG. 1. Schematic of the measurement scheme for a single feedback cycle. A $\pi/2$ Ramsey pulse is applied to the qubit initially in state $|e\rangle$ to prepare the qubit's $|+\rangle$ state. After interaction between the qubit and cavity for a given time (interaction time), the qubit's state, which has now become dependent on the cavity's state that prevailed prior to the interaction, is projected back onto its energy eigenbasis, and subsequently measured. This Ramsey protocol implements a weak measurement of the cavity's state.

and $|\psi\rangle$ is the state of the cavity. Applying a second $\pi/2$ pulse projects the qubit back onto its energy eigenbasis, with the measurement probabilities determined by the phase between the two superposed qubit eigenstates, a procedure known as Ramsey interferometry. In the case the cavity is in one of its eigenstate $|n\rangle$ (number state or Fock state), the state of the qubit after interaction can be written as

$$|q\rangle = \frac{|e\rangle + e^{i(\phi_0 n + \delta\phi)}|g\rangle}{\sqrt{2}}. \tag{7}$$

This can be seen as the cavity state $|n\rangle$ imparting a phase $\phi_0 n + \delta\phi$ to the initial qubit superposition $\frac{(|e\rangle+|g\rangle)}{\sqrt{2}}$. To each cavity eigenstate $|n\rangle$, there corresponds a direction on the equatorial plane of the qubit's Bloch sphere, this direction being specified by the angle $\phi_0 n + \delta\phi$. In this case, Ramsey interferometry allows the weak QND measurement of the number of photons in the cavity [57]. Such weak measurements combined with coherent drive excitation of the cavity was the building block of the iterative measurement-based quantum feedback (MBQFB) control approach pioneered by Haroche *et al.* to prepare cavity Fock states using atoms in Rydberg states as probe qubits [16,17] (also called ancilla or auxiliary qubits in the error-correction literature).

B. Stabilization scheme

Owing to the dispersive interaction between the microwave cavity and the qubit, the backaction on the cavity's state following a measurement of the state of the qubit giving as outcome e or g is obtained through the following measurement operators:

$$M_g = \cos\left(\frac{\phi_0 N - \varphi_R}{2}\right), \tag{8}$$

$$M_e = \sin\left(\frac{\phi_0 N - \varphi_R}{2}\right). \tag{9}$$

Here, $\varphi_R = \phi_R - \delta\phi$, where ϕ_R is the phase of the second $\frac{\pi}{2}$ Ramsey pulse relative to the first one, which is a tunable parameter. More precisely, assuming the cavity to be in state

$|\psi\rangle$ prior to a measurement of the qubit, and given that the qubit is measured to be in state s, with $s = e$ or g, the state of the cavity after the qubit measurement is $|\psi'\rangle = M_s|\psi\rangle$. If the density operator is used instead of the state vector, this becomes

$$\rho' = \frac{M_s \rho M_s^\dagger}{\text{tr}(M_s \rho M_s^\dagger)}, \tag{10}$$

which in superoperator form is written as

$$\rho' = \mathbf{M}_s \rho. \tag{11}$$

To prepare a cavity-Fock-state superposition with an iterative feedback loop protocol in which a measurement operator M_e or M_g affects the cavity's state in each loop, similarly to the protocol of Haroche *et al.* for preparing single Fock states, it is necessary that the targeted-Fock-state superposition be left unchanged by the measurement's backaction so as to keep stable the state reached upon convergence of the feedback loops. This means that the Fock-state superposition must be an eigenstate of both measurement operators. For simplicity, taking as target the two-state superposition

$$|\psi^{\text{target}}\rangle = c_1|n_1\rangle + c_2|n_2\rangle, \tag{12}$$

where without loss of generality $n_2 > n_1$, the following conditions must therefore hold:

$$\begin{aligned} M_g|\psi^{\text{target}}\rangle &= \lambda_g|\psi^{\text{target}}\rangle, \\ M_e|\psi^{\text{target}}\rangle &= \lambda_e|\psi^{\text{target}}\rangle. \end{aligned} \tag{13}$$

These conditions translate to

$$\begin{aligned} \cos\left(\frac{\phi_0 n_1 - \varphi_R}{2}\right) &= \cos\left(\frac{\phi_0 n_2 - \varphi_R}{2}\right), \\ \sin\left(\frac{\phi_0 n_1 - \varphi_R}{2}\right) &= \sin\left(\frac{\phi_0 n_2 - \varphi_R}{2}\right). \end{aligned} \tag{14}$$

Since both cosines and sines must be equal, the arguments must be equal to within $2\pi k$, with k an integer, which imposes that the phase shift per photon be of the following form:

$$\phi_0 = \frac{4\pi k}{\Delta n}, \tag{15}$$

with $\Delta n = n_2 - n_1$ ($k = 1$ is used in the sequel). It is to be noted that only Δn matters in ϕ_0. Furthermore, it is easily seen that the argument for the two-state superposition is readily generalizable to an arbitrary superposition of Fock states whose numbers differ by Δn. Thus, entire subspaces containing Fock states with numbers differing by Δn can be stabilized. Specifically, for a given Δn that can be chosen, leading to a specific value of ϕ_0, states in the following subspaces can be stabilized by the measurement operators:

$$W_0^{\Delta n} = \text{span}\{|0\rangle, |\Delta n\rangle, |2\Delta n\rangle, \ldots\},$$

$$W_1^{\Delta n} = \text{span}\{|1\rangle, |1 + \Delta n\rangle, |1 + 2\Delta n\rangle, \ldots\},$$

$$\vdots$$

$$W_m^{\Delta n} = \text{span}\{|m\rangle, |m + \Delta n\rangle, \ldots |m + l\Delta n\rangle, \ldots\},$$

$$\vdots$$

$$W_{\Delta n-1}^{\Delta n} = \text{span}\{|\Delta n - 1\rangle, |2\Delta n - 1\rangle, |3\Delta n - 1\rangle, \ldots\}. \tag{16}$$
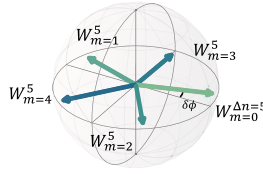
FIG. 2. Example of the mapping between stabilizable subspaces $W_m^{\Delta n}$ and directions in the equatorial plane of the qubit's Bloch sphere in the case $\Delta n = 5$ ($\phi_0 = 4\pi/5$).

These subspaces $W_m^{\Delta n}$, $m = 0, \ldots, \Delta n - 1$, will be called the stabilizable subspaces. It is seen that each basis state in the generic subspace $W_m^{\Delta n}$ contains a number n of photons such that $n \bmod \Delta n = m$. Only superpositions of Fock states living inside each of these subspaces can be prepared using the iterative measurement feedback protocol described above. While the conditions given in Eq. (13) restrict the state superpositions that can be prepared, the modulo nature of the number of photons contained in the stabilizable subspaces is a resource that is exploited in many bosonic codes that allow for error correction [5].

Similarly to the discussion following Eq. (7) about the effect on the qubit-state superposition when the cavity is in a Fock state, here any cavity state in a stabilizable subspace $W_m^{\Delta n}$ imparts the same phase $\Phi^{\Delta n}(m)$ on a qubit superposition. Indeed, the qubit state after interaction with the cavity in such a state is given by

$$\left| q_m^{\Delta n} \right\rangle = \frac{|e\rangle + e^{i\Phi(W_m^{\Delta n})}|g\rangle}{\sqrt{2}}, \tag{17}$$

with $\Phi(W_m^{\Delta n}) \equiv \Phi(m + l\Delta n) = \Phi^{\Delta n}(m) = \frac{4\pi km}{\Delta n} + \delta\phi$ (mod 2π). Hence, each subspace $W_m^{\Delta n}$ is mapped to its own direction in the equatorial plane of the qubit's Bloch sphere, since all basis vectors $|m + l\Delta n\rangle$ in this subspace are mapped to the same angle $\Phi^{\Delta n}(m)$, which does not depend on l. Figure 2 illustrates an example for $\Delta n = 5$. It is to be noted that in the case $\Delta n = 2p$ is even, with p a positive integer, the subspaces with indices m and $m + \Delta n/2$ will be mapped to the same angle $\Phi^{\Delta n}(m)$ (mod 2π) in the equatorial plane. Furthermore, if $\Delta n = 2p$, with p even, that is $\Delta n = 4r$, with r a positive integer ($r = 1, 2, \ldots$), stabilizable subspaces with indices m and $m + \Delta n/4$ will be mapped to angles $\Phi^{\Delta n}(m)$ opposing in the equatorial plane [i.e., angles that differ by π (mod 2π)]. Since the case $\Delta n = 4r$ is a particular case of $\Delta n = 2p$, it means that in the case $\Delta n = 4r$, four subspaces, namely the subspaces with indices m, $m + \Delta n/4$, $m + 2\Delta n/4$, and $m + 3\Delta n/4$ will be mapped on the same angle or opposing angle (in fact two on the same angle and the other two on the opposing angle). The significance of this is that it is not possible to discriminate subspaces, which are mapped to the same angle (mod 2π) through the probabilities of the measurement outcomes e or g when measuring the qubit. However, as discussed below, by an appropriate choice of φ_R, subspaces that are mapped to opposing angles, can be discriminated through the probabilities of the measurement outcomes for the qubit (note that for example if φ_R is taken equal to $\pi/2$, that is when the Ramsey interferometer

operates at midfringe as is often done [16], then one cannot discriminate subspaces with opposing angles).

Odd and even stabilizable subspaces

A 4π factor appears in the numerator of Eq. (15), as opposed to the 2π more often seen in usual parity measurements [58]. This is caused by the necessity of having the same eigenvalue for all superposed states after a measurement, rather than preserving the same measurement probability for these states. Indeed, setting the prefactor as 2π would imply to keep track of an alternating phase between each measurement for the corresponding Fock states, which have nonzero population. Such tracking of the phase is possible in superconducting circuits, as it is always possible to know when a system measurement is carried out.

To determine the value of the parameter ϕ_0 in the measurement operators, this implies that setting a 4π factor for odd subspaces leads to an adequate subspace stabilizer, with no overlapping problem. For even subspaces, the only option is to choose a 2π factor to prevent the overlapping, and then resort to phase tracking.

The second parameter defining the measurement operators, φ_R, is free. Dotsenko *et al.* [16] is followed here, and its value is set at midfringe visibility for odd Δn subspaces. On the Bloch sphere representation, this means the projection axis is perpendicular to the target subspace. For even subspaces, each vector on the Bloch sphere is always opposing another one, which prevents the use of a perpendicular projection axis. Rather, ϕ_R is set at an angle of $\frac{2\pi}{5}$ from the target subspace, a value chosen so that all subspaces can be assigned a different probability of being measured in either $|g\rangle$ or $|e\rangle$.

C. Control decision problem

As mentioned in the introduction, conventional control techniques to stabilize specific superpositions such as Lyapunov-based control are difficult to implement, and are bound to get trapped in local minima, as the measurement backaction stabilizes the whole subspace consisting of superpositions of states with $n \bmod \Delta n$ photons. In this context, designing a Lyapunov function that decreases monotonically towards the target state rather than the target subspace is not straightforward, if possible at all. Such an approach would prevent both large early displacements in the feedback sequence that temporarily move the state further from the target, as well as small displacement corrections within the target subspace. To overcome this difficulty, a reinforcement learning (RL) approach is used herein, which in principle can learn the system dynamics directly in a model-free manner from the experimental apparatus.

While previous work on feedback RL-based quantum control in continuous actions spaces has mainly focused on on-policy architectures [19,52,54], here use is made of the more sample efficient off-policy learning paradigm. Specifically, an actor-critic architecture is resorted to, where the actor network updates its policy by deriving optimal state-action tuples from the Q function learned by the critic. In contrast to Monte Carlo-based methods such as proximal policy optimization (PPO), the critic here bootstraps state-action estimates in the update of its own Q function, which

reduces the variance in the reward estimate. However, this also increases sensitivity to bias in the estimator, which can be problematic in stochastic setting, and lead to overestimation of the Q function [59]. To prevent this, a distributional RL approach [60,61] is used here, which learns the Q function by regressing over the distribution of the returns, rather than their mean. Although it is still unclear what makes such an approach more stable, it is believed to help Q learning when using nonlinear functions approximation [62], by providing richer information about the environment dynamics. Additionally, a recent algorithm has been introduced, called truncated quantile critic (TQC), which builds a distributional version of the Q function on top of a soft-actor critic algorithm. TQC allows for finer control of overestimation bias, and has demonstrated superior performance compared to other state-of-the-art algorithms, particularly in high-dimensional stochastic environments [48]. Here, the TQC implementation from STABLE-BASELINES3 [63] is resorted to, whereas the specific hyperparameters used can be found in Appendix A.

Finally, the reward function must be defined. Its formulation is critical in guiding the agent towards achieving maximum fidelity with respect to the target state, using as few feedback cycles as possible. Similar to Porotti *et al.* [19], it was found, for the present purposes, that a reward consisting of higher powers of the fidelity was most effective to prevent convergence to a local optimum. Defining the fidelity between two density matrices as

$$F(t) = (\mathrm{tr}\sqrt{\sqrt{\rho(t)}\rho^{\mathrm{target}}\sqrt{\rho(t)}})^2, \qquad (18)$$

the reward function is chosen to be

$$r(t) = (F(t))^4 + 4\,(F(t))^{25}. \qquad (19)$$

The choice of the form of the reward function was guided by heuristics, and the specific numerical values of the exponents (4 and 25) and coefficients (1 and 4) were empirically chosen so as to give agents, which reach highest fidelities. The right-hand side of Eq. (19) includes two terms that influence the training process. The first term helps to accelerate training by providing a dense reward to the agent, whereas the second ensures that the maximum reward available is sharply peaked near unit fidelity. This borrows from approaches found in curriculum learning [40,64], where the agent is given guidance in a first step on how to reach the correct subspace, and then in a second step on how to reach the actual target state inside this subspace.

D. Quantum filter

To use a neural network to process the sequence of measurement outcomes e or g, it is necessary to incorporate some form of memory of past inputs into the network. One option is to use a recurrent neural network, which can process sequential data by using feedback connections that allow information to be passed from one time step to the next.

Alternatively, the density matrix of the cavity state can be used as an input vector, which encodes all past information about the system. This approach requires using a quantum filter to estimate the cavity state recursively, using the displacement drive α and measurement outcome as inputs at each time step. A quantum filter, which provides a state estimator

analogous to a Kalman filter in classical control theory, allows obtaining in the computer an estimate of the true state of the physical system of interest in real time.

Here, as input vector, use is made of the vectorized density matrix of the cavity state, separated into two parts to account for both real and imaginary components. For target states that do not contain any imaginary part, only the real components are kept to minimize the network dimensions. The cavity state is therefore estimated recursively using a quantum filter. Following Haroche's group previous work [16,17], and in absence of decoherence, this filter can be expressed in superoperator form as

$$\rho_{t+1} = \mathbf{M}_t\,\mathbf{D}_t\,\rho_t, \qquad (20)$$

with ρ_t being the cavity density matrix estimated at time step t, and ρ_{t+1} being the estimate at the next time step $t+1$. $\mathbf{M}_t$ and $\mathbf{D}_t$ are, respectively, the measurement and displacement superoperators at time step t associated with the measurement operators given in Eqs. (8) and (9) and the displacement operator given in Eq. (3).

For states with a relative phase, the RL agent outputs two actions, corresponding to the real and imaginary parts of the displacement amplitude α, which are limited to be in the interval $[-1, +1]$. In the case of a state without a relative phase, the agent outputs only one action for the real component of the displacement. The cavity is initialized as an educated guess to a coherent state with mean energy similar to that in the target-Fock-state superposition. The initial cavity state for the feedback sequence is then

$$\rho_{t=0} = \mathbf{D}(\alpha_{\mathrm{guess}})|0\rangle\langle0|,$$

where the modulus of α_{guess} is given by the mean photon number of the target state. its exact form is then given by $\alpha_{\mathrm{guess}} = \sqrt{\bar{n}}e^{i\theta}$ with θ the relative phase between the superposed states. After that initial displacement, the agent is then allowed to make an additional one before the measurement sequence in order to fine tune the initial guess.

In the absence of quantum jumps and other decoherence channels, this filtered density matrix corresponds exactly to the cavity state. Below in Sec. III B, this filter, along with the cavity-state evolution, will be modified to account for noise in the system. The general structure of the feedback loop just described is summarized in Fig. 3.

III. RESULTS

A. Idealized case

An idealized case is first considered, in which measurements are assumed perfect and the cavity is not subjected to decoherence such as decay and dephasing. Simulations rely on the stochastic evolution of trajectories, updating the cavity-state density matrix at every feedback cycle, according to the recursive quantum filter of Eq. (20). During training, the maximum number of feedback cycles per episode is limited to 50, and the cavity Hilbert space is truncated to $n = 29$ photons. Whenever the photon number population in the Fock states $n = 28$ or $n = 29$ is above a 2% threshold, the episode is stopped to prevent the RL agent from being biased by Hilbert space truncation. Figure 4 depicts the training curves for three
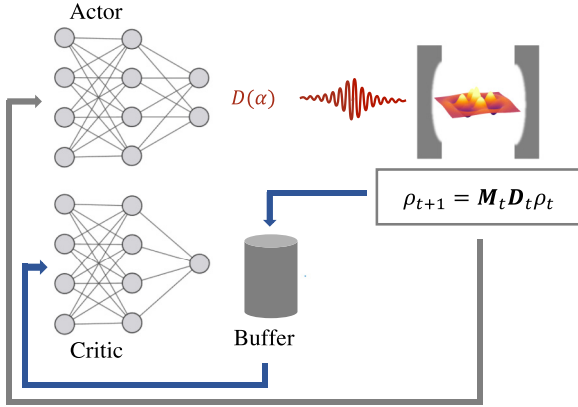
FIG. 3. Schematic of the RL procedure. The cavity state is estimated by a quantum filter F, which is fed as input to the actor. The actor then outputs a displacement amplitude to be applied to the cavity. At every step, the critic samples from a buffer of past experiences to approximate the Q function. During policy iteration, the actor optimizes its policy with the Q function learned by the critic.

different bosonic states, obtained by averaging final states fidelities obtained for 600 trajectories.

The state $|\psi\rangle = (|1\rangle + |4\rangle)/\sqrt{2}$ is significantly harder to learn for the agent, in part because it has support on only two Fock states. As such, it is harder to control leakage to other states inside the stabilized manifold. This is similar to the limitations of a Lyapunov-based controller, which stops once the state reaches the target subspace, rather than the target state. This state will thus be taken as a benchmark to explore the behavior of the feedback protocol further on.

Figure 5 shows two examples of trajectories that reach the target state $|\psi\rangle = (|1\rangle + |4\rangle)/\sqrt{2}$, starting from an initial coherent state of amplitude equal to the square root of the target state's mean photon number as mentioned previously. Both cases converge within about five feedback cycles. The trajectory in the top panel converges monotonically towards
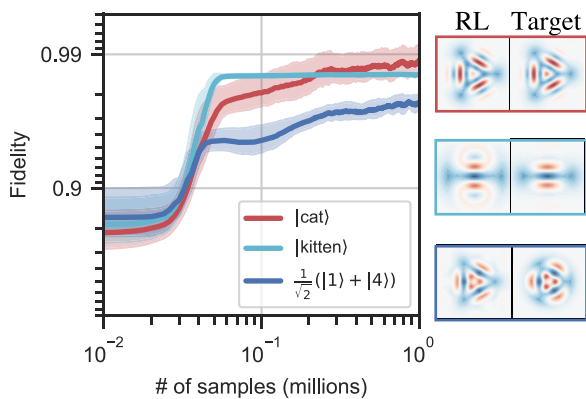

FIG. 4. Training curves of the TQC RL agent for: (i) a three-components cat state with mean photon number $n = 3$ (red), (ii) an equal superposition of the two kitten binomial logical states (light blue), and the state $|\psi\rangle = \frac{1}{\sqrt{2}}(|1\rangle + |4\rangle)$ (dark blue). On the right are the Wigner functions for prepared states, evaluated after training and compared with the target state.
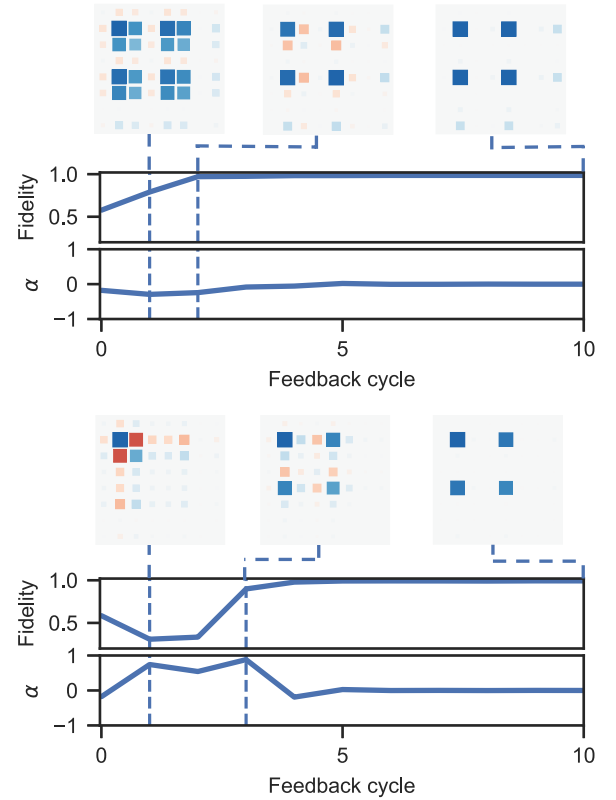

FIG. 5. Trajectory examples for the preparation of the state $|\psi\rangle = (|1\rangle + |4\rangle)/\sqrt{2}$, with Hinton plots of the density matrix at different steps during the preparation sequence. A monotonically increasing fidelity state preparation is shown in the top panel, and the bottom panel shows a trajectory that requires stronger control drives to recover from a sequence of measurements projecting away from the target state.

the target state, although with the presence of some leakage to the $|n = 7\rangle$ Fock state. The trajectory shown in the bottom panel does not show such leakage, but requires larger displacements to converge to the target state. This illustrates a key benefit of reinforcement learning (RL): as it directly learns the control dynamics from experience, it can handle nonlinearities in the control space that are essential for fast convergence, but difficult to handle analytically.

Figure 6(a) shows the time evolution of the fidelity with a fully trained RL agent. It is seen that the mean fidelity increases slowly compared to the median, Also, 75% of the trajectories are above 98% fidelity after about ten feedback cycles. Comparing this RL agent to a Lyapunov function-based controller (see its derivation in Appendix B), Fig. 6(b) shows the fidelity distribution at the end of a 50 cycles sequence. The RL framework outperforms the Lyapunov controller, as its fidelity distribution is clearly above that obtained with Lyapunov control. Remarkably, it also does so without increasing the amount of unsuccessful state preparations. Section IV further explores the disparities in the two approaches.

Performances of fully trained RL agents for the preparation of a variety of bosonic states are shown in Fig. 6(c). While some of the states correspond to an equal superposition of two photon number states, the logical states presented in the
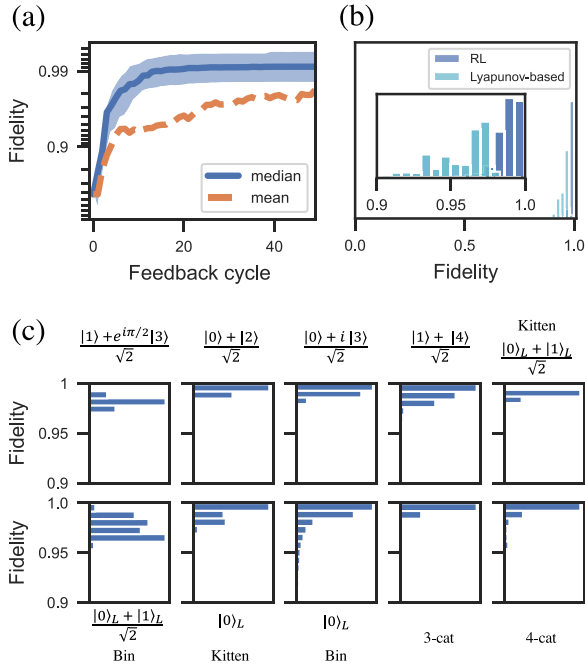
FIG. 6. (a) Time evolution of the state for the preparation of the $(|1\rangle + |4\rangle)/\sqrt{2}$ target state, showing the median with its 25–75 percentile distribution (shaded area) as a function of time, with the mean converging slower than the median to the target state. Most of the state preparation occurs within the first ten feedback cycles. (b) Distribution of the final fidelities in (a) for the RL agent alongside a Lyapunov function-based controller. In the inset, it can be observed that the RL procedure converges to higher fidelities, preventing early stopping due to local optima. (c) Final fidelities distribution for a set of different cavity states. All states have most of their distributions above the 95% mark, even for more difficult binomial encodings. Here, Bin corresponds to the binomial code with support on number states {0, 3, 6, 9}. The multicomponent cat states are defined on the $n \bmod 3 = 0$ number states for the three-cat state and on $n \bmod 4 = 1$ for the four-cat state.

figure all correspond to superpositions of multiple number states with different amplitudes. It is observed that having larger superpositions does not necessarily result in increased difficulties, i.e., lower state preparation fidelities. Based on the specific states targeted, we rather interpret it in that the difficulty lies, in part, in how far the envelope of the target state amplitudes is to that of a coherent state photon distribution. Indeed, in the absence of any feedback, the measurements employed here act as projection operators onto a given subspace. For the case of cat states, applying such projectors to an initial coherent state generates a cat state of corresponding generalized parity. As such, the feedback control from coherent displacements is mainly used, for these cat states, to steer the state towards the right subspace. This could explain why the two studied cat states show the highest fidelities. Furthermore, for target states that differ from a cat state, the controller has to finely control the amplitudes and leakage, which is a harder task due to the limited control of the coherent drive. In all cases, average fidelities are above 96.6%, with the median consistently above the average by about 2%. Similar to the previous case, the distribution is

sharply asymmetric, with only a handful of trajectories failing to converge towards the target state after 50 cycles. In a real experiment, as the fidelity of each specific trajectory is tracked with the quantum filter, one could simply discard any trajectory that fails to converge after a given prefixed number of feedback cycles. Such heralded state preparation has already been realized experimentally [65]. This illustrates a main advantage of measurement-based feedback over open-loop control, and the reason why the median fidelity metric is in this case more representative of the performance than the mean for the proposed scheme.

Unsurprisingly, states with higher mean photon population have lower fidelities, which is likely due to the need for larger displacement drives to prepare these states. These larger drives tend to populate other Fock states within the stabilized subspace. This could be improved by varying the phase shift per photon in time to decimate these in-subspace Fock states, or by directly leaving the measurement parameters as additional controls for the RL agent as in Ref. [66].

It was also observed empirically that a smaller Δn requires less feedback cycles, caused by the larger sensitivity of the measurements to changes in m (i.e., more efficient discrimination between different subspaces). This makes the backaction stronger, hence decimating the population in other subspaces faster. Interestingly, superposition states that contain a $|0\rangle$ component have higher fidelities and less dispersion in their distribution. A possible explanation is that the $|0\rangle$ state is a lower bound of the Hilbert space, in the sense that there cannot be negative numbers of photons. Along with the nature of the coherent drive, which has a decreasing exponential envelop, this possibly lowers the probability of a transition to another subspace. Such features of the control drive and environment may make it easier for the agent to find the best policy for preparing superpositions.

B. Realistic case with decoherence and imperfections

Decoherence adds additional loss channels to the system Hamiltonian, with the master equation in Lindblad form now governing the evolution of the state density operator, and given in the rotating frame by

$$\frac{d\rho}{dt} = \mathbf{L}\rho = \kappa \left(a\rho a^\dagger - \frac{1}{2}(a^\dagger a \rho + \rho a^\dagger a) \right), \qquad (21)$$

where the cavity is assumed to be at zero temperature, so that only photon decay contributes to decoherence. Because of the negligible intrinsic dephasing rate of superconducting cavities, photon loss is therefore the only decoherence channel affecting the cavity considered here [67].

Errors coming from the probe qubit are, however, taken into account, as these impact the resulting backaction on the cavity, given by the updated filter terms in Eq. (20) [see also the upcoming Eq. (25), which takes measurement errors into account]. Since qubit T_2 errors commute with the interaction Hamiltonian, they can be considered as occurring after the interaction. Such errors have an effect that is similar to measurement errors, since both are induced by the σ_z operator. As for qubit T_1 errors, which would dephase the cavity, an error transparency method is considered, where relaxation events do not impact the cavity state [68]. So, it is assumed here that

our protocol is T_1 fault tolerant. In summary, qubit errors are modeled in the simulation as an effective σ_z type error, with the effective decoherence rate being the sum of measurement and T_2 errors.

1. Quantum filter and simulation

As the agent input should consist of the best possible estimate of the true quantum state of the cavity, the filter update equation [Eq. (20)] needs to be adapted to account for the decoherence channels mentioned above. In the simulations, two distinct density matrices are evolved. One evolution concerns the true cavity state, where photon loss events correspond to discrete quantum jumps. This is thus a simulation of the true physical system. The other evolution provides a filtered estimated state of the cavity, which does not have access to these discrete jumps. This filtered estimated state is indeed the only information available about the actual state in practice for control purposes (built from measurements of the true system), since having full information about the state is impossible. The latter evolution estimates the cavity relaxation using a first-order expansion of the Lindblad dissipator given in Eq. (21), which can be added to the filtering equation with a superoperator of the form

$$\mathbf{T}\rho = (\mathbf{1} + T_{\text{cav}}\mathbf{L})\rho, \tag{22}$$

where T_{cav} is the cavity lifetime. In the case of a discrete jump, the filter will then move away from the actual cavity state, while subsequent measurements obtained from the true cavity simulation will update the filter until it converges back towards the real cavity state. This is the beauty of quantum filtering, which is analogous to Kalman filtering in classical control theory, as it is able to improve the estimate of the state as information is accumulated over time, consisting of the known measurements results obtained and the actions performed on the system.

Due to the finite measurement accuracy of the probe qubit, the resulting estimated state is a statistical mixture of the two possible outcomes. This is taken into account with the following superoperators

$$\mathbf{P}_e\rho = (1 - P_{f,e})\mathbf{M}_e\rho + P_{f,e}\mathbf{M}_g\rho, \tag{23}$$

$$\mathbf{P}_g\rho = (1 - P_{f,g})\mathbf{M}_g\rho + P_{f,g}\mathbf{M}_e\rho, \tag{24}$$

where the weights $P_{f,e}$ and $P_{f,g}$ of each measurement operator depend on the erroneous state assignation probabilities $\eta_{e|g}$ and $\eta_{g|e}$ in a measurement. Following Ref. [17], in the case of a measurement that delivers e, whereas it should have been g, the weight is given by $P_{f,g} = \eta_{e|g}P_e/[(1 - \eta_{e|g})P_g + \eta_{e|g}P_e]$, where $P_e = \text{tr}(M_e\rho M_e^\dagger)$ and $P_g = \text{tr}(M_g\rho M_g^\dagger)$. In the other case that a measurement delivers g, whereas it should have been e, the weight is $P_{f,e} = \eta_{g|e}P_g/[(1 - \eta_{g|e})P_e + \eta_{g|e}P_g]$. The estimated state is then updated using the following modified recursive filtering equation:

$$\rho_{t+1} = \mathbf{P}_t \, \mathbf{T} \, \mathbf{D}_t \, \rho_t. \tag{25}$$

It is to be noted that only the filtered state is given as input to the agent, as this would be the only information available in a real experiment.

2. Simulation results

Following a procedure similar to that in Sec. III A, 3000 trajectories are simulated, each consisting of 2000 feedback cycles (the smaller number of total trajectories simulated here compared to than in Sec. III A is due to the need to simulate more feedback cycles for each trajectory in the present case). In these simulations, a time of $1\,\mu s$ is considered for one feedback cycle (300 ns for the cavity-qubit interaction time, while leaving 700 ns for qubit measurement and rotations; according to literature, these values are well within current device performances [14,69]). The cavity lifetime considered here is 1 ms [58,67]. The errors in qubit state measurement assignments are $\eta_{e|g} = 0.01$ and $\eta_{g|e} = 0.02$.

Figure 7 shows state evolution and preparation results with decoherence, with Fig. 7(a) depicting trajectory examples with noticeable photon loss events and subsequent recovery. In this case, the agent is able to recover from photon loss events with a delay depending on the speed at which the filter recognizes the loss event. An interesting behavior is shown in the bottom panel of Fig. 7(a), where no photon loss events are registered; the cavity state instead evolves deterministically inside the stabilized subspace as indicated by the small and slow decay of the fidelity between jumps. This decay is due to the slow population transfer from the $|4\rangle$ state to the $|1\rangle$ state; this comes from the Zeno dynamics induced by the backaction of the measurements [70], whereby the evolution of the state is slowed down by the measurements. The out-of-subspace leakage from the deterministic evolution of the stochastic master equation between quantum jumps is hence being suppressed by the measurements. To correct this slow fidelity decay, the RL agent has limited control. Indeed, between jumps, the controller shows a jittering behavior of growing amplitude lasting over 0.1 ms as the fidelity decays. This is distinct from the behavior following a photon loss, where the correction applied appears as an isolated sharp peak in the control amplitude. This phenomenon will be discussed further in Sec. IV, where it will be shown that small corrections once near the target state are not the primary means to achieve high fidelity. Figure 7(b) shows the behavior of an ensemble of trajectories, for the same loss parameters as above. Unsurprisingly, the mean is significantly lower, as photon loss events, occurring randomly, drag it downwards. The median stabilizes around a fidelity of 95.5%, after having reached fidelities of 98%. This drop can be explained by the RL agent failing to recover fully from photon losses on average.

To compare with the situation corresponding to no control (free evolution), the green (dashed) curve shows the average evolution where perfect state preparation is assumed at time $t = 0$. This shows that the approach proposed here, even with limited control, namely measurement backaction combined with coherent driving in a feedback loop, helps stabilize the target state.

IV. DISCUSSION

A. Robustness to noise

To further analyze the robustness of the TQC agent proposed here, its performance is studied with a fully trained agent for ranges of cavity lifetimes and probe qubit
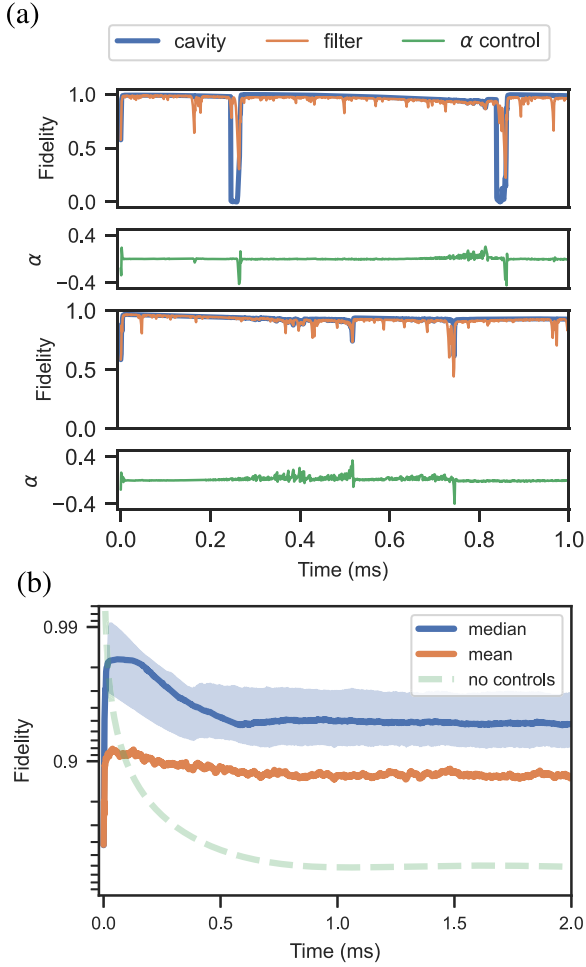
(a)

(b)

FIG. 7. Preparation and stabilization of the $(|1\rangle + |4\rangle)/\sqrt{2}$ state under decoherence. (a) The RL agent is able to recover from photon loss events, as shown in top panel. When no photon loss occurs (bottom panel), Zeno dynamics take place, with population transfer occurring only inside the target subspace (due to measurement backaction suppression of population leakage out of the subspace), but which does not prevent the slow decay of the fidelity. The control amplitude α (green curve) is 0 for most of the sequence, except when the filter [orange (light gray) curve] detects a photon loss, or when the fidelity goes below certain value (bottom panel). (b) Time evolution for a cavity lifetime T_1 of 1 ms and a feedback cycle time of 1 μs. Median fidelities are still able to reach values similar to the ideal case, although they stabilize around about 95% when multiple photon loss events occur. The mean fidelity remains at 90% throughout, pushed down by the momentary photon loss. With perfect initialization at $t = 0$, the average state fidelity would have decayed according to the master equation, as shown by the green curve.

imperfections. Note that all results were obtained using the same RL agent, trained on the ideal model. It was found that such a model usually performs better than one trained on a lossy system. One possible explanation is that the Markovian nature of the quantum filter combined with the exploration properties of the RL training are sufficient to learn an optimal policy. Indeed, as the RL agent learns about the environment, it explores states similar to those resulting from quantum jumps. However, contrary to the case with significant deco-
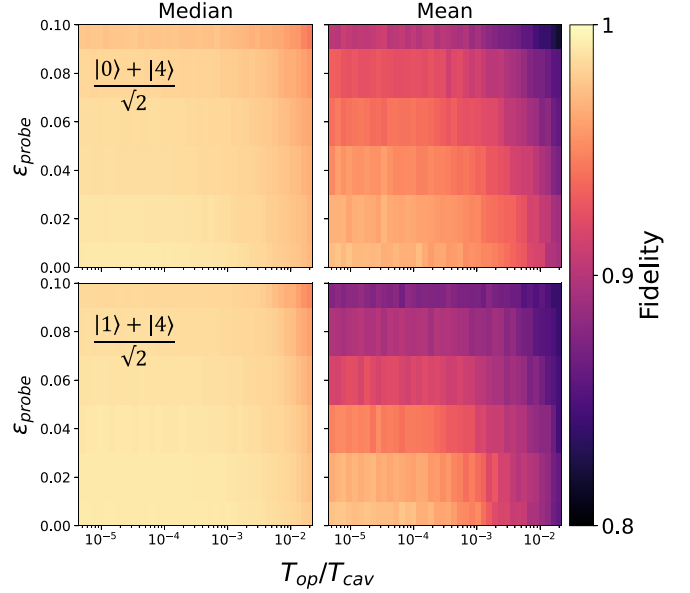

FIG. 8. Maximum fidelities reached by the RL agent during state preparation as a function of decoherence parameters, with the horizontal and vertical axes corresponding, respectively, to the cavity lifetime and the probe qubit's errors.

herence, it is also able to explore high-fidelity states, which then makes it a more complete agent, able to perform well under different system dynamics. In Fig. 8, the lifetimes are expressed as the ratio of the total feedback cycle operation time (1 μs is considered here) and the cavity lifetime (which is varied to include currently experimentally realistic values). Also, as mentioned in Sec. III B, errors from the probe qubit are summarized into an effective error denoted ϵ_{probe}, consisting of the sum of all individual qubit decoherence channels. Figure 8 shows the maximum median and average fidelities attained during a 50-cycles state preparation procedure, as a function of both cavity decay and qubit errors for the $(|1\rangle + |4\rangle)/\sqrt{2}$ and $(|0\rangle + |4\rangle)/\sqrt{2}$ states, which, respectively, have an odd and even Δn.

For the median fidelity metric, it is seen in Fig. 8 that both states are robust to a large range of parameters, with a drop in fidelity being seen in the top right corner corresponding to high error rates in both cavity and qubit. The impact of decoherence is unsurprisingly more pronounced for the mean fidelity metric, indicating that while the majority of state preparation sequences lead to high-fidelity states, a small fraction, however, completely fail to reach high fidelity.

The drop in the median fidelity value as measurement errors are more prominent can be attributed to a state estimation problem for the controller. When measurement errors are high, the quantum filter needs more measurement results to construct an accurate estimate of the true cavity state. On some occasions, the displacement applied to the cavity can increase the deviation between the estimated state and the true cavity state in such a way that the former fails to converge back to the cavity state. Such instances of deviation between the cavity and the estimated states happen more frequently following a photon loss, which could also explain why the state $(|1\rangle + |4\rangle)/\sqrt{2}$ is more subject to decoherence, as it has
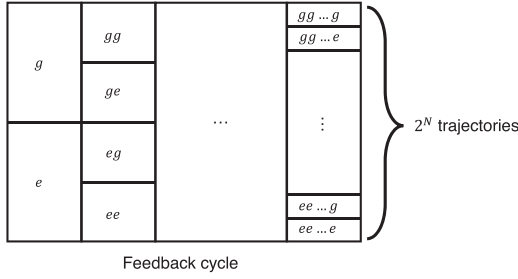
FIG. 9. Structure of the trajectories calculations presented in Fig. 10. At each additional feedback cycle, the binary tree is expanded by a factor of two, corresponding by two possible measurements outcomes in each branch.

a slightly larger mean photon number, and also contains the state $|1\rangle$, which can still decay down to $|0\rangle$ as compared to the state $(|0\rangle + |4\rangle)/\sqrt{2}$ in which $|0\rangle$ cannot further decay down.

B. Understanding the trained policies

Attention will now be turned to the policies learned by different types of agents. The TQC agent proposed herein will be compared with two other approaches: the RL-based PPO and the Lyapunov-based controller mentioned previously. This controller chooses the best displacement drive by performing a line search over a linearization of the Lyapunov fidelity following application of the displacement operator to the current cavity state, see Appendix B.

The policy space has the form of a binary tree with 2^N possible trajectories as each measurement leads to two distinct possibilities in the decision tree. Here, N is the depth of the tree corresponding to the number of feedback cycles, which in this case is chosen to be 10 as a compromise so that the computation time does not become prohibitive since here all trajectories are exhaustively examined. Figure 9 provides details on the procedure and how the results in Fig. 10 are

obtained. Starting from a given initial state, the cavity states of all possible combinations of g and e measurement outcomes are computed at every feedback cycle. The corresponding metric, that is, the fidelity between the cavity state and the target state, is given by

$$F(M_{s_k}D_{s_k}, ..., M_{s_2}D_{s_2}, M_{s_1}D_{s_1}|\psi_0\rangle, |\psi^{\text{target}}\rangle). \quad (26)$$

To each measurement outcome corresponds a probability of occurrence, with the product of these individual probabilities for a given trajectory being the probability of occurrence of the whole trajectory.

Figure 10 shows such trajectory distributions, with the color scale corresponding to the fidelity after a measurement at a given time step (feedback cycle), lighter colors corresponding to higher fidelities. Displaying the fidelity in this manner provides a high-level view of the policies learned, and tells how far from the target state (and thus out of subspace) the agent can go in order to maximize the final fidelity.

1. Starting from a coherent initial state

Trajectories are shown in the top row of Fig. 10, with the initial state of the feedback sequence being the same initial coherent state used during training of the RL agents. There are differences between the agents in the structure of their trajectories. For instance, the Lyapunov controller has well-defined paths in the fidelity space, influenced by previous measurements results. This behavior is similar for the PPO agent, although less pronounced. These branchings are much less apparent for the TQC agent. This indicates a better ability to exploit the combined effect of measurement backaction and coherent drive. In other words, RL agents, and especially the TQC agent, appear to be less influenced by the measurement outcomes than they are exploiting them. It is seen in Fig. 10, especially in the top part corresponding to a first ground-state measurement, that both RL agents learn to do penalizing displacements in the early feedback cycles, that
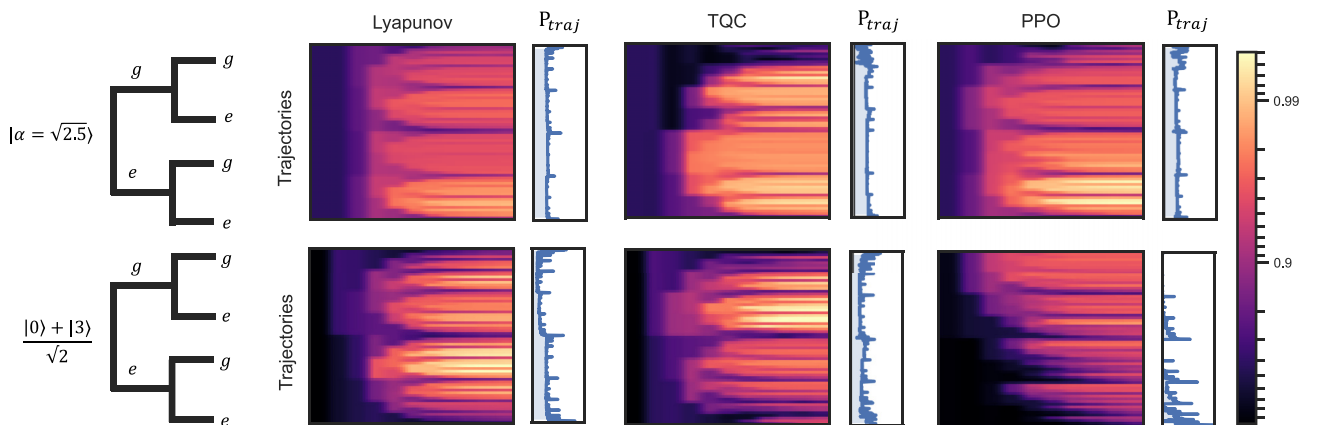

FIG. 10. Comparison of policies learned by different agents to prepare the state $\frac{|1\rangle+|4\rangle}{\sqrt{2}}$. Top row, left: Schematic of the exhaustive trajectory search in the form of a binary tree. The initial state is a coherent state with a mean photon number of 2.5; each branch corresponds to a specific trajectory outcome. The rest of the top row shows the evolution of the fidelity along each of the 2^{10} trajectories, for Lyapunov-based control, the TQC and PPO RL agents. Also shown are the probabilities of occurrence of the different trajectories plotted on a log scale. RL-based methods perform best, owing to their strong initial displacements allowing us to reach higher final fidelities. Bottom row: Same as in top row, but initializing with the state $\frac{|0\rangle+|3\rangle}{\sqrt{2}}$, which is in a stabilizable subspace different than that of the target state. Lyapunov and TQC agents are both able to reach the target subspace. The PPO agent fails at learning any policy, which indicates a lack of exploration during training.
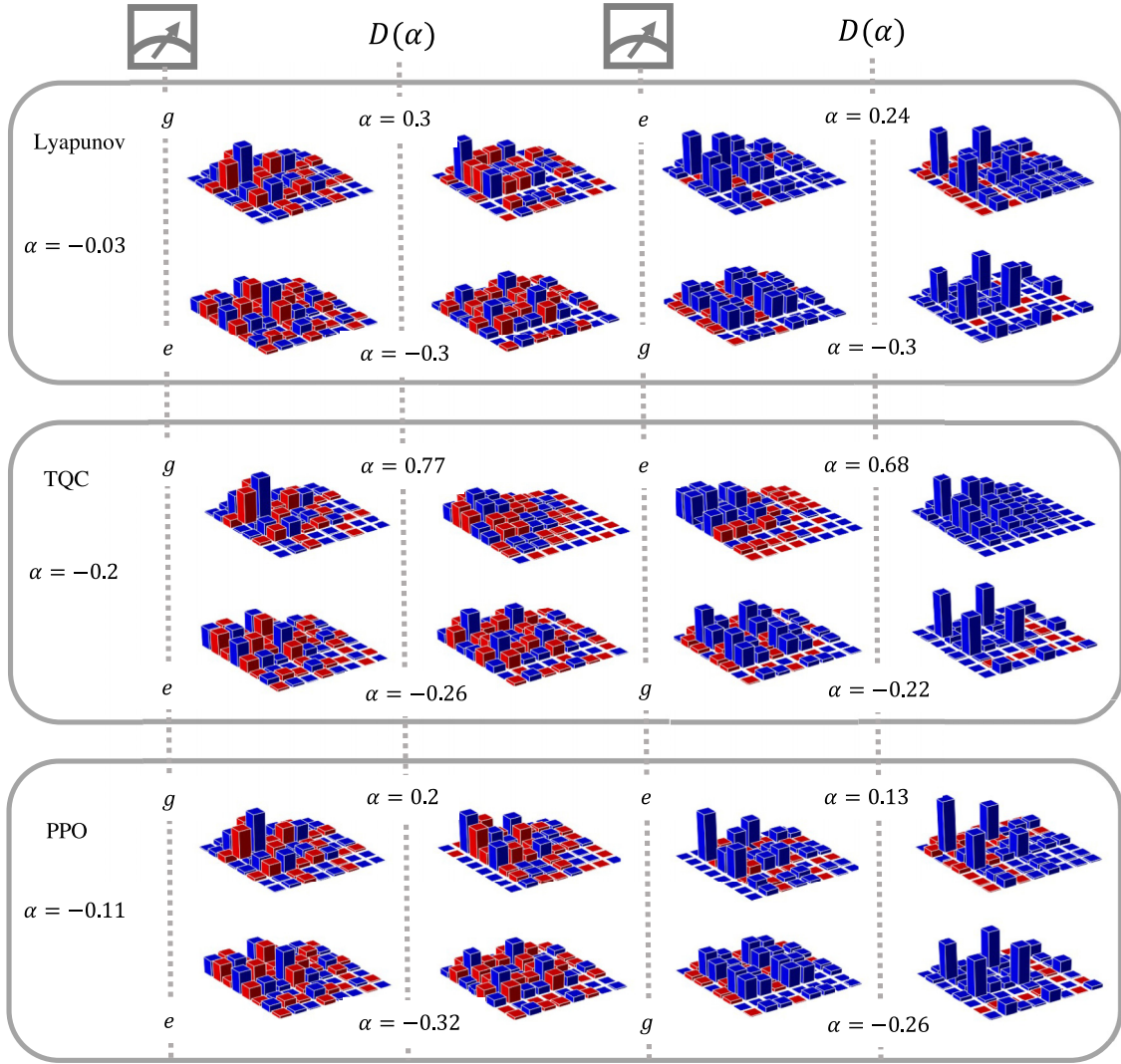
FIG. 11. First feedback cycles for the cavity initialized as a coherent state, showing how the cavity states evolve at each step of measurements and displacements. For each agent, two different trajectories are shown, consisting of a sequence of *g-e* or *e-g* measurement outcomes (respectively, top and bottom row). The α value at the left before the first measurement corresponds to the first control, adjusting the initial α_{guess} value. While the bottom row is similar for all agents, it is seen in the top row that the TQC agent opts for a drastically different strategy by performing a large displacement to reset the cavity in a state similar to a coherent state.

will, however, allow reaching higher fidelities later on, and significantly more so for the TQC agent. Although the first few time steps have lower fidelities compared to the other two approaches, the TQC agent is nevertheless able to reach high-fidelity states early on in the control sequence.

2. First feedback cycles

The aforementioned behavior of the first few feedback cycles is depicted in Fig. 11, where the evolution of the cavity density matrix is shown for all approaches. In situations where the current state is far from the target, a Newton-based line-search method, such as that used in the Lyapunov control approach implemented here (see Appendix B), will not optimally determine the displacement needed to be applied to the cavity state. This is a consequence of the locally convergent behavior of line-search methods. Indeed, it was found that the Lyapunov control was constantly selecting large

displacements in such cases, which prevented further convergence to high-fidelity states. This required us to limit the maximum allowed displacement for the Lyapunov control. This was here done by optimizing over a range of candidate maximum amplitudes, and choosing the one maximizing the median fidelity to the target state. The $\alpha = \pm 0.3$ displacement performed by the Lyapunov control approach shown in Fig. 11 is the result of such optimization. Compared to the RL methods, this is a serious drawback as the choice of the large displacement to be applied early in the control sequence is crucial to adequately balance the amplitudes of the resulting state after the control sequence, and should ideally be conditioned on a specific state, rather than be optimized over all trajectories.

For RL agents, it is indeed found that they select displacements of similar amplitudes applied to the cavity after the first excited state qubit measurement, but which is better adapted as it takes into account the impact it may have in

future feedback cycles. This is even more striking in the case following a ground-state measurement at the beginning of the sequence. Here, the TQC agent performs a large displacement of amplitude 0.77, which effectively does an operation akin to a state reset. On a short time scale, this is penalizing as shown in the states that follow in the sequence, which are farther from the target state than for the other approaches. However, as seen in the top row of Fig. 10, this opens the way to states with high fidelity, and indeed higher fidelities are reached than with the other approaches. The TQC agent is thus better at exploiting the effect of the measurement backaction to reach its target state.

One can also notice the larger initial displacement, in absolute value, taken by the RL agents, correcting the initial α_{guess} value. RL agents are able to infer, simply by retropropagating future fidelities observed, that the initial state, even though it is one with the best overlap with the target state, is not the optimal one when taking into account the measurement dynamics, and when considering the global objective to be reached on a longer horizon.

3. *Out-of-subspace initial state*

Finally, the bottom row of Fig. 10 shows results of a sequence initialized with the state $|\Psi\rangle = (|0\rangle + |3\rangle)/\sqrt{2}$, which has amplitudes similar to those of the target state, but which is in another subspace. This case is perhaps the most interesting, as the different approaches show drastically different behaviors. PPO is not able to learn how to leave the subspace, as shown by the higher probabilities for trajectories that are associated to low-fidelity states. Indeed, it was found that it only performs small displacements of about $\alpha = 0.03$, which are not sufficient to transfer enough population to the target subspace, with the backaction simply annihilating all target subspace populations at every step. In this situation, Lyapunov control performs better. In this case, it applies the maximum allowed displacement $\alpha = 0.3$ at the beginning (this cannot be seen from the figure; it is the value given by the algorithm), which allows it to eventually reach the target subspace. The TQC agent, however, appears to find the best policy ($\alpha = 0.54$ is applied at the first cycle of feedback, value from the algorithm), transferring 40% of subsequent trajectories towards high-fidelity values. It is also able to learn how to bring a large portion of the remaining 60% towards higher fidelities. In fact, as the TQC policy evolved during training, it was seen to opt for a tradeoff. In early stages, it was performing a larger initial displacement, which favored a higher probability for a measurement to bring it back to the target subspace. However, it did so at the expense of the remaining trajectories, which never reached the target state. As training evolved, it lowered the initial displacement, so as to still handle the remaining states resulting from the other measurement outcomes. It should also be emphasized that during training, the agent was never initialized in the $(|0\rangle + |3\rangle)/\sqrt{2}$ state, but only in the coherent state as mentioned above. As such, the policy found by the TQC agent is a result of its higher exploration and generalization capabilities.

This analysis shows that the simple extension of the task of preparing a single Fock state to that of preparing superpositions translates into a qualitatively different control

problem. Whereas the former only needs the state to be steered towards the desired eigenstate of suitably chosen measurement operators, the latter requires a complex interaction between measurement outcomes and displacement operations. In some cases, even something akin to a state reset is necessary in order to maximize the achievable fidelity.

It should also be noted that while an on-policy was compared to an off-policy method here, it is still unclear what is the role of adding a distributional approach on top of a soft-actor critic algorithm. Nevertheless, compared to the standard soft-actor critic implementation [47], it was found in the present work that the distributional version is more robust to hyperparameters tuning and more stable during training. More work would, however, be needed, such as ablation studies, to better understand the role of the distributional procedure in the quantum setting.

V. CONCLUSION

In this work, a measurement-based quantum-feedback protocol was proposed and analyzed to prepare and stabilize superpositions of Fock states in a superconducting cavity. By using a generalization of parity measurements, states in a target subspace with Fock basis states with equally spaced number of photons can be stabilized, but also prepared using a coherent drive as the only control. It was shown that a classical control technique such as Lyapunov-function-based control, which was previously developed for the stabilization of Fock states [16,17], fail to prepare superposition due to their lack of exploitation of the measurement backaction.

Here, using an RL method proved useful in overcoming this limitation. Indeed, by exploring the optimal policies learned by different algorithmic methods, the measurement backaction could be emphasized as a useful resource in itself to create the nonlinearities required to prepare quantum state superpositions. It also highlighted the interaction between control actions with weak measurements in a way that takes into account the impact of the control actions later in the feedback cycle, in order to reach a target state with high fidelity. Because such measurements are the same as those used in some error-correction procedures, our proposed protocol could easily be integrated in the bosonic computation paradigm.

From an RL point of view, the ability of TQC to reuse past experiences, and thus learn a more general policy than, for instance, PPO-like algorithms, might prove useful as future quantum control experiments scale up in complexity. We believe that further exploration of these different behaviors already noticed in the low-complexity settings considered here would be an interesting and potentially fruitful avenue of investigation.

TABLE I. Hyperparameters for TQC.

Hyperparameter	Value
Number of Layers	2
Actor Neurons per Layer	256
Critics Neurons per Layer	512
Discount (γ)	0.95
Batch size	1024
Activation Function	tanh
Entropy coefficient	0.09
Number of critics	5
Learning Rate	0.0001
Target update interval (τ)	0.001

APPENDIX A: NEURAL NETWORKS HYPERPARAMETERS

Tables I and II give the hyperparameter values used for both the TQC and PPO agents. All others hyperparameters not presented here are those by default in the STABLE-BASELINES3 implementation.

APPENDIX B: DERIVATIONS FOR LYAPUNOV-FUNCTION-BASED CONTROL

In Lyapunov-function-based control, a positive-definite function V of the state ρ is used, so that the minimum of V, which is necessarily zero owing to positive definiteness, is reached for a targeted state ρ^{target}. Control actions are performed iteratively in a feedback loop, whereby the goal of each iteration's control action on ρ is to reduce the value of V, ultimately reaching the minimum, hence the targeted state [26]. The essential details of this approach in the present specific context will now be provided.

1. Positive-definite and Lyapunov functions and their significance in control

Given two arbitrary states ρ_1 and ρ_2, a positive-definite function $d(\rho_1, \rho_2)$ will be considered in the sequel, that is with the following properties:

$$d(\rho_1, \rho_2) \begin{cases} > 0 & \text{for } \rho_1 \neq \rho_2, \\ = 0 & \Leftrightarrow \rho_1 = \rho_2. \end{cases} \qquad \text{(B1)}$$

This is akin to a distance function, but it is not required here that d obeys the triangle inequality (which is one of the axioms that a distance function must satisfy). Given such a function

TABLE II. Hyperparameters for PPO.

Hyperparameter	Value
Number of Layers	2
Neurons per Layer	256
Discount (γ)	0.95
Number of steps	2048
Batch size	256
Activation Function	tanh
Learning Rate	0.0001

and a targeted state ρ^{target}, one can in turn define a positive-definite function over the set of states ρ by

$$V^{\text{target}}(\rho) = d(\rho^{\text{target}}, \rho). \qquad \text{(B2)}$$

The significance of such a function in the present context is that if α in Eq. (5) is chosen so that

$$V^{\text{target}}(\rho') < V^{\text{target}}(\rho), \qquad \text{(B3)}$$

then by performing a series of steps $k = 1, 2, \ldots$, whereby at each step this inequality is satisfied, i.e., $V^{\text{target}}(\rho_{k+1}) < V^{\text{target}}(\rho_k)$, then the state will converge to the targeted state since the value of $V^{\text{target}}(\rho_k)$ will eventually reach zero, and, by hypothesis, for $V^{\text{target}}(\rho_k)$ to equal zero the only possibility is that $\rho_k = \rho^{\text{target}}$. Such a function decreasing over the evolution of a system is known in the control literature as a Lyapunov function [71].

2. Fidelity-based positive-definite and Lyapunov functions

One of the simplest positive-definite functions that can be considered in quantum mechanics is based on the Frobenius scalar product between two operators, which for density operators is given by

$$\text{Fr}(\rho_1, \rho_2) = \text{tr}(\rho_1 \rho_2), \qquad \text{(B4)}$$

which for pure states $\rho_1 = |\psi_1\rangle\langle\psi_1|$ and $\rho_2 = |\psi_2\rangle\langle\psi_2|$ amounts to $\text{Fr}(\rho_1, \rho_2) = |\langle\psi_1|\psi_2\rangle|^2$. The Frobenius scalar product is also sometimes simply called the fidelity [16] [note that this fidelity is different from that defined in Eq. (18)]. Since $0 \leqslant \text{Fr}(\rho_1, \rho_2) \leqslant 1$ with $\text{Fr}(\rho_1, \rho_2) = 1 \Leftrightarrow \rho_1 = \rho_2$ and ρ_1 is pure, one can define the positive-definite function

$$d_{\text{Fr}}(\rho_1, \rho_2) = 1 - \text{Fr}(\rho_1, \rho_2), \qquad \text{(B5)}$$

which satisfies the conditions given in Eq. (B1). This will be called the Frobenius distance, although it does not formally satisfies all the axioms of a distance. This leads to the positive-definite function defined over states

$$V_{Fr}^{\text{target}}(\rho) = 1 - \text{Fr}(\rho^{\text{target}}, \rho) = 1 - \text{tr}(\rho^{\text{target}}\rho). \qquad \text{(B6)}$$

Since $\text{tr}\rho = 1$, this can be written as

$$V_{\text{Fr}}^{\text{target}}(\rho) = \text{tr}((\mathbb{I} - \rho^{\text{target}})\rho) = \text{tr}(\Upsilon^{\text{target}}\rho), \qquad \text{(B7)}$$

where $\mathbb{I}$ is the identity operator, and $\Upsilon^{\text{target}} = \mathbb{I} - \rho^{\text{target}}$. Here the superscript "target" on Υ reminds that Υ depends on the targeted state. Such a Lyapunov function will be called a fidelity Lyapunov function, and will be denoted $V_{\Upsilon}^{\text{target}}(\rho)$ in the sequel, hence

$$V_{\Upsilon}^{\text{target}}(\rho) = \text{tr}(\Upsilon^{\text{target}}\rho). \qquad \text{(B8)}$$

3. Second-order expansion of the Lyapunov function with actuator action

It will now be assumed that a Lyapunov function given in the generalized form of Eq. (B8) is defined. The objective of the feedback law can now be stated by requiring that, for

a given state ρ, α must be chosen such that $V_\Upsilon^{\text{target}}(\rho') < V_\Upsilon^{\text{target}}(\rho)$, that is

$$V_\Upsilon^{\text{target}}[D(\alpha)\rho\,D(-\alpha)] < V_\Upsilon^{\text{target}}(\rho). \quad \text{(B9)}$$

Obtaining such a condition is similar to a line search in numerical minimization [72] and a common way is to develop to second order in α the left-hand side of this inequality, and use the approximation thus obtained to find a value of α that will allow satisfying this condition. To do this, Eq. (4) is first expanded to second order in α by resorting to an expansion of $D(\alpha)$ to second order. This leads to

$$\rho' = D(\alpha)\rho\,D(-\alpha)$$
$$\approx \rho + [\alpha a^\dagger - \alpha^* a, \rho] + \frac{1}{2}[[\rho, \alpha a^\dagger - \alpha^* a], \alpha a^\dagger - \alpha^* a]. \quad \text{(B10)}$$

With this, $V_\Upsilon^{\text{target}}(\rho')$ can be approximated to second order by

$$V_\Upsilon^{\text{target}}(\rho') = \text{tr}(\Upsilon^{\text{target}}\rho')$$
$$\approx V_\Upsilon^{\text{target}}(\rho) + T^{(1)}(\alpha) + \frac{1}{2}T^{(2)}(\alpha), \quad \text{(B11)}$$

where $T^{(1)}(\alpha)$ is the first-order term given by

$$T^{(1)}(\alpha) = \text{tr}(\Upsilon^{\text{target}}[\alpha a^\dagger - \alpha^* a, \rho]), \quad \text{(B12)}$$

and $T^{(2)}(\alpha)$ is the second-order term given by

$$T^{(2)}(\alpha) = \text{tr}(\Upsilon^{\text{target}}[[\rho, \alpha a^\dagger - \alpha^* a], \alpha a^\dagger - \alpha^* a]). \quad \text{(B13)}$$

By expanding the commutator and reordering terms, the first-order term can be rewritten as

$$T^{(1)}(\alpha) = -\text{tr}([\alpha a^\dagger - \alpha^* a, \Upsilon^{\text{target}}]\rho). \quad \text{(B14)}$$

This form is more convenient since it leads to commutators that can be precomputed. Indeed, so expressed, the first-order term can be further expanded as

$$T^{(1)}(\alpha) = -\alpha\,\text{tr}([a^\dagger, \Upsilon^{\text{target}}]\rho) + \alpha^*\,\text{tr}([a, \Upsilon^{\text{target}}]\rho). \quad \text{(B15)}$$

Defining

$$B = [a, \Upsilon^{\text{target}}]\rho, \quad \text{(B16)}$$

then

$$B^\dagger = \rho^\dagger[a, \Upsilon^{\text{target}}]^\dagger = -\rho[a^\dagger, \Upsilon^{\text{target}}], \quad \text{(B17)}$$

hence

$$\text{tr}(B^\dagger) = -\text{tr}(\rho[a^\dagger, \Upsilon^{\text{target}}]) = -\text{tr}([a^\dagger, \Upsilon^{\text{target}}]\rho). \quad \text{(B18)}$$

Now, since $\text{tr}(B^\dagger) = (\text{tr}B)^*$, then

$$\text{tr}([a^\dagger, \Upsilon^{\text{target}}]\rho) = -[\text{tr}(B)]^*. \quad \text{(B19)}$$

With these developments, and setting

$$\zeta = \text{tr}B = \text{tr}([a, \Upsilon^{\text{target}}]\rho), \quad \text{(B20)}$$

$T^{(1)}(\alpha)$ can simply be rewritten as

$$T^{(1)}(\alpha) = \alpha\zeta^* + \alpha^*\zeta. \quad \text{(B21)}$$

It is convenient for the sequel to define the commutator

$$C^{\Upsilon^{\text{target}}} = [a, \Upsilon^{\text{target}}], \quad \text{(B22)}$$

which can be precomputed; with this

$$B = C^{\Upsilon^{\text{target}}}\rho \quad \text{(B23)}$$

and

$$\zeta = \text{tr}(C^{\Upsilon^{\text{target}}}\rho). \quad \text{(B24)}$$

Now, by similar reasoning as for $T^{(1)}(\alpha)$, $T^{(2)}(\alpha)$ can be rewritten as

$$T^{(2)}(\alpha) = \text{tr}([\alpha a^\dagger - \alpha^* a, [\alpha a^\dagger - \alpha^* a, \Upsilon^{\text{target}}]]\rho)$$
$$= \text{tr}(K\rho), \quad \text{(B25)}$$

where

$$K = [\alpha a^\dagger - \alpha^* a, [\alpha a^\dagger - \alpha^* a, \Upsilon^{\text{target}}]]. \quad \text{(B26)}$$

Developing this operator, and using Jacobi's identity $[A, [B, C]] + [C, [A, B]] + [B, [C, A]] = 0$ along with $[a, a^\dagger] = \mathbb{I}$, one obtains

$$K = \alpha^2[a^\dagger, [a^\dagger, \Upsilon^{\text{target}}]] + \alpha^{*2}[a, [a, \Upsilon^{\text{target}}]]$$
$$-2|\alpha|^2[a^\dagger, [a, \Upsilon^{\text{target}}]]. \quad \text{(B27)}$$

Setting

$$G^{\Upsilon^{\text{target}}} = [a, [a, \Upsilon^{\text{target}}]] = [a, C^{\Upsilon^{\text{target}}}], \quad \text{(B28)}$$

$$E^{\Upsilon^{\text{target}}} = [a^\dagger, [a, \Upsilon^{\text{target}}]] = [a^\dagger, C^{\Upsilon^{\text{target}}}], \quad \text{(B29)}$$

and using that

$$(G^{\Upsilon^{\text{target}}})^\dagger = [a^\dagger, [a^\dagger, \Upsilon^{\text{target}}]] \quad \text{(B30)}$$

leads to

$$K = \alpha^2(G^{\Upsilon^{\text{target}}})^\dagger + \alpha^{*2}G^{\Upsilon^{\text{target}}} - 2|\alpha|^2E^{\Upsilon^{\text{target}}}, \quad \text{(B31)}$$

and hence

$$T^{(2)}(\alpha) = \text{tr}(K\rho)$$
$$= \alpha^2\text{tr}((G^{\Upsilon^{\text{target}}})^\dagger\rho) + \alpha^{*2}\text{tr}(G^{\Upsilon^{\text{target}}}\rho)$$
$$-2|\alpha|^2\text{tr}(E^{\Upsilon^{\text{target}}}\rho). \quad \text{(B32)}$$

It can be shown that $T^{(2)}(\alpha)$ is a real quantity (the development will not be provided here).

$$\text{tr}((G^{\Upsilon^{\text{target}}})^\dagger\rho) = \text{tr}((G^{\Upsilon^{\text{target}}})^\dagger\rho^\dagger)$$
$$= \text{tr}((\rho G^{\Upsilon^{\text{target}}})^\dagger)$$
$$= \text{tr}((\rho G^{\Upsilon^{\text{target}}}))^*.$$

Defining

$$\gamma = \text{tr}(G^{\Upsilon^{\text{target}}}\rho), \quad \text{(B33)}$$

and

$$\chi = \text{tr}(E^{\Upsilon^{\text{target}}}\rho), \quad \text{(B34)}$$

where χ can be demonstrated to be real, $T^{(2)}(\alpha)$ can be written as

$$T^{(2)}(\alpha) = \alpha^2\gamma^* + \alpha^{*2}\gamma - 2|\alpha|^2\chi. \quad \text{(B35)}$$

With the previous developments, the second-order expansion of the Lyapunov function can be rewritten as [refer back

to Eq. (B11)]

$$V_\Upsilon^{\text{target}}(\rho') \approx V_\Upsilon^{\text{target}}(\rho) + \alpha\zeta^* + \alpha^*\zeta$$
$$+ \frac{1}{2}(\alpha^2\gamma^* + \alpha^{*2}\gamma - 2|\alpha|^2\chi)$$
$$= V_\Upsilon^{\text{target}}(\rho) + q(\alpha), \tag{B36}$$

with $q(\alpha)$ being the following quadratic form:

$$q(\alpha) = \alpha\zeta^* + \alpha^*\zeta + \frac{1}{2}(\alpha^2\gamma^* + \alpha^{*2}\gamma - 2|\alpha|^2\chi). \tag{B37}$$

Recall that it is required to determine α according to the inequality given in Eq. (B9), which means that $q(\alpha)$ must be negative, i.e.,

$$q(\alpha) < 0, \tag{B38}$$

and ideally $q(\alpha)$ shall be as negatively large as possible.

There are different ways in which $q(\alpha) < 0$. One standard possibility is equivalent to gradient steepest descent, and a second one is equivalent to a Newton method in numerical optimization, which is resorted to, as it has faster convergence properties.

4. Newton method

The quadratic form $q(\alpha)$ is first be represented in terms of real quantities. First, $q(\alpha)$ is written as follows:

$$q(\alpha) = 2\operatorname{Re}\alpha\zeta^* + \operatorname{Re}\alpha^2\gamma^* - |\alpha|^2\chi. \tag{B39}$$

The complex quantities appearing in $q(\alpha)$ are written as

$$\alpha = x + iy, \tag{B40}$$

$$\zeta = u + iv, \tag{B41}$$

$$\gamma = g + ih. \tag{B42}$$

This allows writing the quadratic form as

$$q(\alpha) \equiv q(x, y) = \begin{bmatrix} x & y \end{bmatrix} Q \begin{bmatrix} x \\ y \end{bmatrix} + 2L \begin{bmatrix} x \\ y \end{bmatrix}, \tag{B43}$$

with

$$Q = \begin{bmatrix} g - \chi & h \\ h & -(g + \chi) \end{bmatrix}, \quad L = \begin{bmatrix} u & v \end{bmatrix}. \tag{B44}$$

The Newton approach in optimization is to take the direction $[x \quad y]^{\text{T}}$, which minimizes the quadratic form. This leads to

$$Q \begin{bmatrix} x \\ y \end{bmatrix} = -L^{\text{T}},$$

which is equivalent to

$$\begin{bmatrix} x \\ y \end{bmatrix} = -Q^{-1}L^{\text{T}}. \tag{B45}$$

The inverse of Q is easily obtained and given by

$$Q^{-1} = \frac{1}{(g^2 + h^2 - \chi^2)} \begin{bmatrix} g + \chi & h \\ h & -(g - \chi) \end{bmatrix}. \tag{B46}$$

Reexpressing everything in term of the complex quantities ζ, γ, and χ gives

$$\begin{bmatrix} x \\ y \end{bmatrix} = -Q^{-1}L^{\text{T}} = \frac{1}{\chi^2 - |\gamma|^2} \begin{bmatrix} \operatorname{Re}\{\chi\zeta + \gamma\zeta^*\} \\ \operatorname{Im}\{\chi\zeta + \gamma\zeta^*\} \end{bmatrix}, \tag{B47}$$

hence

$$\alpha = \frac{1}{\chi^2 - |\gamma|^2}(\chi\zeta + \gamma\zeta^*). \tag{B48}$$

Equation (B48) is therefore the one defining the update rule for the control amplitude parameter in the feedback procedure. For Newton's method, it is necessary that the matrix Q be positive definite [72]. This is true if both eigenvalues of Q are positive. These eigenvalues are easily found to be

$$\lambda_\pm = -\chi \pm |\gamma|. \tag{B49}$$

Hence, for Q to be positive definite the following must hold:

$$\chi < -|\gamma|. \tag{B50}$$

This condition is numerically verified. For the values of x and y given through Eq. (B45), the value of $q(x, y)$ is found to be

$$q(x, y) = -L\, Q^{-1}L^{\text{T}}, \tag{B51}$$

which is negative whenever Q is positive definite, since in this case Q^{-1} is also positive definite.

---

[1] A. Joshi, K. Noh, and Y. Y. Gao, Quantum information processing with bosonic qubits in circuit QED, Quantum Sci. Technol. **6**, 033001 (2021).

[2] W.-L. Ma, S. Puri, R. J. Schoelkopf, M. H. Devoret, S. Girvin, and L. Jiang, Quantum control of bosonic modes with superconducting circuits, Sci. Bull. **66**, 1789 (2021).

[3] W. Cai, Y. Ma, W. Wang, C.-L. Zou, and L. Sun, Bosonic quantum error correction codes in superconducting quantum circuits, Fund. Res. **1**, 50 (2021).

[4] M. H. Michael, M. Silveri, R. T. Brierley, V. V. Albert, J. Salmilehto, L. Jiang, and S. M. Girvin, New class of quantum error-correcting codes for a bosonic mode, Phys. Rev. X **6**, 031006 (2016).

[5] A. L. Grimsmo, J. Combes, and B. Q. Baragiola, Quantum computing with rotation-symmetric bosonic codes, Phys. Rev. X **10**, 011058 (2020).

[6] A. Grimm, N. E. Frattini, S. Puri, S. O. Mundhada, S. Touzard, M. Mirrahimi, S. M. Girvin, S. Shankar, and M. H. Devoret, Stabilization and operation of a Kerr-cat qubit, Nature (London) **584**, 205 (2020).

[7] Z.-Y. Zhou, C. Gneiting, W. Qin, J. Q. You, and F. Nori, Enhancing dissipative cat-state generation via nonequilibrium pump fields, Phys. Rev. A **106**, 023714 (2022).

[8] M. Mirrahimi, Z. Leghtas, V. V. Albert, S. Touzard, R. J. Schoelkopf, L. Jiang, and M. H. Devoret, Dynamically

protected cat-qubits: A new paradigm for universal quantum computation, New J. Phys. **16**, 045014 (2014).

[9] C. K. Law and J. H. Eberly, Arbitrary control of a quantum electromagnetic field, Phys. Rev. Lett. **76**, 1055 (1996).

[10] M. Hofheinz, E. M. Weig, M. Ansmann, R. C. Bialczak, E. Lucero, M. Neeley, A. D. O'Connell, H. Wang, J. M. Martinis, and A. N. Cleland, Generation of fock states in a superconducting quantum circuit, Nature (London) **454**, 310 (2008).

[11] M. Hofheinz, H. Wang, M. Ansmann, R. C. Bialczak, E. Lucero, M. Neeley, A. D. O'Connell, D. Sank, J. Wenner, J. M. Martinis, and A. N. Cleland, Synthesizing arbitrary quantum states in a superconducting resonator, Nature (London) **459**, 546 (2009).

[12] S. Krastanov, V. V. Albert, C. Shen, C.-L. Zou, R. W. Heeres, B. Vlastakis, R. J. Schoelkopf, and L. Jiang, Universal control of an oscillator with dispersive coupling to a qubit, Phys. Rev. A **92**, 040303(R) (2015).

[13] R. W. Heeres, B. Vlastakis, E. Holland, S. Krastanov, V. V. Albert, L. Frunzio, L. Jiang, and R. J. Schoelkopf, Cavity state manipulation using photon-number selective phase gates, Phys. Rev. Lett. **115**, 137002 (2015).

[14] R. W. Heeres, P. Reinhold, N. Ofek, L. Frunzio, L. Jiang, M. H. Devoret, and R. J. Schoelkopf, Implementing a universal gate set on a logical qubit encoded in an oscillator, Nature Commun. **8**, 94 (2017).

[15] T. Fösel, S. Krastanov, F. Marquardt, and L. Jiang, Efficient cavity control with SNAP gates, arXiv:2004.14256 [quant-ph].

[16] I. Dotsenko, M. Mirrahimi, M. Brune, S. Haroche, J.-M. Raimond, and P. Rouchon, Quantum feedback by discrete quantum nondemolition measurements: Towards on-demand generation of photon-number states, Phys. Rev. A **80**, 013805 (2009).

[17] C. Sayrin, I. Dotsenko, X. Zhou, B. Peaudecerf, T. Rybarczyk, S. Gleyzes, P. Rouchon, M. Mirrahimi, H. Amini, M. Brune, J.-M. Raimond, and S. Haroche, Real-time quantum feedback prepares and stabilizes photon number states, Nature (London) **477**, 73 (2011).

[18] A. Aarab, R. Azouit, V. Reiher, and Y. Bérubé-Lauzière, State initialization of a hot spin qubit in a double quantum dot by measurement-based quantum feedback control, Phys. Rev. B **106**, 235309 (2022).

[19] R. Porotti, A. Essig, B. Huard, and F. Marquardt, Deep reinforcement learning for quantum state preparation with weak nonlinear measurements, Quantum **6**, 747 (2022).

[20] 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 for real-time quantum feedback, Nat. Commun. **14**, 7138 (2023).

[21] A. Essig, Q. Ficheux, T. Peronnin, N. Cottet, R. Lescanne, A. Sarlette, P. Rouchon, Z. Leghtas, and B. Huard, Multiplexed photon number measurement, Phys. Rev. X **11**, 031045 (2021).

[22] A. Blais, R.-S. Huang, A. Wallraff, S. M. Girvin, and R. J. Schoelkopf, Cavity quantum electrodynamics for superconducting electrical circuits: An architecture for quantum computation, Phys. Rev. A **69**, 062320 (2004).

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

[24] A. Wallraff, D. I. Schuster, A. Blais, L. Frunzio, R. S. Huang, J. Majer, S. Kumar, S. M. Girvin, and R. J. Schoelkopf, Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics, Nature (London) **431**, 162 (2004).

[25] H. Amini, R. A. Somaraju, I. Dotsenko, C. Sayrin, M. Mirrahimi, and P. Rouchon, Feedback stabilization of discrete-time quantum systems subject to non-demolition measurements with imperfections and delays, Automatica **49**, 2683 (2013).

[26] M. Mirrahimi and R. Van Handel, Stabilizing feedback controls for quantum systems, SIAM J. Control Optim. **46**, 445 (2007).

[27] R. van Handel, J. Stockton, and H. Mabuchi, Feedback control of quantum state reduction, IEEE Transactions on Automatic Control **50**, 768 (2005).

[28] Y. Liu, S. Kuang, and S. Cong, Lyapunov-based feedback preparation of GHZ entanglement of N qubit systems, IEEE Transactions on Cybernetics **47**, 3827 (2017).

[29] S. Cong and F. Meng, A survey of quantum Lyapunov control methods, Sci. World J. **2013**, 967529 (2013).

[30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An introduction* (MIT Press, Cambridge, 2018).

[31] M. Krenn, J. Landgraf, T. Foesel, and F. Marquardt, Artificial intelligence and machine learning for quantum technologies, Phys. Rev. A **107**, 010101 (2023).

[32] J. Barry, D. T. Barry, and S. Aaronson, Quantum partially observable markov decision processes, Phys. Rev. A **90**, 032311 (2014).

[33] Y. Baum, M. Amico, S. Howell, M. Hush, M. Liuzzi, P. Mundada, T. Merkh, A. R. R. Carvalho, and M. J. Biercuk, Experimental deep reinforcement learning for error-robust gate-set design on a superconducting quantum computer, PRX Quantum **2**, 040324 (2021).

[34] 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**, 01LT02 (2021).

[35] S. Z. Baba, N. Yoshioka, Y. Ashida, and T. Sagawa, Deep reinforcement learning for preparation of thermal and prethermal quantum states, Phys. Rev. Appl. **19**, 014068 (2023).

[36] 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**, 85 (2019).

[37] 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).

[38] 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, Nature (London) **616**, 50 (2023).

[39] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, Universal quantum control through deep reinforcement learning, npj Quantum Inf. **5**, 33 (2019).

[40] H. Ma, D. Dong, S. X. Ding, and C. Chen, Curriculum-based deep reinforcement learning for quantum control, IEEE Trans. Neural Net. Learn. Syst. **34**, 8852 (2022).

[41] Z. An, H.-J. Song, Q.-K. He, and D. L. Zhou, Quantum optimal control of multilevel dissipative quantum systems with reinforcement learning, Phys. Rev. A **103**, 012404 (2021).

[42] P. Peng, X. Huang, C. Yin, L. Joseph, C. Ramanathan, and P. Cappellaro, Deep reinforcement learning for quantum Hamiltonian engineering, Phys. Rev. Appl. **18**, 024033 (2022).

[43] J.-J. Chen and M. Xue, Manipulation of spin dynamics by deep reinforcement learning agent, arXiv:1901.08748 [quant-ph].

[44] M.-Z. Ai, Y. Ding, Y. Ban, J. D. Martín-Guerrero, J. Casanova, J.-M. Cui, Y.-F. Huang, X. Chen, C.-F. Li, and G.-C. Guo, Experimentally realizing efficient quantum control with reinforcement learning, Sci. China Phys. Mech. Astron. **65**, 250312 (2022).

[45] M. August and J. M. Hernández-Lobato, Taking gradients through experiments: LSTMs and memory proximal policy optimization for Black-Box quantum control, in *High Performance Computing* (Springer, Berlin, 2018), pp. 591–613.

[46] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, Continuous control with deep reinforcement learning, arXiv:1509.02971.

[47] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor, in *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 80, edited by J. Dy and A. Krause (PMLR, New York, 2018), pp. 1861–1870.

[48] A. Kuznetsov, P. Shvechikov, A. Grishin, and D. Vetrov, Controlling overestimation bias with truncated mixture of continuous distributional quantile critics, in *International Conference on Machine Learning* (PMLR, New York, 2020), pp. 5556–5566.

[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] A. Fallani, M. A. C. Rossi, D. Tamascelli, and M. G. Genoni, Learning feedback control strategies for quantum metrology, PRX Quantum **3**, 020310 (2022).

[51] Y. Ding, X. Chen, R. Magdalena-Benedito, and J. D. Martín-Guerrero, Closed-loop control of a noisy qubit with reinforcement learning, Mach. Learn.: Sci. Technol. **4**, 025020 (2023).

[52] V. V. Sivak, A. Eickbusch, H. Liu, B. Royer, I. Tsioutsios, and M. H. Devoret, Model-free quantum control with reinforcement learning, Phys. Rev. X **12**, 011059 (2022).

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

[54] 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).

[55] J. Mackeprang, D. B. R. Dasari, and J. Wrachtrup, A reinforcement learning approach for quantum state engineering, Quant. Mach. Intel. **2**, 5 (2020).

[56] D. Dong, C. Chen, T.-J. Tarn, A. Pechen, and H. Rabitz, Incoherent control of quantum systems with wavefunction-controllable subspaces via quantum reinforcement learning, IEEE Trans. Syst. Man Cybernet. B **38**, 957 (2008).

[57] S. Haroche and J.-M. Raimond, *Exploring the Quantum: Atoms, Cavities, and Photons* (Oxford University Press, Oxford, 2006).

[58] L. Sun, A. Petrenko, Z. Leghtas, B. Vlastakis, G. Kirchmair, K. M. Sliwa, A. Narla, M. Hatridge, S. Shankar, J. Blumoff, L. Frunzio, M. Mirrahimi, M. H. Devoret, and R. J. Schoelkopf, Tracking photon jumps with repeated quantum non-demolition parity measurements, Nature (London) **511**, 444 (2014).

[59] S. Thrun and A. Schwartz, Issues in using function approximation for reinforcement learning, in *Proceedings of the 1993 Connectionist Models Summer School* (Erlbaum Associates, 1993).

[60] M. G. Bellemare, W. Dabney, and R. Munos, A distributional perspective on reinforcement learning, in *Proceedings of the 34th International Conference on Machine Learning*, Proceedings of Machine Learning Research, Vol. 70, edited by D. Precup and Y. W. Teh (PMLR, New York, 2017), pp. 449–458.

[61] M. G. Bellemare, W. Dabney, and M. Rowland, *Distributional Reinforcement Learning* (MIT Press, Cambridge, 2023).

[62] C. Lyle, M. G. Bellemare, and P. S. Castro, A comparative analysis of expected and distributional reinforcement learning, Proceedings of the AAAI Conference on Artificial Intelligence **33**, 4504 (2019).

[63] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, Stable-baselines3: Reliable reinforcement learning implementations, JMLR **22**, 1 (2021).

[64] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, Curriculum learning, in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09* (Association for Computing Machinery, New York, 2009), p. 41-48.

[65] Y. Liu, S. Shankar, N. Ofek, M. Hatridge, A. Narla, K. M. Sliwa, L. Frunzio, R. J. Schoelkopf, and M. H. Devoret, Comparing and combining measurement-based and driven-dissipative entanglement stabilization, Phys. Rev. X **6**, 011022 (2016).

[66] R. Porotti, V. Peano, and F. Marquardt, Gradient-ascent pulse engineering with feedback, PRX Quantum **4**, 030305 (2023).

[67] M. Reagor, W. Pfaff, C. Axline, R. W. Heeres, N. Ofek, K. Sliwa, E. Holland, C. Wang, J. Blumoff, K. Chou, M. J. Hatridge, L. Frunzio, M. H. Devoret, L. Jiang, and R. J. Schoelkopf, Quantum memory with millisecond coherence in circuit QED, Phys. Rev. B **94**, 014506 (2016).

[68] S. Rosenblum, P. Reinhold, M. Mirrahimi, L. Jiang, L. Frunzio, and R. J. Schoelkopf, Fault-tolerant detection of a quantum error, Science **361**, 266 (2018).

[69] S. Krinner, N. Lacroix, A. Remm, A. Di Paolo, E. Genois, C. Leroux, C. Hellings, S. Lazar, F. Swiadek, J. Herrmann, G. J. Norris, C. K. Andersen, M. Müller, A. Blais, C. Eichler, and A. Wallraff, Realizing repeated quantum error correction in a distance-three surface code, Nature (London) **605**, 669 (2022).

[70] P. Facchi and S. Pascazio, Quantum Zeno subspaces, Phys. Rev. Lett. **89**, 080401 (2002).

[71] H. Khalil, *Nonlinear Systems*, 3rd ed. (Prentice Hall, Hoboken, 2001).

[72] R. Fletcher, Newton-like methods, in *Practical Methods of Optimization* (John Wiley & Sons, New York, 2000), Chap. 3, pp. 44–79.