# Optimizing measurement-based cooling by reinforcement learning

Jia-shun Yan and Jun Jing*
*School of Physics, Zhejiang University, Hangzhou 310027, Zhejiang, China*

(Received 8 June 2022; accepted 19 September 2022; published 30 September 2022)

Conditional cooling-by-measurement holds a significant advantage over its unconditional (nonselective) counterpart in the average-population-reduction rate. However, it has a clear weakness with respect to the limited success probability of finding the detector in the measured state. In this work, we propose an optimized architecture to cool down a target resonator, which is initialized as a thermal state, using an interpolation of conditional and unconditional measurement strategies. An optimal measurement-interval $\tau_{\rm opt}^{u}$ for unconditional measurement is analytically derived, which is inversely proportional to the collective dominant Rabi frequency $\Omega_d$ as a function of the resonator's population in the end of the last round. A cooling algorithm under global optimization by reinforcement learning results in the maximum value for the cooperative cooling performance, an indicator to measure the comprehensive cooling efficiency for arbitrary cooling-by-measurement architecture. In particular, the average population of the target resonator under only 16 rounds of measurements can be reduced by four orders in magnitude with a success probability of about 30%.

DOI: 10.1103/PhysRevA.106.033124

## I. INTRODUCTION

Cooling mesoscopic and microscopic resonators down to their minimum-energy state is fundamental to observing a classical-quantum transition and exploiting the quantum advantage in nanoscience [1,2]. The ground-state preparation is a crucial and implicit step in quantum information processes, including but not limited to continuous-variable quantum computations [3–6], ultrahigh precision measurements [7,8], and quantum interface constructions [9]. Various strategies have been designed to reach an effective temperature as low as possible in the trapped atom and ion systems [10–12]. In atomic laser cooling, popular strategies consist of laser Doppler cooling [9,13,14], resolved-sideband cooling, and electromagnetically induced transparency (EIT) cooling [15,16].

Beyond the paradigms extracting system energy through dissipative channels based on blueshifted (anti-Stokes) sidebands, a versatile approach to cooling the mechanical states of motion is provided by the interaction with electromagnetic radiation or quantum measurement. Back-action-evading measurement techniques that can surpass the standard quantum limit have attracted enormous interest. Through the pulsed measurement process in optomechanics [17–21], they can dramatically change the mechanical thermal occupation with no initial cooling. A genuine quantum-mechanical cooling engine is proposed [22], whereby the fuel is the energy exchanged with an apparatus performing invasive quantum measurements.

Among these measurement-based techniques, quantum state engineering based on measurements on ancillary systems has been proposed recently in theory [23,24] and demonstrated in experiment [25]. Rather than directly detecting the target system, a net nonunitary propagator is realized by inserting projective measurements on the ground state of the detector system in between the joint unitary-evolution segments of target and detector. The induced postselection of the ground state of the target system (typically modeled as a resonator) reduces its high-energy distribution in the ensemble. In other words, the resonator is gradually steered by the outcomes of the conditional measurement (CM) to its ground state via dynamically filtering out its vibrational modes. Ranging from cooling the nonlinear mechanical resonators [26], cooling by one shot measurement [27], and expanding the cooling range by an external driving [28], to accelerating the cooling rate by optimized measurement intervals [29], an unexplored weakness of the CM strategies is their limited success probability inherited from the projective operation. The amount of experimental overhead increases unavoidably with more samples in the ensemble. In sharp contrast to CM, the unconditional measurement (UM) strategy is used to perform a nonselective and impulsive measurement in all the bases of the bare Hamiltonian of the detector at the end of each round of the joint evolution [30,31]. Using this strategy, we are more likely to realize a unit-success-probability cooling, but it suffers from a much slower cooling rate than CM, indicating a much higher number of measurements toward the ground-state cooling. To compromise between the cooling rate and the success probability, the interpolating configuration of conditional and unconditional measurements constitutes an optimization problem.

The integration of a small-scale quantum circuit with a classical optimizer, e.g., the neural network, provides a paradigm by designing a sequence of parametrized quantum operations that are well suited to implement robust and high-fidelity algorithms. Many reinforcement learning (RL) algorithms constructed by the neural network, which

*jingjun@zju.edu.cn

 

demonstrated remarkable capabilities in board games and video games [32–35], have substantiated a wide and timely interest in studying several areas of quantum physics [36], including quantum error correction [37,38], quantum simulation [39,40], and quantum state preparation [41–43], to name a few. The proximal policy optimization (PPO) algorithm, as a typical RL algorithm with a significant sample complexity, scalability, and robustness for hyperparameters, has proven to be a fruitful tool in quantum optimization control [44–46].

In this work, we propose a measurement-based cooling architecture as a hybrid sequence of UM and CM strategies. It involves a double optimization: for each step along the sequence, either UM or CM can be improved considerably by using a local optimized measurement interval, and for the global efficiency of the sequence, its arrangement can be separably optimized through reinforcement learning. In particular, in a typical measurement-based cooling model, i.e., the Jaynes-Cummings (JC) model, where a mechanical resonator (the target system) is coupled to a qubit (the detector system), conditional and unconditional measurements are alternatively performed to cool down the resonator to its ground state. A feedback scheme is triggered upon calling a CM to determine whether or not to launch the next round of evolution-and-measurement according to the measurement outcome. Analogous to the optimized measurement-interval obtained for CM [29], we derive analytically an optimized interval for UM. Then the free-evolution intervals between any neighboring measurements, either UM or CM, can be optimized for cooling. The global sequence of measurements or the implementing order of UM and CM can be further optimized with reinforcement learning. The optimizer is fed with the cooperative cooling performance, a function of the average population of the resonator, the success probability of the detector in the measured subspace, and the fidelity of the resonator in the ground state. Eventually, we find an optimal sequence holding an overwhelming advantage over all the others.

The rest of this work is structured as follows. We briefly revisit the general framework for the cooling protocols based on conditional and unconditional measurements in Secs. II A and II B, respectively. In Sec. II B, an analytical expression of the optimized measurement interval is obtained for UM. In Sec. III, we introduce the interpolation diagram for the cooling architecture based on these two measurements. On the definition of the cooperative cooling performance to comprehensively quantify various strategies, we present the optimized result through reinforcement learning. The PPO algorithm and the optimal-control procedure are provided in Appendixes A and B, respectively. The whole work is discussed and summarized in Sec. IV.

## II. CONDITIONAL AND UNCONDITIONAL MEASUREMENTS

### A. Conditional measurement

Consider a JC model used for cooling-by-measurement protocols, whose Hamiltonian in the rotating frame with respect to $H_0 = \omega_a(|e\rangle\langle e| + a^\dagger a)$ reads

$$H = \Delta|e\rangle\langle e| + g(a^\dagger\sigma_- + a\sigma_+). \tag{1}$$

Here $\Delta \equiv \omega_e - \omega_a$ is the detuning between the level-spacing of the atomic detector $\omega_e$ and the frequency of the target resonator $\omega_a$ and $|\Delta| \ll \omega_e, \omega_a$. $g$ is the coupling strength between the detector (qubit) and the target resonator. Pauli matrices $\sigma_-$ and $\sigma_+$ denote the transition operators of the qubit, and $a$ ($a^\dagger$) represents the annihilation (creation) operator of the resonator.

The conditional measurement-based cooling is described by a sequence of piecewise joint evolutions of the resonator and the detector, which are interrupted by instantaneous projective measurements on a particular subspace of the detector. Initially, the resonator is in a thermal-equilibrium state $\rho_a^{\rm th}$ with a finite temperature $T$, and the detector qubit starts from the ground state. Then the overall initial state has the form $\rho_{\rm tot}(0) = |g\rangle\langle g| \otimes \rho_a^{\rm th}$. To cool down the resonator, a conditional or selective measurement $M_g = |g\rangle\langle g|$ is implemented on the detector after the free evolution with an interval $\tau$, when the overall state becomes $\rho_{\rm tot}(\tau) = \exp(-iH\tau)\rho_{\rm tot}(0)\exp(iH\tau)$. And then conditional measurement yields a probabilistic result:

$$\rho_a(\tau) = \frac{\langle g|\rho_{\rm tot}(\tau)|g\rangle}{{\rm Tr}[\langle g|\rho_{\rm tot}(\tau)|g\rangle]}. \tag{2}$$

Based on the time dependence of the interval $\tau$, conditional cooling protocols can be categorized into the equal-time-spacing and unequal-time-spacing strategies [24,29]. The unequal-time-spacing strategy has demonstrated a dramatic cooling efficiency by setting the measurement interval as the inverse of the time-evolved thermal Rabi frequency $\tau_{\rm opt}^c(t) = 1/\Omega_{\rm th}(t)$, where $\Omega_{\rm th}(t) \equiv g\sqrt{\bar{n}(t)} = g\sqrt{\sum_n n p_n(t)}$, with $p_n(t)$ denoting the current population of the resonator on the Fock state $|n\rangle$. To optimize the cooling performance, our cooling architecture in this work employs the unequal-time-spacing strategy. After $N$ rounds of free-evolution and instantaneous-measurement described by an ordered time sequence $\{\tau_1(t_1), \tau_2(t_2), \ldots, \tau_N(t_N)\}$ with $t_{i>1} = \sum_{j=1}^{j=i-1}\tau_j$ and $\tau_1 \equiv 1/[g\sqrt{{\rm Tr}(\hat{n}\rho_a^{\rm th})}]$, the resonator state becomes

$$\rho_a\left(t = \sum_{i=1}^{N}\tau_i\right) = \frac{\sum_n \prod_{i=1}^N |\alpha_n(\tau_i)|^2 p_n|n\rangle\langle n|}{P_g(N)}, \tag{3}$$

where

$$p_n = \frac{e^{-n\hbar\omega_a/k_BT}}{Z}, \quad Z \equiv \frac{1}{1-e^{-\hbar\omega_a/k_BT}} \tag{4}$$

is the initial population,

$$P_g(N) = \sum_n \prod_{i=1}^N |\alpha_n(\tau_i)|^2 p_n \tag{5}$$

is the survival or success probability of CM, and

$$|\alpha_n(\tau_i)|^2 = \frac{\Omega_n^2 - g^2 n\sin^2(\Omega_n\tau_i)}{\Omega_n^2} \tag{6}$$

is the cooling coefficient, with $\Omega_n = \sqrt{g^2 n + \Delta^2/4}$ denoting the $n$-photon Rabi frequency. The cooling coefficient in Eq. (3) determines the average population

$$\bar{n}(t) = {\rm Tr}[\hat{n}\rho_a(t)], \quad \hat{n} \equiv a^\dagger a, \tag{7}$$

by reshaping the population distributions over all the Fock states. Note in Eq. (6) the cooling coefficient for $|0\rangle$ is unit, $|\alpha_0(\tau_i)|^2 = 1$, meaning that the ground-state population is always under protection during the cooling process. The populations on high-occupied Fock states are gradually reduced by $|\alpha_n(\tau_i)|^N < 1$ with increasing $N$ unless $\sin(\Omega_n\tau_i) = 0$ or $\Omega_n\tau_i = j\pi$ with integer $j$.

### B. Unconditional measurement

Unconditional-measurement cooling is a statistical mixture of the conditional-measurement counterpart achieved by expanding the measurement subspace to the whole space of the detector system. After a period of joint unitary evolution under the Hamiltonian (1), the overall state can be written as

$$\rho_{\rm tot}(\tau) = \bigoplus_n p_n \begin{pmatrix} |\alpha_n(\tau)|^2 & \chi_n(\tau) \\ \chi_n^*(\tau) & |\beta_n(\tau)|^2 \end{pmatrix}, \tag{8}$$

where

$$\chi_n(\tau) \equiv \frac{-g\sqrt{n}[\Delta \sin^2(\Omega_n\tau) - i\Omega_n \sin(2\Omega_n\tau)]}{2\Omega_n^2},$$

$$|\beta_n(\tau)|^2 \equiv \frac{g^2 n \sin^2(\Omega_n\tau)}{\Omega_n^2}.$$

UM can be implemented by tracing out the degrees of freedom of the detector $\mathrm{Tr}_d[\rho_{\rm tot}(\tau)]$. Then the resonator state reads

$$\rho_a(\tau) = \sum_{n\geqslant 0}[|\alpha_n(\tau)|^2 p_n + |\beta_{n+1}(\tau)|^2 p_{n+1}]|n\rangle\langle n|. \tag{9}$$

So after a nonselective measurement, i.e., a measurement without recording the result, a population transfer in the target resonator occurs as

$$p_n \to |\alpha_n(\tau)|^2 p_n + |\beta_{n+1}(\tau)|^2 p_{n+1}. \tag{10}$$

In contrast to the CM strategy that is characterized by a single cooling coefficient $|\alpha_n|^2$ in Eq. (6), the UM strategy depends subtly on an extra cooling coefficient $|\beta_n|^2$. According to Eq. (10), the initial population on the ground state $p_0$ becomes $|\alpha_0(\tau)|^2 p_0 + |\beta_1(\tau)|^2 p_1 = p_0 + |\beta_1(\tau)|^2 p_1$, indicating that a part of the population on the first excited state is transferred to the ground state. Under rounds of nonselective measurements, it is intuitive to expect that the populations on the higher states of the resonator keep moving to the lower states and eventually to the ground state. In practice, however, the cooling is constrained and even invertible since the populations on certain excited states can be fixed or enhanced when $|\alpha_n(\tau)|^2 = 1$ and $|\beta_{n+1}(\tau)|^2 \geqslant 0$, i.e., $\Omega_n\tau = 1$ and $\Omega_{n+1}\tau \geqslant 0$. This problem can be addressed by employing the unequal-time-spacing strategy. A time-varying $\tau$ could ensure that populations on all excited states are gradually reduced.

The cooling efficiency of the UM strategy depends strongly on the choice of $\tau$ spacing neighboring measurements, analogous to that of CM [29]. That could be observed in Fig. 1 by the average population of the resonator $\bar{n}$ under one measurement on the detector. The $\tau$-dependence of $\bar{n}$ demonstrates similar patterns across four orders in a scale of initial temperature. It is found that the average population declines gradually to a minimal point (the relative reduction becomes smaller with increasing temperature) at an optimized measurement-interval $\tau_{\rm opt}^u$, then it rebounds quickly and ends up with a random fluctuation around a value slightly lower than its initial thermal occupation $\bar{n}_{\rm th} \equiv \mathrm{Tr}(\hat{n}\rho_a^{\rm th})$.

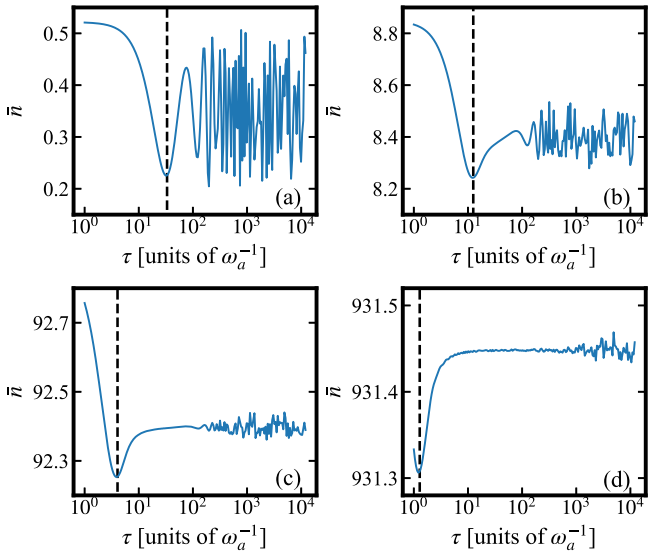


FIG. 1. Average population of the resonator after a single unconditional measurement as a function of the measurement-interval $\tau$ under various initial temperatures. (a) $T = 0.01\,\mathrm{K}$, (b) $T = 0.1\,\mathrm{K}$, (c) $T = 1.0\,\mathrm{K}$, and (d) $T = 10\,\mathrm{K}$. The vertical black-dashed lines indicate the analytical results for the optimized intervals given by Eq. (14). The parameters for the blue-solid curves are set as $g = 0.04\omega_a$ and $\Delta = 0.01\omega_a$.

To make full use of the cooling strategy, it is desired to analytically find the optimized interval $\tau_{\rm opt}^u$ as a functional of the current state and the model parameters. By virtue of Eq. (9) and under the resonant condition, the average population after a single unconditional measurement reads

$$\bar{n} = \sum_{n\geqslant 0} n(p_n \cos^2\Omega_n\tau + p_{n+1}\sin^2\Omega_{n+1}\tau)$$

$$= \eta + \frac{1}{2Z}\sum_{n\geqslant 0} n e^{-nx}(\cos 2\Omega_n\tau - e^{-x}\cos 2\Omega_{n+1}\tau), \tag{11}$$

where $\eta \equiv (\bar{n}_{\rm th} + 2\bar{n}_{\rm th}^2)/(2 + 2\bar{n}_{\rm th})$ and $x \equiv \hbar\omega_a/k_BT$. Since the weight function $ne^{-nx}$ in Eq. (11) is dominant around $n_d \equiv k_BT/\hbar\omega_a = 1/x$, the variables $\Omega_n$ and $\Omega_{n+1}$ could thus be expanded around $n = n_d$. To the first order of $n - n_d$, we have

$$\cos 2\Omega_n\tau - e^{-x}\cos 2\Omega_{n+1}\tau$$

$$\approx \cos 2\Omega_d\tau - e^{-x}\cos 2\Omega_{d+1}\tau + (n - n_d)$$

$$\times\left(-\frac{\Omega_d\tau\sin 2\Omega_d\tau}{n_d} + e^{-x}\frac{\Omega_{d+1}\tau\sin 2\Omega_{d+1}\tau}{n_d+1}\right),$$

where

$$\Omega_d \equiv g\sqrt{n_d}, \quad \Omega_{d+1} \equiv g\sqrt{n_d+1} \tag{12}$$

define the dominant Rabi frequencies. Under the approximations that are appropriate for a moderate temperature $e^{-x} = \bar{n}_{\rm th}/(\bar{n}_{\rm th}+1) \approx 1$ and $\Omega_{d+1}/(n_d+1) \approx \Omega_d/n_d$, the average

population in Eq. (11) can be expressed by

$$\bar{n} \approx \eta + \sin\Omega_- \tau(\bar{n}_{\rm th}\sin\Omega_+\tau + \eta'\Omega_d\tau\cos\Omega_+\tau), \quad (13)$$

where $\Omega_\pm \equiv \Omega_{d+1} \pm \Omega_d$ and $\eta' \equiv \bar{n}_{\rm th}(1+2\bar{n}_{\rm th}-n_d)/n_d$. Note that we have used the formulas about the geometric series $\sum_{n=0}^{\infty} ne^{-nx} = e^x/(e^x-1)^2$ and $\sum_{n=0}^{\infty} n^2e^{-nx} = e^x(1+e^x)/(e^x-1)^3$. Within a moderate time step $\tau$, Eq. (13) depends predominantly on the high-frequency terms characterized by $\Omega_+$. In the regime of $T \sim 0.1$–$10$ K, the term weighted by $\eta'\Omega_d\tau$ overwhelms that weighted by $\bar{n}_{\rm th}$. And as evidenced by Fig. 1, this advantage expands with a larger $\tau^u_{\rm opt}$ when the initial or effective temperature of the resonator becomes lower. We can therefore focus on the last term in Eq. (13) to minimize $\bar{n}$. Subsequently, $\cos\Omega_+\tau = -1$ yields

$$\tau^u_{\rm opt} = \frac{\pi}{\Omega_d + \Omega_{d+1}}. \quad (14)$$

This result can be extended to the near-resonant situation by modifying the definition of $\Omega_d$ in Eq. (12) to $\sqrt{g^2 n_d + \Delta^2/4}$. The vertical black-dashed lines in Fig. 1 denote the measurement-intervals optimized by Eq. (14). It is found that the analytical expression is well suited to estimate the minimum values of the average population in a wide range of temperature. As demonstrated by both analytical and numerical results, a shorter measurement interval is required to cool down a higher-temperature resonator. In the JC-like models, coupling a qubit to a high-temperature resonator induces a faster transition between the ground state and the excited state of the qubit. Although a quick measurement would interrupt this process, an inappropriate time interval would have a negative effect on cooling [30].

Similar to the optimized interval $\tau^c_{\rm opt}(t)$ for the conditional-measurement strategy [29], here $\tau^u_{\rm opt}$ is also updatable by substituting time-varied $\Omega_d$ and $\Omega_{d+1}$ into Eq. (14). The dominant Fock-state-number $n_d$ determining $\Omega_d$ in Eq. (12) could be understood as a function of the effective temperature during the cooling procedure, which relies uniquely on $\bar{n}(t)$ or $p_n(t)$.

## III. MEASUREMENT OPTIMIZATION

A thermal resonator could be steadily yet slowly cooled down by an unconditional measurement strategy equipped with an optimized measurement interval in Eq. (14). And this strategy is performed with a unit probability in the absence of postselection over the measurement outcome. In sharp contrast, a conditional measurement strategy is a more efficient cooling protocol but with a poor success probability. It is therefore desired to find an optimized sequence of measurements as a hybrid of UM and CM to maintain a great performance, taking both cooling efficiency and experimental overhead into account. In this section, we present an algorithm that employs the reinforcement learning to generate the optimized control sequence indicating when and which measurement is performed.

The performance of any cooling-by-measurement strategy can be characterized or evaluated by the cooling ratio $\bar{n}(t)/\bar{n}_{\rm th}$, the success probability $P_g$ of the detector in the measured subspace, and the fidelity of the resonator in its ground state, $F = \langle n=0|\rho_a(t)|n=0\rangle$ [24]. To compare various interpolation sequences of UM and CM in cooling performance and to

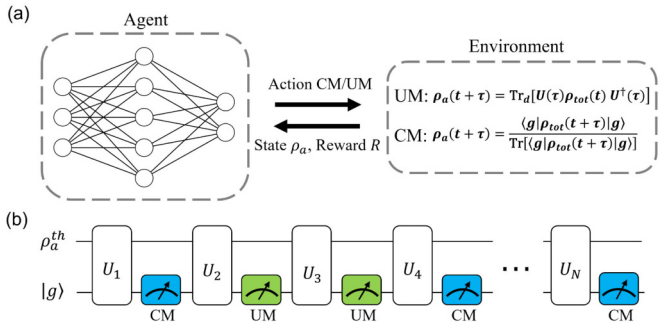


FIG. 2. (a) RL-optimization diagram on cooling by measurement. An agent constructed by the neural network interacts with an environment. The agent chooses an action (CM or UM strategy) according to the current state of the resonator. Then the environment would take this action and return both the state under the measurement and the reward $R$ based on the cooperative cooling performance $\mathcal{C}$ in Eq. (15). (b) Circuit model for our cooling algorithm based on the local-optimized UM and CM strategies. Starting from a thermal state, the resonator (the upper line) would be gradually cooled down to its ground state with implementation of measurement on the detector (the lower line), which starts from the ground state. The measurement sequence can be obtained by reinforcement learning.

evaluate the figure of merit for the reinforcement learning, we can define a cooperative cooling quantifier as

$$\mathcal{C} = FP_g \log_{10}\frac{\bar{n}_{\rm th}}{\bar{n}(t)}. \quad (15)$$

Notably, the logarithm function is used to obtain a positive value with almost the same order as $F$ and $P_g$ in magnitude. Then $\bar{n}(t)$, $P_g$, and $F$ could be considered in a balanced manner. In fact, the average population could be reduced by several (normally less than 10) orders of magnitude under an efficient cooling protocol. In the EIT cooling [47], $\log_{10}[\bar{n}_{\rm th}/\bar{n}(t)] \sim (2,3)$; and in the resolved sideband cooling [48], $\log_{10}[\bar{n}_{\rm th}/\bar{n}(t)] \sim (4,5)$. Although Eq. (15) is not a unique choice, it is instructive to find that a lower average population, a larger success probability, and a higher ground-state fidelity yield a better cooling performance.

The RL optimization is shown in Fig. 2(a). It is constituted by the "agent" part based on a series of neural networks and the "environment" part performing the cooling-by-measurement actions on a quantum system. In the reinforcement learning, the agent has a cluster of parameters, which would be learned and trained using the data collected through its interaction with the environment. In our architecture, the agent would choose an action, i.e., conditional or unconditional measurement, on the resonator, given its current state. Then the environment takes this action and returns the updated resonator-state $\rho_a$ and a "reward" $R$ after the measurement. The reward is generated by the indicator in Eq. (15) to estimate whether the action is good or bad, which would be used to update the agent's parameters. During one "episode", the agent would interact with the environment for $N$ times, i.e., the number of measurements during the whole sequence, which has been fixed from the beginning. A total reward is eventually counted. And the agent is trained to maximize the total reward through artificial episodes until it converges.

Then the agent could provide a realistic control sequence of the measurement strategies with their own (optimized) measurement intervals. The cooling-by-measurement sequence can be realized in a circuit model in Fig. 2(b). Rounds of free evolutions and measurements are successively arranged. The evolution time between two neighboring measurements depends on the measurement strategy and the resonator state at the end of the last round. We follow the PPO algorithm in the agent structure, the data-collecting methods, and the updating parameters, whose details can be found in Appendix A. The interpolation algorithm of UM and CM and the implementation of the measurement sequence are illustrated by a pseudocode in Appendix B.

We consider cooling down a mechanical microresonator in gigahertz [49,50] with various interpolation sequences of UM and CM. Using the resonator frequency, $\omega_a = 1.4$ GHz; the coupling strength between the resonator and the detector, $g = 0.04\omega_a$; and the initial temperature of the resonator, $T = 0.1$ K; it is found that the average population starts from $\bar{n}_{\rm th} = 8.85$. The cooling performances under the sequences entirely consisting of UM and CM are shown by the blue solid lines with circle markers and the orange dotted lines in Figs. 3(a)–3(d), labeled by $S_u$ and $S_c$, respectively. It is found that under the conditional measurement strategy with $N = 16$, the average population $\bar{n}$ is reduced by five orders of magnitude [see Fig. 3(a)], and the ground-state fidelity is over $F > 0.9999$ [see Fig. 3(b)] with less than 10% of the success probability [see Fig. 3(c)]. In sharp contrast, under the same number of unconditional measurements, $\bar{n}$ is merely reduced to $\bar{n} \approx 3.36$ and the ground-state fidelity $F \approx 0.78$, despite having a unit success probability. In terms of all the individual quantifiers, i.e., $\bar{n}$, $F$, and $P_g$, the results under the hybrid sequences of UM and CM labeled by $S_k$, $k = 1, 2, 4$, are among the former two limits $S_u$ and $S_c$. As illustrated by Figs. 3(e), 3(f), and 3(g), the three sequences start from a CM (indicated by 1), switch to the UM (indicated by 0) after $k$ rounds of free evolution and measurement, switch back to CM after a single round, and then the preceding arrangement is repeated. In comparison to the entire UM sequence, the interpolation with CM promotes the cooling efficiency in $\bar{n}$. A larger $k$ gives rise to a smaller proportion of the unconditional measurements and a lower probability $P_g$ that the detector remains in its measured subspace.

With respect to the cooperative cooling performance given by Eq. (15), it is found [see Fig. 3(d)] that $\mathcal{C}(S_1) > \mathcal{C}(S_2) > \mathcal{C}(S_4) > \mathcal{C}(S_u)$ and yet $\mathcal{C}(S_2) \approx \mathcal{C}(S_c)$, such that a regular interpolation sequence could therefore have a better cooperative cooling performance than the entire CM sequence. However, the dependence of $\mathcal{C}$ for an arbitrary hybrid sequence on its proportion of CM strategies might not be monotonic. We are then motivated to find an optimized sequence by virtue of the PPO algorithm. A typical RL-optimized sequence of cooling strategies labeled by $S_{\rm opt}$ is described in Fig. 3(h). With four orders reduction in the average population (close to the cooling efficiency provided by $S_c$), an almost unit ground-state fidelity $F > 0.9999$, and a moderate success probability $P_g \approx 30\%$ (much larger than that by $S_c$), the optimized sequence achieves an overwhelming cooperative cooling performance $\mathcal{C}(S_{\rm opt}) = 2.73$ according to Eq. (15) over all the other measurement sequences. Therefore, we have

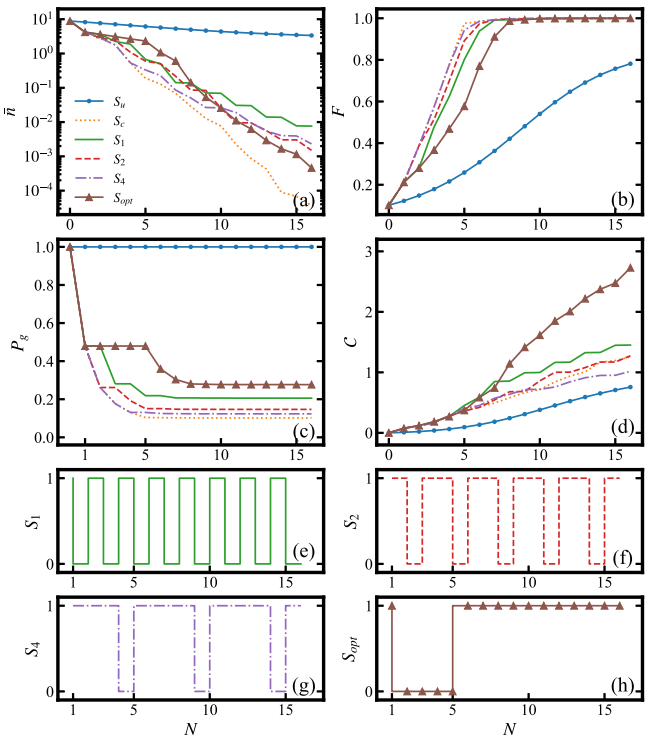


FIG. 3. (a) Average population, (b) fidelity of the resonator in its ground state, (c) success probability, and (d) cooperative cooling performance under various sequences of cooling-by-measurement. The blue solid lines with circle markers labeled by $S_u$ and the orange dotted lines labeled by $S_c$ indicate the sequences entirely consisting of UM and CM strategies, respectively. The green solid lines, red dashed lines, and purple dot-dashed lines describe the hybrid sequences shown in (e), (f), and (g), and labeled by $S_1$, $S_2$, and $S_4$, respectively. The brown solid lines with triangle markers labeled by $S_{\rm opt}$ denote the RL-optimized sequence presented in (h). For all the sequences in (e), (f), (g), and (h), 1 and 0 indicate CM and UM strategies, respectively. The parameters are set as $\omega_a = 1.4$ GHz, $T = 0.1$ K, $g = 0.04\omega_a$, and $\Delta = 0.01\omega_a$.

achieved a compromise of the cooling rate and the success probability through the reinforcement leaning method with much less overhead than the brute-force searching. The RL-optimized sequence is not unique, yet the current results of $\bar{n}$, $F$, $P_g$, and $\mathcal{C}$ in Fig. 3 are almost invariant as long as there is one CM in the first several rounds.

The RL-optimized algorithm applies to a wide range of initial temperature for the resonator. Starting from various $\bar{n}_{\rm th}$ determined by the temperature, the average populations could be reduced by three to five orders of magnitude under the optimized measurement sequences, as demonstrated in Fig. 4(a). It is found that under a higher temperature, it is harder to suppress the transitions between the ground state and the excited states of the detector. Then both the relative magnitude in the population reduction [see Fig. 4(a)] and the cooperative cooling performance [see Fig. 4(b)] manifest a monotonically decreasing behavior as temperature increases.

Similar to Fig. 3(h), here we present in Figs. 4(c), 4(d), 4(e), and 4(f) the optimized sequences fully determined by the PPO algorithm, which still outperform any regular interpolated sequence in the cooling quantifier $\mathcal{C}$. Comparing these

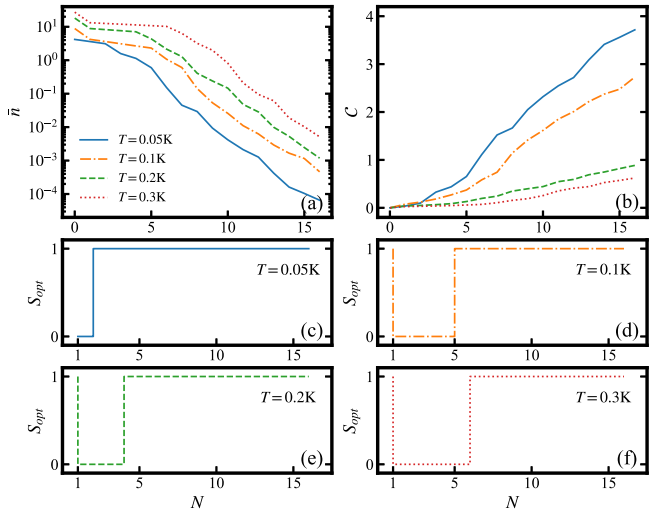


FIG. 4. (a) Average populations and (b) cooperative cooling performance under the RL-optimized cooling algorithm with various initial temperatures. Parts (c), (d), (e), and (f) describe the optimized sequences of UM and CM with $T = 0.05$, 0.1, 0.2, and 0.3 K, respectively. The other parameters are the same as those in Fig. 3.

four subfigures corresponding to various temperatures, it is interesting to find that a larger portion of the unconditional measurements is required along the optimized sequence for a higher temperature. This is consistent with the fact that under CM the success probability $P_g$ to find a detector in its ground state decreases exponentially with increasing temperature of the target resonator. Then more UMs are used to save a rapidly declining $P_g$ for obtaining a larger $\mathcal{C}$. In addition, for $T > 0.05$ K, the RL-optimized sequence always starts from a conditional measurement, which is an important part of having a significant cooling rate for $\bar{n}$ during the first several rounds of the whole sequence.

The profiles shown in Fig. 3(h) and Figs. 4(d), 4(e), and 4(f) manifest a common pattern for all the RL-optimized sequences. It is found in the previous several rounds that a conditional or projective measurement should be performed on the detector, when the resonator is normally in a comparatively high-temperature state, and several unconditional measurements ensued before further cooling. This pattern is consistent with the variations of both energy and entropy in nonunitary controls [51]. The energy variation induced by a projective measurement is $k_B T H(\rho)$ on average, where $H(\rho)$ is the Shannon entropy of the whole system after a free evolution. Then in the end of the first round, a projective measurement is desired to cut down as much energy as it could, which is followed by several rounds of unconditional measurements to save the success probability. Thus in general we anticipate seeing more UMs than CMs in the first several rounds and more CMs than UMs in the remaining rounds.

## IV. DISCUSSION AND CONCLUSION

The preceding analysis over the cooling performance neglects environment-induced dissipation. We now consider the cooling process in an open-quantum-system scenario, in which the free evolution between neighboring measurements

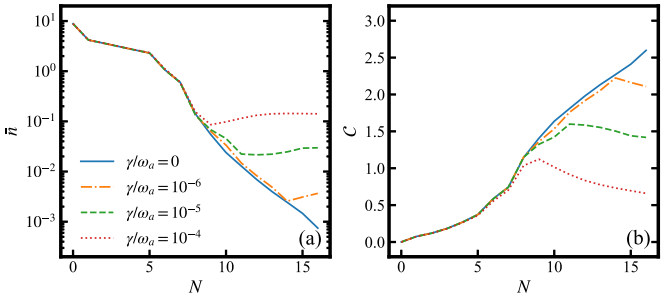


FIG. 5. (a) Average population and (b) cooperative cooling performance of the resonator coupled to a thermal environment under the optimized cooling strategy with various dissipative rates. The dissipation-free results are those labeled by $S_{\rm opt}$ in Figs. 3(a) and 3(d).

is influenced by a finite-temperature environment. The dynamics is then described by the master equation

$$\dot{\rho}(t) = -i[H, \rho(t)] + \gamma(\bar{n}_{\rm th} + 1)\mathcal{D}[a]\rho(t) + \gamma\bar{n}_{\rm th}\mathcal{D}[a^\dagger]\rho(t), \tag{16}$$

where $\mathcal{D}[A]$ represents the Lindblad superoperator

$$\mathcal{D}[A]\rho(t) \equiv A\rho(t)A^\dagger - \tfrac{1}{2}\{A^\dagger A, \rho(t)\}. \tag{17}$$

In Figs. 5(a) and 5(b), we present the average population $\bar{n}$ and the cooperative cooling performance $\mathcal{C}$, respectively, with various dissipation rates. To compare the cooling performances in the presence of thermal decoherence to the dissipation-free situation, we apply the RL-optimized sequence provided in Fig. 3(h). It is found that a larger dissipation rate gives rise to a weaker cooling performance in terms of both $\bar{n}$ and $\mathcal{C}$, exhibiting the struggle between cooling effects by measurement and the accumulated heating effects by environment. Nevertheless, for typical mechanical resonators in gigahertz with $\gamma/\omega_a \sim 10^{-5}$ [49,50], our optimized cooling protocol is still capable of reducing $\bar{n}$ by three orders of magnitude with about $N = 10$ measurements [see the green dashed line in Fig. 5(a)]. In the mean time, the asymptotic value of $\mathcal{C}$ still overwhelms the CM strategy labeled by $S_c$ in Fig. 3(d).

Even in the absence of thermal decoherence, $\bar{n}$ does not keep decreasing. Fundamentally, it is under the constraint of the third law of thermodynamics that absolute zero cannot be attained within a finite number of operations. Actually, either $\tau^c_{\rm opt}$ or $\tau^u_{\rm opt}$ approaches infinity as $\bar{n} \to 0$, which indicates that the whole cooling process has to be truncated by a maximum timescale.

We emphasize again that the preceding hybrid cooling sequences based on the conditional and unconditional measurements are optimized in both global and local perspectives. Globally, we use the reinforcement learning to find the optimized order for UM and CM. The local optimization depends on the selected measurement interval to obtain a minimum average-population $\bar{n}$ under one measurement. For UM in Eq. (14), $\tau^u_{\rm opt}(t)$ is not necessarily obtained by an instant feedback mechanism during a realistic practice. The measurement sequence $\{\tau_1(t_1), \tau_2(t_2), \ldots, \tau_N(t_N)\}$ can actually be obtained prior to the cooling measurements. $\tau_1(t_1)$ depends on the initial population-distribution $p_n$, and $\tau_k(t_k)$, $k \geqslant 2$, can be calculated on the effective temperature that is uniquely

determined by the dynamics of $p_n(t)$ through Eq. (12). In other words, we can avoid the feedback error and imprecision induced by detecting the resonator states during the experiment.

In summary, we present an optimized cooling architecture on a sequential arrangement of both conditional and unconditional measurements. We analyze and compare the advantages and disadvantages of both CM and UM on cooling rate and success probability. We obtain analytically an analytical expression for the optimized unconditional measurement interval $\tau^u_{\rm opt} = \pi/(\Omega_d + \Omega_{d+1})$ in parallel to that for conditional measurement [29]. Here the dominant Rabi frequency $\Omega_d$ depends on the dominant distribution of the resonator in its Fock state with $n_d = k_B T/(\hbar\omega_a)$ and the coupling strength between target and detector. The combination of the advantages of both measurement strategies gives rise to an optimized hybrid cooling algorithm assisted by reinforcement learning. It is justified by the cooperative cooling performance that we defined to quantify the comprehensive cooling efficiency for an arbitrary cooling-by-measurement strategy. Our work, therefore, pushes the cooling-by-measurement to an unattained degree with regard to efficiency and feasibility. It offers an appealing interdisciplinary application of quantum control and artificial intelligence.

## ACKNOWLEDGMENTS

We acknowledge financial support from the National Science Foundation of China (Grants No. 11974311 and No. U1801661).

## APPENDIX A: PROXIMAL POLICY OPTIMIZATION

This Appendix provides more details on proximal policy optimization, a typical reinforcement learning algorithm that we use to optimize the measurement sequence for cooling. The PPO algorithm follows an "actor-critic" frame, in which actors receive the current state as an input, and then they output an action according to an updatable policy, and a critic evaluates this action to determine whether the action should be encouraged or not. In the following, we do not discriminate "actor" and "policy" for simplicity.

As shown in Fig. 6, the PPO algorithm has two actors (policies) $\pi_{\rm old}(\{\theta\})$ and $\pi_{\rm new}(\{\theta'\})$ and one critic. Any of them is of an agent constructed by the neural networks (see Fig. 2) feathered with a set of parameters $\{\theta\}$. The two policies have the same structures in PPO. The old policy collects the sampling data through interaction with the environment, and the new one would use these data stored in a buffer to update $\{\theta\}$ to be $\{\theta'\}$. At first, the environment would initialize and deliver the state $s_1$ of the target system to the old policy $\pi_{\rm old}(\{\theta\})$; then the old policy generates an action $a_1$ according to $s_1$ and $\{\theta\}$. In the environment, the action $a_1$ is taken and the system state becomes $s_2$. The environment also provides a reward $R_1$ indicating how good the action is. The reward is generated by a task-specified reward function. At this stage, an interaction between the policy and the environment is completed, and one set of "trajectory" or return $\{s_1, a_1, R_1\}$ is collected. $N$ trajectories are collected in one episode, where $N$ amounts to the number of actions required to complete the task. The critic

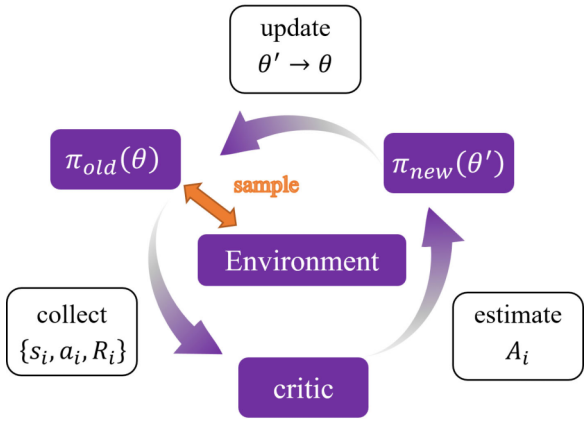


FIG. 6. Diagram of the proximal policy optimization algorithm.

takes both actions and states as input and outputs an advantage $A_i$ representing the contribution of the current action $a_i$ on the current state $s_i$. After collecting a sufficient amount of data, the critic would estimate the actions' contribution as precisely as possible. In the mean time, according to the advantages to maximize a clipped surrogate objective function $L^{\rm CLIP}(\{\theta\})$ [52], the new policy would transfer its parameters $\{\theta'\}$ to the old one.

In our application for optimizing the cooling sequence, the allowed inputs of the system states are defined as the populations in the Fock states, i.e., the diagonal elements of the target resonator $\rho_a$,

$$s_i = \{p_0(t), p_1(t), p_2(t), \ldots, p_{n_c}(t)\}, \tag{A1}$$

where $n_c$ indicates the cutoff Fock state for the resonator. The actions taken by the environment are selected from the set

$$a_i \in \{0, 1\}, \tag{A2}$$

where 0 and 1 represent unconditional and conditional measurements, respectively. Two policies are used to decide which type of measurement to be performed due to the current state of the resonator. The environment represents the quantum devices performing measurements, obtaining the updated states, and returning the rewards. When an action is selected and sent to the environment, the optimized measurement interval is calculated according to the measurement type. After unitary evolution lasting $\tau_{\rm opt} \in \{\tau^c_{\rm opt}, \tau^u_{\rm opt}\}$, measurement is performed on the detector. Then the average population $\bar{n}$, the ground-state fidelity $F$, and the success probability $P_g$ are obtained to calculate the cooperative cooling performance $\mathcal{C}$ given by Eq. (15). The reward function is set as a certain multiple of $\mathcal{C}$, $R_i(s_i, a_i) = 100 \times \mathcal{C}(s_i, a_i)$. After measurement, the environment then returns the resonator state and the reward to the policies. When the training is completed, a policy $\pi(\{\theta_{\rm opt}\})$ with a set of optimized parameters is achieved. The neutral network equipped with $\{\theta_{\rm opt}\}$ could then be used to generate the optimized actions to cool down the current state.

## APPENDIX B: GENERATION OF AN OPTIMIZED SEQUENCE

Both the order of measurements and the sequence of measurement intervals could be regarded as output of our

RL-optimized cooling algorithm as shown in Algorithm 1. The input information is the initial temperature $T$, fully determining the thermal state of the resonator. When the reinforcement learning process was completed by the PPO algorithm (see Appendix A), the parameters $\{\theta\}$ of the neural network (policy $\pi$) had been trained to be capable of selecting one of the two measurement strategies for the current state, which maximizes the cooperative cooling performance, and then the cooling procedure is formally launched. We run the policy $\pi(\{\theta_{\rm opt}\})$ on $\rho_a(0)=\rho_a^{\rm th}$, which generates the first measurement strategy $\mathcal{M}_1$, $\mathcal{M}_1\in\{0,1\}$. Here 0 and 1 indicate UM and CM, respectively. If $\mathcal{M}_1=0$, then $\tau_{\rm opt}(t_1)=\tau_{\rm opt}^u$ in Eq. (14), which could be obtained by the effective temperature $T_{\rm eff}$ of the resonator (initially $T_{\rm eff}=T$, and it is updated by the current state of the last round). Subsequently, the cooling coefficients $|\alpha_n|^2$ and $|\beta_n|^2$ are calculated and the resonator state is modified according to Eq. (9). Otherwise if $\mathcal{M}_1=1$, a conditional measurement will be implemented after an interval $\tau_{\rm opt}(t_1)=\tau_{\rm opt}^c=1/\Omega_{\rm th}(t)$ and the resonator state is modified according to Eq. (3). In the end of this round, one can calculate $T_{\rm eff}$ by the current $p_n(t)$ and then go to the next round. After $N$ iterations, the optimized measurement sequence characterized by $S_{\rm opt}=\{\mathcal{M}_1,\mathcal{M}_2,\ldots,\mathcal{M}_N\}$ and $\mathcal{T}=\{\tau_{\rm opt}(t_1),\tau_{\rm opt}(t_2),\ldots,\tau_{\rm opt}(t_N)\}$ appears as described in Fig. 3(h) and Figs. 4(c), 4(d), 4(e), and 4(f), respectively.

**Algorithm 1.** RL-optimized cooling procedure.

**Output:** $S_{\rm opt}=\{\mathcal{M}_1,\mathcal{M}_2,\ldots,\mathcal{M}_N\}$ and $\mathcal{T}=\{\tau_{\rm opt}(t_1),\tau_{\rm opt}(t_2),\ldots,\tau_{\rm opt}(t_N)\}$
**Input:** Temperature $T$
Initialize the thermal state $\rho_a=\sum_n p_n|n\rangle\langle n|$ with $T$ Use PPO to train an optimized policy $\pi(\{\theta_{\rm opt}\})$
**for** $i=1,2,\ldots,N$ **do**
  Run the policy $\pi(\{\theta_{\rm opt}\})$ on $\rho_a$ to generate $\mathcal{M}_i$
  Attain $T_{\rm eff}=\hbar\omega_a/[k_B\ln(1+1/\bar{n})]$ on $\bar{n}(\rho_a)$
  **if** $\mathcal{M}_i=0$ **then**
    Calculate $\tau_{\rm opt}(t_i)=\pi/(\Omega_d+\Omega_{d+1})$ on $T_{\rm eff}$
    Get the cooling coefficients $|\alpha_n|^2$ and $|\beta_n|^2$
    UM: $\rho_a\leftarrow\sum_n(|\alpha_n|^2p_n+|\beta_{n+1}|^2p_{n+1})|n\rangle\langle n|$
  **end**
**else if** $\mathcal{M}_i=1$ **then**
    Calculate $\tau_{\rm opt}(t_i)=1/\Omega_{\rm th}$ on $T_{\rm eff}$
    Get the cooling coefficients $|\alpha_n|^2$
    CM: $\rho_a\leftarrow\sum_n|\alpha_n|^2p_n|n\rangle\langle n|/(\sum_n|\alpha_n|^2p_n)$
  **end**
**end**

In practical implementations, the measurements by $S_{\rm opt}$ and $\mathcal{T}$ can be acted on the detector without knowledge of the target-resonator state.


[1] G. J. Milburn and M. J. Woolley, Quantum nanoscience, Contemp. Phys. **49**, 413 (2008).
[2] M. Aspelmeyer, T. J. Kippenberg, and F. Marquardt, Cavity optomechanics, Rev. Mod. Phys. **86**, 1391 (2014).
[3] S. Lloyd and S. L. Braunstein, Quantum Computation over Continuous Variables, Phys. Rev. Lett. **82**, 1784 (1999).
[4] J. Q. You and F. Nori, Atomic physics and quantum optics using superconducting circuits, Nature (London) **474**, 589 (2011).
[5] K. Toyoda, R. Hiji, A. Noguchi, and S. Urabe, Hong–ou–mandel interference of two phonons in trapped ions, Nature (London) **527**, 74 (2015).
[6] M. Um, J. Zhang, D. Lv, Y. Lu, S. An, J.-N. Zhang, H. Nha, M. S. Kim, and K. Kim, Phonon arithmetic in a trapped ion system, Nat. Commun. **7**, 11410 (2016).
[7] M. F. Bocko and R. Onofrio, On the measurement of a weak classical force coupled to a harmonic oscillator: experimental progress, Rev. Mod. Phys. **68**, 755 (1996).
[8] C. M. Caves, K. S. Thorne, R. W. P. Drever, V. D. Sandberg, and M. Zimmermann, On the measurement of a weak classical force coupled to a quantum-mechanical oscillator. i. issues of principle, Rev. Mod. Phys. **52**, 341 (1980).
[9] S. Sharma, Y. M. Blanter, and G. E. W. Bauer, Optical Cooling of Magnons, Phys. Rev. Lett. **121**, 087205 (2018).
[10] I. Wilson-Rae, N. Nooshi, W. Zwerger, and T. J. Kippenberg, Theory of Ground State Cooling of a Mechanical Oscillator Using Dynamical Backaction, Phys. Rev. Lett. **99**, 093901 (2007).
[11] S. Gigan, H. R. Böhm, M. Paternostro, F. Blaser, G. Langer, J. B. Hertzberg, K. C. Schwab, D. Bäuerle, M. Aspelmeyer, and A. Zeilinger, Self-cooling of a micromirror by radiation pressure, Nature (London) **444**, 67 (2006).
[12] X. Wang, S. Vinjanampathy, F. W. Strauch, and K. Jacobs, Ultraefficient Cooling of Resonators: Beating Sideband Cooling with Quantum Control, Phys. Rev. Lett. **107**, 177204 (2011).
[13] J. Zhang, D. Li, R. Chen, and Q. Xiong, Laser cooling of a semiconductor by 40 kelvin, Nature (London) **493**, 504 (2013).
[14] R. I. Epstein, M. I. Buchwald, B. C. Edwards, T. R. Gosnell, and C. E. Mungan, Observation of laser-induced fluorescent cooling of a solid, Nature (London) **377**, 500 (1995).
[15] G. Morigi, J. Eschner, and C. H. Keitel, Ground State Laser Cooling Using Electromagnetically Induced Transparency, Phys. Rev. Lett. **85**, 4458 (2000).
[16] C. F. Roos, D. Leibfried, A. Mundt, F. Schmidt-Kaler, J. Eschner, and R. Blatt, Experimental Demonstration of Ground State Laser Cooling with Electromagnetically Induced Transparency, Phys. Rev. Lett. **85**, 5547 (2000).
[17] M. R. Vanner, I. Pikovski, G. D. Cole, M. S. Kim, C. Brukner, K. Hammerer, G. J. Milburn, and M. Aspelmeyer, Pulsed quantum optomechanics, Proc. Natl. Acad. Sci. USA **108**, 16182 (2011).
[18] M. R. Vanner, J. Hofer, G. D. Cole, and M. Aspelmeyer, Cooling-by-measurement and mechanical state tomography via pulsed optomechanics, Nat. Commun. **4**, 2295 (2013).
[19] J. S. Bennett, K. Khosla, L. S. Madsen, M. R. Vanner, H. Rubinsztein-Dunlop, and W. P. Bowen, A quantum optomechanical interface beyond the resolved sideband limit, New J. Phys. **18**, 053030 (2016).
[20] M. Rossi, D. Mason, J. Chen, Y. Tsaturyan, and A. Schliesser, Measurement-based quantum control of mechanical motion, Nature (London) **563**, 53 (2018).

[21] M. Brunelli, D. Malz, A. Schliesser, and A. Nunnenkamp, Stroboscopic quantum optomechanics, Phys. Rev. Res. **2**, 023241 (2020).

[22] L. Buffoni, A. Solfanelli, P. Verrucchi, A. Cuccoli, and M. Campisi, Quantum Measurement Cooling, Phys. Rev. Lett. **122**, 070603 (2019).

[23] H. Nakazato, T. Takazawa, and K. Yuasa, Purification through Zeno-Like Measurements, Phys. Rev. Lett. **90**, 060401 (2003).

[24] Y. Li, L.-A. Wu, Y.-D. Wang, and L.-P. Yang, Nondeterministic ultrafast ground-state cooling of a mechanical resonator, Phys. Rev. B **84**, 094502 (2011).

[25] J.-S. Xu, M.-H. Yung, X.-Y. Xu, S. Boixo, Z.-W. Zhou, C.-F. Li, A. Aspuru-Guzik, and G.-C. Guo, Demon-like algorithmic quantum cooling and its realization with quantum optics, Nat. Photon. **8**, 113 (2014).

[26] R. Puebla, O. Abah, and M. Paternostro, Measurement-based cooling of a nonlinear mechanical resonator, Phys. Rev. B **101**, 245410 (2020).

[27] P. V. Pyshkin, D.-W. Luo, J. Q. You, and L.-A. Wu, Ground-state cooling of quantum systems via a one-shot measurement, Phys. Rev. A **93**, 032120 (2016).

[28] J.-s. Yan and J. Jing, External-level assisted cooling by measurement, Phys. Rev. A **104**, 063105 (2021).

[29] J.-s. Yan and J. Jing, Simultaneous cooling by measuring one ancillary system, Phys. Rev. A **105**, 052607 (2022).

[30] J.-M. Zhang, J. Jing, L.-A. Wu, L.-G. Wang, and S.-Y. Zhu, Measurement-induced cooling of a qubit in structured environments, Phys. Rev. A **100**, 022107 (2019).

[31] G. Harel and G. Kurizki, Fock-state preparation from thermal cavity fields by measurements on resonant atoms, Phys. Rev. A **54**, 5410 (1996).

[32] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, Mastering the game of go with deep neural networks and tree search, Nature (London) **529**, 484 (2016).

[33] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, Mastering the game of go without human knowledge, Nature (London) **550**, 354 (2017).

[34] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, A general reinforcement learning algorithm that masters chess, shogi, and go through self-play, Science **362**, 1140 (2018).

[35] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, Human-level control through deep reinforcement learning, Nature (London) **518**, 529 (2015).

[36] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[37] I. Convy, H. Liao, S. Zhang, S. Patel, W. P. Livingston, H. N. Nguyen, I. Siddiqi, and K. B. Whaley, Machine learning for continuous quantum error correction on superconducting qubits, New J. Phys. **24**, 063019 (2022).

[38] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Reinforcement Learning with Neural Networks for Quantum Feedback, Phys. Rev. X **8**, 031084 (2018).

[39] A. Bolens and M. Heyl, Reinforcement Learning for Digital Quantum Simulation, Phys. Rev. Lett. **127**, 110502 (2021).

[40] X. Yuan, J. Sun, J. Liu, Q. Zhao, and Y. Zhou, Quantum Simulation with Hybrid Tensor Networks, Phys. Rev. Lett. **127**, 040501 (2021).

[41] S.-F. Guo, F. Chen, Q. Liu, M. Xue, J.-J. Chen, J.-H. Cao, T.-W. Mao, M. K. Tey, and L. You, Faster State Preparation across Quantum Phase Transition Assisted by Reinforcement Learning, Phys. Rev. Lett. **126**, 060401 (2021).

[42] 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).

[43] 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).

[44] 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).

[45] D.-K. Kim and H. Jeong, Deep reinforcement learning for feedback control in a collective flashing ratchet, Phys. Rev. Res. **3**, L022002 (2021).

[46] J. Yao, L. Lin, and M. Bukov, Reinforcement Learning for Many-Body Ground-State Preparation Inspired by Counterdiabatic Driving, Phys. Rev. X **11**, 031070 (2021).

[47] L. Feng, W. L. Tan, A. De, A. Menon, A. Chu, G. Pagano, and C. Monroe, Efficient Ground-State Cooling of Large Trapped-Ion Chains with an Electromagnetically-Induced-Transparency Tripod Scheme, Phys. Rev. Lett. **125**, 053001 (2020).

[48] J. F. Triana, A. F. Estrada, and L. A. Pachón, Ultrafast Optimal Sideband Cooling under Non-Markovian Evolution, Phys. Rev. Lett. **116**, 183602 (2016).

[49] L. Ding, C. Baker, P. Senellart, A. Lemaitre, S. Ducci, G. Leo, and I. Favero, Wavelength-sized gaas optomechanical resonators with gigahertz frequency, Appl. Phys. Lett. **98**, 113108 (2011).

[50] J. Chan, T. P. M. Alegre, A. H. Safavi-Naeini, J. T. Hill, A. Krause, S. Gröblacher, M. Aspelmeyer, and O. Painter, Laser cooling of a nanomechanical oscillator into its quantum ground state, Nature (London) **478**, 89 (2011).

[51] S. Gherardini, F. Campaioli, F. Caruso, and F. C. Binder, Stabilizing open quantum batteries by sequential measurements, Phys. Rev. Res. **2**, 013095 (2020).

[52] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, Proximal policy optimization algorithms, arXiv:1707.06347.