Optimize cooling-by-measurement by reinforcement learning

Jia-shun Yan and Jun Jing

School of Physics, Zhejiang University, Hangzhou 310027, Zhejiang, China

(Dated: June 2, 2022)

Cooling by the conditional measurement demonstrates a transparent advantage over that by the unconditional counterpart on the average-population-reduction rate. This advantage, however, is blemished by few percentage of the successful probability of finding the detector system 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 the conditional and unconditional measurement strategies. Analogous to the conditional measurement, an optimal measurement-interval $\tau_{\text{opt}}$ for the unconditional (nonselective) measurement is analytically found for the first time, which is inversely proportional to the collective dominant Rabi frequency $\Omega_d$ as a function of the resonator’s population at the end of the last round. A cooling algorithm under the global optimization by the reinforcement learning results in the maximum value for the cooperative cooling performance, an indicator function to quantify 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 over four orders in magnitude with a successful probability about 30%.

I. INTRODUCTION

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

Beyond the paradigms extracting the system energy through the designed dissipation channels based on the blue-shifted (anti-Stokes) sidebands, nondeterministic methods of measurement-based cooling have been proposed in theory [17, 18] and demonstrated in experiment [19]. Rather than providing a unidirectional decay channel for the target system, the nonunitary evolution induced by repeated measurements of the ground state of the detector system leads to the postselection of the ground state of the target system (typically modelled as a resonator) and the reduction of its high-energy distribution in the ensemble. In another word, 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 [20], cooling by one shot measurement [21], expanding cooling range by an external driving [22], to accelerating cooling rate by optimizing the measurement intervals [23], an unexplored weakness of all these CM strategies is their small successful probability inherited from the projective operation (postselection). An inevitable amount of cost rises with more samples in ensemble. The unconditional measurement (UM) strategy, in sharp contrast to CM, performs a nonselective and impulsive measurement in all the bases of the target bare Hamiltonian at the end of each round of the joint free evolution of the target and the detector systems [24, 25]. It is more likely to realize a unit-success-probability cooling but suffers from a much slower cooling-rate than CM, indicating more number of measurements towards the ground-state cooling. To compromise the cooling rate and the successful probability, the interpolating-configuration of the conditional and unconditional measurements becomes an optimization problem.

The integration of a small-scale quantum circuit with a classical optimizer (e.g., the neural network) provides a design paradigm by specifying 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, that demonstrated remarkable capabilities in the board and video games [26–29], have substantiated widely and timely interest in studying quantum physics [30], such as quantum error correction [31, 32], quantum simulation [33, 34], and quantum state preparation [35–37], 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 [38–40].

In this work, we propose a measurement-based cooling architecture as a hybrid sequence of UM and CM strategies. It involves a double optimization: on each local step along the sequence, either UM or CM can be con-

* Email address: jingjun@zju.edu.cn
considerably improved by using the timely-optimized measurement intervals; and on the global efficiency of the sequence, its arrangement can be separably optimized through reinforcement learning. Particularly, 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), the conditional and unconditional measurements are alternatively performed to cool down the resonator to its ground state. In parallel to the optimized measurements obtained for CM \[23\], we here analytically derive the optimized interval of 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 successful probability of the detector in the measured subspace, and the fidelity of the resonator in the ground state, as we defined to rank the comprehensive efficiencies of various sequences of measurements. 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 the 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 the unconditional measurement. In Sec. III, we introduce the interpolation diagram for the cooling architecture based on these two measurements, define the cooperative cooling performance to comprehensively quantify various strategies, and 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

The cooling-by-measurement framework is typically established on the JC model, whose Hamiltonian in the rotating frame with respect to \(H_0 = \hbar \Delta |e\rangle \langle e| + \hbar a^\dagger a\) reads

\[
H = \Delta |e\rangle \langle e| + \hbar (a^\dagger \sigma_- + a \sigma_+).
\]

Here \(\Delta \equiv \omega_r - \omega_a\) is the detuning between the atomic level-spacing \(\omega_r\) and the to-be-cooled resonator frequency \(\omega_a\); and \(|\Delta| \ll \omega_r, \omega_a\). \(g\) 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 annihilation (creation) operator of the resonator. The cooling process is described by a sequence of piecewise joint evolutions of the resonator and the detector, that are interrupted by instantaneous projective measurements on a particular subspace of the detector.

The conditional measurement-based cooling is characterized by fixing the subspace as the ground state \(|g\rangle\). Initially, the resonator is in a thermal-equilibrium state \(\rho_a^{\text{th}}\) with a finite temperature \(T\), while the detector qubit starts from the ground state. Then the overall initial state has the form of

\[
\rho_{\text{tot}}(0) = |g\rangle \langle g| \otimes \rho_a^{\text{th}}.
\]

To cool down the resonator, a conditional or selective measurement \(M_g = \langle g| \langle g|\) is implemented on the detector after a free-evolution with interval \(\tau\), when the overall state becomes \(\rho_{\text{tot}}(\tau) = \exp(-iH\tau)\rho_{\text{tot}}(0)\exp(iH\tau)\). And then the conditional measurement yields a non-deterministic result:

\[
\rho_a(\tau) = \frac{\langle g| \rho_{\text{tot}}(\tau) |g\rangle}{\text{Tr}[\langle g| \rho_{\text{tot}}(\tau) |g\rangle]}.
\]

In regard of the time-dependence of the interval \(\tau\), the conditional cooling protocols can be categorized into the equal-time-spacing and unequal-time-spacing strategies \(18, 23\). The unequal-time-spacing strategy has demonstrated a dramatic advantage on the cooling performance by setting the measurement interval as the inverse of the thermal Rabi frequency \(\tau_{\text{th}}(t) = 1/\Omega_{\text{th}}(t)\), where \(\Omega_{\text{th}}(t) \equiv g\sqrt{n} \left(t\right) = g\sqrt{\sum_\alpha n_{\alpha}(t)}\) with \(n_{\alpha}(t)\) denoting the current population of the resonator on the Fock state \(|n\rangle\).

To attain an optimal 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 \(t_{i+1} = \sum_{j=1}^{i-1} \tau_j\) and \(\tau_1 \equiv 1/\sqrt{\text{Tr}(\hat{n}\rho_a^{\text{th}})}\), the resonator state becomes

\[
\rho_a\left(t = \sum_{i=1}^{N} \tau_i\right) = \frac{\sum_\alpha \prod_{i=1}^{N} |\alpha_n(\tau_i)|^2 p_n |n\rangle \langle n|}{P_g(N)},
\]

where \(p_n = \exp(-n\hbar\omega_a/k_B T)/Z\) with \(Z \equiv 1/[1 - \exp(-\hbar\omega_a/k_B T)]\) is the initial population,

\[
P_g(N) = \sum_\alpha \prod_{i=1}^{N} |\alpha_n(\tau_i)|^2 p_n
\]

is the survival or successful probability of CM, and

\[
|\alpha_n(\tau_i)|^2 = \frac{\Omega_n^2 - g^2 n \sin^2(\Omega_n \tau_i)}{\Omega_n^2}
\]

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. (4) determines the average population

\[
\bar{n}(t) = \text{Tr}[\hat{n}\rho_a(t)], \quad \bar{n} \equiv a^\dagger a,
\]
by reshaping the population distributions over all the Fock states. Note in Eq. (6), the 0th cooling coefficient is unit, $|a_0(\tau)|^2 = 1$, meaning that the ground-state population is always under protection during the cooling process. The populations on the high-occupied Fock states are gradually reduced by $|a_n(\tau)|^2/N < 1$ with increasing $N$ unless $\sin(\Omega_n \tau) = 0$ or $\Omega_n \tau = j\pi$ with integer $j$.

**B. Unconditional Measurement**

Expanding the measurement subspace to the whole space of the detector system, we can transform a conditional-measurement strategy into its unconditional-measurement counterpart. After a period of joint unitary evolution under the Hamiltonian (1), the overall state can be written as

$$\rho_{\text{tot}}(\tau) = \sum_p p_n \left( \frac{|a_n(\tau)|^2}{\chi_n^*(\tau)} \frac{\chi_n(\tau)}{|\beta_n(\tau)|^2} \right), \quad (8)$$

where

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

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

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

$$\rho_a(\tau) = \sum_{n\geq0} \left( |a_n(\tau)|^2 p_n + |\beta_{n+1}(\tau)|^2 p_{n+1} \right) \langle n | n \rangle \langle \text{Tr}_d[\rho_{\text{tot}}(\tau)] \rangle. \quad (9)$$

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

$$p_n \rightarrow |a_n(\tau)|^2 p_n + |\beta_{n+1}(\tau)|^2 p_{n+1}. \quad (10)$$

In contrast to the CM strategy that is characterized by a single cooling coefficient $a_n$ in Eq. (6), the UM strategy depends subtly on an extra cooling coefficients $\beta_n$. According to Eq. (10), the initial population of the ground state $p_0$ becomes $|a_0(\tau)|^2 p_0 + |\beta_1(\tau)|^2 |p_1| = p_0 + |\beta_1(\tau)|^2 |p_1|$. It indicates that partial population on the first excited state of the resonator is transferred to the ground state, whose original population is left untouched. Under rounds of nonselective measurements, it is intuitively to expect that the populations on the higher excited states of the resonator will keep moving to the lower states and eventually to the ground state. In practice, the cooling efficiency is however limited since the population on certain high excited states can also be fixed and even enhanced when $|a_n(\tau)|^2 = 1$ and $|\beta_{n+1}(\tau)|^2 \geq 0$, i.e., $\Omega_n \tau = 1$ and $\Omega_{n+1} \tau \geq 0$. This problem can be addressed by employing the unequal-time-spacing strategy. A time-varying $\tau$ could ensure that the populations on all the excited states are gradually reduced.

**FIG. 1.** The average populations of the resonator after a single unconditional measurement as a function of the measurement-interval $\tau$ under various initial temperatures. (a) $T = 0.01$ K, (b) $T = 0.1$ K, (c) $T = 1.0$ K and (d) $T = 10$ 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 $\Delta = 0.01\omega_a$.

Analogous to the CM case [23], the cooling efficiency of the UM strategy depends severely on the choice of $\tau$ between neighboring measurements. That could be observed in Fig. 1 by the average populations of the resonator $n$ under a single measurement on the detector. With the initial temperatures across four orders in magnitude, the $\tau$-dependence of $n$ demonstrates similar patterns. It is found that the average population declines gradually to a minimal point (the magnitude of the relative reduction decreases with increasing temperature) at an optimized measurement-interval $\tau^*_u$, and then rebounds quickly and ends up with a stochastic fluctuation around a value slightly lower than its initial population $n_{th} \equiv \text{Tr}(\hat{\rho}_u^0)$.

To make full use of the cooling strategy, it is desired to analytically find the optimized interval $\tau^*_u$ as a functional of the current state and the mode parameters. By virtue of Eq. (9) and under the resonant situation, the average population after a single unconditional measurement reads

$$\bar{n} = \sum_{n\geq0} n \left( p_n \cos^2 \Omega_n \tau + p_{n+1} \sin^2 \Omega_{n+1} \tau \right) = \frac{1}{2} \sum_{n\geq0} ne^{-nx} \left( \cos 2\Omega_n \tau - e^{-x} \cos 2\Omega_{n+1} \tau \right),$$

where $\eta \equiv (\bar{n}_{th} + 2\bar{n}_{th})/(2 + 2\bar{n}_{th})$ and $x \equiv h\omega_a/k_BT$. Since the weight function $ne^{-nx}$ in Eq. (11) is dominant around $n_d \equiv k_BT/h\omega_a$, the variables $\Omega_n$ and $\Omega_{n+1}$ could thus be expanded around $n = n_d$. To the first order of
\[ n - n_d, \text{ we have} \]
\[
\cos 2\Omega \tau - e^{-x} \cos 2\Omega_{n+1} \tau \\
\approx \cos 2\Omega \tau - e^{-x} \cos 2\Omega_{d+1} \tau + (n - n_d) \\
\times \left( -\frac{\Omega \tau \sin 2\Omega \tau}{n_d} + e^{-x} \frac{\Omega_{d+1} \tau \sin 2\Omega_{d+1} \tau}{n_d + 1} \right),
\]
\[
\Omega_d \equiv g\sqrt{n_d}, \quad \Omega_{d+1} \equiv g\sqrt{n_d + 1}. \quad (12)
\]
define the dominant Rabi frequencies. Under the approximations that \( e^{-x} \approx \tilde{n}_{th}/(\tilde{n}_{th} + 1) \approx 1 \) and \( \Omega_{d+1}/(n_d + 1) \approx \Omega_d/n_d \) appropriate for a moderate temperature, the average population in Eq. \((11)\) can be expressed by
\[
\tilde{n} \approx \eta + \sin \Omega \tau (\tilde{n}_{th} \sin \Omega + \eta' \Omega \tau \cos \Omega + \tau), \quad (13)
\]
where \( \Omega_{d+1} \equiv \Omega_d + \Omega_d \) and \( \eta' \equiv \tilde{n}_{th}(1 + 2\tilde{n}_{th} - n_d)/n_d \).

Note we have applied the formulas about the geometric series \( \sum_{x=0}^{\infty} xe^{-nx} = e^x/(e^x - 1)^2 \) and \( \sum_{n=1}^{\infty} n^2 e^{-nx} = e^x(1 + x^2)/(e^x - 1)^3 \). \( \tilde{n} \) in Eq. \((13)\) depends predominantly on the fast-frequency terms characterized by \( \Omega_d \) in a moderate time step \( \tau \). In the regime of \( T \sim 0.1 - 10 \) K, the term weighted by \( \eta' \Omega\tau \) overwhelms that weighted by \( \tilde{n}_{th} \). And this advantage expands with a larger \( \tau_{opt} \) given the initial or effective temperature of the resonator becomes lower, as evidenced by Fig. 1. We can therefore focus on the last term in Eq. \((13)\) to determine how to minimize \( \tilde{n} \). Consequently, we have
\[
\tau_{opt} = \frac{\pi}{\Omega_d + \Omega_{d+1}}. \quad (14)
\]
This result can be extended to the off-resonant situation by modifying the definition of \( \Omega_d \) in Eq. \((12)\) to be \( \sqrt{g^2 n_d + \Delta^2/4} \). The vertical black-dashed lines in Fig. 1 denote the optimal measurement-intervals given by Eq. \((14)\). It is found that the analytical expression is well suited to estimate the minimum values of average population in a wide temperature range. As demonstrated by both analytical and numerical results, a shorter measurement-interval is demanded to cool down a higher-temperature resonator. In the JC-like models, coupling a high-temperature resonator to a qubit could induce a faster transition from the ground state to the excited state of the qubit. A quick measurement would interrupt this process that has a negative effect on cooling.

Similar to the optimized interval \( \tau_{opt} = \tau(t) \) for the conditional measurement strategy [23], here \( \tau_{opt} \) is also updatable by substituting the time-evolved \( \Omega_{d} \) and \( \Omega_{d+1} \) to 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 the time-varying \( \tilde{n}(t) \) or \( p_n(t) \).

III. MEASUREMENT OPTIMIZATION

The thermal resonator could be steadily cooled down by the unconditional measurement strategy armed with the optimized measurement-interval in Eq. \((14)\). And the cooling is performed with a unit successful probability in the absence of the postselection over the measurement outcome. As demonstrated by the blue-solid line with circle markers and the orange-dotted line in Figs. 3(a) and 3(c), the complete UM and CM cooling strategies dominate in the cooling rate and the successful probability, respectively. It is therefore desired to find an optimized sequence of measurements, which is a hybrid of UM and CM, to compromise the cooling efficiency and the experimental cost. In this section, we present an algorithm that is assisted by the reinforcement learning to generate the optimized control sequence indicating when and which measurement is performed.

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 \( R \) from the cooperative cooling performance \( C \) in Eq. \((15)\). (b) Circuit model for our cooling algorithm based on the optimized UM and CM measurements. Starting from the thermal state, the resonator (the upper line) would be gradually cooled down to its ground state with implementation of the measurement on the detector (the lower line), which starts from the ground state. The measurement sequence can be obtained by the reinforcement learning.

The RL-optimization diagram is shown in Fig. 2(a), consisting of the “agent” part based on a series of neural network and the “environment” part performing the cooling-by-measurement actions on the quantum system. In the reinforcement learning, the agent is fed with a cluster of to-be-trained parameters, which would be learned and updated using the data collected through its interaction with the environment. In our architecture, the agent would choose an action, i.e., the conditional or the unconditional measurement, on the resonator, given its current state. Then the environment takes this action and returns the updated resonator-state \( p_n \) and a “reward” \( R \) after the measurement. The reward is generated by a function to estimate whether the action is good or bad, that would be used to update the parameters of the agent. During one “episode”, the agent would interact with the environment for \( N \) times, i.e., the number of measurements during the global 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 the total reward 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 demonstrated by 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 parameters updating, 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.

Notably, the logarithm function is used to obtain a positive value with almost the same order as $F$ and $P_g$ in magnitude. Then $\bar{n}(t)/\bar{n}_{th}$, $P_g$, and $F$ could be considered in a balanced manner. In fact, the average population could be reduced by several (normally less than 10) orders in magnitude under an efficient cooling protocol. In the EIT cooling [41], $\log_{10}[\bar{n}(t)/\bar{n}_{th}] \sim (2,3)$; and in the resolved sideband cooling [42], $\log_{10}[\bar{n}(t)/\bar{n}_{th}] \sim (4,5)$. It is instructive by Eq. (15) to find that a lower average population, a larger successful probability, and a higher ground-state fidelity yield a better cooling performance.

We consider to cool down a mechanical microresonator in gigahertz [43, 44] 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 resonator $T = 0.1$ K, it is found that the average population starts from $\bar{n}_{th} = 8.85$. The cooling performances under the sequences completely consisting of UM and CM are shown by the blue-solid lines with circle markers and the orange-dotted lines in Figs. 3(a)-(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 in magnitude [see Fig. 3(a)] and the ground-state fidelity is over $F > 0.9999$ [see Fig. 3(b)] at the cost of less than 10% of the successful probability [see Fig. 3(c)]. In sharp contrast, under $N = 16$ unconditional measurements, $\bar{n}$ is merely reduced to $\bar{n} \approx 3.36$ with a moderate fidelity $F \approx 0.78$, despite with a unit successful probability. With respect to all the individual quantifiers, i.e., $\bar{n}$, $F$, and $P_g$, the results under the hybrid sequences of UM and CM labelled by $S_k$, $k = 1, 2, 4$, are among the former two limits $S_u$ and $S_c$. As illustrated by Figs. 3(e), (f), and (g), the three sequences start from a CM strategies (indicated by 1), switch to UM (indicated by 0) after $k$ rounds of free-evolution and measurement, switch back to CM after a single round, and then continue the periodical arrangement. In comparison to the completed 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 the less probability $P_g$ that the detector remains in its measured subspace.

In terms of the cooperative cooling performance indicated by $C$ [see Fig. 3(d)], it is found that $C(S_1) > C(S_2) > C(S_4) > C(S_0)$ and yet $C(S_2) \approx C(S_0)$. The interpolation sequences could therefore have better cooperative cooling performance than the pure conditional-

\[
C = FP_g \log_{10} \frac{\bar{n}_{th}}{\bar{n}(t)}. \tag{15}
\]

The performance of any cooling-by-measurement strategy can be characterized or evaluated by the cooling ratio $\bar{n}(t)/\bar{n}_{th}$, the successful probability $P_g$ of the detector in the measured subspace, and the fidelity of the resonator in its ground state $F = \langle n = 0 | \rho_a(t) | n = 0 \rangle$ [18]. To compare various interpolation sequences of UM and CM in cooling performance and to evaluate the figure of merit for the reinforcement learning, we here define a cooperative cooling quantifier as

\[
C = FP_g \log_{10} \frac{\bar{n}_{th}}{\bar{n}(t)}. \tag{15}
\]

![Fig. 3. (a) Average population, (b) Fidelity of the resonator in its ground state, (c) Successful 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 completely consisting of UM and CM strategies, respectively. The green-solid lines, the red-dashed lines, and the purple-dot-dashed lines describe the hybrid sequences shown in (e), (f), and (g), respectively. The brown-solid lines with triangle markers labeled by $S_{opt}$ is the RL-optimized sequence presented in (h). The parameters are set as $\omega_a = 1.4$ GHz, $T = 0.1$ K, $g = 0.04\omega_a$, and $\Delta = 0.01\omega_a$.](image)
measurement-based protocol. The dependence of $C$ of arbitrary hybrid sequence on its proportion of CM strategies might not be monotonic. We are then motivated to find an optimized sequence with the help of the PPO algorithm. A typical optimized sequence of cooling strategies labeled by $S_{opt}$ 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 successful probability $P_g \approx 30\%$ (much larger than that by $S_c$), the optimized sequence achieves an overwhelming cooperative cooling performance $C = 2.73$ according to Eq. (15) over all the other measurement sequences. In other words, we attained the compromise of the cooling rate and the successful probability through the reinforcement learning. Notably, the optimized sequence is not unique, yet the current results of $\bar{n}$, $F$, $P_g$, and $C$ are almost invariant after dozen measurements as long as there is one CM in the first 5 rounds.

The RL-optimized algorithm applies to a wide range of the initial temperature of the resonator. Starting from different $n_{th}$ determined by the temperature, the average populations could be reduced by three to five orders in 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 resonator. Then both the relative magnitude in population reduction [see Fig. 4(a)] and the cooperative cooling performance [see Fig. 4(b)] manifest a monotonically decreasing behavior with increasing temperatures. Similar to Fig. 3(h), here we present in Figs. 4(c), (d), (e), and (f) the optimized sequences fully determined by the PPO algorithm, which still outperform any regular interpolated sequence in cooling quantifier $C$. Comparing these four sub-figures corresponding to various temperatures, it is interesting to find that a larger portion of the unconditional measurements is required in the optimized sequence for a higher temperature. It is consistent with the fact that under CM the successful probability $P_g$ to find a detector in its ground state decreases exponentially with increase of the temperature of the target resonator. Then more UMs are used to save a rapidly declining $P_g$ for obtaining a moderate $C$. In addition, for $T > 0.05$ K, the RL-optimized sequence always starts from a conditional measurement, which is important to have a significant cooling rate for $\bar{n}$ during the first several rounds of the whole sequence.

In our cooling algorithm, both UM and CM are performed on the detector following an optimized interval of unitary evolution. The nonunitary evolution induced by the conditional measurements could remarkably reduce the average population of the resonator, at the cost of a low successful probability [e.g., see the first several rounds in Fig 3(c)]. The unconditional measurements could reduce the average population with a much slower rate yet with a unit successful probability. Thus in general we anticipate to see more UMs than CMs in the first several rounds in an RL-optimized sequence, and more CMs than UMs in the remaining rounds.

IV. DISCUSSION AND CONCLUSION

We emphasised again that the preceding hybrid cooling sequences based on the conditional and unconditional measurements could be optimized in both global and local manners. Globally, we use the reinforcement learning to find the optimized order for UM and CM. The local optimization depends on the selected measurement intervals to obtain a minimum average-population $\bar{n}$ under a single measurement. Taking the unconditional case in Eq. (14) as an example, $\tau_{opt}(t)$ is not necessarily obtained by an instant feedback mechanism during a realistic practice. The measurement sequence $\{\tau_1(t_1), \tau_2(t_2), \cdots, \tau_N(t_N)\}$ can be actually obtained prior to the performance of the cooling measurements. $\tau_1(t_1)$ depends on the initial temperature as well as the population distribution $p_n$ and $\tau_k(t_k), k \geq 2$, 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 the sequential arrangement of the conditional and unconditional measurements. We analyse and compare the advantages and disadvantages of both CM and UM on the cooling rate and the success probability. We obtain analytically for the first time an expression of the optimized unconditional measurement-interval $\tau_{opt}^u = \pi/\Omega_d + \Omega_{d+1}$ in parallel to that of the conditional measurement [23]. Here the dominant Rabi frequency $\Omega_d$
depends on the dominant distribution of the resonator in the Fock state of \( n_d = \frac{k_B T}{h\omega_d} \) and the coupling strength between the target system and the detector. The compromise of the advantages of both measurement strategies gives rise to an optimized hybrid cooling algorithm assisted by the reinforcement learning. That is justified by the cooperative cooling performance as we defined to quantify the comprehensive cooling efficiency for arbitrary cooling-by-measurement strategy. Our work therefore pushes the cooling-by-measurement to an unattained degree in regard of efficiency and feasibility. Also, 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 contributes to more details of the proximal policy optimization, a typical reinforcement learning algorithm that we use to optimize the measurement sequence for cooling. PPO algorithm follows an “actor-critic” frame, in which actors receive the current state as an input and then outputs 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 differentiate “actor” and “policy” for simplicity.

As shown in Fig. 5, PPO algorithm has two actors (policies) \( \pi_{\text{old}}(\theta) \) and \( \pi_{\text{new}}(\theta') \) and one critic. Any of them is of an agent constructed by the neural networks (see Fig. 2) feathered with parameters \( \{\theta\} \). The two policies have the same structures in PPO. The old policy collects the sampling data through the 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 \( \pi_{\text{old}}(\theta) \); 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. There are \( N \) trajectories are collected in one episode, where \( N \) amounts to the number of actions required to complete the task. The critic takes both actions and states as input and outputs an advantage \( A_t \) 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 be trained to estimate the contributions of actions as precise as possible. In the mean time, according to the advantages to maximize a clipped surrogate objective function \( L_{\text{CLIP}}(\{\theta\}) \) [45], the new policy would update 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 \( \rho_d \)

\[
s_i = \{ p_0(t), p_1(t), p_2(t), \cdots, p_{n_e}(t) \},
\]

where \( n_e \) indicates the cutoff dimension for the resonator. The actions taken by the environment are selected from the set

\[
a_i \in \{0, 1\},
\]

where 0 and 1 represent the 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 the measurements, attaining 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 the unitary evolution lasting \( \tau_{\text{opt}} \in \{\tau_{\text{opt}}^{\dagger}, \tau_{\text{opt}}\} \), the measurement is performed on the detector. Then the average population \( \bar{\rho} \), the ground-state fidelity \( F \), and the successful probability \( P_s \) are obtained to calculate the cooperative cooling performance \( C \) by Eq. (15).

The reward function is set as a certain multiple of \( C \), \( R_d(s_i, a_i) = 100 \times C(s_i, a_i) \). After the measurement, the environment then returns the resonator state and the reward to the policies. When the training is completed, a policy \( \pi(\{\theta_{\text{opt}}\}) \) with a set of optimized parameters is achieved. The neural network armed with \( \{\theta_{\text{opt}}\} \) could then be used to generate optimized actions to cool down the current state.
Appendix B: Generation of optimized sequence

Algorithm 1: RL-optimized cooling procedure

Output: $S_{opt} = \{M_1, M_2, \ldots, M_N\}$ and $T = (\tau_{opt}(t_1), \tau_{opt}(t_2), \ldots, \tau_{opt}(t_N))$

Input: Temperature $T$

Initialize the thermal state $\rho_0 = \sum_n p_n |n\rangle \langle n|$ with $T$. Use PPO to train an optimized policy $\pi(\{\theta_{opt}\})$. For $i = 1, 2, \ldots, N$ do

Run the policy $\pi(\{\theta_{opt}\})$ on $\rho_0$ to generate $M_i$.

If $M_i = 0$, then

Calculate $\tau_{opt}(t_i) = \pi/(\Omega_d + \Omega_d+1)$ on $T_{eff}$. Get the cooling coefficients $|\alpha_n|^2$ and $|\beta_n|^2$.

UM: $\rho_a \left\langle \sum_n |\alpha_n|^2 p_n + |\beta_{n+1}|^2 p_{n+1}\rangle |n\rangle \langle n|$

Else if $M_i = 1$, then

Calculate $\tau_{opt}(t_i) = 1/\Omega_{th}$ on $T_{eff}$. Get the cooling coefficients $|\alpha_n|^2$.

CM: $\rho_a \left\langle \sum_n |\alpha_n|^2 p_n |n\rangle \langle n|/\left(\sum_n |\alpha_n|^2 p_n\right)$

end

Both the measurement order and the measurement-interval sequences could be regarded as the outputs 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 $\pi$) have been trained to be capable to select one of the two measurement strategies for the current state, which maximizes the cooperative cooling performance. And then the cooling procedure is formally launched. First, we run the policy $\pi(\{\theta_{opt}\})$ on $\rho_0 = \rho_{0h}$, which generates the first measurement strategy $M_1$, $M_1 \in \{0, 1\}$. Here 0 and 1 indicate UM and CM, respectively. If $M_1 = 0$, then $\tau_{opt}(t_1) = \tau_{opt}$ in Eq. (14) that could be obtained by the effective temperature $T_{eff}$ of the resonator (initially $T_{eff} = T$, 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 $M_1 = 1$, a conditional measurement will be implemented after an interval $\tau_{opt}(t_1) = \tau_{opt} = 1/\Omega_{th}(t)$ and the resonator state is modified according to Eq. (4). In the end of this round, we can find $T_{eff}$ by the current $p_n(t)$ and then go to the next round. After $N$ iterations, the optimized measurement sequence characterized by $S_{opt} = \{M_1, M_2, \ldots, M_N\}$ and $T = (\tau_{opt}(t_1), \tau_{opt}(t_2), \ldots, \tau_{opt}(t_N))$ appear as described in Fig. 3(h) and Figs. 4(e), (d), (c), (f). In practical implementations, these measurements according to $S_{opt}$ and $T$ can be acted on the detector without knowledge of the target-resonator state through any feedback mechanism.

[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 474, 589 (2011).
[5] K. Toyoda, R. Hijii, A. Noguchi, and S. Urabe, Hong-ou-mandel interference of two phonons in trapped ions, Nature 527, 74 (2015).
[6] M. Um, J. Zhang, D. L. 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. Gigon, 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 micro-mirror by radiation pressure, Nature 444, 67 (2006).
[12] X. Wang, S. Vinjanampathy, F. W. Strauch, and K. Jacobs, Ultracold 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 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 377, 500 (1995).
[15] G. Morigi, J. Essner, 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. Essner, and R. Blatt, Experimental demonstration of ground state laser cooling with electromagnetically induced transparency, Phys. Rev. Lett. 85, 5547 (2000).
[17] H. Nakazato, T. Takazawa, and K. Yuasa, Purification through zeno-like measurements, Phys. Rev. Lett. 90, 060401 (2003).
[18] Y. Li, L.-A. Wu, Y.-D. Wang, and L.-P. Yang, Nonde-
terministic ultrafast ground-state cooling of a mechanical resonator, Phys. Rev. B 84, 094502 (2011).

[19] 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. Photonics 8, 113 (2014).

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

[21] 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).

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

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

[24] 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).

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

[26] 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 529, 484 (2016).

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

[28] 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).

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

[30] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeb, Learning for continuous quantum error correction on superconducting qubits, arXiv, 2110.10378 (2022).

[31] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Reinforcement learning with neural networks for quantum feedback, Phys. Rev. X 8, 031084 (2018).

[32] A. Bolens and M. Heyl, Reinforcement learning for digital quantum simulation, Phys. Rev. Lett. 127, 110502 (2021).

[33] X. Yuan, J. Sun, J. Liu, Q. Zhao, and Y. Zhou, Quantum simulation with hybrid tensor networks, Phys. Rev. Lett. 127, 040501 (2021).

[34] 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).

[35] 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).

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

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

[39] 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).

[40] 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).

[41] J. F. Triana, A. F. Estrada, and L. A. Pachón, Ultrafast optimal sideband cooling under non-markovian evolution, Phys. Rev. Lett. 116, 183602 (2016).

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

[43] 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 478, 89 (2011).

[44] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, Proximal policy optimization algorithms, arXiv, 1707.06347 (2017).