Deep Reinforcement Learning Control of Quantum Cartpoles

Zhikang T. Wang,[1,*] Yuto Ashida,[2] and Masahito Ueda[1,3]

[1]*Department of Physics and Institute for Physics of Intelligence, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*
[2]*Department of Applied Physics, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan*
[3]*RIKEN Center for Emergent Matter Science (CEMS), Wako, Saitama 351-0198, Japan*

(Dated: October 25, 2019)

We generalize a standard benchmark of reinforcement learning, the classical cartpole balancing problem, to the quantum regime by stabilizing a particle in an unstable potential through measurement and feedback. We use the state-of-the-art deep reinforcement learning to stabilize the quantum cartpoles and find that our deep learning approach performs comparably to or better than other strategies in standard control theory. Our approach also applies to measurement-feedback cooling of quantum oscillators, showing the applicability of deep learning to general continuous-space quantum control.

Introduction — With unprecedented success of deep-learning-based artificial intelligence (AI) [1–4], researchers have started to consider machine learning as a viable tool to tackle existing problems [5–10]. Deep learning has begun to be used in physics recently in the form of either supervised learning, i.e. learning from existing data to predict physical quantities [11–14], or reinforcement learning (RL), i.e. learning from trial and error to pursue an intended target [15–18]. Reinforcement learning has been applied to quantum control problems that are not amenable to analytical methods [16, 18–20].

Meanwhile, quantum control has attracted increasing attention over the last few decades [21, 22] due to rapid experimental developments. Compared with classical control problems, there are far fewer results known for quantum cases due to the intrinsic complexity of quantum mechanics that makes analytic approaches difficult except for some simple cases [23, 24]. In real circumstances, numerical methods are used to find appropriate controls, such as GRAPE, QOCT and CRAB [25–28]. However, these methods are gradient-based and only guarantee the local optimality of their strategies [29], and they mostly work for isolated quantum systems which are unitarily deterministic. For stochastic systems, there is no known generally applicable approach. Thus, it is desirable to explore alternative generic methods for quantum control, and this is where machine learning is expected to be effective.

Deep reinforcement learning (RL) is a cutting-edge machine learning strategy that uses deep learning in its RL system. It is model-free and requires no prior knowledge, and often achieves the state-of-the-art performance. Recently, deep RL has been applied to a few quantum control problems, including manipulation of spin and spin chains [16], finding noise-robust control of qubits [17, 20] and designing quantum error correcting gates [18, 30]. In most of these cases, deep RL achieves success by demonstrating performance comparable or superior to conventional methods [16, 17], or it deals with some problems that are intractable by conventional means [18, 20, 30]. Despite its success, there appear to be not many existing works so far, and all of them have only considered discrete systems. However, there are continuous-space systems that need to be controlled, such as superconducting circuits and cavity optomechanical systems [31, 32]. Since continuous systems are typically more difficult to control, it is still unclear whether deep RL can deal with the continuous case. In this Letter, as a proof of principle, we demonstrate that deep RL can indeed solve simple continuous-space control problems, even in the presence of measurement backaction noise. To the best of our knowledge, this is the first time that deep RL is applied to continuous-space quantum control.

We construct quantum analogues to the classical cartpole balancing problem [33–35], which is arguably the best known RL benchmark, and apply deep RL to them. In the original classical cartpole problem, a controller moves a cart properly to prevent the pole from falling down (see Fig. 1(a)). It is a prototypical example of controlled stabilization of an unstable system. We consider measurement-feedback control of an unstable quantum system which is analogous to the classical cartpole (see Fig. 1(b)). The quantum cartpole is more difficult to control than the classical one, because measurement is needed to localize the cartpole state but it also generates backaction noise that perturbs the state, resulting in an inevitable tradeoff. Adopting the state-of-the-art deep Q-learning strategy [3, 36], we find that deep RL can indeed learn to stabilize the quantum cartpoles just like the classical one. In addition, when quantum-mechanical behaviour is significant in the system, deep RL outperforms other known control strategies. We also apply our strategy to the problems of measurement-

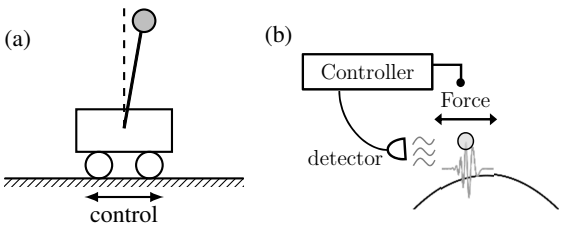

FIG. 1. (a) Classical cartpole system. A controller moves the cart to prevent the pole, which is an inverted pendulum, from falling down. (b) Generalized quantum cartpole system. A controller monitors a particle and applies forces to keep it at the top of a potential.

feedback cooling of oscillators. These results show a great potential of deep RL for continuous-space and stochastic quantum control, and from the perspective of machine learning, the quantum control problems themselves can also be a new type of hard benchmark task for reinforcement learning AI.

Deep Q-Learning — We use deep Q-learning as our RL strategy, implementing the deep Q-network (DQN) algorithm. In the following, we briefly introduce RL and DQN.

Given some reward r, RL aims to find a policy π that maximizes the accumulated total reward $\sum_t r_t$ in discretized time. The reward could be the obtained score in a game, artificially assigned ± 1 corresponding to a success/failure, or some control target in a control problem, e.g. the minus system energy for a cooling problem. In general, a RL AI interacts with a controllable environment, and the reward $r(s)$ is a function of the state s of the environment. RL starts with no prior knowledge and learns from trial and error to maximize $\sum_t r(s_t)$, by taking actions $a_t = \pi(s_t)$ that influence the time evolution $s_t \to s_{t+1}$, where π is the learned action policy.

Q-learning realizes RL through a *Q function* that represents the expected total future reward of a policy π, defined by [37]

$$Q^{\pi}(s_t, a_t) = r(s_t) + \mathbb{E}_{\{(s_{t+i}, a_{t+i})|\pi\}_{i=1}^{\infty}}\left[\sum_{i=1}^{\infty} \gamma_q^i\, r(s_{t+i})\right], \tag{1}$$

where γ_q is a discount factor (i.e., the future reward will be discounted by this factor) and satisfies $0 < \gamma_q < 1$, and the expectation is taken over trajectories of the action and the state of the environment. The parameter γ_q is manually set to ensure the convergence and is close to 1. For the optimal policy π^* which maximizes Q^{π^*}, Q^{π^*} satisfies the following Bellman equation [3, 38]

$$Q^{\pi^*}(s_t, a_t) = r(s_t) + \gamma_q\, \mathbb{E}_{s_{t+1}}\left[\max_{a_{t+1}} Q^{\pi^*}(s_{t+1}, a_{t+1})\right]. \tag{2}$$

In addition, the function Q^{π^*} that satisfies the Bellman equation is generally unique. Therefore, if we find a function that satisfies Eq. (2), we effectively obtain Q^{π^*} and can therefore derive the optimal policy π^*. This is the main idea of Q-learning.

Deep Q-learning uses a deep feedforward neural network as a universal function approximator to approximate Q^{π^*} [1, 4], and the network is named a deep Q-network (DQN) and denoted by $f_{\boldsymbol{\theta}}(s, a)$, where $\boldsymbol{\theta}$ is its internal parameters and the environment state s is the network input. Assuming that the space of actions a is discrete, the network outputs a value for each a at its output layer. The function $f_{\boldsymbol{\theta}}(s, a)$ approximates Q^{π^*} by modifying its internal parameters $\boldsymbol{\theta}$ through gradient descent, so that the squared difference between the two sides of Eq. (2) is minimized. In this way, $f_{\boldsymbol{\theta}}(s_t, a_t)$ approximately becomes the solution of Eq. (2) and is used to represent Q^{π^*} and derive the policy π^*.

As suggested in Ref. [36], we incorporate the state-of-the-art technical improvements of deep Q-learning, which include prioritized sampling, noisy networks, double Q-learning and the duel network structure [39–42]. See Ref. [36] for details.

Quantum Cartpole — As in Fig. 1(a), the classical cartpole is a simple system that can be stabilized by an external control. Instead of quantizing this two-body cartpole, we consider a one-body one-dimensional system that reproduces its stability. Specifically, we put a particle at the top of a potential and try to keep it at that unstable position by applying appropriate external forces. The Hamiltonian is

$$\hat{H}(F) = \frac{\hat{p}^2}{2m} + V(\hat{x}) - F\hat{x}, \tag{3}$$

where V is the potential and F is a controllable time-dependent force. We require V to be symmetric about 0 and $V \to -\infty$ for $x \to \pm\infty$. Also we require $|F|$ to be bounded from above by a constant $F_{\max}$. It is clear that under unitary evolution, the wavefunction cannot be stabilized at the top of the potential. To counteract delocalization, we impose a continuous position measurement on the particle [43], so that its wavefunction contracts due to state reduction and can be stabilized if controlled properly by means of measurement feedback control. The time-evolution equation of the particle state ρ following the Itô stochastic calculus is given by

$$d\rho = -\frac{i}{\hbar}[\hat{H}, \rho]dt - \frac{\gamma}{4}[\hat{x}, [\hat{x}, \rho]]dt + \sqrt{\frac{\gamma\eta}{2}}\{\hat{x} - \langle\hat{x}\rangle, \rho\}dW, \tag{4}$$

where dW is a Wiener increment sampled from the Gaussian distribution $\mathcal{N}(0, dt)$, γ is the measurement strength, $\eta \in [0, 1]$ is the measurement efficiency, and $\{\cdot, \cdot\}$ is the anti-commutator.

In the classical cartpole problem, the pole is judged to have fallen down if the tilting angle of the pole exceeds a prescribed threshold. We judge that the stabilization of the quantum cartpole fails if more than 50% of the probability distribution of the particle lies outside boundaries $\pm x_{\text{th}}$. Then, a controller tries to keep the cartpole stabilized without satisfying this failure criterion for as long time as possible. Due to random measurement backaction, open-loop control generally fails, and measurement-feedback control is necessary. The controller may either take the raw measurement outcomes or the post-measurement state as its input. In the deep RL case, this input becomes the input s of the neural network.

To implement Q-learning, we discretize the continuous force interval $[-F_{\max}, F_{\max}]$ into 21 equispaced forces that are used as a set of possible actions a, and we discretize the time into control steps. At each control step, the controller decides a force F to apply, and the force is kept constant until the next control step. Concerning the RL reward, we follow the original cartpole problem and set $r = 1$ if the system is stable and $Q = r = 0$ if it fails, so that the RL controller aims at an infinitely long-time stabilization. This is a general specification of our RL quantum cartpole task.

As a proof of principle, we investigate two simplest cases of V, i.e., an inverted harmonic potential ($V = -\frac{k}{2}\hat{x}^2$) and a quartic potential ($V = -\lambda\hat{x}^4$). A particle under continuous position measurement in a quadratic potential is well

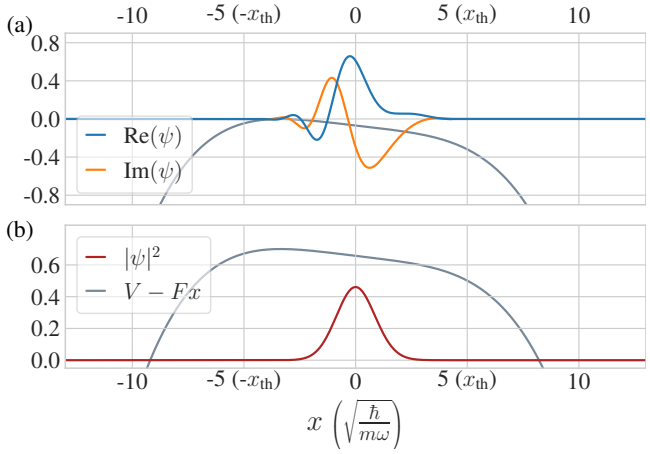

FIG. 2. Snapshots of the controlled quartic cartpole system. The blue and the orange curves are the real and imaginary parts of the wavefunction, and the grey and the red curves are the controlled potential $V - Fx$ and the probability density. The potential is only schematic and does not scale. See supplementary videos [47].

described by a Gaussian Wigner distribution [44, 45]. This is because the shape of the Wigner distribution is preserved by its time evolution, and the Gaussianity of the distribution increases with the Gaussian measurement. The state is fully characterized by the means $\langle\hat{x}\rangle, \langle\hat{p}\rangle$ and the covariances C_{xx}, C_{pp}, C_{xp}, where $C_{xx} := \langle\hat{x}^2\rangle - \langle\hat{x}\rangle^2$, $C_{pp} := \langle\hat{p}^2\rangle - \langle\hat{p}\rangle^2$ and $C_{xp} := \frac{1}{2}\langle\hat{x}\hat{p} + \hat{p}\hat{x}\rangle - \langle\hat{x}\rangle\langle\hat{p}\rangle$. As the covariances converge to steady-state values during the time evolution [24], the state becomes effectively described by two variables $\langle\hat{x}\rangle$ and $\langle\hat{p}\rangle$ only, and the effective time-evolution equations are

$$d\langle\hat{x}\rangle = \frac{\langle\hat{p}\rangle}{m}dt + \sqrt{2\gamma\eta}\, C_{xx}\, dW, \tag{5a}$$

$$d\langle\hat{p}\rangle = (-k\langle\hat{x}\rangle + F)dt + \sqrt{2\gamma\eta}\, C_{xp}\, dW, \tag{5b}$$

where dW is a Wiener increment as in Eq. (4). In contrast, for a quartic potential, a Gaussian state is not preserved; moreover, we have $d\langle\hat{p}\rangle = -4\lambda\langle\hat{x}^3\rangle dt \neq -4\lambda\langle\hat{x}\rangle^3 dt$, while a classical particle should satisfy $dp = -4\lambda x^3 dt$. Thus, the system exhibits intrinsic quantum-mechanical behaviour. In fact, it is known that the quartic system corresponds to the ϕ^4 theory in quantum field theory [46], and it is interesting to investigate how the quartic system can be controlled by RL.

Now we determine the system parameters to numerically simulate the systems and train the RL controller. For simplicity, we assume the perfect measurement efficiency $\eta = 1$ and only consider pure states. With complete measurement information, we can purify an arbitrary initially mixed state through repeated measurements. Therefore, we assume that the wavefunction is available. The time evolution of the state $|\psi\rangle$ is described by

$$d|\psi\rangle = \left[\left(-\frac{i}{\hbar}\hat{H} - \frac{\gamma}{4}(\hat{x} - \langle\hat{x}\rangle)^2\right)dt + \sqrt{\frac{\gamma}{2}}(\hat{x} - \langle\hat{x}\rangle)dW\right]|\psi\rangle. \tag{6}$$

The measurement strength γ is determined so that the wavefunction has a width comparable to the ground state of the non-inverted potential. The maximal control strength $F_{\rm max}$ is determined so that the peak of the controlled potential $V - Fx$ is allowed to move within several times the width of the wavefunction, and the failure threshold $x_{\rm th}$ is set to be the peak position of $V + F_{\rm max}x$, as demonstrated in Fig. 2. The number of control steps is roughly 30 per one oscillation period of the non-inverted potential. These settings ensure that the cartpole problems are non-trivial and there are non-vanishing probabilities of stabilization failure due to measurement backaction. The choice of RL hyperparameters mainly follows Ref. [36]. Specific parameter values and detailed settings of RL including the neural network specifications are given in Supplementary Material. To specify the input s of the neural network $f_{\boldsymbol{\theta}}$, and we test three cases: (i) the sequence of measurement outcomes in time; (ii) the wavefunction $|\psi\rangle$; (iii) the distribution moments $\langle\hat{x}\rangle, \langle\hat{p}\rangle, C_{xx}, \langle(\hat{x} - \langle\hat{x}\rangle)^3\rangle$, etc. The length of the measurement outcome sequence and the high-order cutoff for the distribution moments are determined so that the RL controller can learn. Specifically, the time for the measurement outcome sequence is 2 or 3 oscillation periods, and the distribution moments include up to the second moment for the harmonic system and up to the fifth moment for the quartic system.

To benchmark the above RL quantum control scheme by comparing it with known control protocols, we first test it on the measurement-feedback cooling problems with the corresponding non-inverted potentials ($+\frac{k}{2}x^2$ and $+\lambda x^4$). The RL reward becomes the minus system energy, and all other settings remain the same.

Results — We train the RL AI on the measurement-feedback cooling problems of quantum harmonic/quartic oscillators and on the stabilization problems of quantum harmonic/quartic cartpoles, using the deep Q-learning method, simulating the systems for 5×10^5 oscillation periods. The resulting dynamics of the controlled quantum systems are plotted and recorded as videos in Supplementary Material [47], clearly showing that the RL control is non-trivial. To evaluate the quality of its control, we benchmark and compare it with other control strategies. All the controls are discretized in the same manner to allow for fair comparison.

For the measurement-feedback cooling of harmonic oscillators, using the Gaussian approximation, the optimal control is rigorously given by the standard linear-quadratic-Gaussian (LQG) control theory [24, 48]. Therefore, we benchmark our deep RL control with this known optimal control, comparing the average energy of their controlled systems. As the LQG controller minimizes the squared position and momentum of a particle, it can also stabilize a cartpole system. Thus, we compare with this controller on the cartpole problem as well. The results are presented in Table I. Since the performance of the deep RL using the measurement outcomes as input is clearly worse, we do not consider the case of direct input of measurement outcomes below.

For quartic oscillators, the optimal control strategy is not

controller	input	cooling ($\langle\hat{n}\rangle$)	cartpole control (T)
deep RL	distribution moments	0.325 ± 0.002	42.124 ± 0.667
	wavefunction	0.327 ± 0.004	37.891 ± 0.598
	measurement outcomes	0.349 ± 0.005	24.789 ± 0.546
LQG control		0.326 ± 0.003	41.564 ± 0.658
no control		heats up to ∞	0.515 ± 0.006

TABLE I. Results of deep RL control of harmonic systems, i.e. the cooling of harmonic oscillators and stabilization of inverted harmonic cartpoles, compared with the linear-quadratic-Gaussian (LQG) control. The performance in cooling and that in cartpole stabilization are measured in terms of expected excitation numbers $\langle\hat{n}\rangle$ and average time before failure, in units of the oscillation periods T, with the estimation error shown.

controller	input	cooling ($\langle\hat{H}\rangle - E_0$)	cartpole (T)
deep RL	distribution moments	0.0068 ± 0.0001	14.455 ± 0.224
	wavefunction	0.0098 ± 0.0002	13.769 ± 0.210
controlled damping		0.0171 ± 0.0008	4.826 ± 0.272
LQG control		0.0329 ± 0.0010	14.138 ± 0.218
semiclassical control		0.0122 ± 0.0003	5.266 ± 0.374
no control		heats up to ∞	0.810 ± 0.007

TABLE II. Results of deep RL control of quartic systems compared with other control strategies. The performance in cooling is measured in terms of energies $\langle\hat{H}\rangle$ with $F = 0$, subtracted by the ground energy E_0. The quartic systems are constructed to be comparable to the harmonic ones. The units are the same as those in Table I.

known and we only compare with simple strategies, including controlled damping, the standard LQG control, and a semiclassical approximation control designed by us. The damping and the LQG coefficients are determined by grid search to obtain their best results. The control based on the semiclassical approximation assumes the Gaussianity and a fixed variance C_{xx} of the wavefunction so as to map $\langle\hat{x}\rangle$ and $\langle\hat{p}\rangle$ to the position x and momentum p of a classical particle, and it uses the optimal control for the derived classical particle to minimize the system energy (see Supplementary Material). The obtained results are presented in Table II.

As shown in Table I and II, our deep RL strategy successfully solves the quantum control problems. It can match the performance of the optimal control, and when no optimal control is known, it can outperform other strategies, demonstrating itself as a strong candidate for quantum control.

Discussion — We have demonstrated that the LQG control performs well for most of the cases. This is because the state is well-localized and Gaussian-like due to the harmonic potential or the strong continuous measurement in the cartpole systems, which allows for effective classical LQG control. For the cooling problem of quartic oscillators, with significant non-Gaussianity of the state, none of the conventional controllers performs well. In contrast, the model-free deep RL deals with arbitrary general potentials and handles both Gaussian and non-Gaussian states well, showing clear superiority over the other approaches.

Although deep RL can achieve good results, the performance depends on neural network inputs. This is presumably because the neural network cannot handle its input information precisely. For measurement outcome inputs, the neural network does not distinguish between recent and distant measurement outcomes and learn all of them equally, which may hinder its learning. For the wavefunction inputs, since physical observables are of the form $\langle\psi|\hat{O}|\psi\rangle$ that is quadratic in $|\psi\rangle$, the network using linear mappings and rectified linear units may not accurately evaluate relevant physical quantities. Therefore, it seems that within the current framework of deep learning, AI can be helpful to achieve the best performance on these problems.

An important future extension of this work is to consider mixed states. When the numerical experiments on mixed states are successful, the controller may be carried over to real experiments to control the systems, for example, in cavity optomechanical systems as in Ref. [49]. On the other hand, as discussed previously, the most important direction would be to use the deep RL strategy to deal with more complicated and realistic systems, such as superconducting circuits, which may improve the performance of realistic quantum devices.

From the perspective of RL, the quantum control problems can serve as quantum RL benchmarks. These problems are qualitatively different from most of the existing RL tasks. The quantum ones are difficult, stochastic, and of practical significance, while most of the current RL benchmarks are deterministic toy problems or Atari video games [35, 50].

Conclusion — We have constructed quantum analogues to the classical cartpole balancing problem using inverted harmonic and quartic potentials. We have used deep RL to tackle the quantum cartpole problems and measurement-feedback cooling of the corresponding quantum oscillators. The systems are numerically simulated and are stochastic and continuous-space. We have demonstrated that the deep RL can match or outperform other strategies in these problems, showing a great potential of deep RL for general continuous-space quantum control, and these quantum control tasks may also serve as RL benchmarks by themselves.

Acknowledgements — The authors thank Ryusuke Hamazaki for fruitful discussions. This work was supported by KAKENHI Grant No. JP18H01145 and a Grant-in-Aid for Scientific Research on Innovative Areas "Topological Materials Science (KAKENHI Grant No. JP15H05855) from the Japan Society for the Promotion of Science. Z. T. W. is supported by Global Science Graduate Course (GSGC) program of the University of Tokyo.

* wang@cat.phys.s.u-tokyo.ac.jp

[1] Y. LeCun, Y. Bengio, and G. Hinton, nature **521**, 436 (2015).
[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton,

et al., Nature **550**, 354 (2017).
[3] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, Nature **518**, 529 (2015).
[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016) http://www.deeplearningbook.org.
[5] C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, Molecular systems biology **12** (2016).
[6] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, A. A. Kalinin, B. T. Do, G. P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M. M. Hoffman, *et al.*, Journal of The Royal Society Interface **15**, 20170387 (2018).
[7] N. Ball and R. Brunner, International Journal of Modern Physics D **19** (2009), 10.1142/S0218271810017160.
[8] L. Wei and E. Roberts, Scientific Reports **8**, 7313 (2018).
[9] D. Zhang, X. Han, and C. Deng, CSEE Journal of Power and Energy Systems **4**, 362 (2018).
[10] P. Palittapongarnpim, P. Wittek, E. Zahedinejad, S. Vedaie, and B. C. Sanders, Neurocomputing **268**, 116 (2017).
[11] Y. Fujimoto, K. Fukushima, and K. Murase, Phys. Rev. D **98**, 023019 (2018).
[12] T. Mano and T. Ohtsuki, Journal of the Physical Society of Japan **86**, 113704 (2017).
[13] E. Barberio, B. Le, E. Richter-Was, Z. Was, J. Zaremba, and D. Zanzi, Phys. Rev. D **96**, 073002 (2017).
[14] T. Weiss and O. Romero-Isart, arXiv preprint arXiv:1906.08133 (2019).
[15] Y. Liu, T. Zhao, W. Ju, and S. Shi, Journal of Materiomics **3**, 159 (2017).
[16] M. Bukov, A. G. R. Day, D. Sels, P. Weinberg, A. Polkovnikov, and P. Mehta, Phys. Rev. X **8**, 031086 (2018).
[17] M. Y. Niu, S. Boixo, V. N. Smelyanskiy, and H. Neven, npj Quantum Information **5**, 33 (2019).
[18] T. Fösel, P. Tighineanu, T. Weiss, and F. Marquardt, Phys. Rev. X **8**, 031084 (2018).
[19] Z. An and D. L. Zhou, EPL (Europhysics Letters) **126**, 60002 (2019).
[20] R. Porotti, D. Tamascelli, M. Restelli, and E. Prati, Communications Physics **2**, 61 (2019).
[21] D. Dong and I. R. Petersen, IET Control Theory & Applications **4**, 2651 (2010).
[22] W. S. Warren, H. Rabitz, and M. Dahleh, Science **259**, 1581 (1993).
[23] N. Khaneja, R. Brockett, and S. J. Glaser, Phys. Rev. A **63**, 032308 (2001).
[24] A. C. Doherty and K. Jacobs, Phys. Rev. A **60**, 2700 (1999).
[25] N. Khaneja, T. Reiss, C. Kehlet, T. Schulte-Herbrggen, and S. J. Glaser, Journal of Magnetic Resonance **172**, 296 (2005).
[26] A. P. Peirce, M. A. Dahleh, and H. Rabitz, Phys. Rev. A **37**, 4950 (1988).
[27] J. Werschnik and E. K. U. Gross, Journal of Physics B: Atomic, Molecular and Optical Physics **40**, R175 (2007).
[28] P. Doria, T. Calarco, and S. Montangero, Phys. Rev. Lett. **106**, 190501 (2011).
[29] E. Zahedinejad, S. Schirmer, and B. C. Sanders, Phys. Rev. A **90**, 032310 (2014).
[30] H. Poulsen Nautrup, N. Delfosse, V. Dunjko, H. J. Briegel, and N. Friis, arXiv e-prints (2018), arXiv:1812.08451 [quant-ph].
[31] F. Motzoi, J. M. Gambetta, P. Rebentrost, and F. K. Wilhelm, Phys. Rev. Lett. **103**, 110501 (2009).
[32] M. Aspelmeyer, T. J. Kippenberg, and F. Marquardt, Rev. Mod. Phys. **86**, 1391 (2014).
[33] D. Michie and R. A. Chambers, Machine intelligence **2**, 137 (1968).
[34] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, in *International Conference on Machine Learning* (2016) pp. 1329–1338.
[35] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI Gym," (2016), arXiv:1606.01540.
[36] M. Hessel, J. Modayil, H. van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, arXiv e-prints , arXiv:1710.02298 (2017), arXiv:1710.02298 [cs.AI].
[37] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction* (MIT press, 2018).
[38] R. Bellman, Proceedings of the National Academy of Sciences **38**, 716 (1952), https://www.pnas.org/content/38/8/716.full.pdf.
[39] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, arXiv e-prints (2015), arXiv:1511.05952 [cs.LG].
[40] M. Fortunato, M. Gheshlaghi Azar, B. Piot, J. Menick, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, C. Blundell, and S. Legg, arXiv e-prints (2017), arXiv:1706.10295 [cs.LG].
[41] H. Van Hasselt, A. Guez, and D. Silver, in *Thirtieth AAAI conference on artificial intelligence* (2016).
[42] Z. Wang, T. Schaul, M. Hessel, H. van Hasselt, M. Lanctot, and N. de Freitas, arXiv e-prints (2015), arXiv:1511.06581 [cs.LG].
[43] L. Disi, Physics Letters A **129**, 419 (1988).
[44] W. H. Zurek, S. Habib, and J. P. Paz, Phys. Rev. Lett. **70**, 1187 (1993).
[45] K. Jacobs and P. L. Knight, Phys. Rev. A **57**, 2301 (1998).
[46] C. M. Bender and T. T. Wu, Phys. Rev. **184**, 1231 (1969).
[47] See supplementary videos at https://github.com/Z-T-WANG/DeepReinforcementLearningControlOfQuantumCartpol
[48] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods* (Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1990).
[49] M. Rossi, D. Mason, J. Chen, Y. Tsaturyan, and A. Schliesser, Nature **563**, 53 (2018).
[50] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, Journal of Artificial Intelligence Research **47**, 253 (2013).
[51] T. Salimans and D. P. Kingma, arXiv e-prints , arXiv:1602.07868 (2016), arXiv:1602.07868 [cs.LG].
[52] T. Tieleman and G. E. Hinton, "Neural networks for machine learning: Lecture 6a – overview of mini-batch gradient descent," http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf (2012).
[53] D. Kingma and J. Ba, International Conference on Learning Representations (2014).
[54] L. Ziyin, (2019), unpublished work.
[55] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, in *NIPS-W* (2017).

Supplementary Material

Parameters Used in Numerical Experiments

The parameters used in our numerical experiments are listed in Tables III and IV. Units thereof are shown in the parentheses, where a reference mass m_c and a reference angular frequency ω_c are used. The potentials of the harmonic system and the quartic system are respectively given by $V = \frac{k}{2}x^2$ and $V = \lambda x^4$. The period T of the harmonic oscillator is $\frac{2}{\omega_c}$. The control boundaries x_{th} of the quadratic cartpole and the quartic cartpole systems are 8 and 5 $\left(\sqrt{\frac{\hbar}{m_c\omega_c}}\right)$, respectively, and the measurement efficiency η is always set to 1.

	ω (ω_c)	m (m_c)	k ($m_c\omega_c^2$)	γ ($\frac{m_c\omega_c^2}{\hbar}$)	F_{max} ($\sqrt{\hbar m_c\omega_c^3}$)
harmonic oscillator	π	$\frac{1}{\pi}$	π	π	5π
quadratic cartpole	N/A	$\frac{1}{\pi}$	$-\pi$	2π	8π

TABLE III. Parameters of the quadratic systems used in the numerical experiments.

	m (m_c)	λ $\left(\frac{m_c^2\omega_c^3}{\hbar}\right)$	γ ($\frac{m_c\omega_c^2}{\hbar}$)	F_{max} ($\sqrt{\hbar m_c\omega_c^3}$)
quartic oscillator	$\frac{1}{\pi}$	$\frac{\pi}{25}$	$\frac{\pi}{100}$	5π
quartic cartpole	$\frac{1}{\pi}$	$-\frac{\pi}{100}$	π	5π

TABLE IV. Parameters of the quartic systems used in the numerical experiments.

Details of Implemented Controllers

The control force is tuned stepwise in every $\frac{T}{36}$ for each run of time interval T. During each step, the force is kept constant. The force is chosen from the range $[-F_{\text{max}}, F_{\text{max}}]$ which is discretized into 21 equispaced forces.

The linear-quadratic-Gaussian (LQG) controller takes two input values x and p at the beginning of a control step and decides a force in an attempt to satisfy the condition $p = -\sqrt{m|k|}\,x$ at the end of the control step. This controller effectively minimizes the quadratic control loss $\int(\frac{p^2}{2m} + \frac{kx^2}{2})\,dt$ and is known to be optimal in the presence of a Gaussian noise in the continuous limit. Here we have removed the control loss associated with its output control force to make a fair comparison with our reinforcement learning controller. For quantum systems, the expectation values $\langle\hat{x}\rangle$ and $\langle\hat{p}\rangle$ are used in place of classical x and p.

The damping controller takes p as an input value and attempts to exponentially reduce p in time. It aims to change p to $(1-\zeta)p$ at every control step.

The semiclassical controller assumes the Gaussianity of the state. Gaussianity implies the zero skewness $\langle(\hat{x} - \langle\hat{x}\rangle)^3\rangle = 0$ and the zero excess kurtosis $\frac{\langle(\hat{x}-\langle\hat{x}\rangle)^4\rangle}{C_{xx}^2} - 3 = 0$, where $C_{xx} := \langle\hat{x}^2\rangle - \langle\hat{x}\rangle^2$. Therefore, for the quartic systems, we have

$$\lambda\langle\hat{x}^4\rangle = \lambda(6C_{xx}\langle\hat{x}\rangle^2 + \langle\hat{x}\rangle^4 + 3C_{xx}^2)$$

Assuming a fixed C_{xx}, we replace $\langle\hat{x}\rangle$ and $\langle\hat{p}\rangle$ with classical position x and momentum p and set $V = 6C_{xx}\lambda x^2 + \lambda x^4$. Then we seek to minimize $\int(\frac{p^2}{2m} + |V|)\,dt$, which is the average energy of a quartic oscillator. If the system is deterministic and free from noise, this minimization is achieved if the condition $p = -\sqrt{|2m(6\lambda C_{xx} + \lambda x^2)|}x$ is satisfied. Thus, the semiclassical controller aims to make the controlled system satisfy this above condition at each control step.

When the LQG controller is applied to quadratic systems, k and m are simply the quadratic system parameters. The same holds true for the semiclassical controller on quartic systems. However, when we use the LQG controller and the damping controller on quartic systems, we do trial and error to find the optimal parameters ζ and k for the controllers to perform well, and we obtain the best results among all trials of parameters.

Settings of Reinforcement Learning

The deep-Q networks used in this research are the standard feedforward neural networks with linear connections and rectified linear unit (ReLU) activations. The neural networks that input distribution moments or wavefunction data are composed of 4 fully connected layers, where the last two layers are separated into two branches following Ref. [42]. The numbers of neurons are 512, 512, 256+128, and 21+1. The networks that input measurement outcomes are composed of 3 one-dimensional convolutional layers and 3 fully connected layers. The convolution kernel sizes and strides are (13,5), (11,4) and (9,4), and the number of filters are 32, 64 and 64. The neurons in the fully connected layers are 256, 256+128 and 21+1. The last two layers are always noisy layers with factorized noise (see Ref. [40]), and all fully connected layers learn normalized weight matrices as suggested in Ref. [51].

For the cooling tasks, the reinforcement learning reward is twice the minus energies of the controlled systems, namely, the energies rescaled by a factor of -2. For the cartpole tasks, the reward is always 10 when the system does not fail, and when the system fails, the expected future reward Q is set to zero. To ensure that the neural networks have output values of a moderate size, we rescale the received reward by a factor of $1-\gamma_q$ when we train the network.

During reinforcement learning, when training DQNs on cooling tasks, we discard the experience associated with high system energies; otherwise the network may not learn. This is because a high-energy experience typically results in a large training loss, which disturbs the learning of appropriate control at low energies especially at an early stage of training. The energy cutoff we employ is $\langle\hat{n}\rangle = 10$ for the harmonic oscillator and $20\hbar\omega_c$ for the quartic oscillator.

The networks that input measurement outcomes for the harmonic-oscillator problem and the quadratic-cartpole problem respectively input time sequences of measurements of lengths $\frac{3}{2}T$ and T, taking one measurement outcome at each numerical simulation step of the quantum system. In our case, this corresponds to 4320 and 5760 numbers. The networks that input wavefunction data simply separate the complex wavefunctions into real and imaginary parts and use them as network inputs.

We follow Ref. [39] to implement prioritized memory replay buffers for reinforcement learning. The prioritization parameter α is 0.4 except for the inverted harmonic cartpole task, where α is 0.6. The size of the memory replay buffer is about 3×10^5T for the cases of distribution moments and wavefunction inputs, and for the case of measurement outcome inputs, the size is about 3×10^4T. The total simulated time is about 7×10^5T. When the memory replay buffer becomes full, we discard previous experience using the first-in first-out method. We find that if we only discard experience associated with low training loss, the final performance of learning would deteriorate.

The reinforcement learning actors adopt the ϵ-greedy strategy to take actions and gather experiences for training [3]. The training algorithm uses the double Q learning strategy as in Ref. [41], and the update period of target networks is set to be 300 times the gradient descent step. Specifically, we set the update period to be 30 at the beginning of training, and when it starts to learn, we gradually change it to 300. The gradient descent algorithm we adopt is a slightly modified version of RMSprop, which implements a momentum-invariant update step size and incorporates the bias correction terms as in Adam [52–54]. The batch size is 512, and each experience is learned 8 times on average. The learning rate starts from 4×10^{-4} and decreases down to 1×10^{-6}, except for the case of wavefunction inputs where the learning rate starts from 4×10^{-3}. We use Pytorch as our deep learning library and use its default initialization for network parameters [55]. The Q-learning parameter γ_q is 0.99, and it is changed to 0.998 after the decay of the learning rate, and for the harmonic cartpole problem γ_q is further increased to 0.9996. We find that fine-tunings of the RL strategy and the parameters often cause a change of performance, suggesting that the quantum control problems are useful RL benchmarks that provide meaningful and insightful results.

Evaluation of Performances

The quantum systems are initialized as small Gaussian wave packets with zero momentum at the center, except for the quartic oscillator system. For the quartic oscillator, we implement a two-stage initialization by first initializing a Gaussian state with a small momentum and then letting it evolve for $10T$. In this manner, we obtain a sufficiently non-Gaussian state, and we use this state as an initial state of the quartic oscillator that is to be controlled.

After the systems are initialized, we apply different controls and record the system behaviour. For a cartpole system, we simply record how much time elapses before the system fails, and we repeat about 4000 times to obtain an estimate for the average control time and use it as a measure of the performance of the control. We find that the recorded time approximately follows an exponential distribution, and the variances of the estimates are large. For the cooling problems, we estimate the energies of the controlled systems as a measure of the control performance. To alleviate the effect of initialization, for the harmonic oscillator, we start to sample the system energy at $20T$ after initialization, and we start to sample the quartic system energy at $30T$. To reduce the correlations among the samples, a new sample is taken every $5T$ for the harmonic oscillator and every $10T$ for the quartic oscillator. The systems are simulated up to $50T$ and then reinitialized.

The ground-state energy E_0 of the quartic oscillator shown in Table II of the main text is approximately $0.2285\hbar\omega$, and the first and the second excited state energies are $0.8186\hbar\omega$ and $1.6063\hbar\omega$, which are obtained by exact diagonalization. Because the measurement squeezes the wavefunction in position space, it is impossible for the state to stay at the ground state energy, and the lowest possible energy under control is always larger than E_0.