Robust two-qubit gate with reinforcement learning and dropout

Tian-Niu Xu [1], Yongcheng Ding [2,3], José D. Martín-Guerrero [4,5,*] and Xi Chen [6,†]

[1] *College of Information Science and Engineering, Qilu Normal University, Jinan 250013, China*

[2] *Department of Physical Chemistry, University of the Basque Country UPV/EHU, Apartado 644, 48080 Bilbao, Spain*

[3] *Institute for Science and Technology, Department of Physics, Shanghai University, Shanghai 200444, China*

[4] *IDAL, Electronic Engineering Department, ETSE-UV, University of Valencia, Avinguda Universitat s/n, 46100 Burjassot, Valencia, Spain*

[5] *Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), 46022 Valencia, Spain*

[6] *Instituto de Ciencia de Materiales de Madrid (CSIC), Cantoblanco, E-28049 Madrid, Spain*

(Received 12 December 2023; revised 31 July 2024; accepted 30 August 2024; published 13 September 2024)

In the realm of quantum control, reinforcement learning, a prominent branch of machine learning, emerges as a competitive candidate for computer-assisted optimal experiment design. This paper investigates the extent to which guidance from human experts is necessary for effectively implementing reinforcement learning in the design of quantum control protocols. Specifically, our focus lies on engineering a robust two-qubit gate, utilizing a combination of analytical solutions as prior knowledge and techniques from computer science. Through thorough benchmarking of various models, we identify dropout—a widely used method for mitigating overfitting in machine learning—as a particularly robust approach. Our findings demonstrate the potential of integrating computer science concepts to propel the development of advanced quantum technologies.

DOI: 10.1103/PhysRevA.110.032614

I. INTRODUCTION

With advancements in algorithms, computer-designed experiments [1,2] have disrupted the traditional notion that human experts are solely responsible for proposing new experiments. This raises the fundamental question: How much information must human experts provide to algorithms for efficient setup searching? This question becomes particularly pertinent when considering computer-assisted experiment design as a simplified version, where one is aware of the experimental setup but seeks specific protocols to achieve desired goals. In this scenario, reinforcement learning (RL) emerges as a natural choice, allowing agents to explore solutions by interacting with the environment. RL [3] has seen significant applications in studying physics problems over the past decade [4], primarily focusing on quantum control problems [5–13]. Meanwhile, its collaboration with artificial neural networks (ANNs) has been employed to solve pulse design for quantum state preparation [14–16], gate operation [17–20], and the quantum Szilard engine [21]. Furthermore, RL has been utilized in information retrieval, controlling the measurement process in quantum metrology, and extracting real-time quantum information for closed-loop quantum control [22,23]. Additionally, quantum RL has also been developed by replacing the classical models with quantum systems due to the model-free nature [24].

To explore the potential of RL in the presence of limited physical knowledge, a targeted selection of a quantum control task is imperative. In light of compelling motivations, our focus is directed towards the design of robust quantum gates. Simulating the dynamics of quantum systems for building the RL environments becomes exponentially complex as the system size increases. However, by numerically calculating the propagator of a Hamiltonian with only a few energy levels per episode, we can present fair comparisons among various models with affordable computational resources. From the physical perspective, numerous approaches to achieving robustness have been proposed, such as adiabatic quantum control [25], pulse-shaping engineering [26–28], composite pulses [29–33], and shortcuts to adiabaticity [34,35]. Among them, geometric quantum computing, which leverages topological properties to cancel the effects of systematic errors, has shown promise across various quantum systems [36–42], and has been combined with machine learning (ML) methods [43–45]. It is indeed a well-studied example that provides sufficient information for encoding in the environment to explore the necessity of physics knowledge. Based on our research background, we believe it is possible to design robust quantum gates without considering explicit methods, e.g., accumulating geometric phase or canceling error sensitivity. To accomplish this, we have the option to select from various quantum platforms that are susceptible to noise and systematic errors. While other quantum systems could also be of interest, for the purposes of this paper, we intentionally focus on liquid nuclear magnetic resonance (NMR) and Josephson charge qubits as used in Refs. [37–39].

The paper is structured as follows. In Sec. II, we propose an RL model that encodes pertinent information, including operation time and tunable parameter ranges, along with the analytical solution for the geometric two-qubit gate, serving as an example of robust quantum gate design based on physical knowledge. Sections III and IV focus on benchmarking the

*Contact author: jose.d.martin@uv.es

†Contact author: xi.chen@csic.es

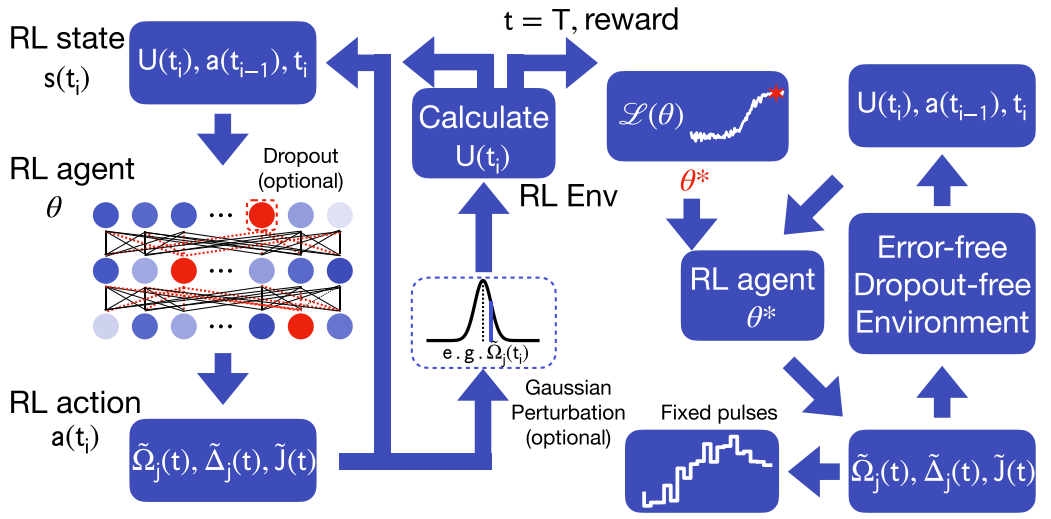

FIG. 1. Schematic diagram of the DRL approach to the robust two-qubit gate. The state encodes the real part and imaginary part of elements in the propagator $U(t_i)$, the last action $a(t_{i-1})$, and the normalized systematic time $t_i = i/N_{\max} i\delta T$. The state is sent to the input layer of the ANN as the DRL agent θ, outputting action to control the system for the next time step. The agent is trained to accumulate maximum reward, targeting the two-qubit gate $R_{YY}(\pi/4)$ at the end of each episode. In our protocol, we use Gaussian perturbation on the action and dropout in the ANN to obtain robustness against systematic errors. Once the training process converges by maximizing the reward-dependent objective function $\mathcal{L}(\theta)$, the optimal model θ^* is tested in the error-free and drop-out-free environment for obtaining the pulses for robust quantum gates. The pulses are fixed and later evaluated under systematic errors.

performances of RL models trained under different settings and methods from computer science, such as perturbation on nodes or dropout. We find out that the dropout itself leads the model to robustness against systematic errors without guidance from human experts. Finally, we provide concluding remarks in Sec. V, discussing potential avenues for further research and the implications of our findings for practical quantum information processing and quantum computing.

II. DEEP REINFORCEMENT LEARNING

In this section, we shall first outline our task and present the model, which includes the Hamiltonian and its preliminaries. Our primary objective is to design the robust entangling gate $R_{YY}(\pi/4)$ for two qubits; see review [46]. One well-known approach is the geometric gate, known for its robustness mechanism induced by topological protection. Acting as a perfect entangling gate that transforms separable states into a maximally entangled state, it serves the same critical character as the controlled-NOT (CNOT) gate does in constructing the universal gate set for quantum computing. The gate exploits the geometric phase acquired by a quantum state during its cyclic evolution in parameter space, thereby offering inherent resilience against systematic errors on the controlling parameters. The design of Hamiltonians plays a crucial role in shaping the geometric properties of the system.

In this context, the general Hamiltonian appears in the study of Josephson charge qubits or liquid-state NMR ($\hbar = 1$) [37]:

$$H = H_1 \otimes I_2 + I_1 \otimes H_2 + \frac{J}{2} Z_1 \otimes Z_2,$$
$$H_{1,2} = \tfrac{1}{2}[\Omega_{1,2}\cos(\omega t)X_{1,2} + \Omega_{1,2}\sin(\omega t)Y_{1,2} + \Delta Z_{1,2}], \quad (1)$$

where Ω_i, Δ_i, X_i, Y_i, and Z_i represent the Rabi frequency, detuning, and Pauli operators on the ith qubit, respectively, and J denotes the exchange energy between two qubits. In general, two-qubit Hamiltonians are elements of su(4), which makes an analytic approach challenging. However, the Hamiltonian (1) is restricted to a subalgebra of su(4) [47], systematically reducing the complexity. Thus, the analytical analysis can be found in Appendix A by utilizing dynamical invariants [47,48], and the fast nonadiabatic gate can be designed using inverse engineering accordingly.

With this knowledge, we turn to deep reinforcement learning (DRL), as shown in its workflow (Fig. 1). The insight is that a RL model can mimic the behavior of creatures that interact with the environment, being educated by reward to alter its decision based on its observation. In this way, an environment should be defined, specifying the task and its relevant state and action spaces, which can be either continuous or discrete. Next, an agent equipped with a deep neural network is trained to interact with the environment, making sequential decisions and receiving feedback in the form of rewards. Through an iterative process of exploration and exploitation, the agent refines its policy, using techniques like value-based methods or policy gradients, to maximize long-term cumulative rewards. After the training process, the agent can be deployed to perform the desired task autonomously, exhibiting learned behaviors that demonstrate the efficacy of DRL.

To be more specific, our goal is to train a neural network as an agent to design the robust entangling gate $R_{YY}(\pi/4) = \exp(i\pi Y_1 \otimes Y_2/4)$ for two qubits. We aim to benchmark the effects of physical priors or methods from the RL community that induce robustness. To explore all possibilities, we unbound the constraints on X_i and Y_i in Eq. (1), resulting in the most general Hamiltonian for the agent to

explore:

$$H_{\mathrm{DRL}} = \sum_{j=1}^{2} H_j + \frac{J(t)}{2} Z_1 \otimes Z_2, \tag{2}$$

where

$$H_j = \Omega_j(t)\cos(\omega t)X_j + \Omega_j(t)\sin(\omega t)Y_j + \Delta_j(t)Z_j,$$

and the tunable ranges of the parameters are $\Omega_j(t) \in [0, \Omega_{\max}]$, $\Delta_j(t) \in [-\Delta_{\max}, \Delta_{\max}]$, and $J(t) \in [0, J_{\max}]$. We then normalize these control parameters to [0,1] for ANN encoding in the action or state. Once the total operation time T is set, we obtain the length of each time step $\delta T = T/N_{\max}$ by bounding the maximum time steps per episode. At the ith time step, the accumulated propagator reads

$$U(t_i, t_0) = \mathcal{T} \prod_{j=0}^{i-1} U(t_{j+1}, t_j) = \mathcal{T} \prod_{j=0}^{i-1} \exp[-iH_{\mathrm{DRL}}(j\delta T)\delta T], \tag{3}$$

which should be close enough to $\exp(-i\pi Y_1 \otimes Y_2)$ at the last time step after the training. The neural network acts as the general function approximator representing the policy $\pi(a|s)$, where the input layer observes the state $s(t_i)$, and the output layer provides the action $a(t_i)$ for evolving the state to the next time step. The state consists of the ordered elements of the propagator at the present, the last action, and the normalized systematic time as

$$s(t_i) = \{\mathrm{Re}[U_{j,k}(t_i)], \mathrm{Im}[U_{j,k}(t_i)], a(t_{i-1}), t_i\}, \tag{4}$$

with $t_i = i/N_{\max}$ and $j, k \in \{1, 2, 3, 4\}$. The action contains the renormalized control parameters in Eq. (2) as

$$a(t_i) = \{\tilde{\Omega}_1(t_i), \tilde{\Delta}_1(t_i), \tilde{\Omega}_2(t_i), \tilde{\Delta}_2(t_i), \tilde{J}(t_i)\}, \tag{5}$$

where these parameters should be restored as $\tilde{\Omega}_j(t_i) = \Omega_j(t_i)/\Omega_{\max}$, $\tilde{\Delta}_j(t_i) = (\Delta_j(t_i) + \Delta_{\max})/2\Delta_{\max}$, and $\tilde{J}(t_i) = J(t_i)/J_{\max}$ to simulate the evolution of the environment to the next time step. The performance and training of RL are highly related to the reward function or pretraining. For a fair comparison among all models and to avoid cherry picking, we do not perform any pretraining and define a very simple reward function by rewarding the agent with $r = -\log_{10}(1 - F)$ in terms of logarithmic infidelity at the end of each episode, where the gate fidelity is defined as

$$F = \left| \frac{\mathrm{Tr}[\exp(-i\pi Y_1 \otimes Y_2/4)U(T, 0)]}{\dim(U)} \right|^2. \tag{6}$$

Using the baseline proximal policy optimization (PPO) algorithm [49], we train the agent to maximize the accumulated reward, aiming for the fast nonadiabatic yet robust quantum gates we seek.

III. NUMERICAL EXPERIMENTS

Now we present the specific settings for our numerical experiments and subsequently showcase our findings while analyzing the gate fidelities. Initially, we employ the most general Hamiltonian depicted in (2) to evolve the quantum states, with the primary objective of assessing the capabilities of DRL in scenarios with limited prior physical knowledge.

The first setting is designed to efficiently obtain the gate operation. In this setting, we reward the agent with $r = -\log_{10}(1 - F)$ after each time step in one episode, and conditionally end the episode to calculate the total reward if the maximum time step is met or $F > 0.99$ is satisfied after any time step. Together with the discount rate γ on the reward $\gamma^i r(t_i)$, this design encourages the agent to achieve the target more quickly by placing higher value on short-term rewards. Meanwhile, we clip the reward function by a value of 1 if the gate fidelity falls within the range [0.95,0.99), and provide an additional bonus of 10 if it reaches the threshold of $F = 0.99$. The clipped reward function can expedite training with the PPO algorithm without affecting our objectives. We halt the training and evaluate the model once it exceeds the fidelity threshold.

In Fig. 2(a), we illustrate the gate's robustness against over-rotating errors $[\Omega_{1,2} \to \Omega_{1,2}(1 + \delta\Omega)]$ and off-resonance errors $[\Delta_{1,2} \to \Delta_{1,2}(1 + \delta\Delta)]$ under such a setting using a heat map. As expected, the region enclosed by the contour $F = 0.99$ is relatively small. However, this solution is not the time-optimal solution with minimal robustness due to the batch method employed in training the DRL agent. It can be easily verified that a time-optimal solution consists of single-qubit resonant pulses that exchange Y and Z and a two-qubit pulse of ZZ. In other words, DRL naturally introduces robustness into quantum control through the standard batch method, which is a commonly used technique in ML, being applicable to gradient-based optimization as well [50,51]. This result serves as the baseline for robustness to study the effects of other methods and settings.

A. Gaussian perturbation on nodes

To enhance robustness, we engineer the environment illustrated in Fig. 2(a) with minimal human expert input. Instead of modifying the reward function by introducing Lagrangian multipliers for error sensitivity, we investigate whether perturbations applied to nodes of the ANN can achieve a similar effect. These perturbations are exclusively imposed on nodes within the output layer due to the inherent lack of interpretability of the network. The underlying idea is that if the DRL agent can be trained effectively in such an environment, it should exhibit greater robustness against over-rotating and off-resonance errors, as the environment simulates their effects by perturbing the agent's actions. In essence, we expect the agent to accumulate more rewards by successfully navigating this customized challenge.

We introduce perturbations to the action nodes using random Gaussian variables: $\tilde{\Omega} \to \tilde{\Omega}[1 + N(\mu = 0, \sigma = 0.1)]$ and $\tilde{\Delta} \to \tilde{\Delta}[1 + N(\mu = 0, \sigma = 0.1)]$; see Eq. (5). In the entire context, we omit the subscript j and the dependence t_i for brevity. It is worthwhile to mention that we perform miscellaneous numerical experiments with different settings of standard deviation $\sigma = 0.02, 0.05, 0.1, 0.2$, and observe the tradeoff between the convergence of the model and robustness. The less the standard deviation is, the easier it is to obtain a converged model but the robustness is also reduced, and vice versa. Thus, we select the standard deviation $\sigma = 0.1$ as a trade-off. In each episode, once the Gaussian perturbation is randomized at the initial time step, it remains unchanged until

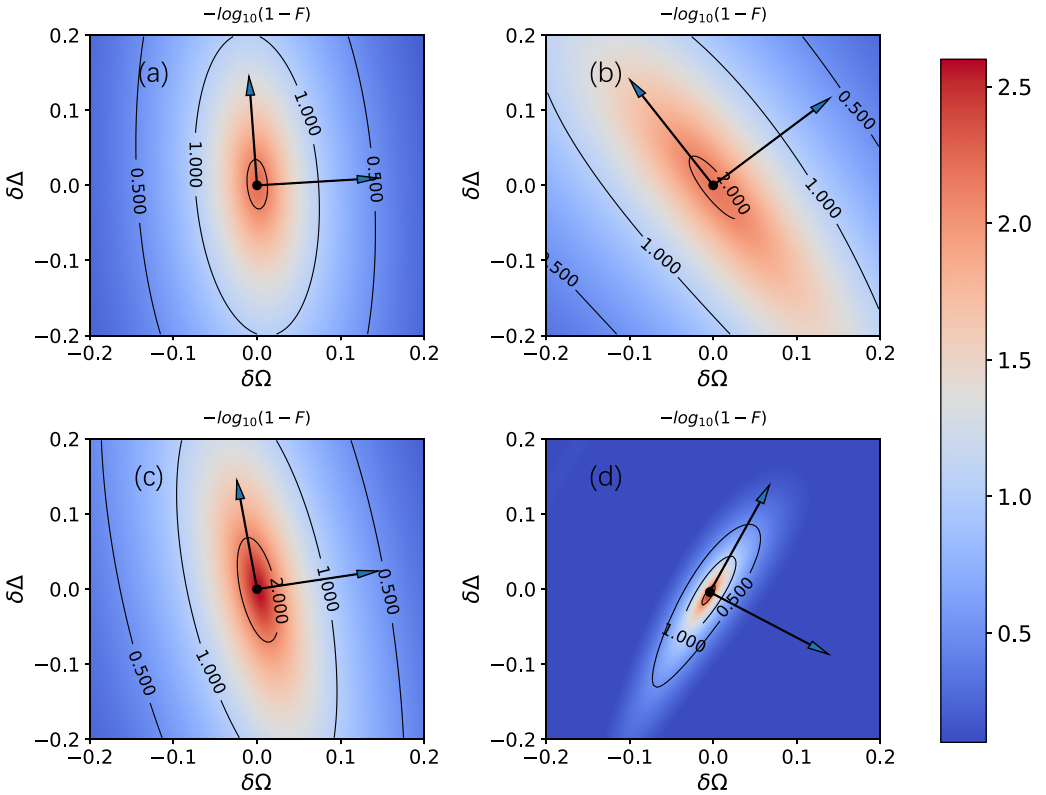

FIG. 2. Robustness of the entangling gate $R_{YY}(\pi/4)$ against over-rotating errors ($\Omega_{1,2} \to \Omega_{1,2} + \Omega_{\max}\delta\Omega$) and off-resonance errors ($\Delta_{1,2} \to \Delta_{1,2} + \Delta_{\max}\delta\Delta$). The gate fidelity is defined as $F = |\frac{1}{4}\mathrm{Tr}[\exp(-i\pi Y_1 \otimes Y_2/4)U(T,0)]|^2$, where $U(T,0)$ represents the propagator under these errors. To emphasize the impact of systematic errors, we present a heat map displaying the logarithmic infidelity. Models in (a)–(c) are trained in an environment featuring the general two-qubit Hamiltonian depicted in (2). These models are rewarded with logarithmic infidelity at the end of each episode, and they operate without prior human expert knowledge about robustness. We employ three different techniques from computer science: the standard batch method in (a), Gaussian perturbation on the nodes in the output layer in (b), and dropout in (c). (d) Evaluated by operating the analytical nonadiabatic geometric gate; see Appendix A. The fidelities at zero noise $F_{(a)} = 0.9926$, $F_{(b)} = 0.9917$, $F_{(c)} = 0.9967$, and $F_{(d)} = 0.9962$ are calculated and compared. Parameters: Random Gaussian perturbations on nodes following $\tilde{\Omega} \to \tilde{\Omega}[1 + N(\mu = 0, \sigma = 0.1)]$ and $\tilde{\Delta} \to \tilde{\Delta}[1 + N(\mu = 0, \sigma = 0.1)]$, and a dropout rate of $\kappa = 0.1$.

the end. Notably, we observe that the DRL agent converges, as evidenced by the activation of the additional bonus once the gate fidelity surpasses the threshold under these perturbations. Consequently, we obtain control pulses by running the model in an error-free environment. Although the corresponding gate fidelity does not exceed $F = 0.99$ due to the model converging to a solution with systematic errors induced by statistical fluctuations and the batch method, the characteristic of robustness remains intact.

In Fig. 2(b), we rectify the deviated model by identifying the maximum fidelity point on the original heat map and adjusting the control parameters to relocate the maximum fidelity to the center $(\delta\Omega, \delta\Delta) = (0, 0)$ for illustrative purposes. This adjustment reveals a significantly larger area enclosed by the contour of $F = 0.99$, demonstrating that Gaussian perturbations applied to nodes provide additional robustness in the design of quantum control using DRL. Moreover, the principal axis of the ellipse aligns with the diagonal direction because both types of systematic errors are included during the training process.

B. Dropout

Dropout is a common technique to mitigate overfitting during the training of ANNs [52]. Larger weights in the network tend to overfit the training data more easily [53]. The concept behind dropout is that probabilistically disconnecting nodes in the network serves as a simple regularization method to reduce weight magnitudes and perform model averaging. While this method was originally proposed for classical supervised learning, physicists have suggested its quantum analog [54] to reduce circuit complexity in quantum algorithms, aligning with the technique's underlying philosophy. This insight motivates us to investigate its performance in quantum control with DRL. The key idea is that we can emulate various systematic errors by randomly disconnecting nodes during training. Although we lack specific information about the correlation between a particular node and robustness against over-rotating or off-resonance errors, we believe that averaging the network across weights and biases should enhance gate robustness. The training process should converge when the dropout rate is appropriately set based on the network size and connectivity.

Instead of applying Gaussian perturbations to the output layer, we randomly disconnect each node with a dropout rate of $\kappa = 0.1$. The setting of κ is empirical and optimized, which has a similar effect on the model convergence and robustness as the standard deviation of Gaussian perturbation. While the ANN remains a black box, the lack of interpretability of specific nodes does not hinder our approach. Once the model converges, we disable dropout to obtain the control pulse, which is expected to be robust against all types of errors in the output layer, including variations in the length and magnitude of ZZ interactions. After centralizing the heat map using the same method, we achieve the best gate performance, as depicted in Fig. 2(c). Based on the results of all our experiments, we conclude that DRL can explore solutions with robustness without the need for human expert intervention, relying instead on techniques from computer science. Particularly noteworthy is the fact that robustness can be introduced simply by implementing dropout, a regularization method, rather than modeling systematic errors in the environment or encoding criteria for robustness, such as error amplitudes or fidelity sensitivity, into the reward function. In experiments with dropout, the RL scheme no longer has knowledge on any interpretable error. The network disconnects nodes randomly as a black box.

After demonstrating the effectiveness of DRL for generating robust two-qubit gates without human guidance, we now proceed to compare the results obtained from different techniques, including the analytical solution derived from nonadiabatic geometric theory, which involves the cancellation of dynamical phases [38,39]. The gate fidelities at zero noise for the DRL models are $F_{(a)} = 0.9926$, $F_{(b)} = 0.9917$, and $F_{(c)} = 0.9967$, corresponding to the models evaluated using the standard batch method, Gaussian perturbation on the nodes, and dropout, respectively, as depicted in Figs. 2(a)–2(c). As indicated in Fig. 2(d), the heat map presents the robustness of the fast nonadiabatic geometric gate, yielding $F_{(d)} = 0.9962$ error free for comparison.

To present the quantitative study on robustness, we define the differences in fidelity as

$$\mathcal{A} = (\mathcal{A}_{//}, \mathcal{A}_{\perp}) = \left(\frac{F(0) - F(\mathbf{r}_{//})}{|\mathbf{r}|}, \frac{F(0) - F(\mathbf{r}_{\perp})}{|\mathbf{r}|} \right), \quad (7)$$

where $\mathbf{r}$ is the vector from the maximum of fidelity (as the pivot) to an arbitrary point in the parameter space of systematic errors. $\mathbf{r}_{//}$ and $\mathbf{r}_{\perp}$ denote the direction along the major and minor axes of the elliptical contour, respectively. Therefore, the smaller $\mathcal{A}_{//}$ and $\mathcal{A}_{\perp}$ are, the more robust the corresponding control protocol is. We calculate the fidelity differences in Figs. 2(a)–2(d): $\mathcal{A}(a) = (0.6773, 0.088)$, $\mathcal{A}(b) = (0.3039, 0.032)$, $\mathcal{A}(c) = (0.3526, 0.037)$, and $\mathcal{A}(d) = (3.626, 1.453)$, showcasing that one can obtain robust quantum gates by both perturbing the nodes and dropout compared to the standard batch method as RL baseline and explicit geometric quantum gates with a distance of $|\mathbf{r}| = 0.04$. To maintain objectivity and prevent cherry picking, we train five models with random seeds ranging from 1 to 5 under each setting (see Appendix B) and average the figures of merit to derive a general conclusion. Although the Gaussian perturbation model exhibits slightly better robustness than dropout, the dropout-trained model does not necessitate prior knowledge of error types. Therefore, these results reveal that DRL with dropout achieves superior performance over explicit geometric quantum gates in addressing systematic errors and optimizing gate performance in quantum systems.

The standard batch method, Gaussian perturbation on nodes, and dropout, used in Figs. 2(a)–2(c), represent distinct optimization techniques for deriving control pulses. Each approach tackles the task of finding optimal control pulses differently. The batch method relies on gradient-based optimization, while Gaussian perturbation on nodes and dropout introduce stochasticity into the optimization process. The incorporation of randomness in Gaussian perturbation on nodes and dropout can lead to exploration of different regions of the parameter space, potentially resulting in diverse optimal control pulses. In Fig. 3, we present normalized control pulses for all tunable parameters obtained through DRL exploration. These pulses result from running the trained DRL models in an error-free environment. Notably, the adjustment of single qubit pulses aims to centralize the maximum gate fidelity at $(\delta\Omega, \delta\Delta) = (0, 0)$ within the respective parameter space. The pulse amplitudes, as outputs of the DRL, exhibit continuity in the decision layer but are discretized into stepwise functions, rendering them interpretable and amenable to smoothing if necessary. It is worth emphasizing the broad applicability of these control pulses across various physical systems. Dropout, with its focus on preventing overfitting through regularization, may produce pulses more resilient to variations in the quantum system or experimental conditions. Moreover, to implement the robust two-qubit gate, operational time calculation depends on the tunable parameter ranges. While in certain scenarios, like the weak-coupling limit between qubits, retraining the model with modified tunable coupling ranges may be necessary, the methodology remains fundamentally applicable. Furthermore, pulse smoothness is crucial due to the notable similarities in consecutive time-step actions within the PPO algorithm. This consistency underscores insights into the stability and predictability of the agent's decision-making process.

IV. DISCUSSION

In this section, we shall have insight into the more technical aspects after conducting a thorough analysis and evaluation of DRL's performance based on our previous numerical experiments. As one might discern, rewarding the agent with logarithmic infidelity at the end of each episode can inspire the agent to discover more efficient control pulses. In essence, it motivates the agent to explore an "as-soon-as-possible" solution, aiming to achieve the gate fidelity threshold by the Nth time step and conclude the episode before reaching the maximum time step $N_{\max}$. We use "as-soon-as-possible" instead of "time-optimal" in the description, as an explicit objective function that encompasses terms related to robustness or energy cost is not provided in the problem statement. Notably, robustness in our approach is introduced through techniques from computer science, such as the batch method, perturbations on nodes, or dropout, rather than by engineering the reward function to incorporate error sensitivity.

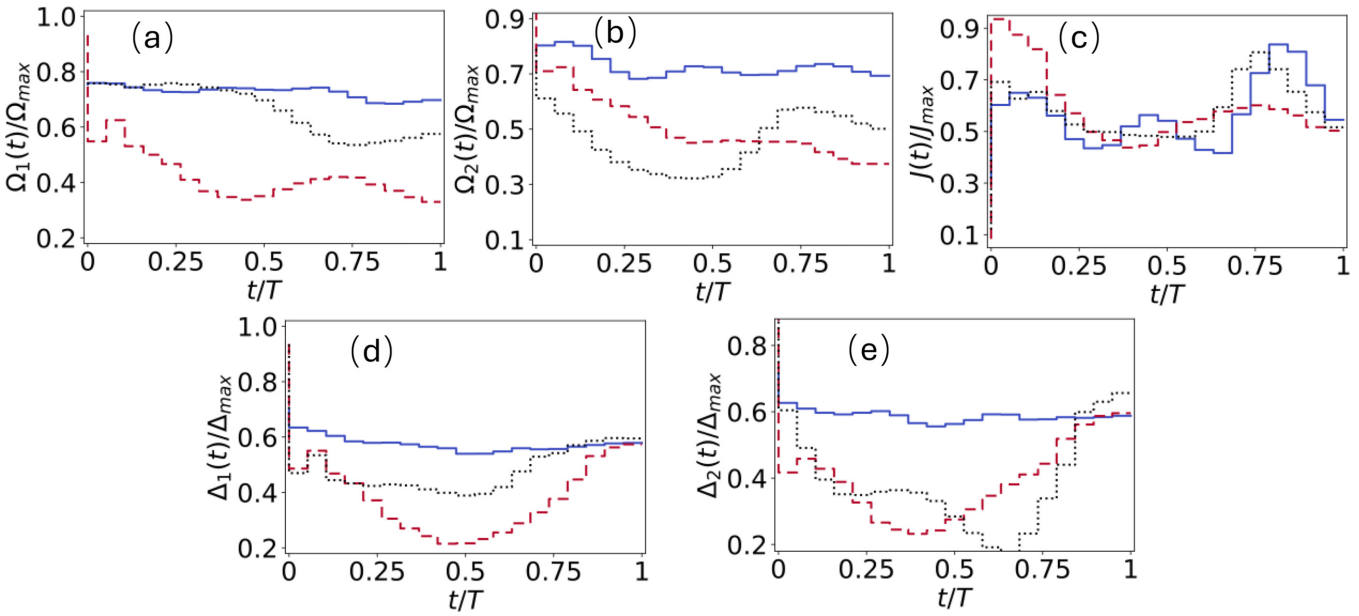

FIG. 3. Control pulses obtained from the models evaluated in Figs. 2(a)–2(c) without the guidance of human experts. (a) Rabi frequency on the first qubit. (b) Rabi frequency on the second qubit. (c) Magnitude of the ZZ interaction between the two qubits. (d) Detuning on the first qubit. (e) Detuning on the second qubit. The solid blue lines represent pulses obtained from the standard batch method, the dashed red lines from Gaussian perturbation on nodes, and the dot-dashed black lines from dropout. Parameters: Dimensionless tunable Rabi frequency $\Omega \in [0, \Omega_{\max}]$, $\Omega_{\max} = 2\pi$, detuning $\Delta \in [-\Delta_{\max}, \Delta_{\max}]$, $\Delta_{\max} = 2\pi$, magnitude of ZZ interaction $J \in [0, J_{\max}]$, $J_{\max} = 2\pi$, operation time $T = 2$, and maximal time step $N_{\max} = 20$.

Another critical consideration is that the DRL agent should not seek the global optimal solution that maximizes accumulated rewards in each episode. For example, in a scenario in which the agent can learn a gate operation with fidelity surpassing 0.99 within a maximal time step $N < N_{\max}$, it could exploit a loophole to gain extra rewards by deactivating all pulses once the clipped reward $r = 1 - \log_{10}(1 - F)$ is triggered. The episode would continue since the gate fidelity does not meet the threshold. Finally, the agent could reactivate the pulses $a(t_N)$ at the last time step of the episode to trigger the $r = 10 - \log_{10}(1 - F)$ bonus. Through this method, the agent would earn an additional reward of

$$\begin{aligned}\Delta r = &\sum_{i=N-1}^{N_{\max}-1} \gamma^i [1 - \log_{10}(1 - F_{N-1})] \\ &+\gamma^{N_{\max}} [10 - \log_{10}(1 - F_{N_{\max}})] \\ &-\gamma^{N} [10 - \log_{10}(1 - F_{N_{\max}})], \end{aligned} \tag{8}$$

which is positive when N is sufficiently small and requires large tunable ranges of control parameters. This would result in a sudden drop in the total reward during the training process, even if the model appears to be converging. In our numerical experiments, neither of these phenomena occurred because we set the tunable ranges of control parameters within reasonable bounds, preventing the model from initially converging to an as-soon-as-possible solution. However, it should be noted that such solutions, as well as the cheating solution described, could potentially be discovered with certain hyperparameters after training for more episodes. In essence, what we performed in the numerical experiments, without guidance from human experts, corresponds to the early-stopping method, a common technique to prevent overfitting in supervised learning.

The absence of geometric gates as robust solutions within specialized Hamiltonians in the environment is related to the initialization and exploration of the agent. None of the techniques from computer science led to geometric gates (at least not after training a moderate number of episodes) because the mechanism of robustness is not restricted to the property of topological protection. Our assumption is that DRL algorithms can discover geometric gates under settings with well-tuned hyperparameters and a more extensive number of episodes. Considering that the robustness performance shown in Figs. 2(d)–2(f) is already quite promising compared to geometric gates, extensive effort would be required to uncover such solutions. However, we provide a trick to obtain geometric gates from the agent if one insists on pursuing this approach. One can evaluate the dynamical phases γ_n^d at the end of each episode and design a reward function as follows: $r = -\log_{10}(1 - F) - \lambda(\gamma_n^d)^2$, where $\lambda > 0$ is a tunable coefficient for the Lagrangian multiplier. This additional term penalizes the agent by the square of dynamical phases if they are not canceled. Nevertheless, the requirement for this additional physical knowledge may discourage us from employing DRL for such a task, especially when the calculation of dynamical phases has already led to explicit solutions.

V. CONCLUSION

After thoroughly investigating the capabilities of DRL for exploring robust two-qubit gates without the guidance of human experts, we have conducted extensive numerical experiments to explore the potential of DRL in quantum control.

Our findings demonstrate that DRL, when equipped with techniques from computer science such as dropout, can effectively navigate the complex parameter space and identify robust quantum control strategies. Specifically, dropout, which introduces stochasticity by randomly disconnecting nodes in the DRL agent during training, has emerged as a promising method for achieving satisfactory levels of robustness.

Interestingly, our experiments suggest that there is no necessity to incorporate additional physical knowledge, such as operation time or Hamiltonian details, into the DRL framework. In fact, the inclusion of such information might hinder the agent's ability to explore and discover optimal solutions. Instead, a general Hamiltonian with appropriately set tunable ranges, coupled with the utilization of dropout, suffices for robust quantum control exploration using DRL.

We have selected the entangling two-qubit gate of $R_{YY}(\pi/4)$ as our focus due to its maximal entanglement capability and ease of realization in liquid NMR and Josephson charge qubits [37–39]. To show the advantage, we have benchmarked our results against the analytical solution of the nonadiabatic geometric gate [38]. Although other platforms may employ different two-qubit gates, such as the Mølmer-Sørensen or CPHASE gate for trapped ions [55–57], we believe that the general Hamiltonian and dropout techniques can facilitate the learning of various gates across diverse quantum platforms.

In summary, our paper highlights the potential of DRL as a powerful tool for exploring robust quantum control strategies. By leveraging techniques from computer science and embracing the inherent flexibility of general Hamiltonians, DRL shows promise in addressing complex quantum control challenges without the need for explicit human guidance. Looking ahead, further research could explore the scalability of DRL techniques to larger quantum systems and investigate their applicability in real-world experimental settings. Additionally, refining the incorporation of physical constraints and domain knowledge into the DRL framework may enhance its performance and broaden its applicability in quantum optimal control tasks.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (Grants No. 12075145 and No. 12211540002), STCSM (Grant No. 2019SHZDZX01-ZX04), the Innovation Program for Quantum Science and Technology (Grant No. 2021ZD0302302), EU FET Open Grant EPIQUS (Grant No. 899368), HORIZON-CL4-2022-QUANTUM-01-SGA Project No. 101113946 OpenSuperQPlus100 of the EU Flagship on Quantum Technologies, the Basque Government (Grant No. IT1470-22), Grant No. PID2021-126273NB-I00 funded by MCIN/AEI/10.13039/501100011033, "ERDFA Way of Making Europe," "ERDF Invest in Your Future," Nanoscale NMR and Complex Systems (Grant No. PID2021-126694NB-C21), the Valencian Government Grant (Grant No. CIAICO/2021/184), the Spanish Ministry of Economic Affairs and Digital Transformation through the QUANTUM ENIA project call—Quantum Spain project, and the European Union through the Recovery, Transformation, and Resilience Plan—NextGenerationEU within the framework of the Digital Spain 2026 Agenda.

APPENDIX A: ANALYTICAL SOLUTION OF THE NONADIABATIC GEOMETRIC GATE

In this Appendix, we provide an analytical solution for the robust entangling gate $R_{YY}(\pi/4)$ described by the Hamiltonian (1). While the nonadiabatic geometric theory has been previously proposed in Refs. [37–39,48], we repeat the procedure here for the sake of completeness.

To obtain the dynamical invariant, we deactivate the local pulses on the first qubit and simplify the two-qubit Hamiltonian (2) into two parts:

$$H^{\pm}(t)=\frac{1}{2}[\Omega\cos(\omega t)G_x^{\pm}+\Omega\sin(\omega t)G_y^{\pm}+\Delta^{\pm}G_z^{\pm}], \quad \text{(A1)}$$

where $G_i^{\pm}=(I_1\pm Z_1)/2\otimes(i\in\{X,Y,Z\})$ are effective Pauli operators and $\Delta^{\pm}=\Delta\pm J$ is the effective detuning. The dynamical invariant of $H^{\pm}$ is then given by [58]

$$I^{\pm}(t)=\Omega\cos(\omega t)G_x^{\pm}+\Omega\sin(\omega t)G_y^{\pm}+(\Delta^{\pm}-\omega)G_z^{\pm}. \quad \text{(A2)}$$

The instantaneous eigenstates of the Hamiltonian $H^{\pm}$ are superpositions of the eigenvectors of the invariant $I^{\pm}$, expressed as $|\Psi(t)\rangle=\sum_n c_n e^{i\alpha_n(t)}|\phi(t)\rangle$, where $\alpha_n(t)=\int_0^T dt'\langle\phi_n(t')|i\partial_{t'}-H(t')|\phi_n(t')\rangle$ represents the Lewis-Riesenfeld phases incorporating both geometric and dynamical phases. A more general case involving a four-level system can be found in Ref. [47]. By canceling the dynamical phases $\gamma_n^d=-\int_0^T dt'\langle\phi_n(t')|H(t')|\phi_n(t')\rangle$, we obtain the geometric gate:

$$U(t,0)=\sum_n e^{i\alpha_n(t)}|\phi_n(t)\rangle\langle\phi_n(0)|, \quad \text{(A3)}$$

governed solely by geometric phases after an operation time of T. The eigensystem of $I^{\pm}$ yields eigenvalues $\lambda^{\pm}=\sqrt{(\Delta^{\pm}-\omega)^2+\Omega^2}$ and corresponding eigenstates:

$$E_{\pm}^{+}=\pm\lambda^{+},\ |\phi_{\pm}^{+}(t)\rangle=(\cos\theta_{\pm}^{+}e^{-i\omega t},\ -\sin\theta_{\pm}^{+},\ 0,\ 0)^T, \quad \text{(A4)}$$

$$E_{\pm}^{-}=\pm\lambda^{-},\ |\phi_{\pm}^{-}(t)\rangle=(0,\ 0,\ \cos\theta_{\pm}^{-}e^{-i\omega t},\ -\sin\theta_{\pm}^{-})^T, \quad \text{(A5)}$$

with $\sin\theta_{\pm}^{\pm}=1/\sqrt{\xi_{\pm}^{\pm 2}+1}$, $\cos\theta_{\pm}^{\pm}=\xi_{\pm}^{\pm}/\sqrt{\xi_{\pm}^{\pm 2}+1}$, and $\xi_{\pm}^{\pm}=\Omega/(\Delta^{\pm}\mp\lambda^{\pm}-\omega)$. In this case, the Lewis-Riesenfeld phases are obtained as $\alpha_{\pm}^{\pm}=(\lambda^{\pm}\mp\omega)t/2$. The parameter settings $\Delta=\omega/2$ and $\Omega=\pm\sqrt{\omega^2-4J^2}/2$ in the nonadiabatic regime give the two-qubit geometric gate, $V_U=e^{i\gamma_{\pm}^{+}}|\phi_{\pm}^{+}(0)\rangle\langle\phi_{\pm}^{+}(0)|+e^{i\gamma_{\pm}^{-}}|\phi_{\pm}^{-}(0)\rangle\langle\phi_{\pm}^{-}(0)|$, yielding

$$V_U=-e^{i\pi a_-(a_-G_z^{+}-a_+G_x^{+})}e^{i\pi a_+(a_+G_z^{-}-a_-G_x^{-})}, \quad \text{(A6)}$$

where $a_{\pm}=\sqrt{J/\omega\pm 1/2}$. To ensure that the gate V_U allows maximum entanglement, we calculate the singular values $D_{\pm}$ of the matrix D, the elements of which are $D_{ij}=\mathrm{Tr}(V_U i\otimes j)/4$ with $i,j\in\{I,X,Y,Z\}$, and ensure that it equals to $D_{\pm}^{\mathrm{CNOT}}=\sqrt{1/2}$ or other perfect entanglers.

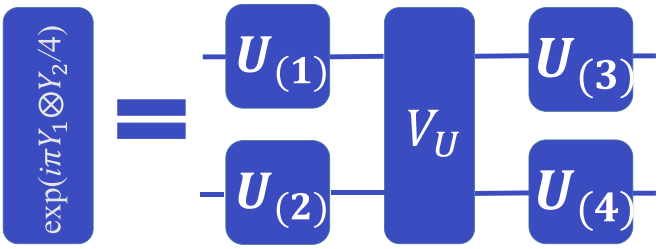

FIG. 4. Decomposition of a robust two-qubit entangling gate, $R_{YY}(\pi/4)$, highlighting the topological protection mechanism against systematic errors. The entangling gate is decomposed into universal geometric qubit gates, which are analytically solvable by canceling dynamical phases. The gate operation requires a total time of $\max(T(U_{(1)}), T(U_{(2)})) + T(V_U) + \max(T(U_{(3)}), T(U_{(4)}))$. Information gained from the analytical solution, such as the operation time, switching time, and form of Hamiltonian, is utilized to design the environment for investigating the capabilities of DRL.

Accordingly, we find the parameter setting $J/\omega = \pm 0.3187$ that corresponds to the gate time of V_U as $T = 2\pi/\omega$. With Cartan decomposition, one can verify that the entangling part of V_U is exactly $\exp(i\pi Y_1 \otimes Y_2/4)$. For the standard $R_{YY}(\pi/4)$ gate, we have local single qubit geometric gates on each qubit before and after V_U as shown in Fig. 4, which are all robust against systematic errors on local Rabi frequency and detuning. The method for designing the single qubit gate is similar to the two-qubit gate. The dynamical invariant of $H_{1,2}$ shares the same structure as $I^{\pm}$ but with effective operators $G_i^{\pm}$ and effective detuning $\Delta^{\pm}$ replaced by the standard version X_i, Y_i, Z_i, and Δ, respectively. We also have the nonadiabatic condition for canceling the dynamical phases as $\Omega^2 + \Delta(\Delta - \omega) = 0$, yielding the parametrized single qubit gate of gate time $U_i(\beta_j) = -\exp[i\pi \sin\beta_j(-\cos\beta_j X_i + \sin\beta_j Z_i)]$, where $\cos^2\beta_j = \Delta/\omega_j$ with the gate time $T = 2\pi/\omega_j$. By a sequence of β_j on the ith qubit gate, one achieves the universal single qubit rotation gate with three Euler angles. To simplify, we minimize the Frobenius norm $||U - R_{YY}(\pi/4)||$ between the matrix expression of the target gate $R_{YY}(\pi/4) = \exp(-i\pi Y_1 \otimes Y_2/4)$ and $U = (U_{(3)} \otimes U_{(4)})V_U(U_{(1)} \otimes U_{(2)})$, resulting in optimal parameters obtained through sequential least-squares programming:

$$
\begin{aligned}
U_{(1)} &= U_1(0.13)U_1(0.91)U_1(0.29)U_1(0.52),\\
U_{(2)} &= U_2(0.46)U_2(0.31)U_2(0.90)U_2(0.3)U_2(0.69)\\
&\quad \times U_2(0.23)U_2(0.48),\\
U_{(3)} &= U_1(0.24)U_1(0.56)U_1(0.29)U_1(0.24)U_1(0.81)\\
&\quad \times U_1(0.29)U_1(0.81),\\
U_{(4)} &= U_2(1.11)U_2(0.27)U_2(0.90)U_2(0.16)U_2(0.62).
\end{aligned}
$$

APPENDIX B: NUMERICAL RESULTS AND HYPERPARAMETERS

In the following Appendix, we explore the detailed robustness of the general Hamiltonian (2) using three distinct models: the standard batch method, Gaussian perturbation to nodes, and dropout, as illustrated in Fig. 2. To ensure objectivity and prevent bias, we adopt a systematic approach. Specifically, we train five models for each setting, utilizing random seeds spanning from 1 to 5, as outlined in Figs. 5–7. By averaging the figures of merit obtained from these models, our analysis provides a comprehensive comparison of the robustness performance of each model, enabling us to draw meaningful conclusions with an unbiased assessment.

Expanding on the detailed robustness demonstration of the general Hamiltonian (2) for three different models, we first examine the performance of the entangling gate $R_{YY}(\pi/4)$ against various types of errors. These errors include overrotating errors $[\Omega_{1,2} \to \Omega_{1,2}(1 + \delta\Omega)]$ and off-resonance

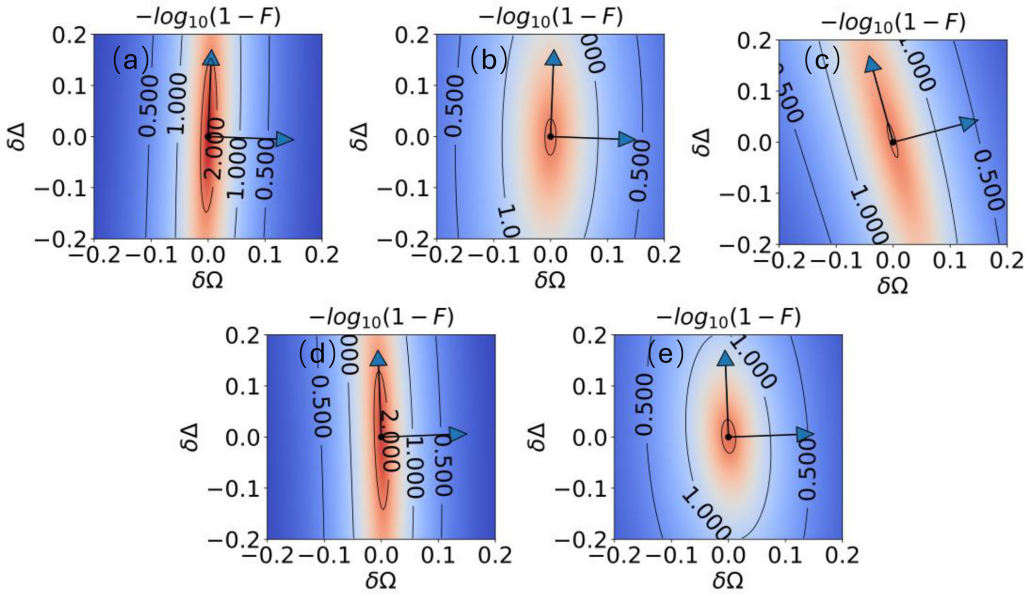

FIG. 5. Heat maps illustrating the logarithmic infidelity are plotted, being similar to Fig. 2(a). Here, the evaluated models labeled (a)–(e) are trained with random seeds ranging from 1 to 5 using the standard batch method.

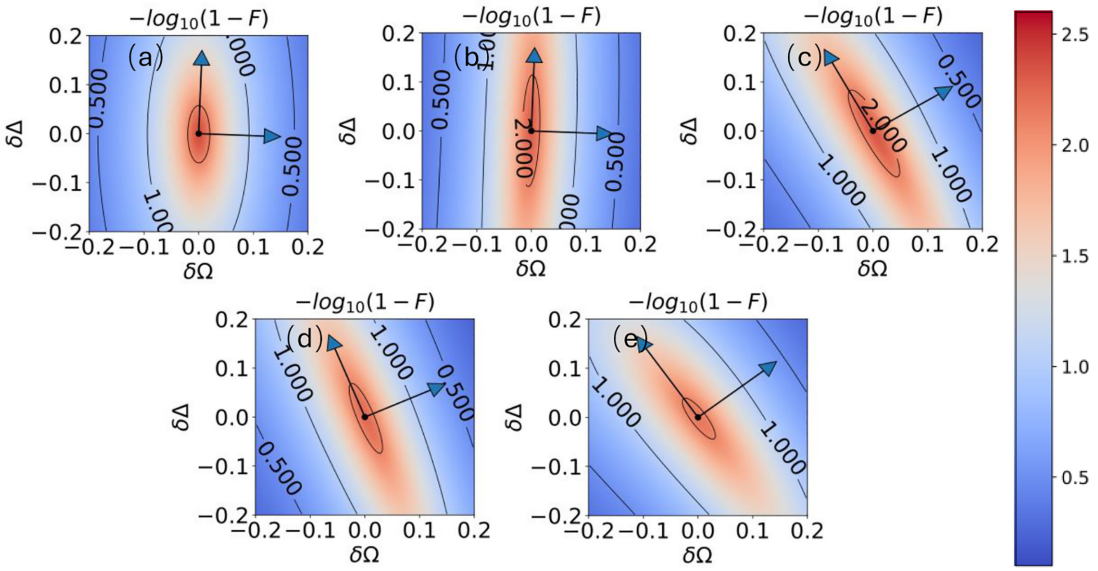

FIG. 6. Heat maps illustrating the logarithmic infidelity are plotted, being similar to Fig. 2(b), where Gaussian perturbation is applied on the nodes in the output layer. Here, the evaluated models labeled (a)–(e) are trained with random seeds ranging from 1 to 5. The parameters of random Gaussian perturbations on nodes are $\tilde{\Omega} \to \tilde{\Omega}[1 + N(\mu = 0, \sigma = 0.1)]$ and $\tilde{\Delta} \to \tilde{\Delta}[1 + N(\mu = 0, \sigma = 0.1)]$, respectively.

errors $[\Delta_{1,2} \to \Delta_{1,2}(1 + \delta\Delta)]$. The gate fidelity, denoted as F, is evaluated as $|\frac{1}{4}\mathrm{Tr}[\exp(-i\pi Y_1 \otimes Y_2/4)U(T, 0)]|^2$, where $U(T, 0)$ represents the propagator under these errors. This fidelity metric highlights the impact of systematic errors through a heat map displaying the logarithmic infidelity. The models under assessment are trained in an environment governed by the general two-qubit Hamiltonian depicted in (2). These models are rewarded with random seeds ranging from 1 to 5, and importantly, they operate without prior human expert knowledge about robustness.

First, we explore the robustness of the gate under the standard batch method. Similar to previous experiments in Fig. 5, we introduce systematic errors such as over-rotation and off-resonance errors to assess the gate fidelity. The standard batch method, a conventional approach in machine learning, involves using the entire dataset in each iteration of the optimization algorithm, typically gradient descent. In the realm of training neural networks for quantum gate optimization, this method entails feeding the entire training dataset into the model at once and updating model parameters based on gradients computed from the entire dataset. To ensure a comprehensive evaluation, we train five models for each configuration, utilizing random seeds ranging from 1 to 5, as depicted in Figs. 5(a)–5(e). This practice of varying the random seeds during initialization ensures exploration of different regions of the parameter space, guarding against the

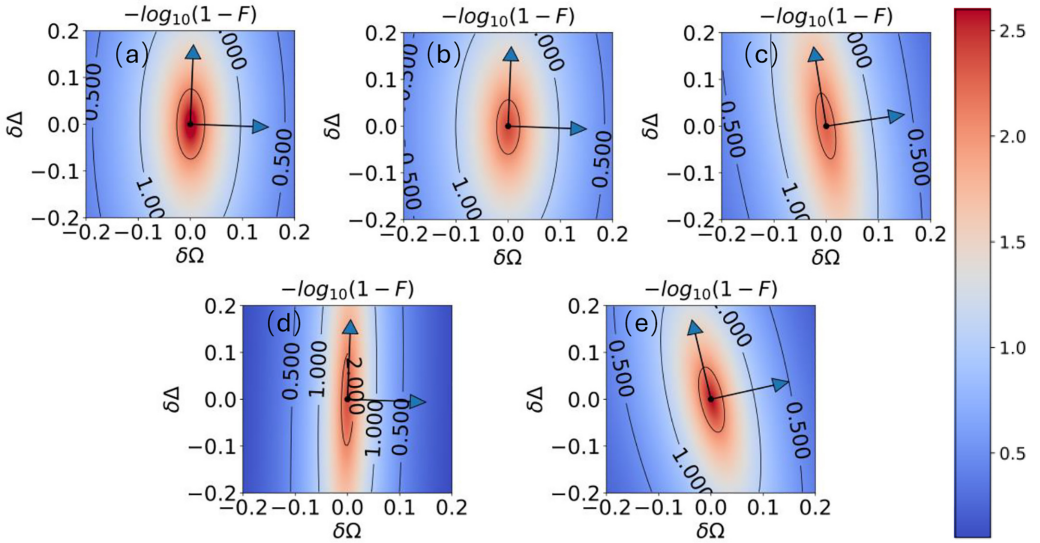

FIG. 7. Heat maps illustrating the logarithmic infidelity are plotted, being similar to Fig. 2(c), where the evaluated models in (a)–(e) are trained with random seed from 1 to 5. Dropout with a rate of $\kappa = 0.1$ is used.

TABLE I. Hyperparameters of three models in Fig. 2.

	General Hamiltonian						
Figure	Neurons in hidden layers	Batch size	Learning rate	Episodes	T	$N_{\max}$	Complementary
Fig. 2(a)	{64, 64, 64}	32	10^{-4}	10^5	2	20	Default
Fig. 2(b)	{64, 64, 64}	32	10^{-4}	2×10^5	2	20	$\tilde{\Omega}, \tilde{\Delta} \to \tilde{\Omega}[1 + N(\mu, \sigma)], \tilde{\Delta}[1 + N(\mu, \sigma)]$
Fig. 2(c)	{64, 64, 64}	32	10^{-4}	2×10^5	2	20	Dropout rate $\kappa = 0.1$

optimization algorithm becoming trapped in local minima. Training multiple models with distinct random seeds offers a more exhaustive assessment of the model's performance and robustness. By aggregating the results obtained from training these five models, we derive a more dependable estimate of the model's performance and robustness. Averaging serves to mitigate the variance in results stemming from the randomness in initialization and training. This approach yields a more stable and representative evaluation, thereby bolstering the credibility of the conclusions drawn from our analysis.

Similar to the standard batch method, when applying Gaussian perturbation, we also train five models for each configuration using random seeds ranging from 1 to 5; see Figs. 6(a)–6(e). Gaussian perturbation refers to the process of introducing random noise to the parameters of a model according to a Gaussian distribution. In the context of our simulation, this involves adding random Gaussian noise to the parameters $\tilde{\Omega}$ and $\tilde{\Delta}$ in the output layer of the DRL models. The noise is characterized by a mean (μ) of zero and a standard deviation (σ) of 0.1, and it is added independently to each parameter. This perturbation incorporates randomness into the model's parameters, simulating variations or uncertainties in the system being modeled. The main difference between Gaussian perturbation and the standard batch method lies in how the noise is introduced. In Gaussian perturbation, noise is explicitly added to the actions, simulating uncertainty or variability. In the standard batch method, a batch of experiences is used for policy updates, which helps to average out the variability and stabilize learning without explicitly introducing noise. Additionally, the standard batch method involves training the models using the entire dataset in each iteration of the optimization algorithm. Meanwhile, Gaussian perturbation is an additional step in the training process, integrated into the environment after each time step, which is not explicitly shown in the objective function for policy optimization. As a consequence, the results obtained from applying Gaussian perturbation to the DRL models are likely to differ from those obtained using the standard batch method, leading to different patterns of errors and variations in the gate fidelity.

Comparing the three models mentioned—the standard batch method, Gaussian perturbation on nodes, and dropout—they each have their own advantages and limitations. Finally, we present the results for five models trained with random seeds ranging from 1 to 5, as shown in Figs. 7(a)–7(e), where dropout with a rate of $\kappa = 0.1$ is utilized to assess its effectiveness in enhancing the gate's robustness. Dropout is a regularization technique commonly used in neural network training, including in the context of DRL. It works by randomly "dropping out" a proportion of the neurons in the network during each training iteration. This process helps prevent overfitting by introducing noise and reducing the reliance of the network on specific neurons or features. When training neural networks for quantum gate optimization, dropout can be used to enhance the robustness of the gate by introducing variability and preventing the network from memorizing noise in the training data. By encouraging the network to learn more general and robust representations, dropout can improve the model's ability to generalize to unseen data and mitigate the impact of systematic errors.

When random seeds ranging from 1 to 5 are used to train the models, it means that each model is initialized with different random seeds for the weights and biases. This ensures that each model explores different regions of the parameter space during training, which can help prevent us from cherry picking on a certain result with better performance. Upon the the results from models trained with random seeds ranging from 1 to 5 under each setting, we can average the figures of merit to derive a conclusion. For three RL methods toward robustness, the averaged differences are obtained as $\mathcal{A}_{\text{std}} = (0.851, 0.040)$, $\mathcal{A}_{\text{Gauss}} = (0.407, 0.030)$, and $\mathcal{A}_{\text{dropout}} = (0.543, 0.031)$, with a distance of $|\mathbf{r}| = 0.04$. Although the Gaussian-perturbated model exhibits slightly greater robustness than dropout, the dropout-trained model does not require prior knowledge of the types of errors. As mentioned above, dropout regularization provides a mechanism for preventing overfitting and improving generalization performance in neural networks by randomly dropping units during training, potentially resulting in improved robustness and better adaptation to unseen data.

Besides the above information for reproduction (see Table I), we make all codes, including environments, models, and scripts, open access [59]. Other hyperparameters of the PPO algorithm are set to their default values if they are not explicitly mentioned. We utilized the open-source library TENSORFORCE V0.5.3 for the implementation.

[1] M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz, and A. Zeilinger, Automated search for new quantum experiments, Phys. Rev. Lett. **116**, 090405 (2016).

[2] M. Krenn, M. Erhard, and A. Zeilinger, Computer-inspired quantum experiments, Nat. Rev. Phys. **2**, 649 (2020).

[3] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, An introduction to deep reinforcement learning, Foundations and Trends in Machine Learning **11**, 219 (2018).

[4] J. D. Martín-Guerrero and L. Lamata, Reinforcement learning and physics, Appl. Sci. **11**, 8589 (2021).

[5] 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).

[6] 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).

[7] R. Porotti, D. Tamascelli, M. Restelli, and E. Prati, Coherent transport of quantum states by deep reinforcement learning, Commun. Phys. **2**, 61 (2019).

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

[9] M. Dalgaard, F. Motzoi, J. J. Sørensen, and J. Sherson, Global optimization of quantum dynamics with AlphaZero deep exploration, npj Quantum Inf. **6**, 6 (2020).

[10] X.-M. Zhang, Z.-W. Cui, X. Wang, and M.-H. Yung, Automatic spin-chain learning to explore the quantum speed limit, Phys. Rev. A **97**, 052333 (2018).

[11] R.-B. Wu, H. Ding, D. Dong, and X. Wang, Learning robust and high-precision quantum controls, Phys. Rev. A **99**, 042327 (2019).

[12] M. Ostaszewski, J. Miszczak, L. Banchi, and P. Sadowski, Approximation of quantum control correction scheme using deep neural networks, Quantum Inf. Process. **18**, 126 (2019).

[13] C. Jiang, Y. Pan, Z.-G. Wu, Q. Gao, and D. Dong, Robust optimization for quantum reinforcement learning control using partial observations, Phys. Rev. A **105**, 062443 (2022).

[14] 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).

[15] T. Haug, W.-K. Mok, J.-B. You, W. Zhang, C. E. Png, and L.-C. Kwek, Classifying global state preparation via deep reinforcement learning, Mach. Learn.: Sci. Technol. **2**, 01LT02 (2021).

[16] C. Messikh and A. Messikh, Robust stimulated Raman shortcuts to adiabatic passage with deep learning, Europhys. Lett. **140**, 48003 (2022).

[17] Z. An and D. Zhou, Deep reinforcement learning for quantum gate control, Europhys. Lett. **126**, 60002 (2019).

[18] Y. Ding, Y. Ban, J. D. Martín-Guerrero, E. Solano, J. Casanova, and X. Chen, Breaking adiabatic quantum control with deep learning, Phys. Rev. A **103**, L040401 (2021).

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

[20] O. Shindi, Q. Yu, P. Girdhar, and D. Dong, Model-free quantum gate design and calibration using deep reinforcement learning, arXiv:2302.02371.

[21] V. B. Sørdal and J. Bergli, Deep reinforcement learning for quantum Szilard engine optimization, Phys. Rev. A **100**, 042314 (2019).

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

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

[24] D. Dong, C. Chen, H. Li, and T. J. Tarn, Quantum reinforcement learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **38**, 1207 (2008).

[25] P. Král, I. Thanopulos, and M. Shapiro, Colloquium: Coherently controlled adiabatic passage, Rev. Mod. Phys. **79**, 53 (2007).

[26] M. Steffen and R. H. Koch, Shaped pulses for quantum computing, Phys. Rev. A **75**, 062326 (2007).

[27] E. Barnes and S. Das Sarma, Analytically solvable driven time-dependent two-level quantum systems, Phys. Rev. Lett. **109**, 060401 (2012).

[28] D. Daems, A. Ruschhaupt, D. Sugny, and S. Guérin, Robust quantum control by a single-shot shaped pulse, Phys. Rev. Lett. **111**, 050404 (2013).

[29] K. R. Brown, A. W. Harrow, and I. L. Chuang, Arbitrarily accurate composite pulse sequences, Phys. Rev. A **70**, 052318 (2004).

[30] B. T. Torosov, S. Guérin, and N. V. Vitanov, High-fidelity adiabatic passage by composite sequences of chirped pulses, Phys. Rev. Lett. **106**, 233001 (2011).

[31] X. Rong, J. Geng, F. Shi, Y. Liu, K. Xu, W. Ma, F. Kong, Z. Jiang, Y. Wu, and J. Du, Experimental fault-tolerant universal quantum gates with solid-state spins under ambient conditions, Nat. Commun. **6**, 8748 (2015).

[32] H.-N. Wu, C. Zhang, J. Song, Y. Xia, and Z.-C. Shi, Composite pulses for optimal robust control in two-level systems, Phys. Rev. A **107**, 023103 (2023).

[33] Z.-C. Shi, J.-T. Ding, Y.-H. Chen, J. Song, Y. Xia, X. X. Yi, and F. Nori, Supervised learning for robust quantum control in composite-pulse systems, Phys. Rev. Appl. **21**, 044012 (2024).

[34] D. Guéry-Odelin, A. Ruschhaupt, A. Kiely, E. Torrontegui, S. Martínez-Garaot, and J. G. Muga, Shortcuts to adiabaticity: Concepts, methods, and applications, Rev. Mod. Phys. **91**, 045001 (2019).

[35] X. Chen, A. Ruschhaupt, S. Schmidt, A. del Campo, D. Guéry-Odelin, and J. G. Muga, Fast optimal frictionless atom cooling in harmonic traps: Shortcut to adiabaticity, Phys. Rev. Lett. **104**, 063002 (2010).

[36] D. Leibfried, B. DeMarco, V. Meyer, D. Lucas, M. Barrett, J. Britton, W. M. Itano, B. Jelenković, C. Langer, T. Rosenband, and D. J. Wineland, Experimental demonstration of a robust, high-fidelity geometric two ion-qubit phase gate, Nature (London) **422**, 412 (2003).

[37] S.-L. Zhu and Z. D. Wang, Implementation of universal quantum gates based on nonadiabatic geometric phases, Phys. Rev. Lett. **89**, 097902 (2002).

[38] S.-L. Zhu and Z. D. Wang, Unconventional geometric quantum computation, Phys. Rev. Lett. **91**, 187902 (2003).

[39] Y. Ota, Y. Goto, Y. Kondo, and M. Nakahara, Geometric quantum gates in liquid-state NMR based on a cancellation of dynamical phases, Phys. Rev. A **80**, 052311 (2009).

[40] A. A. Abdumalikov, Jr., J. M. Fink, K. Juliusson, M. Pechal, S. Berger, A. Wallraff, and S. Filipp, Experimental realization of non-Abelian geometric gates, Nature (London) **496**, 482 (2013).

[41] C. Zu, W.-B. Wang, L. He, W.-G. Zhang, C.-Y. Dai, F. Wang, L.-M. Duan, Experimental realization of universal geometric quantum gates with solid-state spins, Nature (London) **514**, 72 (2014).

[42] C. Zhang, T. Chen, S. Li, X. Wang, and Z.-Y. Xue, High-fidelity geometric gate for silicon-based spin qubits, Phys. Rev. A **101**, 052302 (2020).

[43] B.-J. Liu, X.-K. Song, Z.-Y. Xue, X. Wang, and M.-H. Yung, Plug-and-play approach to nonadiabatic geometric quantum gates, Phys. Rev. Lett. **123**, 100501 (2019).

[44] G. Dridi, K. Liu, and S. Guérin, Optimal robust quantum control by inverse geometric optimization, Phys. Rev. Lett. **125**, 250403 (2020).

[45] M.-Y. Mao, Z. Cheng, Y. Xia, A. M. Oleś, and W.-L. You, Neural-network-based optimal quantum control of nonadiabatic geometric quantum computation via reverse engineering, Phys. Rev. A **108**, 032616 (2023).

[46] M. A. Nielsen and I. L. Chuang: *Quantum Computation and Quantum Information* (Cambridge University, Cambridge, England, 2000).

[47] U. Güngördü, Y. Wan, M. A. Fasihi, and M. Nakahara, Dynamical invariants for quantum control of four-level systems, Phys. Rev. A **86**, 062312 (2012).

[48] U. Güngördü, Y. Wan, and M. Nakahara, Non-adiabatic universal holonomic quantum gates based on Abelian holonomies, J. Phys. Soc. Jpn. **83**, 034001 (2014).

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

[50] M. Kang, Q. Liang, B. Zhang, S. Huang, Y. Wang, C. Fang, J. Kim, and K. R. Brown, Batch optimization of frequency-modulated pulses for robust two-qubit gates in ion chains, Phys. Rev. Appl. **16**, 024039 (2021).

[51] N. Heimann, L. Broers, N. Pintul, T. Peterson, K. Sponselee, A. Ilin, C. Becker, and L. Mathey, Quantum gate optimization for Rydberg architectures in the weak-coupling limit, arXiv:2306.08691.

[52] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. **15**, 1929 (2014).

[53] R. Russell and R. J. Marks II, *Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks* (MIT, Cambridge, MA, 1999).

[54] Z. Wang, P.-L. Zheng, B. Wu, and Y. Zhang, Quantum dropout: On and over the hardness of quantum approximate optimization algorithm, Phys. Rev. Res. **5**, 023171 (2023).

[55] A. Sørensen and K. Mølmer, Quantum computation with ions in thermal motion, Phys. Rev. Lett. **82**, 1971 (1999).

[56] J. I. Cirac and P. Zoller, Quantum computations with cold trapped ions, Phys. Rev. Lett. **74**, 4091 (1995).

[57] F. Schmidt-Kaler, H. Häffner, M. Riebe, S. Gulde, G. P. T. Lancaster, T. Deuschle, C. Becher, C. F. Roos, J. Eschner, and R. Blatt, Realization of the Cirac-Zoller controlled-NOT quantum gate, Nature (London) **422**, 408 (2003).

[58] X. Chen, E. Torrontegui, and J. G. Muga, Lewis-Riesenfeld invariants and transitionless quantum driving, Phys. Rev. A **83**, 062116 (2011).

[59] All codes and models for this paper are accessible via this Github repository: https://github.com/QLNUxu/DRL-for-two-qubit-gate