Closed-loop control of a noisy qubit with reinforcement learning

IOP Publishing Mach. Learn.: Sci. Technol. 4 (2023) 025020 https://doi.org/10.1088/2632-2153/acd048

MACHINE LEARNING
Science and Technology

OPEN ACCESS

RECEIVED
2 January 2023

REVISED
7 April 2023

ACCEPTED FOR PUBLICATION
25 April 2023

PUBLISHED
5 May 2023

PAPER

Closed-loop control of a noisy qubit with reinforcement learning

Yongcheng Ding[1,2], **Xi Chen**[1,3,*], **Rafael Magdalena-Benedito**[4] **and José D Martín-Guerrero**[4,5]

[1] Department of Physical Chemistry, University of the Basque Country UPV/EHU, Apartado 644, 48080 Bilbao, Spain
[2] International Center of Quantum Artificial Intelligence for Science and Technology (QuArtist) and Department of Physics, Shanghai University, 200444 Shanghai, People's Republic of China
[3] EHU Quantum Center, University of the Basque Country UPV/EHU, Barrio Sarriena, s/n, 48940 Leioa, Spain
[4] IDAL, Electronic Engineering Department, ETSE-UV, University of Valencia, Avgda. Universitat s/n, 46100 Burjassot, Valencia, Spain
[5] Valencian Graduate School and Research Network of Artificial Intelligence (ValgrAI), Valencia, Spain
[*] Author to whom any correspondence should be addressed.

E-mail: chenxi1979cn@gmail.com

Keywords: deep reinforcement learning, noisy qubit, closed-loop quantum control

Abstract

The exotic nature of quantum mechanics differentiates machine learning applications in the quantum realm from classical ones. Stream learning is a powerful approach that can be applied to extract knowledge continuously from quantum systems in a wide range of tasks. In this paper, we propose a deep reinforcement learning method that uses streaming data from a continuously measured qubit in the presence of detuning, dephasing, and relaxation. The model receives streaming quantum information for learning and decision-making, providing instant feedback on the quantum system. We also explore the agent's adaptability to other quantum noise patterns through transfer learning. Our protocol offers insights into closed-loop quantum control, potentially advancing the development of quantum technologies.

1. Introduction

Quantum computation and quantum information [1] are no longer just promising research fields, but they have become current realities with increasing applicability in the next decade, in particular, that related to computational intelligence [2, 3]. Quantum computation relies on quantum bits (qubits), which are the quantum generalization of classical bits. The two basic states of a qubit are $|0\rangle$ and $|1\rangle$, corresponding with the states zero and one, respectively, of a classical bit. However, a qubit $|\Psi\rangle$ has the unique feature of allowing states formed by the superposition of $|0\rangle$ and $|1\rangle$, namely, $|\Psi\rangle = \alpha|0\rangle + \beta|1\rangle$, where α and β are complex coefficients. When a qubit is in a superposition state, its measurement will collapse it to one of its basic states, but it is impossible to determine in advance which one. The only available information is that the probability of $|0\rangle$ is $|\alpha|^2$ and the probability of $|1\rangle$ is $|\beta|^2$, hence, $|\alpha|^2 + |\beta|^2 = 1$. The primary operation when dealing with qubits is the unitary transformation U. Applying U to a superposition state results in another superposition state that superposes all basis vectors, which is known as quantum parallelism. This feature can be employed to evaluate the different values of a function $f(x)$ for a given input x at the same time. The unitary transformation $U(t, t_0)$ evolves a qubit state $|\Psi(t_0)\rangle$ to $|\Psi(t)\rangle = U(t, t_0)|\Psi(t_0)\rangle$ $= \mathcal{T}\exp[-(i/\hbar)\int_{t_0}^{t} H(t')\mathrm{d}t']|\Psi(t_0)\rangle$, where $H(t')$ is its Hamiltonian. Consequently, quantum control arises as the most critical problem in realizing quantum computation. Its goal is to design a time-dependent Hamiltonian $H(t)$, which drives the qubit to its target by a unitary transform. While simple solutions like the quantum NOT gate, which can be achieved with a resonant pulse $H = \hbar\Omega\sigma_x/2$ and an operation time $T = \hbar\pi/2\Omega$, exist, they are not robust and far from optimal. Any slight systematic error $T \to T + \delta T$ or equivalently $\Omega \to \Omega + \delta\Omega$ will lead to fidelity loss. Furthermore, qubits cannot be perfectly isolated from the external environment, where quantum noises induce decoherence. Therefore, optimal quantum control is necessary to achieve high-fidelity and high-robustness gate operations, which are the milestones in fault-tolerant universal quantum [4–6].

Physicists have proposed several protocols for achieving quantum control objectives, such as adiabatic quantum evolution [7], composite pulses [8–10], pulse-shaping engineering [11–14], shortcuts to

adiabaticity [15–17]. Particularly, machine learning (ML) algorithms can be combined with them for further optimizations [18–22]. It is also natural to consider applying reinforcement learning (RL) individually for quantum control tasks [22–33]. In recent years, deep reinforcement learning (DRL) has successfully addressed pulse design for fast and robust quantum state preparation [34–36], gate operation [37], and quantum Szilard engine [38]. However, as highlighted in our previous research [20, 21], the full potential of RL for quantum control has yet to be realized due to the challenge of quantum measurement. RL requires the observation of states for outputting an action, which conflicts with quantum mechanics' fundamental feature that the state is destroyed after direct quantum measurement. Most RL models for quantum control are trained in numerical environments instead of real quantum devices, to save resources, followed by fixed pulses after observation and evaluation. These fixed pulses hardly prevail over gradient-based optimization methods. Another approach involves combining the model with a quantum environment for evaluation. This approach allows the RL model to output an instant action after observing a state, even though the state is destroyed by direct measurement. However, this approach suffers from inefficiencies when historical actions are stored for repetitive operation of $n(n+1)/2$ steps, retrieving the last destroyed state. Here, n is the maximum number of time steps in each episode.

In this work, we present a RL approach to quantum control by employing the RL algorithm for closed-loop quantum control. In this paradigm, qubit's wave functions are no longer destroyed but slightly perturbed after information extraction via weak measurement. The model observes the state, which contains weak values as the partial information of the qubit with less confidence, resulting in an action to evolve the quantum environment to the next timestep. Our scheme reflects the spirit of stream learning once the length of each timestep is sufficiently small, resembling the dynamics of continuous measurement. It also enables transfer learning by adapting the model to the environment during the evaluation while external noises patterns are changing. We reckon that our protocol enhances the performance of quantum computing and quantum information processing in real-time experiments, accelerating its development from noisy intermediate-scale devices to the next level.

2. Physics models

2.1. Open quantum system

The dynamics of isolated quantum systems are governed by Schrödinger equation $i\hbar\partial_t|\psi(t)\rangle = H(t)|\psi(t)\rangle$, where the operator $H(t)$ represents the Hamiltonian of the quantum system, with its expectation be in the unit of energy. The von Neumann equation, $\dot{\rho}(t) = -(i/\hbar)[H(t), \rho(t)]$, is equivalent to the Schrödinger equation, where the pure state wave function $|\psi(t)\rangle$ is extended to density matrix $\rho(t) = |\psi(t)\rangle\langle\psi(t)|$. However, isolated quantum system are theoretical constructs that do not exist in the real world. External environments always affect the quantum system by coupling themselves to it, inducing undesired dynamics, such as decoherence.

Generally speaking, one can always write down the total Hamiltonian $H_{\mathrm{T}} = H_{\mathrm{S}} + H_{\mathrm{E}} + H_{\mathrm{I}}$, including the system Hamiltonian H_{S}, the environment Hamiltonian H_{E}, and the coupling interaction Hamiltonian H_{I}. The dynamics of the new system are described by the von Neumann equation, and one retrieves information about the original system by tracing out the environmental subsystem $\rho = \mathrm{Tr}_{\mathrm{E}}(\rho_{\mathrm{T}})$, resulting in the Lindblad master equation

$$\dot{\rho}(t) = -\frac{i}{\hbar}[H(t), \rho(t)] + \sum_n \frac{1}{2}\left[2C_n\rho(t)C_n^\dagger - \rho(t)C_n^\dagger C_n - C_n^\dagger C_n\rho(t)\right], \tag{1}$$

where $C_n = \sqrt{\gamma_n}A_n$ are the collapse operators, A_n are the operators that couples the system to environment in H_{I}, and γ_n are the corresponding rates. The density matrix is assumed to be initially in the product state as $\rho_{\mathrm{T}}(0) = \rho(0) \otimes \rho_{\mathrm{E}}(0)$, i.e. the original system and the environment are not correlated at $t = 0$. They still remains separable $\rho_{\mathrm{T}}(t) \approx \rho(t) \otimes \rho_{\mathrm{E}}$ during the evolution since the environment does not evolve significantly. The environment is considered to be Markovian, requiring the fast decays of its correlation functions than those of the system. it is also worthwhile to mention that the evolution of an open quantum system is no longer unitary. Therefore, the density matrix of the quantum system is not a pure state, but a mixed state $\rho = \sum_n p_n|\psi_n\rangle\langle\psi_n|$ instead, with p_n be the classical probability of being in $|\psi_n\rangle$ state.

2.2. Weak measurement and continuous measurement

One of the major difficulties in applying ML algorithms in the quantum regime is caused by measurement, which is usually costless in the classical realm. The act of measurement in the quantum system destroys it, being projected to an eigenstate once quantum information is extracted. Measuring a wave function by operator $\hat{A}$ outputs eigenvalues, whose expectation follows $\langle\hat{A}\rangle = \langle\psi|\hat{A}|\psi\rangle$. It can also be expressed in the

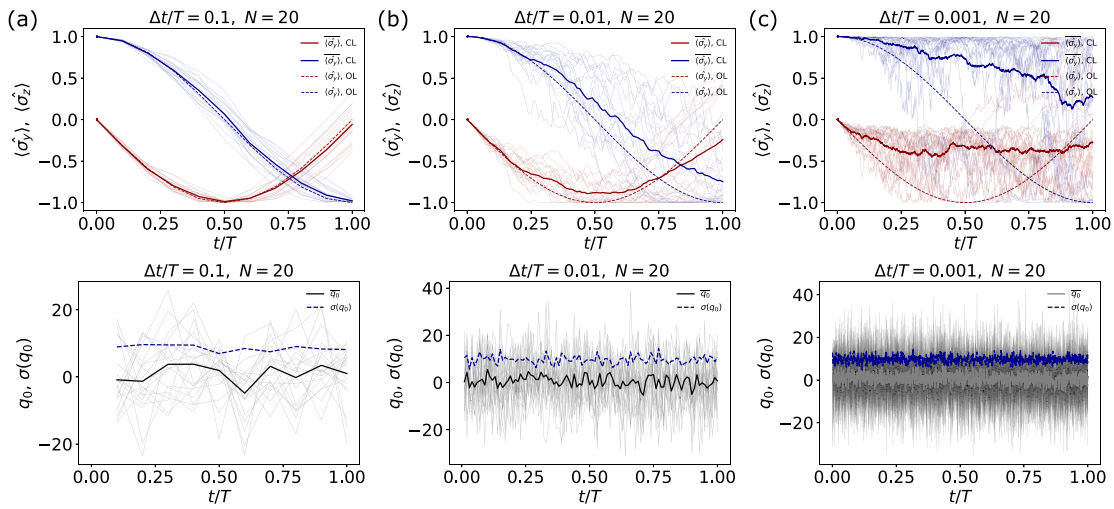

Figure 1. The expectations on Y (red) and Z (blue) directions of repetitively measured qubits, when a weak measurement takes place per (a) $\Delta t/T = 0.1$, (b) $\Delta t/T = 0.01$, and (c) $\Delta t/T = 0.001$. The qubit is driven by a resonant π-pulse, whose measurement-free dynamics is plotted in dashed curves as a textbook open-loop quantum control. By averaging over the trajectories $\langle\hat{\sigma}_i\rangle$ in the ensemble of $N = 20$ qubits (shaded curves), we have the average trajectories $\overline{\langle\hat{\sigma}_i\rangle}$ plotted in solid curves. For the closed-loop quantum control, the characteristics of dynamics vary from each other by the scale of the measurement interval. The corresponding weak value feedbacks q_0 from measuring the inaccurate Gaussian pointer $\sigma = 10$ are also recorded and averaged as $\bar{q}_0$, showing a low signal-to-noise ratio by comparing to the standard deviation $\sigma(q_0)$.

language of density matrix as $\langle\hat{A}\rangle = \mathrm{Tr}(\rho\hat{A})$. Aharonov's work [39] proposed an extension that extracts partial information from the quantum system without destroying it. The weak value $A_\mathrm{w} = \langle\psi_f|\hat{A}|\psi_i\rangle/\langle\psi_f|\psi_i\rangle$ is no longer real eigenvalues of the operator, but exotic values instead or even complex, where $|\psi_i\rangle$ and $|\psi_f\rangle$ are pre/post-selected states. The post-selection operation does not always succeed, and the wave function is discarded once the operation fails. To address this issue, we couple the quantum system to a pointer for entanglement and measure the pointer projectively for a weak value, which is actually the original framework proposed by Aharonov, instead of the later developed pre/post-selection formalism. Specifically, a Gaussian pointer $|\Phi\rangle = \int(2\pi\sigma^2)^{-1/4}\exp(-q^2/4\sigma^2)|q\rangle\mathrm{d}q$ is coupled to the qubit $|\Psi\rangle = [\cos(\alpha/2), \sin(\alpha/2)]^\mathrm{T}$, following the interaction Hamiltonian $H_\mathrm{int} = g(t)p\otimes\hat{A}$, where σ is the standard deviation of the pointer's position, p is its conjugate momentum operator, and $g(t)$ is the coupling strength. A non-correlated initial state $|\Phi(q)\rangle\otimes|\Psi\rangle$ is evolved by the Hamiltonian, entangling as $\cos(\alpha/2)|\Phi(q-a_1)\rangle\otimes|a_1\rangle$ $+\sin(\alpha/2)|\Phi(q-a_2)\rangle\otimes|a_2\rangle$, where a_i and $|a_i\rangle$ are the eigenvalues and eigenstates of the operator $\hat{A}$ to be weakly measured, respectively, if $\int_0^{t_0} g(t)\mathrm{d}t = 1$. For example, if one aims at performing a weak measurement on the Z direction, i.e. the Pauli-Z operator $\hat{A} = \hat{\sigma}_z$, the measurement outputs of the pointer's position follow the probability distribution

$$P(q) \approx \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(q-\cos\alpha)^2}{2\sigma^2}\right], \tag{2}$$

shifting a displacement of the expectation $\langle\Psi|\hat{\sigma}_z|\Psi\rangle = \cos\alpha$. Correspondingly, the wave function of the qubit is slightly perturbed as

$$|\tilde{\Psi}\rangle \propto \frac{1}{(2\pi\sigma^2)^{1/4}}\left\{\cos\left(\frac{\alpha}{2}\right)\mathrm{e}^{-\frac{(q_0-1)^2}{4\sigma^2}}|0\rangle + \sin\left(\frac{\alpha}{2}\right)\mathrm{e}^{-\frac{(q_0+1)^2}{4\sigma^2}}|1\rangle\right\}, \tag{3}$$

if the weak value q_0 is the measurement feedback of the projective measurement.

Additionally, quantum information can be continuously extracted from the quantum system, allowing for continuous measurement as the information obtained per measurement approaches zero. In this framework, the total operation time is divided into intervals of timestep Δt, so that a weak measurement is performed in each interval. The limit $\Delta t \to 0$ results in continuous measurement, with stochastic differential equations governing its dynamics [40, 41]. In figure 1, we illustrate the dynamics of stochastic Schrödinger equations, used to flip a qubit with a fixed resonant π-pulse, with varying scales of time interval Δt. It can be observed that if one continuously measures the qubit weakly, without taking any action based on the feedback, and evolves a resonant π-pulse, the final state is more likely to deviate from the target state, which is given by the open-loop quantum control with a π-pulse as its time-optimal solution. The more frequently we measure, the larger the expected deviation. Hence, for closed-loop quantum control, feedback must be

exploited to control the system. We explore this idea further by studying the design of an RL algorithm to solve the closed-loop control of a noisy qubit.

3. Numerical experiments

3.1. Physical system and task

In section 2, we introduced Lindblad master equations as the governing equations for quantum systems under noise. For pure dephasing, the diagonal Lindblad operators are given by $C_n = \sqrt{\gamma_n}|n\rangle\langle n|$, yielding the master equation

$$\dot{\rho} = -\frac{i}{\hbar}[H(t), \rho(t)] + \Gamma\left[\mathrm{diag}(\rho) - \rho\right], \tag{4}$$

with $\gamma_0 = \gamma_1$, affecting the coherence by reducing the off-diagonal elements of the density matrix. For relaxation, we consider the energy dissipation from the qubit to the external environment on the X direction, modeled by $C = \sqrt{\gamma}\hat{\sigma}_x$. The non-unitary evolution due to the Lindblad terms in master equation leads to a mixed state density matrix, where the classical probability p_n of being in $|\psi_n\rangle$ cannot be retrieved. Therefore, the perturbed system after weak measurement cannot be analytically calculated by equation (3) [42]. To extend our analysis to the case of the density matrix, we consider a Gaussian pointer of pure state $\rho_p = |\Phi\rangle\langle\Phi|$, coupled to a two-level system of mixed state ρ through the interaction Hamiltonian $H_{\mathrm{int}} = g\delta(t-t')p\otimes\hat{\sigma}_z$. The collective system is evolved from the initial state $\rho_{\mathrm{ini}} = \rho_{\mathrm{p}}\otimes\rho$ to ρ_{fin} after the coupling by

$$\rho_{\mathrm{fin}} = \exp(-igp\otimes\hat{\sigma}_z)\rho_{\mathrm{ini}}\exp(igp\otimes\hat{\sigma}_z), \tag{5}$$

shifting the pointer by $\langle\hat{\sigma}_z\rangle = \mathrm{Tr}(\rho\hat{\sigma}_z)$ when $g = 1$. One retrieves the wave pointer after the coupling by tracing out the qubit. The measurement of the pointer's position projects the pointer to its eigenstate $|q_0\rangle$, where the projection operator of the collective system reads $\hat{P} = |q_0\rangle\langle q_0|\otimes I$. In this way, we have the qubit's density matrix after the weak value feedback of q_0 by the projection operator and tracing out the pointer.

After clarifying the calculation of state perturbation in terms of the density matrix, we can now formulate the specific task to be studied by RL. We aim to study the optimal control of a continuously measured qubit within operation time T by ML algorithm. The goal is to flip the qubit from the state $|0\rangle$ to $|1\rangle$ using a sequence of pulses on the X direction. Each pulse lasts a small interval of Δt, being described by the driving Hamiltonian $H = \Omega\hat{\sigma}_x$, followed by a weak measurement on the Z direction. We assume that the measurement process is impulsive, meaning that the coupling and projective measurement on the pointer are instant and independent of the dynamical evolution. Meanwhile, the control pulses may also be imprecise, including slight detuning $H = \Omega\hat{\sigma}_x + \Delta\hat{\sigma}_z$ and amplitude error $\Omega \to \Omega + \delta\Omega$. The weak value and the last pulse amplitude are fed to the ML model as streaming data. Accordingly, the model's instant feedback then controls the quantum system for the next timestep.

3.2. Numerical setup

We apply the DRL method to our task for the RL approach. The environment consists of a qubit that is continuously measured, perturbed for weak values, and controlled by the agent's pulses. The agent is implemented as an artificial neural network (ANN) that takes in the qubit state as input, and outputs an action for the control problem. The ANN is trained by deep learning algorithms to approximate the optimal policy function $\pi(a|s)$. Upon receiving the action from the agent, the environment evolves to the next timestep, computes the new RL state, and provides a corresponding reward. It is worth noting that the environment in the quantum realm is different from other physical environments. In the RL environment, quantum information, e.g. density matrix elements or fidelity, is encoded in the RL state, requiring the numerical simulation. Unlike other physical environments, the density matrix elements cannot be directly obtained from the qubit without destroying it. Hence, one has to compute the density matrix based on the weak value and the control pulses, making the quantum environment non-trivial and computationally demanding.

In our practice, we set the tunable range of the Rabi frequency (pulse amplitude) as the action $\Omega \in [0, 3\pi]$ in dimensionless units, which is then renormalized to $\tilde{\Omega} \in [0, 1]$ for fitting the neuron. Total operation time $T = 1$ is uniformly separated into $n = 100$ control pulses, with each pulse driving the qubit for a time interval of $\Delta t = 0.01$. To save the computational resources, we limit the position space of the pointer to $q \in [-50, 50]$, with uniform separation by $\Delta x = 1$. Consequently, the momentum operator p is constructed by $[q, p] = i\hbar$ with boundary conditions. The density matrix of the collective quantum matrix has a size of 202×202. The coarse grained position space leads to weak values q_0 of integer number, which is renormalized to

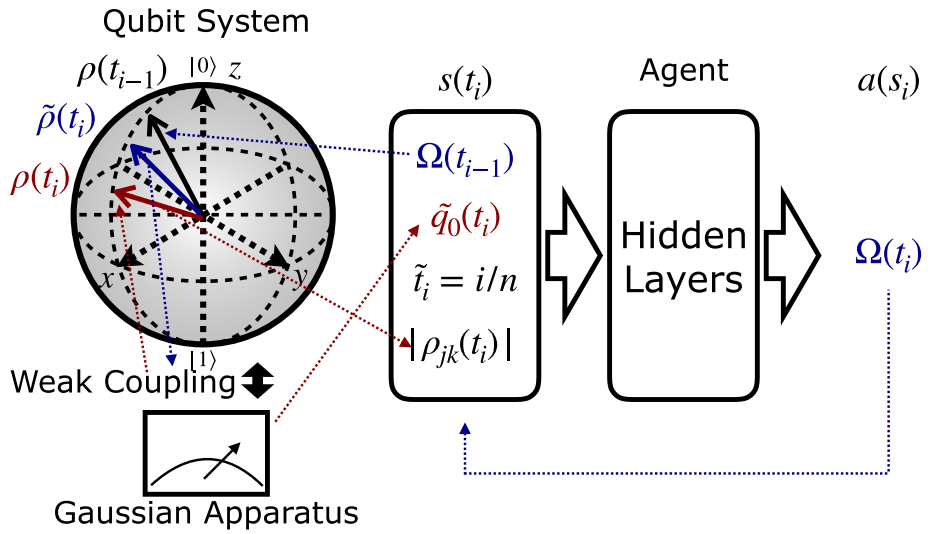

Figure 2. Schematic diagram of the environment without a reward function. The qubit is repetitively coupled to the apparatus weakly for information extraction. Its last state after the measurement is characterized by the density matrix $\rho(t_{i-1})$, being driven to a new unperturbed state by the last action $\Omega(t_i)$ as $\tilde{\rho}(t_i)$. The apparatus weakly measures the qubit for feedback of $q_0(t_i)$, and perturbs it to the state $\rho(t_i)$. The absolute values of its elements, together with the last action $\Omega(t_i)$, feedback $q_0(t_i)$, and renormalized system time $t_i = i/n$, are defined to be the RL State $s(t_i)$, being observed by the RL Agent (an artificial neural network), resulting in the corresponding action $a(s_i)$ for driving the qubit to its next state.

$\tilde{q}_0 = (q_0 + 50)/100 \in [0, 1]$. Thereby, the state is defined as $s(t_i) = \{\tilde{\Omega}(t_{i-1}), \tilde{q}_0(t_i), i/n, |\rho_{11}(t_i)|, |\rho_{12}(t_i)|, |\rho_{21}(t_i)|, |\rho_{22}(t_i)|\}$, including the last action as renormalized pulse, renormalized weak value, current system time, and elements of the density matrix. The RL state is observed by the agent, an ANN with three fully connected hidden layers of 64 neurons activated by ReLU, evolving to the next state by the numerical simulation part of the environment, which receives an action from the agent. We show the schematic flow diagram of the RL environment in figure 2 for a better understanding.

3.3. Training of the agent and results

We train three separate models for driving the qubit in the presence of detuning, dephasing, and relaxation on the X direction, respectively. The agents approximate the optimal policy, which maximizes accumulated artificial rewards. We keep the generality in the design of reward functions since we have no specific preference for any pulse shape. For the task of flipping the qubit, $|0\rangle \rightarrow |1\rangle$, we reward the agent by $r(t_i) = |\rho_{22}(t_i) - 1|$ per timestep as a negative value, aiming at a fast flipping operation. The agent receives an extra reward of 1000 if ρ_{22} exceeds the threshold of $|\rho_{22}| > 0.99$, and terminates the episode for calculating the total reward early. We also notice that punishment of 100 if $|\rho_{11}| > 0.05$ at the final timestep helps the convergence of the model.

Figure 3 shows the high-fidelity closed-loop quantum control under various errors or noises. For relaxation on the X direction, we modify the terminal condition to $|\rho_{22}(t_i)| > 0.99$ for four neighboring timesteps, to prevent the model from converging on trivial resonant π-pulses. We use the Proximal Policy Optimization (PPO) method [43] to train the agent, with the learning rate being 1×10^{-3} and a batch size of 20. PPO is the well-known baseline algorithm for DRL, which guarantees the convergence in most cases. All other hyperparameters are set to the default values in Tensorforce v0.5.3 [44]. Moreover, we introduce a random error on the action, characterized by a centered Gaussian distribution with a standard deviation of 0.02, which emulates the time-varying systematic error in the quantum system. The models give control pulses that are robust against systematic errors. It is important to note that a trade-off between fidelity and robustness often exists. We obtain the models in figure 3 after about 2000, 3000, and 8000 episodes for controlling the system under detuning, dephasing, and relaxation on the X direction, respectively.

3.4. Transfer of the agent

A ML model is online for service after being trained for a particular task, such as flipping a qubit under $\hat{\sigma}_x$ relaxation as figure 3(c) does. One can evaluate the model by querying the information from the environment to check its validity after it is online. The flipped qubit can go for further tasks, which are independent of the model's duty. In this way, the performance of the model can be evaluated by checking the results of additional tasks without querying the environment or the model. If the performance of a well-trained model deviates from its expected behavior, one can conclude that the qubit in the environment

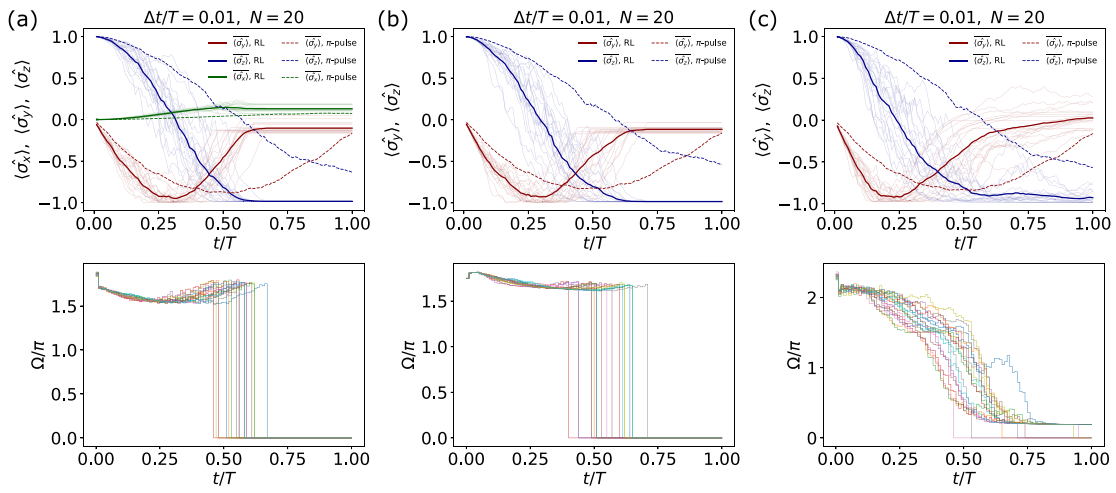

Figure 3. The expectations on X (green), Y (red), and Z (blue) directions of repetitively measured qubits, which are driven by trained DRL agents under (a) detuning, (b) dephasing, and (c) $\hat{\sigma}_x$ relaxation, respectively. The dynamics of each qubit are plotted by shaded curves $\langle\hat{\sigma}_i\rangle$, being averaged for the solid curve $\overline{\langle\hat{\sigma}_i\rangle}$. Fidelities are calculated by $F=\left(\mathrm{Tr}\sqrt{\sqrt{\rho}|1\rangle\langle 1|\sqrt{\rho}}\right)^2$, leading to average fidelities of 0.9922, 0.9919, and 0.9636 for each case, with the standard deviation of 0.0012, 0.0012, and 0.0361. Control pulses provided by the agent as step functions are also plotted in different colors. We set the detuning strength $\Delta=0.05\,\Omega$, the dephasing rate $\Gamma=0.05$, and the relaxation rate $\gamma=0.05$. Other parameters are the same as those in previous figures, which are listed in the main text. Dashed curves, as baselines, are derived by averaging the expectations of qubits under π-pulse control.

has changed, and quantum errors or noises have shifted to other patterns. It is then necessary to develop another model for precise control in the new environment. Instead of discarding the current model and training a new one, which would be inefficient, the agent can be transferred to the new environment in order to explore its capability to adapt to the new conditions with minimal effort.

We test the proposal by starting from the trained agent in figure 3(c). By directly evaluating the agent in a new environment in the presence of detuning $\Delta=0.1\,\Omega$, dephasing rate $\Gamma=0.05$, and $\hat{\sigma}_x$ relaxation rate $\gamma=0.05$, the average final state deviates significantly from the previous result (cf figure 4(a)), resulting in a decrease in fidelity as well. To recover the performance, we train the agent with the same setting for about 2000 episodes, and within an additional 20% of the total episodes, the agent retrieves its performance before the environment shift happens (cf figure 4(b)).

4. Discussion

Based on the numerical experiment in section 3, we have demonstrated that DRL can be employed to investigate the closed-loop quantum control. The fidelity can be further improved by fine-tuning the DRL agent in another training environment, with different thresholds and reward function designs. Interestingly, we have found out that the optimized policy from the agent is interpretable to some extent, as figures 3(c) and 4 shown. Specifically, we have observed that the maximal tunable Rabi frequency is 3π, which is 1.5 times the π-pulse for an operation time of $T=1$. The agent drives the qubit with a relatively high frequency, for reaching a large ρ_{22} as quickly as possible. It is understandable since continuous measurement can be described in the language of superoperators, affecting the dynamics like quantum noises, which can be effectively suppressed by reduced operation time. Later on, the pulse strength decreases significantly once ρ_{22} is large enough, converging to a small constant value for more precise operations. Accordingly, the weak measurement predominantly governs the state evolution instead of the control pulse. This behavior is similar to the quantum Zeno effect, which locks a wave function on its eigenstate by repeatedly performing projective measurements.

Now we further discuss this topic after analyzing the results above. In section 3, we explained that the RL environment consists of a qubit and a numerical simulation part. The qubit can be physical, e.g. constructed in superconducting circuits, trapped ions, photonics, etc or simulated by classical computers as we performed in numerical experiments. Here we emphasize again that the numerical simulation for calculating the qubit dynamics is compulsory if we include quantum information ρ_{ii} or fidelity in the RL state and reward. Although we can perform the weak measurement, extracting partial quantum information and converting it to weak value q_0 without destroying the quantum state, it is still impossible to retrieve the total

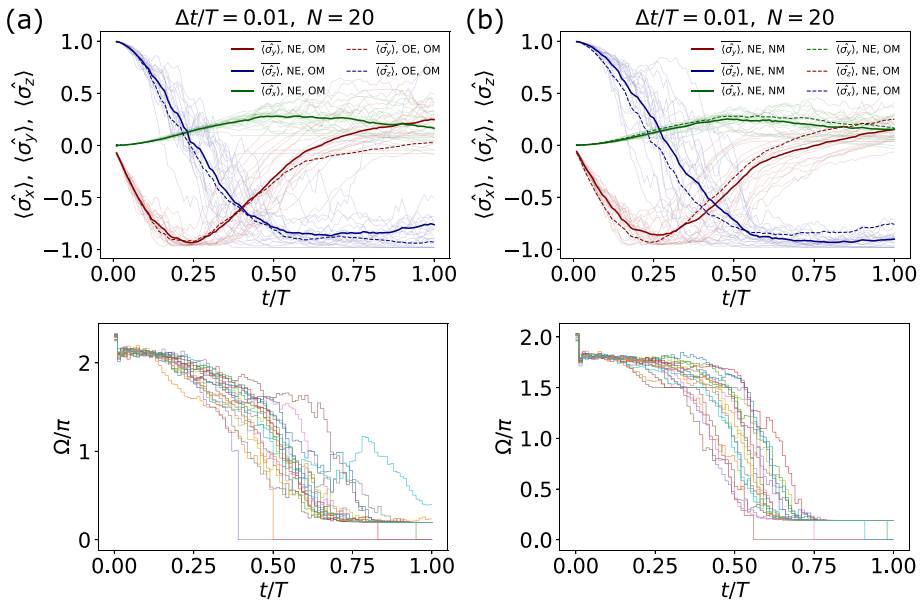

Figure 4. (a) The expectations on X (green), Y (red), and Z (blue) directions of repetitively measured qubits once the agent in figure 3(c) (Old Model trained in the Old Environment) is employed for driving the qubits in the presence of hybrid detuning errors and quantum noises (New Environment). The solid curves in figure 3(c) are plotted in dashed curves as the benchmark. The average fidelity decreases from 0.9636 to 0.8805 with the standard deviation of 0.1874. Imprecise control pulses outputted by the agent as step functions are plotted in different colors. (b) The expectations after training for about 2000 episodes. The solid curves in figure (a) (Old Model, New Environment) are plotted in dashed curves for the benchmark. The agent (New Model) fits the New Environment, retrieving an average fidelity of 0.9513 with the standard deviation of 0.0353. Corrected control pulses outputted by the agent as step functions are plotted in different colors.

information of the density matrix by a single shot of measurement. We cannot treat the qubit as a black box, as we usually do in other classical scenarios, where the observation of the RL state is instant and cost-less. By contrast, we have to calculate the qubit dynamics based on the actions and feedback, deducing ρ_{ii} without operating on the qubit. It becomes a setback when one performs stream learning in the quantum realm since simulating the quantum dynamics is time-consuming, e.g. about 7 s for an episode in our numerical experiment. However, the implementation in real quantum devices requires the simulation speedup of about 10^5 times (compared to the T1 time of state-of-the-art superconducting qubit). A possible solution is to train another ANN to mimic the dynamics of the quantum system, with available information as input, outputting the quantum information to be deduced without measurements. The training of such ANN needs plenty of training data and adequate training methods, which goes beyond the scope of this work.

Another method to avoid the black-box problem is to exclude the quantum information in the RL state. The RL state may contain the weak value q_0 and other classical information such as last action, the system time, etc. However, this approach comes with challenges. Since the threshold criteria for early termination are no longer available in this paradigm, the training environment only rewards the agent by a constant at the end of each episode once a projective measurement on the target state succeeds. The agent struggles to learn the precise control due to the low signal-to-noise ratio of q_0. The reward criteria also needs a large ensemble (batch size) to evaluate the fidelities of final quantum states. In this way, the problem becomes more difficult, which can be applied for evaluating RL algorithms.

5. Conclusion

In summary, we have studied the closed-loop quantum control of a noisy qubit using DRL. We have employed a Gaussian apparatus to extract the quantum information from the qubit through weak coupling. In the presence of detuning, dephasing, and relaxation, which are typical systematic error and quantum noises, we have developed the corresponding models for the bit-flipping task with high fidelity. Moreover, we have proved that transfer learning can be used to adapt a model to a new noise pattern instead of training from scratch, once the performance decay resulting from changing the noises and errors is observed. To facilitate reproducibility, we have made all source codes for the simulation of quantum dynamics, ML models, and evaluation scripts available on an open-source platform.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/yongchengding/closed-loop-qubit.

Acknowledgments

This work was financially supported by EU FET Open Grant EPIQUS (899368); the Basque Government through Grant No. IT1470-22; the Valencian Government Grant with Reference Number 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 2025 Agenda; the Project Grant PID2021-126273NB-I00 funded by MCIN/AEI/10.13039/501100011033 and by 'ERDF A way of making Europe' and 'ERDF Invest in your Future'; NSFC (12075145), STCSM (Grant No. 2019SHZDZX01-ZX04), and QUANTEK Project (KK-2021/00070). X C acknowledges 'Ayudas para contratos Ramón y Cajal'—2015–2020 (RYC-2017-22482).

ORCID iDs

Yongcheng Ding https://orcid.org/0000-0002-6008-0001
Xi Chen https://orcid.org/0000-0003-4221-4288
José D Martín-Guerrero https://orcid.org/0000-0001-9378-0285

References

[1] Nielsen M A and Chuang I 2010 *Quantum Computation and Quantum Information: 10th Anniversary Edition* (Cambridge: Cambridge University Press)
[2] Manju A and Nigam M J 2014 *Artif. Intell. Rev.* **42** 79–156
[3] Nguyen N H, Behrman E C, Moustafa M A and Steck J E 2020 *IEEE Trans. Neural Netw. Learn. Syst.* **31** 2522
[4] Shor P W 1996 *Proc. 37th Conf. on Foundations of Computer Science* (IEEE) pp 56–65
[5] Preskill J 1998 *Introduction to Quantum Computation and Information* (Singapore: World Scientific) pp 213–69
[6] Gottesman D 2010 *Proc. Symp. in Applied Mathematics* vol 68 pp 13–58
[7] Král P, Thanopulos I and Shapiro M 2007 *Rev. Mod. Phys.* **79** 53
[8] Brown K R, Harrow A W and Chuang I L 2004 *Phys. Rev.* A **70** 052318
[9] Torosov B T, Guérin S and Vitanov N V 2011 *Phys. Rev. Lett.* **106** 233001
[10] Rong X, Geng J, Shi F, Liu Y, Xu K, Ma W, Kong F, Jiang Z, Wu Y and Du J 2015 *Nat. Commun.* **6** 1
[11] Steffen M and Koch R H 2007 *Phys. Rev.* A **75** 062326
[12] Barnes E and Das Sarma S 2012 *Phys. Rev. Lett.* **109** 060401
[13] Daems D, Ruschhaupt A, Sugny D and Guérin S 2013 *Phys. Rev. Lett.* **111** 050404
[14] Dridi G, Liu K and Guérin S 2020 *Phys. Rev. Lett.* **125** 250403
[15] Guéry-Odelin D, Ruschhaupt A, Kiely A, Torrontegui E, Martínez-Garaot S and Muga J G 2019 *Rev. Mod. Phys.* **91** 045001
[16] Torrontegui E, Ibánez S, Martínez-Garaot S, Modugno M, Campo A D, Guéry-Odelin D, Ruschhaupt A, Chen X and Muga J G 2013 *Adv. At. Mol. Opt.* **62** 117
[17] Chen X, Ruschhaupt A, Schmidt S, Campo A D, Guéry-Odelin D and Muga J G 2010 *Phys. Rev. Lett.* **104** 063002
[18] Zahedinejad E, Ghosh J and Sanders B C 2016 *Phys. Rev. Appl.* **6** 054005
[19] Liu B-J, Song X-K, Xue Z-Y, Wang X and Yung M-H 2019 *Phys. Rev. Lett.* **123** 100501
[20] Ding Y, Ban Y, Martín-Guerrero J D, Solano E, Casanova J and Chen X 2021 *Phys. Rev.* A **103** L040401
[21] Ai M Z, Ding Y, Ban Y, Martín-Guerrero J D, Casanova J, Cui J M, Huang Y F, Chen X, Li C F and Guo G C 2022 *Sci. China: Phys. Mech. Astron.* **65** 1
[22] Yao J, Lin L and Bukov M 2021 *Phys. Rev.* X **11** 031070
[23] Bukov M, Day A G R, Sels D, Weinberg P, Polkovnikov A and Mehta P 2018 *Phys. Rev.* X **8** 031086
[24] Porotti R, Tamascelli D, Restelli M and Prati E 2019 *Commun. Phys.* **2** 1
[25] Niu M Y, Boixo S, Smelyanskiy V N and Neven H 2019 *npj Quantum Inf.* **5** 1
[26] Dalgaard M, Motzoi F, Sørensen J J and Sherson J 2020 *npj Quantum Inf.* **6** 1
[27] Zhang X-M, Cui Z-W, Wang X and Yung M-H 2018 *Phys. Rev.* A **97** 052333
[28] Wu R-B, Ding H, Dong D and Wang X 2019 *Phys. Rev.* A **99** 042327
[29] Ostaszewski M, Miszczak J, Banchi L and Sadowski P 2019 *Quantum Inf. Process.* **18** 1
[30] Borah S, Sarma B, Kewming M, Milburn G J and Twamley J 2021 *Phys. Rev. Lett.* **127** 190403
[31] Chen C, Dong D, Li H-X, Chu J and Tarn T-J 2014 *IEEE Trans. Neural Netw. Learn. Syst.* **25** 920
[32] Martín-Guerrero J D and Lamata L 2022 *Neurocomputing* **470** 457
[33] Martín-Guerrero J D and Lamata L 2021 *Appl. Sci.* **11** 8589
[34] Henson B M, Shin D K, Thomas K F, Ross J A, Hush M R, Hodgman S S and Truscott A G 2018 *Proc. Natl Acad. Sci.* **115** 13216
[35] Zhang X M, Wei Z, Asad R, Yang X C and Wang X 2019 *npj Quantum Inf.* **5** 1
[36] Haug T, Mok W K, You J B, Zhang W, Png C E and Kwek L C 2020 *Mach. Learn.: Sci. Technol.* **2** 01LT02
[37] An Z and Zhou D 2019 *Europhys. Lett.* **126** 60002
[38] Sørdal V B and Bergli J 2019 *Phys. Rev.* A **100** 042314
[39] Aharonov Y, Albert D Z and Vaidman L 1998 *Phys. Rev. Lett.* **60** 1351

[40] Gross J A, Caves C M, Milburn G J and Combes J 2018 *Quantum Sci. Technol.* **3** 024005
[41] Jacobs K and Steck D A 2006 *Contemp. Phys.* **47** 279
[42] Ding Y, Martín-Guerrero J D, Sanz M, Magdalena-Benedicto R, Chen X and Solano E 2020 *Phys. Rev. Lett.* **124** 140504
[43] Schulman J, Wolski F, Dhariwal P, Radford A and Klimov O 2017 arXiv:1707.06347
[44] Kuhnle A, Schaarschmidt M and Fricke K 2017 Tensorforce: a tensorflow library for applied reinforcement learning (available at: https://github.com/tensorforce/tensorforce)