NISQ State Preparation with Reinforcement Learning

Muhammad Mustafa Umar Gondel, Syed Muhammad Abuzar Rizvi, Uman Khalid, and Hyundong Shin
Department of Electronics and Information Convergence Engineering, Kyung Hee University, Korea
Email: hshin@khu.ac.kr

***Abstract*—Quantum state preparation (QSP) is a fundamental task in quantum computing, essential for initializing algorithms and protocols with high fidelity. However, achieving reliable initial state on Noisy Intermediate-Scale Quantum (NISQ) devices remains challenging due to decoherence and operational noise. In this work, we formulate QSP as a reinforcement learning (RL) problem and investigate the effectiveness of two RL algorithms—Policy Gradient (PG) and Deep Q-Network (DQN)—in preparing a target $|+\rangle$ state from the $|0\rangle$ state under dephasing and depolarizing noise. Experimental results show that DQN achieves higher fidelity and stronger robustness to noise, highlighting its effectiveness for quantum control in realistic, noisy environments.**

I. Introduction

Quantum state preparation (QSP) is an essential prerequisite for many quantum technologies, including quantum computing, communication, and metrology [1], [2], [3]. The ability to initialize a quantum system into a desired state with high fidelity underpins algorithms and protocols across these domains [4]. However, achieving reliable state preparation on real hardware is notoriously challenging due to the presence of quantum noise and decoherence in Noisy Intermediate-Scale Quantum (NISQ) devices [5]. Common noise processes such as dephasing and depolarizing errors can drastically reduce the fidelity of prepared states. These noise channels are widely used to model realistic qubit decoherence and gate errors [6]. Overcoming such noise-induced errors is crucial for unlocking the full potential of QSP in practical devices.

Conventional quantum control methods (e.g., gradient-based pulse shaping or optimal control) often assume an accurate system model and yield open-loop control sequences that do not adapt to noise fluctuations, leading to suboptimal performance in the presence of unmodeled disturbances [7]. To address these limitations, this work investigates reinforcement learning (RL) as a model-free, noise-resilient alternative for QSP, enabling agents to learn robust control policies directly from interaction with a noisy quantum environment.

II. Methodology

QSP is formulated as a RL problem in which an agent learns a control policy to transform an initial qubit state into a target state. The qubit is represented by a density matrix $\rho \in \mathbb{C}^{2\times 2}$, with fidelity to the target state ρ_{target} used as the reward signal. The agent interacts with the system over discrete time steps, and each episode ends when the fidelity exceeds a set threshold or after a maximum number of steps T.

At each time step t, the agent selects an action corresponding to a unitary operation U_a from a discrete action set. This operation evolves the state as

$$\rho_{t+1} = \mathcal{E}_\gamma \left(U_a \rho_t U_a^\dagger \right), \tag{1}$$

where $\mathcal{E}_\gamma$ is a noise channel parameterized by noise level γ. The fidelity between the evolved state and the target is used as a reward signal and is defined as

$$F(\rho_t, \rho_{\text{target}}) = \left(\text{Tr} \left[\sqrt{\sqrt{\rho_t}\rho_{\text{target}}\sqrt{\rho_t}} \right] \right)^2. \tag{2}$$

The environment uses the Bloch sphere representation for state observation and includes four discrete control actions:

$$\left\{ e^{-i\delta t I},\ e^{-i\delta t \sigma_x},\ e^{-i\delta t \sigma_y},\ e^{-i\delta t \sigma_z} \right\}, \tag{3}$$

where $\delta t = \frac{2\pi}{N}$, and N is the number of time steps in an episode. These unitary operations represent rotations around the respective Pauli axes and the identity, enabling the agent to explore the state space.

Two quantum noise models are implemented to simulate real-world decoherence effects. The first is a dephasing channel, modeling phase noise, defined as

$$\mathcal{E}_{\text{deph}}(\rho) = \left(1 - \frac{\gamma}{2}\right)\rho + \frac{\gamma}{2}\sigma_z \rho \sigma_z. \tag{4}$$

The second is a depolarizing channel, which introduces isotropic noise by randomly applying Pauli operations, and is given by

$$\mathcal{E}_{\text{depol}}(\rho) = \left(1 - \frac{3\gamma}{4}\right)\rho + \frac{\gamma}{4}\left(\sigma_x \rho \sigma_x + \sigma_y \rho \sigma_y + \sigma_z \rho \sigma_z\right). \tag{5}$$

Two RL algorithms are employed for this task. The first is a Policy Gradient (PG) method using the REINFORCE algorithm, where the policy $\pi_\theta(a \mid s)$ is optimized to maximize expected cumulative reward. The gradient update is computed as

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t \mid s_t) R_t \right], \tag{6}$$

where θ are the parameters of the policy network.

The second approach is a Deep Q-Network (DQN), which approximates the optimal action-value function $Q(s, a)$ using a neural network. The Q-values are updated by minimizing the temporal difference (TD) error:

$$L = \mathbb{E}\left[\left(r + \gamma \max_{a'} Q_{\text{target}}(s', a') - Q(s, a) \right)^2 \right]. \tag{7}$$

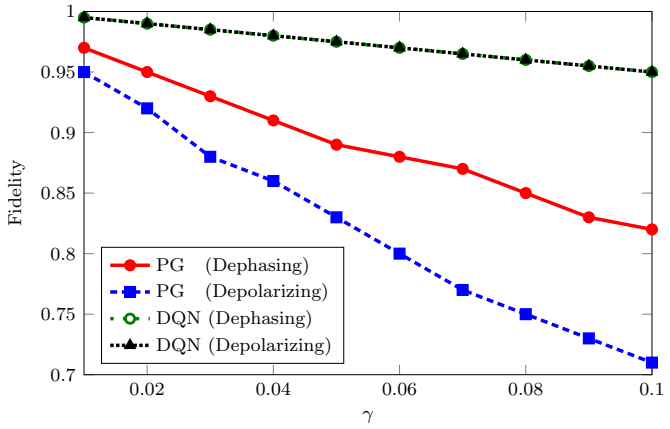

Figure 1. Comparison of maximum fidelity achieved by PG ($T = 2^3$) and DQN ($T = 2^0$) versus varying dephasing and depolarizing noise.

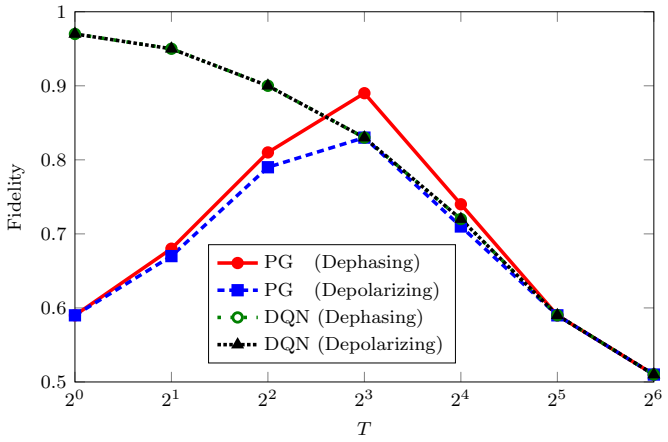

Figure 2. Comparison of fidelity achieved by PG and DQN under dephasing and depolarizing noises (γ = 0.05) along maximum time steps T.

The agent selects actions using an ϵ-greedy policy during training and leverages a target network and experience replay for stability.

III. Results

We evaluate the performance of trained RL agents; PG and DQN for the task of QSP, specifically transforming the initial $|0\rangle$ state into the target $|+\rangle$ state, under varying time steps and noise levels. All models were trained for 1200 episodes with 64 time steps per episode using the Adam optimizer and a learning rate of 0.001. For DQN, exploration was managed using an ϵ-greedy strategy with an initial $\epsilon = 1.0$, decayed by a factor of 0.995 per episode to a minimum of 0.01. Performance during evaluation was measured over 100 test trajectories, using a fixed initial quantum state across all noise levels to ensure consistency.

Analyzing the impact of noise strength γ, varied between 0.01 and 0.1 as shown in Fig. 1, DQN consistently demonstrates stronger robustness to decoherence across both dephasing and depolarizing channels. Its fidelity remains above 95% even at the highest noise level tested ($\gamma = 0.1$), indicating effective generalization under noisy conditions. In contrast, PG exhibits greater sensitivity to noise, especially under depolarizing noise, with a notable drop in fidelity at higher values of γ.

The comparison of average fidelity across increasing maximum time steps T with fixed noise ($\gamma = 0.05$), shown in Fig. 2, reveals differences in the convergence behavior of the two methods. DQN achieves rapid improvement, reaching a fidelity of 97% at early time steps under both noise models. PG, on the other hand, shows a slower but steady improvement, attaining maximum fidelities of 89% under dephasing and 83% under depolarizing noise at $T = 2^3$. These trends suggest that DQN can achieve high fidelity under minimal time steps,while PG requires more time steps to reach its peak performance and is less effective in noisy environments.

IV. Conclusion

This study explored QSP in noisy environments using RL, comparing the performance of PG and DQN algorithms. Through extensive training and evaluation across varying time steps and noise levels, we observed that DQN consistently outperformed PG, achieving higher fidelities more rapidly and maintaining strong robustness under both dephasing and depolarizing noise. These results suggest that value-based methods like DQN are more effective for fast and robust QSP in noisy environments.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under RS-2025-00556064 and by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2025-2021-0-02046) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), and by a grant from Kyung Hee University in 2023 (KHU-20233663).

References

[1] U. Khalid, M. S. Ulum, A. Farooq, T. Q. Duong, O. A. Dobre, and H. Shin, "Quantum semantic communications for metaverse: Principles and challenges," *IEEE Trans. Wireless Commun.*, vol. 30, no. 4, pp. 26–36, Sep. 2023.

[2] R. C. Farrell, M. Illa, A. N. Ciavarella, and M. J. Savage, "Scalable circuits for preparing ground states on digital quantum computers: The schwinger model vacuum on 100 qubits," *PRX Quantum*, vol. 5, no. 2, p. 020315, Apr. 2024.

[3] U. Khalid, J. ur Rehman, H. Jung, T. Q. Duong, O. A. Dobre, and H. Shin, "Quantum property learning for nisq networks: Universal quantum witness machines," vol. 73, no. 4, pp. 2207–2221, Apr. 2024.

[4] V. V. Sivak, A. Eickbusch, H. Liu, B. Royer, I. Tsioutsios, and M. H. Devoret, "Model-free quantum control with reinforcement learning," *Phys. Rev. X*, vol. 12, no. 1, p. 011059, Mar. 2022.

[5] U. I. Paracha, S. M. A. Rizvi, M. M. U. Gondel, W. Park, and H. Shin, "Adaptive quantum readout error mitigation with transfer learning," in *2024 15th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, Oct. 2024, pp. 506–511.

[6] V. Tripathi, D. Kowsari, K. Saurav, H. Zhang, E. M. Levenson-Falk, and D. A. Lidar, "Benchmarking quantum gates and circuits," *Chem. Rev.*, vol. 125, no. 9, pp. 4567–4598, May 2025.

[7] 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 Inform.*, vol. 5, no. 1, p. 85, Oct. 2019.