Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ31187

Deep Q-network in the Iterated Prisoner's Dilemma under Noise

Yixiao Chen

Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore
yixiao.chen@u.nus.edu

Abstract. In game theory, there is a fundamental challenge about maintaining cooperation among selfish players, especially under practical noise. This study applies a noisy Iterated Prisoner's Dilemma (IPD) model to investigate how learning strategies perform against classical strategies when players may receive false or misleading signals due to random observation errors. More specifically, this study compares Deep Q-Network (DQN) agents with basic Q-learning (QL) and several classical strategies such as Tit-for-Tat, Win-Stay-Lose-Shift, and Grudger. The experiment results show that when noise emerges, DQN agents not only achieve higher cumulative rewards than other strategies but also maintain more stability, adaptability, and resilience across repeated interactions. DQN agents' deep neural structure helps them to capture long-term temporal dependencies, effectively differentiate accidental defections from intentional ones, and recover cooperation after disturbances by noise. These findings indicate that deep reinforcement learning is effective in noisy and imperfect settings. The findings also offers valuable insights for understanding the emergence of cooperation and for designing robust multi-agent decision-making mechanisms in noisy or uncertain environments.

Keywords: Deep Q-Network, Game Theory, Prisoner's Dilemma, Noisy Environment

1. Introduction

The Prisoner's Dilemma (PD) is one of the best-known models in game theory. It shows the basic conflict between cooperation and defection [1]. Cooperation gives the highest total reward, but acting selfishly often brings a better individual outcome. This will become a dilemma that appears in many areas such as economics [2], biology [3], and social science [4]. The Iterated Prisoner's Dilemma (IPD) repeats the same game between the same players. Because the players remember past rounds, they can change their choices over time. The IPD, therefore, helps researchers study long-term decision making and the growth of cooperation [5].

Most studies of the IPD assume that players can see each other's actions clearly and know the outcome right away. Real situations are usually more complex. People can misread signals, make judgment errors, or simply fail to notice what really happened [6]. In these situations, a friendly move can look like betrayal, and also a short delay or bit of noise can change how the other side

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ31187

reacts. Once this kind of confusion appears, trust starts to fade and cooperation often breaks down, which will also influence the long-term actions among each players in practical scenarios.

The noisy version of the IPD was introduced to capture these practical problems. It adds random errors to what players observe so that uncertainty is built into the game itself. The question is whether cooperation can still be learned when players cannot fully rely on what they see.

Classic strategies like Tit-for-Tat and Grim Trigger perform well without noise but rapidly fail when noise is introduced [7]. This is because they rely on relatively precise simulation of each other's behavior. This means that even minor errors can trigger a chain reaction of unnecessary revenge. Traditional reinforcement learning methods face similar challenges: due to their reliance on shallow neural networks, they cannot process longer historical information [8]. These shortcomings highlight the need for a learning method to be able to adapt and restore stability in noisy environments.

Three main contributions are made here. Firstly, a simulation setup is built with observation noise. Secondly, an advanced DQN design is developed that learns strategies to resist misinformation. Thirdly, experiments show how noise changes agents' behavior and how different strategies perform.

Overall, the results show that deep reinforcement learning can achieve better rewards even when observations are sometimes fake. The findings may be useful for social dilemmas and distributed systems, where misunderstanding and noise are normal parts of interaction.

2. Preliminary

2.1. IPD

The classical PD shows how two players must each choose between cooperating to defecting without knowing what the other player will do. Mutual cooperation helps both sides, while mutual defection leaves them worse off. If one defects but the other cooperates, the defector gains the most and the cooperator gets the least. This simple game captures the conflict between acting for oneself and acting for the common good. In a single play, defection always seems the rational choice. Yet when the same players meet again and again, cooperation can still appear through memory, trust, and the threat of future punishment.

Through multiple rounds of interaction, IPD offers players the opportunity to adjust their tactics on the basis of experience. This repeat process lets players observe the value of dynamic benefits, and then move towards balancing short-term against long-term benefits. Simple strategies such as Tit-for-Tat show that cooperation can endure if each player reacts similarly to the other's response. Such an interaction pattern creates knowability of trust and tolerance over time, demonstrating that cooperation can indeed persist over time even when competition never totally disappears.

2.2. Noise in repeated interactions

Many theoretical methods for IPD therefore start from a perfect assumption: every player knows perfectly how the opponent has played and gets all the payoffs he or she is supposed to get. Interaction in the world, however, is not quite so clean. Communication breaks down, signals may be noisy, perceptions are sometimes imperfect. Cooperation is an act that can look like a defection and provoke unwarranted retaliation, whereas defection can look genuine and trick partners into taking less credible punishments. These small mistakes repeated over and over can do away with the balance that keeps cooperation flourishing.

DOI: 10.54254/2755-2721/2026.TJ31187

The noisy IPD incorporates this uncertainty by introducing stochastic errors into the observation process. Specifically, after each action is chosen, the opponent's observed move may be flipped with some probability ε, representing observation noise. This simple modification dramatically changes the dynamics of the game: even when both players intend to cooperate, noise can generate apparent defections, which may spiral into sustained mutual punishment. The central challenge is therefore to design strategies that are robust to such misperceptions, maintaining cooperation when possible, while avoiding exploitation when noise obscures intentions.

2.3. Baseline strategies

In addition to learning-based agents, fixed strategies play a crucial role as benchmarks. These strategies are deterministic or stochastic rules that do not adapt during play but provide interpretable behaviors against which learning agents can be tested. A diverse set of baselines is selected to capture different archetypes of cooperative and defecting behavior [9].

• Always Cooperate: A naive strategy that cooperates in every round, regardless of history. Serves as an optimistic upper bound for cooperative potential, but is easily exploited by defectors.

• Always Defect: The opposite extreme, defecting in every round. Represents the individually rational choice in one-shot games and acts as a pessimistic baseline.

• Tit-for-Tat (TFT): Begins with cooperation and then replicates the opponent's last move. Known to sustain cooperation in noise-free IPD but highly sensitive to misperceptions under noise. Generous variants occasionally forgive defections, making them more robust.

• Grudger: Cooperates until the opponent defects once, after which it defects forever. Extremely effective in enforcing cooperation under perfect observation but brittle in noisy settings.

• Win-Stay-Lose-Shift (WSLS): Repeats the previous action if the payoff was a win, and changes when it is a loss. Often considered a strong strategy in noisy environments.

• Random: Chooses cooperation or defection with equal probability. Provides a non-strategic baseline for comparison.

Together, these strategies span the spectrum from fully cooperative to fully defecting, from rigid to adaptive, and from forgiving to punitive. They enable a systematic evaluation of learning agents against diverse behavioral patterns, thereby highlighting both strengths and weaknesses.

2.4. Reinforcement learning

Reinforcement learning (RL) provides a natural framework for studying adaptive behavior in repeated games. In RL, an agent learns by interacting with an environment, updating its policy to maximize long-term cumulative reward. In the context of the noisy IPD, the environment consists of the opponent's actions and the stochastic observation process, while the reward corresponds to the payoff of each round. Two forms of RL are particularly relevant: tabular Q-learning and DQN.

Q-learning is a model-free algorithm that learns a value function, representing the expected return of taking action in a state. The update rule is defined as

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)] \tag{1}$$

In IPD-like games, the state is typically defined by the recent history of moves, and the update rule balances immediate rewards with discounted future returns. Although Q-learning can converge to effective policies in simple environments, its reliance on tabular representations limits its

DOI: 10.54254/2755-2721/2026.TJ31187

scalability. As the history length increases or noise complicates the mapping between actions and outcomes, the state space grows exponentially, and tabular approaches become infeasible.

In contrast, DQN extends Q-learning by approximating the value function with a neural network. Instead of storing a separate entry for each state-action pair, the network generalizes across similar histories, enabling learning in large or continuous state spaces [10]. In the noisy IPD, this allows the agent to process extended interaction histories and infer patterns despite imperfect observations. Moreover, advanced DQN variants incorporate mechanisms such as replay memory, target networks, and adaptive exploration schedules, which improve stability and efficiency.

The contrast between Q-learning and advanced DQN thus provides a meaningful axis of comparison in the experiments: the former represents a classical baseline, while the latter embodies a modern deep reinforcement learning approach capable of handling noise and complexity.

3. Method

The performance and robustness of the proposed method were evaluated using the noisy IPD with different agents (strategies). In this setting, two agents repeatedly interact over a lot of rounds, and each player's observation at the opponent's previous action is influenced by stochastic observation noise, denoted as ε. This configuration emulates realistic social or multi-agent environments, where limited information, signal corruption, and misinterpretation frequently occur.

Each match consisted of 1,000 rounds, and the final results were averaged over multiple independent runs to minimize random variation. The payoff matrix followed the conventional IPD setup: 5, 3, 1, 0, corresponding respectively to temptation, reward, punishment, and sucker's payoff. Each agent aimed to maximize its cumulative payoff across the entire episode. The variation of ε from 0.0 to 0.2 allowed evaluation of how increasing observation uncertainty influences learning dynamics and strategy adaptation.

Both DQN and QL agents shared identical reward definitions and exploration mechanisms to ensure fairness. The DQN utilized a fully connected neural network with one hidden layer of 128 ReLU units, with a learning rate at 0.01, and a discount factor $\gamma = 0.9$ unless otherwise stated. To stabilize learning, experience replay was employed with a buffer size of 10,000 and online updates of size 1. The Q-learning (QL) agent used the same set of hyperparameters but relied on a tabular Q-value representation instead of a neural network. At the start of training, all Q-values were initialized to zero, and both agents followed an ε-greedy exploration rule in which ε decreased linearly from 0.9 to 0.05 across episodes.

This design helps us to understand the impact of the function approximation ability of deep neural networks on learning efficiency and stability in a noisy environment alone. By keeping all other conditions consistent, performance differences can be attributed to the model structure itself, rather than from parameter tuning.

4. Experiment

4.1. Training and evaluation

During training, each learning agent played repeatedly against classical strategies chosen randomly. Faced with such diverse strategies, learners encounter both cooperative and deceitful behaviors, allowing adaptability to gradually emerge over time.

After every match, the DQN adjusted its parameters using stochastic gradient descent on the temporal-difference (TD) error, whereas the Q-learning agent updated its table entries following the

DOI: 10.54254/2755-2721/2026.TJ31187

Bellman rule. Two evaluation metrics were used to assess performance:

• Average Points per Game (APG): The long-term mean reward per round, representing efficiency and stability of learned behavior [11].

• Training Stability (TS): Measured by the smoothness and convergence of loss curves across episodes, reflecting the consistency of policy updates under noise.

These metrics collectively provided a balanced view of both quantitative performance and qualitative behavioral tendencies.

4.2. Results

Figure 1 presents the average points per game achieved by all strategies under moderate noise (10%). The DQN agent achieved the highest mean payoff (2.55), outperforming QL (2.46) and all fixed baselines. This demonstrates the effectiveness of the deep function approximator in capturing complex state-action dependencies even under imperfect observation.

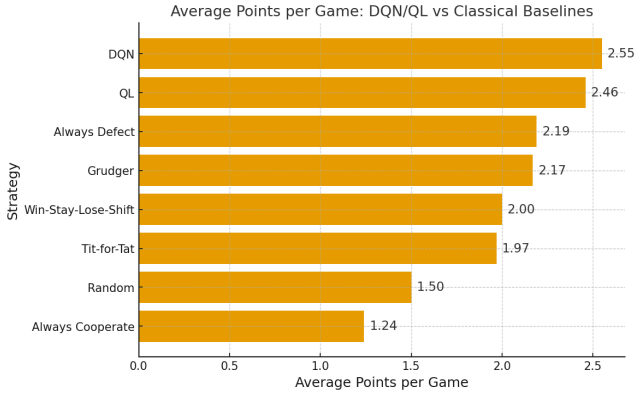

Figure 1. Average points per game of eight strategies under 10% noise

Among classical strategies, Grudger and WSLS achieved competitive results under low noise but exhibited rapid performance degradation as noise increased. Their reactive and deterministic nature caused misinterpretations of accidental defections, triggering unnecessary punishment loops. The Tit-for-Tat family, despite its elegance in ideal conditions, similarly suffered from noise sensitivity, as a single misread action could cascade into mutual defection. Always Cooperate and Random served as weak baselines: the former was heavily exploited by defectors, while the latter lacked any consistent adaptation mechanism, yielding near-random payoffs.

The QL agent displayed moderate adaptability in early stages but proved fragile under higher noise levels (noise > 10%). Its reliance on discrete state representations caused overfitting to local observations and overreaction to spurious defections. In contrast, the DQN's neural network captured more generalizable representations of historical interaction sequences, enabling it to infer intent more effectively and maintain stable cooperation.

In qualitative terms, DQN agents exhibited “forgiveness” behavior: occasional tolerance toward perceived defections, thereby preventing retaliation spirals. This behavior emerged naturally through training, without explicit rule-based encoding. QL agents, however, lacked such flexibility, often switching to persistent defection after a few noisy interactions.

Although the current study emphasizes mean payoff results, preliminary analysis of training dynamics also supports this conclusion. The DQN’s TD loss consistently declined with low variance, indicating stable convergence, whereas the QL loss oscillated heavily across episodes,

DOI: 10.54254/2755-2721/2026.TJ31187

especially in environments with higher noise. Figure 2 illustrates the training loss curve of the DQN agent, showing a rapid decrease in loss during early epochs followed by gradual stabilization, confirming effective learning and convergence. Future work will further quantify these convergence patterns to strengthen the comparative analysis of learning stability.

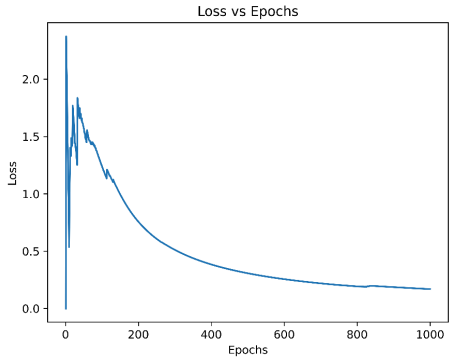

Figure 2. Loss curve of the DQN agent

5. Conclusion

These experiments verify that after 600 epochs, deep reinforcement learning learns highly efficiently on noisy signals, maintaining an easy combination of cooperation in key steps even in the presence of uncertainty, where a high average score per game of DQN was the best for all strategies, yielding better results than QL and traditionally even beating traditions. That is why DQN agents gradually develop an intrinsic understanding of strategy dynamics and errors through trade-off exploration and exploitation. This allows the DQN agents to well-discriminate true betrayal from random mistakes in interpreting opponents strategy but avoid meaningless retaliation and retain long-term trust when studying opponents. By contrast, tabular QL agents are constrained by shallow representations and cannot adapt to changes in the environment. Therefore, their generalization capability is limited, and they often lead to brittle or plateaued strategies that are incapable of recovering from misunderstandings caused by noise.

Overall, the results of our research suggest that deep reinforcement learning methods can not only attain higher returns, but also enhance sustain of cooperation in noisy I-PD. Thus, the reinforcement learning approach presented here will provide strong and flexible possibilities. Going beyond these quantifiable performance improvements, DQN agents can show human-like behavior, with agents possessing feelings of tolerance and forgiveness, and showing day-to-day coping when making mistakes instead of behavior being suicidal (doing revenge blindness). This further sophistication implies that future work can extend this method to larger multi-agent environments and study how cooperation results, settles, and evolves in uncertain and partially-known situations.

References

[1] Axelrod, R. (1984). The evolution of cooperation. Basic Books.
[2] Dal Bó, P., & Fréchette, G. R. (2019). Strategy choice in the infinitely repeated prisoner's dilemma. American Economic Review, 109(11), 3929–3952.
[3] Imhof, L. A., Fudenberg, D., & Nowak, M. A. (2005). Evolutionary cycles of cooperation and defection. Proceedings of the National Academy of Sciences, 102(31), 10797–10800.
[4] Majeski, S. J. (1984). Arms races as iterated prisoner's dilemma games. Mathematical Social Sciences, 7(3), 253–266.

DOI: 10.54254/2755-2721/2026.TJ31187

[5] Nowak, M. A. (2006). Five rules for the evolution of cooperation. Science, 314(5805), 1560–1563.
[6] Wu, J., & Axelrod, R. (1995). How to cope with noise in the iterated prisoner's dilemma. Journal of Conflict Resolution, 39(1), 183–189.
[7] Rapoport, A. (1989). Prisoner's dilemma. In J. Eatwell, M. Milgate, & P. Newman (Eds.), Game theory (pp. 199–204). Palgrave Macmillan UK. https: //doi.org/10.1007/978-1-349-20181-5_23
[8] Cohen, J., & Holland, B. (2024). Deep learning and the prisoner's dilemma: A strategic evaluation. The National High School Journal of Science.
[9] Kuhn, S. (2024, Winter). Prisoner's dilemma. In E. N. Zalta & U. Nodelman (Eds.), The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, Stanford University.
[10] Fan, J., Wang, Z., Xie, Y., & Yang, Z. (2020). A theoretical analysis of deep Q-learning. In Learning for Dynamics and Control (pp. 486–489). PMLR.
[11] Sandholm, T. W., & Crites, R. H. (1996). Multiagent reinforcement learning in the iterated prisoner's dilemma. Biosystems, 37(1–2), 147–166.