Towards Optimal Adversarial Robust Q-learning with Bellman Infinity-error

Haoran Li [1] **Zicheng Zhang** [1] **Wang Luo** [1] **Congying Han** [1] **Yudong Hu** [1] **Tiande Guo** [1] **Shichen Liao** [1]

Abstract

Establishing robust policies is essential to counter attacks or disturbances affecting deep reinforcement learning (DRL) agents. Recent studies explore state-adversarial robustness and suggest the potential lack of an optimal robust policy (ORP), posing challenges in setting strict robustness constraints. This work further investigates ORP: At first, we introduce a consistency assumption of policy (CAP) stating that optimal actions in the Markov decision process remain consistent with minor perturbations, supported by empirical and theoretical evidence. Building upon CAP, we crucially prove the existence of a deterministic and stationary ORP that aligns with the Bellman optimal policy. Furthermore, we illustrate the necessity of L^∞-norm when minimizing Bellman error to attain ORP. This finding clarifies the vulnerability of prior DRL algorithms that target the Bellman optimal policy with L^1-norm and motivates us to train a Consistent Adversarial Robust Deep Q-Network (CAR-DQN) by minimizing a surrogate of Bellman Infinity-error. The top-tier performance of CAR-DQN across various benchmarks validates its practical effectiveness and reinforces the soundness of our theoretical analysis. Our code is available at https://github.com/leoranlmia/CAR-DQN.

1. Introduction

Deep reinforcement learning (DRL) has shown remarkable success in solving intricate problems (Mnih et al., 2015; Lillicrap et al., 2015; Silver et al., 2016) and holds promise across diverse practical domains (Ibarz et al., 2021; Kiran et al., 2021; Yu et al., 2021; Zheng et al., 2018). Nevertheless, subtle perturbations in state observations can severely degrade well-trained DRL agents (Huang et al., 2017; Behzadan & Munir, 2017a; Lin et al., 2017; Ilahi et al., 2021), which limits their trustworthy real-world deployment and emphasizes the crucial need for developing robust DRL algorithms against adversarial attacks.

Pioneering work by Zhang et al. (2020) introduced the state-adversarial paradigm in DRL by formulating a modified Markov decision process (MDP), termed SA-MDP. Here, the underlying true state remains invariant while the observed state is subjected to disturbances. They also highlighted the uncertain existence of an optimal robust policy (ORP) within SA-MDP, indicating a potential conflict between robustness and optimal policy. Consequently, existing methods built upon SA-MDP often seek a trade-off between robust and optimal policies through various regularizations (Oikarinen et al., 2021; Liang et al., 2022) or adversarial training (Zhang et al., 2021; Sun et al., 2021). While enhancing robustness, these methods lack theoretical guarantees and completely neglect the study of ORP.

In this paper, we primarily focus on investigating ORP within SA-MDP. We suppose that only a few exceptional states lack an ORP, thus it is key for theoretical clarity to eliminate these states. Therefore, our work starts with a consistency assumption of policy (CAP), where optimal actions of *all* states within the MDP exhibit consistency despite adversarial disturbance. This implies that, from a decision-making perspective, adversaries cannot alter the essence of state observations, which we call the intrinsic state. Despite seeming implausible, we support the rationality of CAP through theoretical analysis and empirical experiments against strong adversarial attacks like FGSM (Goodfellow et al., 2014) and PGD (Madry et al., 2017), showcasing that the state set violating CAP is nearly empty.

Building upon CAP, we demonstrate that there always exists a stationary and deterministic adversarial ORP coinciding with the Bellman optimal policy, derived from the Bellman optimality equations. Remarkably, this objective has been widely employed in previous DRL algorithms (Silver et al., 2014; Schulman et al., 2015; Wang et al., 2016; Mnih et al., 2016; Schulman et al., 2017) to maximize the natural returns, lacking robust capabilities. Hence, our findings incredibly unveil that *the Bellman optimal policy doubles as the ORP*, and improving the robustness of DRL agents need not com-

[1]School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China. Correspondence to: Congying Han <hancy@ucas.ac.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

promise their performance in natural environments. This insight holds significant value for establishing DRL agents with Bellman optimality equations in real-world scenarios where strong adversarial attacks are relatively rare.

In pursuit of ORP, we delve into understanding *why conventional DRL algorithms, which target the Bellman optimal policy, fail to guarantee adversarial robustness.* By analyzing the theoretical properties of $\|Q_\theta - Q^*\|_p$ and Bellman error $\|\mathcal{T}_B Q_\theta - Q_\theta\|_p$ across diverse Banach spaces, we identify the substantial impact of the parameter p on adversarial robustness. Specifically, achieving ORP corresponds to minimizing the Bellman Infinity-error, *i.e.*, $p = \infty$, whereas conventional algorithms are typically linked to $p = 1$. To address the computational challenges arising from L^∞-norm, we propose the Consistent Adversarial Robust Deep Q-Network (CAR-DQN), utilizing a surrogate objective of Bellman Infinity-error for robust policy learning.

To summarize, our paper makes the following contributions:

(**1**). We propose the rational CAP for SA-MDP, confirm the existence of deterministic and stationary ORP, and demonstrate its strict alignment with the Bellman optimal policy, which is a significant advancement over prior research.

(**2**). We underscore the necessity of employing the L^∞-norm to minimize Bellman error for theoretical ORP attainment. This stands in contrast to conventional DRL algorithms, which lack robustness due to the use of an L^1-norm.

(**3**). We devise CAR-DQN solely utilizing a surrogate objective based on the L^∞-norm to learn both natural return and robustness. We conduct comparative evaluations of CAR-DQN against state-of-the-art approaches across various benchmarks, validating its practical effectiveness and reinforcing our theoretical soundness.

2. Related Work

Adversarial attacks on DRL agents. The seminal work of Huang et al. (2017) first exposed the vulnerability of DRL agents to FGSM attacks (Goodfellow et al., 2014) on state observations in Atari games. Subsequently, Lin et al. (2017); Kos & Song (2017) introduced limited-steps attacks to deceive DRL policies. Pattanaik et al. (2017) employed a critic action-value function and gradient descent to degrade DRL performance. Additionally, Behzadan & Munir (2017a) proposed black-box attacks on DQN and verified the transferability of adversarial examples, while Inkawhich et al. (2019) showed that even a constrained adversary with access only to action and reward signals could launch highly effective and damaging attacks. Kiourti et al. (2020); Wang et al. (2021); Bharti et al. (2022); Guo et al. (2023) investigated backdoor attacks in reinforcement learning. Besides, Gleave et al. (2019) have studied the adversarial policy within multi-agent scenarios, and Zhang et al. (2021); Sun et al. (2021) developed learned adversaries by training attackers as RL agents. Lu et al. (2023) suggested an adversarial cheap talk setting and trained an adversary through meta-learning. Korkmaz (2023) analyzed adversarial directions in the Arcade Learning Environment, and found that even state-of-the-art robust agents (Zhang et al., 2020; Oikarinen et al., 2021) are susceptible to policy-independent sensitivity directions.

Robust discrete action for DRL agents. Earlier works like Kos & Song (2017); Behzadan & Munir (2017b) incorporated adversarial states into the replay buffer during training in Atari environments, leading to limited robustness. Fischer et al. (2019) proposed to separate DQN architecture into a Q network and a policy network, and robustly trained the policy network with generated adversarial states and provably robust bounds. Recently, Zhang et al. (2020) characterized state-adversarial RL as SA-MDP, and revealed the potential non-existence of ORP. They addressed the challenge by considering a balance between robustness and natural returns through a KL-based regularization. Oikarinen et al. (2021) controlled robustness certification bounds to minimize the overlap between perturbed Q values of the current action and others. Liang et al. (2022) estimated the worst-case value estimation and combined it with the classic Temporal Difference (TD)-target, resulting in higher training efficiency compared to prior methods. The latest work by Nie et al. (2023) built the DRL architecture upon SortNet (Zhang et al., 2022), enabling global Lipschitz continuity, thus reducing the need for training extra attackers or finding adversaries. He et al. (2023) proposed robust multi-agent Q-learning to solve the robust equilibrium in discrete state and action two-player games. Bukharin et al. (2023) considered a sub-optimal Lipschitz policy in smooth environments and extended the robustness regularization (Shen et al., 2020; Zhang et al., 2020) to multi-agent settings. Prior methods heuristically constrain local smoothness or invariance to achieve commendable robustness, while compromising natural performance. In contrast, our approach seeks optimal robust policies with strict theoretical guarantees, simultaneously improving natural and robust performance.

3. Preliminaries

Markov decision process (MDP) is defined by a tuple $(\mathcal{S}, \mathcal{A}, r, \mathbb{P}, \gamma, \mu_0)$, where $\mathcal{S}$ represents the state space, $\mathcal{A}$ denotes the action space, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ stands for the transition dynamics where $\Delta(\mathcal{S})$ is the probability space over $\mathcal{S}$, $\gamma \in [0, 1)$ represents the discount factor, and $\mu_0 \in \Delta(\mathcal{S})$ is the initial state distribution. In this paper, we consider the setting where the continuous state space $\mathcal{S} \subset \mathbb{R}^d$ is a compact set and the action space $\mathcal{A}$ is a finite set. Given an MDP, we define the state value function $V^\pi(s) = \mathbb{E}_{\pi,\mathbb{P}}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s]$

and the Q (or action-value) function $Q^\pi(s,a) = \mathbb{E}_{\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0=s,a_0=a\right]$ for every policy π. MDPs exhibit a notable property: there exists a stationary, deterministic policy that maximizes both $V^\pi(s)$ and $Q^\pi(s,a)$ for all $s\in\mathcal{S}$ and $a\in\mathcal{A}$. Additionally, the optimal Q function $Q^*(s,a)=\sup_{\pi\in\Pi}Q^\pi(s,a)$, satisfies the Bellman optimality equation

$$Q^*(s,a)=r(s,a)+\gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}}Q^*\left(s',a'\right)\right].$$

State-Adversarial MDP (SA-MDP) allows an adversary ν to perturb the observed state s into $s_\nu\in B(s)$, where $B(s)$ is the adversary perturbation set. Let $\pi\circ\nu$ denote the policy under perturbations, the adversarial value and Q functions are $V^{\pi\circ\nu}(s)=\mathbb{E}_{\pi\circ\nu,\mathbb{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0=s\right]$ and $Q^{\pi\circ\nu}(s,a) = \mathbb{E}_{\pi\circ\nu,\mathbb{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0=s,a_0=a\right]$, respectively. There always exists a strongest adversary $\nu^*(\pi)=\arg\min_\nu V^{\pi\circ\nu}$ for any policy π. An optimal robust policy (ORP) π^* should satisfy $V^{\pi^*\circ\nu^*(\pi^*)}(s)=\max_\pi V^{\pi\circ\nu^*(\pi)}(s)$ for all $s\in\mathcal{S}$.

4. Optimal Adversarial Robustness

In this section, we delve into the exploration of ORP within SA-MDP. While Zhang et al. (2020) noted that ORP does not necessarily exist for a general adversary, we discover that only a few states lack ORP, and the set of these abnormal states has a measure close to zero in complicated tasks. For theoretical clarity, we first propose a consistency assumption of policy (CAP) to eliminate these states. Then, we devise a novel consistent adversarial robust operator $\mathcal{T}_{car}$ for computing the adversarial Q function, and under CAP, we identify its fixed point as exactly Q^*, thereby proving the existence of a deterministic and stationary ORP.

4.1. Consistency Assumption of Policy

Given a general adversary, we observe the true state s and the perturbed observation s_ν have the same optimal action in practice. Some examples are shown in Figure 1 and Appendix E. This enlightens Bellman optimal policy is robust, motivating us to consider how the adversary affects the optimal action in theory. We assume that the adversary perturbation set is a ϵ-neighbourhood $B_\epsilon(s)=\{\|s'-s\|\le\epsilon\}$ for convenience of description and first define the intrinsic state neighborhood where the optimal action is consistent.

Definition 4.1 (Intrinsic State Neighborhood)**.** Given an SA-MDP, we define the intrinsic state ϵ-neighbourhood for any state s as

$$B^*_\epsilon(s):=\{s'\in\mathcal{S}|s'\in B_\epsilon(s), \arg\max_a Q^*(s',a)=\arg\max_a Q^*(s,a)\}.$$

Further, we characterize the states where the state neighborhood is distinct from the intrinsic one and find their little in real environments, which lays the groundwork for the consistency assumption of policy we develop later.

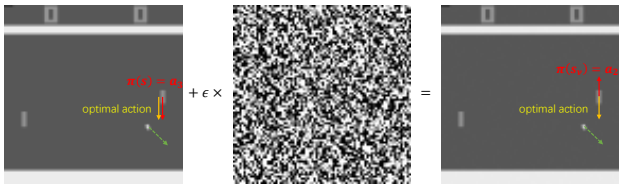

Figure 1. An example of state adversary in DQN. While the adversary disrupts the policy performed by DQN, it does not impact the optimal action dictated by the Bellman optimal policy. This observation prompts the study of two key issues: whether the Bellman optimal policy serves as the ORP, and why vanilla DQN trained with Bellman error fails to ensure robustness.

Theorem 4.2 (Rationality of CAP)**.** *Given $\epsilon>0$, let $\mathcal{S}_{nu}=\{s\in\mathcal{S}|\arg\max_a Q^*(s,a)$ is not a singleton$\}$, and $\mathcal{S}_{nin}=\{s\in\mathcal{S}|B_\epsilon(s)\neq B^*_\epsilon(s)\}$. If $Q^*(\cdot,a)$ is continuous almost everywhere in $\mathcal{S}$ for all $a\in\mathcal{A}$, we have that $\mathcal{S}_{nin}\subseteq\mathcal{S}_{nu}\cup\mathcal{S}_0+B_\epsilon$, where $\mathcal{S}_0$ is a zero measure set.*

Actually, $\mathcal{S}_{nu}$ is also close to an empty set in most practical complex environments, and $\mathcal{S}_0$ is a set of special and rare discontinuous points of Q^* (as shown in our proof). Theorem 4.2 essentially shows, for complicated tasks, $\mathcal{S}_{nin}$ is a quite small set and the magnitude of $\mu(\mathcal{S}_{nin})$ is around $O(\epsilon^d)$, where $\mu(A)$ represents the measure of set A and d is the dimension of state space. The Corollary in Appendix A.1 illustrates better conclusions with stronger conditions.

Motivated by Theorem 4.2 and the above analysis, we assume that all states have a consistent intrinsic state neighborhood.

Assumption 4.3 (Consistency Assumption of Policy (CAP))**.** For all $s\in\mathcal{S}$, its adversary ϵ-perturbation set is the same as the intrinsic state ϵ-neighbourhood, *i.e.*, $B_\epsilon(s)=B^*_\epsilon(s)$.

4.2. Consistent Optimal Robust Policy

To establish the relation between the optimal Q function before and after the perturbation, we propose a consistent adversarial robust (CAR) operator.

Definition 4.4 (CAR Operator $\mathcal{T}_{car}$)**.** Given an SA-MDP, the CAR operator is $\mathcal{T}_{car}:L^p(\mathcal{S}\times\mathcal{A})\to L^p(\mathcal{S}\times\mathcal{A})$,

$$\left(\mathcal{T}_{car}Q\right)(s,a)=r(s,a) + \gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\min_{s'_\nu\in B_\epsilon(s')}Q\left(s',\arg\max_{a_{s'_\nu}}Q\left(s'_\nu,a_{s'_\nu}\right)\right)\right].$$

Although $\mathcal{T}_{car}$ is not contractive (see Appendix A.2.2), Theorem 4.5 shows that under CAP, $\mathcal{T}_{car}$ has a fixed point, which corresponds to the optimal adversarial Q function.

Theorem 4.5 (Relation between Q^* and $Q^{\pi^* \circ \nu^*(\pi^*)}$)**.**

(1). If the optimal adversarial Q function $Q^{\pi^* \circ \nu^*(\pi^*)}$ exists for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then it is the fixed point of $\mathcal{T}_{car}$.

(2). If the CAP holds, then Q^* is the fixed point of $\mathcal{T}_{car}$ and it is also the optimal adversarial Q function, i.e., $Q^*(s,a) = Q^{\pi^* \circ \nu^*(\pi^*)}(s,a)$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Remark 4.6. Note that the min and max operations in Definition 4.4 are not a normal minimax problem because the minimum and maximum objectives are different. It is a bilevel optimization problem. The min and max can not swapped under the general setting. However, they can swap when $\arg\max_{a_{s'_\nu}} Q\left(s'_\nu, a_{s'_\nu}\right)$ is a singleton for all $s'_\nu \in B_\epsilon(s')$, which is a mild condition in our training. Further, we validate that Q^* is also the fixed point of the operator after swapping.

We have demonstrated the convergence of $\mathcal{T}_{car}$ in a smooth environment (see Appendix A.2.3), stating the fixed point iterations of $\mathcal{T}_{car}$ at least converge to a sub-optimal solution close to Q^*. On this basis, it can be derived from Theorem 4.5 that the greedy policy $\pi^*(s) := \arg\max_a Q^*(s,a)$, for all $s \in \mathcal{S}$, is exactly the ORP.

Corollary 4.7 (Existence of ORP)**.** *If the CAP holds, there exists a deterministic and stationary policy π^* which satisfies $V^{\pi^* \circ \nu^*(\pi^*)}(s) \geq V^{\pi \circ \nu^*(\pi)}(s)$ and $Q^{\pi^* \circ \nu^*(\pi^*)}(s,a) \geq Q^{\pi \circ \nu^*(\pi)}(s,a)$, for all $\pi \in \Pi$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

The above theorems indicate that under the CAP, the ORP against the strongest adversary is actually the Bellman optimal policy derived from the Bellman optimality equations. This also suggests that the objectives in natural and adversarial environments are aligned, which is supported by our experiment results in Sec. 7.2.

5. Policy Robustness under Bellman p-error

As the Bellman optimal policy is ORP, we further consider the reasons for the vulnerability of DRL agents: Although former methods, such as Q-learning, essentially take the Bellman optimal policy as a goal, why do they exhibit rather poor robustness? We approach this issue by examining the stability of policy across diverse Banach spaces.

Let Q_θ denote a parameterized Q function. The value-based RL training theoretically requires minimizing $\|Q_\theta - Q^*\|_{\mathcal{B}}$, where $\mathcal{B}$ is a Banach space. In practice, it is hard to make the distance between Q_θ and Q^* vanish due to some limitations, such as the characterization capabilities of neural networks and the convergence of optimization algorithms. Therefore, we analyze the properties of Q_θ when the $\|Q_\theta - Q^*\|_{\mathcal{B}}$ is a small positive value on different spaces $\mathcal{B}$.

5.1. Necessity of L^∞-norm for Adversarial Robustness

We study the adversarial robustness of the greedy policy derived from Q when $0 < \|Q - Q^*\|_p = \delta \ll 1$ for different L^p spaces. Given a function Q and perturbation budget ϵ, let $\mathcal{S}^Q_{sub} = \{s | Q^*(s, \arg\max_a Q(s,a)) < \max_a Q^*(s,a)\}$ denote the set of states where the greedy policy according to Q is suboptimal, and $\mathcal{S}^Q_{adv}$ denote the set of states in whose ϵ-neighbourhood there exists the adversarial state, *i.e.*,

$$\mathcal{S}^Q_{adv} = \{s | \exists s_\nu \in B_\epsilon(s), \text{ s.t. } Q^*(s, \arg\max_a Q(s_\nu, a)) < \max_a Q^*(s,a)\}.$$

Theorem 5.1 (Necessity of L^∞-norm)**.** *There exists an MDP instance such that the following statements hold.*

(1). For any $1 \leq p < \infty$ and $\delta > 0$, there exists a function $Q \in L^p(\mathcal{S} \times \mathcal{A})$ satisfying $\|Q - Q^*\|_p \leq \delta$ such that $\mu\left(\mathcal{S}^Q_{sub}\right) = O(\delta)$ yet $\mu\left(\mathcal{S}^Q_{adv}\right) = \mu(\mathcal{S})$.

(2). There exists a $\bar{\delta} > 0$ such that for any $0 < \delta \leq \bar{\delta}$, and any function $Q \in L^\infty(\mathcal{S} \times \mathcal{A})$ satisfying $\|Q - Q^*\|_\infty \leq \delta$, we have that $\mu\left(\mathcal{S}^Q_{sub}\right) = O(\delta)$ and $\mu\left(\mathcal{S}^Q_{adv}\right) = 2\epsilon + O(\delta)$.

The first statement indicates that when $\|Q - Q^*\|_p$ is a small value for $1 \leq p < \infty$, there always exist adversarial examples near almost all states, resulting in quite poor robustness, while the policy might exhibit excellent performance in a natural environment without adversarial attacks. This observation sheds light on the vulnerability of DRL agents, aligning with findings across various studies (Huang et al., 2017; Ilahi et al., 2021). Importantly, the second statement points out that through minimizing $\|Q - Q^*\|$ in the L^∞-norm space, we can avoid the vulnerability and attain a policy with both natural and robust capabilities. This inspires to optimize $\|Q_\theta - Q^*\|_\infty$ in DRL algorithms. Intuitive examples of Theorem 5.1 are shown in Figure 2.

5.2. Stability of Bellman Optimality Equations

Unfortunately, it is infeasible to measure $\|Q_\theta - Q^*\|$ within a practical DRL procedure due to the unknown nature of Q^*. Most methods train Q_θ via optimizing the Bellman error $\|\mathcal{T}_B Q_\theta - Q_\theta\|_{\mathcal{B}'}$, where $\mathcal{T}_B$ is the Bellman optimal operator

$$(\mathcal{T}_B Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}\left[\max_{a' \in \mathcal{A}} Q(s', a')\right].$$

Similar to Theorem 5.1, we need to figure out which Banach space $\mathcal{B}'$ is the best to train DRL agents which can keep the fewest adversarial states. To analyze this issue, we introduce a concept of functional equations stability drawing on relevant research about physics-informed neural networks for partial differential equations (Wang et al., 2022).

Definition 5.2 (Stability of Functional Equations)**.** Given two Banach spaces $\mathcal{B}_1$ and $\mathcal{B}_2$, if there exist $\delta > 0$ and $C >$

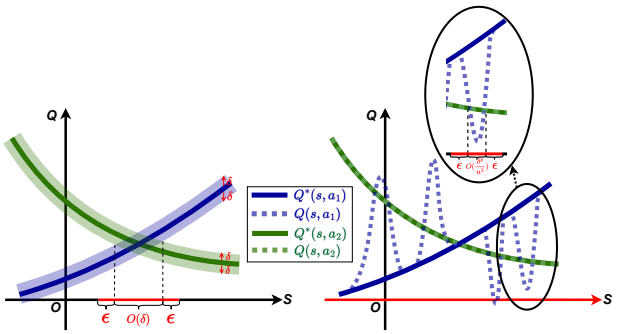

Figure 2. Examples of adversarial robustness for Q satisfying $\|Q - Q^*\|_p \leq \delta$. Given a perturbation radius ϵ, the red line represents the set $\mathcal{S}^Q_{adv}$, in which states have adversarial states. The left panel depicts the case of $p = \infty$, where all Q is distributed in the shadow area with the measure of $\mathcal{S}^Q_{adv}$ being a small value $2\epsilon + O(\delta)$. The right panel shows that for $1 \leq p < \infty$, there always exists Q such that $\mathcal{S}^Q_{adv} = \mathcal{S}$, indicating poor robustness.

0 such that for all $Q \in \mathcal{B}_1 \cap \mathcal{B}_2$ satisfying $\|\mathcal{T}Q - Q\|_{\mathcal{B}_1} < \delta$, we have that $\|Q - Q^*\|_{\mathcal{B}_2} < C\|\mathcal{T}Q - Q\|_{\mathcal{B}_1}$, where Q^* is the exact solution of this functional equation, then we say a nonlinear functional equation $\mathcal{T}Q = Q$ is $(\mathcal{B}_1, \mathcal{B}_2)$-stable. For simplicity, we call that functional $\mathcal{T}$ is $(\mathcal{B}_1, \mathcal{B}_2)$-stable.

The above Definition implies that if there exists a space $\mathcal{B}'$ such that $\mathcal{T}_B$ is $(\mathcal{B}', L^\infty(\mathcal{S} \times \mathcal{A}))$-stable, then $\|Q_\theta - Q^*\|_\infty$ will be controlled when minimizing the Bellman error in $\mathcal{B}'$ space, making DRL agents robust according to Theorem 5.1.(2). The following theorems illustrate the conditions that affect the stability of $\mathcal{T}_B$ and guide for selecting a suitable $\mathcal{B}'$.

Theorem 5.3 (Stable Properties of $\mathcal{T}_B$ in L^p Spaces)**.**

(1). *There exists an MDP such that for all $1 \leq q < p \leq \infty$, the Bellman optimality equations $\mathcal{T}_B Q = Q$ is not $(L^q(\mathcal{S} \times \mathcal{A}), L^p(\mathcal{S} \times \mathcal{A}))$-stable.*

(2). *For any MDP, if the following conditions hold:*

$$C_{\mathbb{P},p} < \frac{1}{\gamma}; \quad p \leq q \leq \infty;$$

$$p \geq \max\left\{1, \frac{\log(|\mathcal{A}|) + \log(\mu(\mathcal{S}))}{\log \frac{1}{\gamma C_{\mathbb{P},p}}}\right\},$$

where $C_{\mathbb{P},p} := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\mathbb{P}(\cdot \mid s,a)\|_{L^{\frac{p}{p-1}}(\mathcal{S})}$, then we have that the Bellman optimality equations $\mathcal{T}_B Q = Q$ is $(L^q(\mathcal{S} \times \mathcal{A}), L^p(\mathcal{S} \times \mathcal{A}))$-stable.

We note that $\lim_{p\to\infty} C_{\mathbb{P},p} = 1 < \frac{1}{\gamma}$ and thus the first condition holds when p is large enough. The second condition suggests it is available for stability that q is larger than p, and the last condition reveals that p is relevant to the size of the state and action spaces. Further, we have that $\mathcal{T}_B$ is $(L^\infty(\mathcal{S} \times \mathcal{A}), L^\infty(\mathcal{S} \times \mathcal{A}))$-stable and thus we can optimize DRL agents in space $\mathcal{B}' = L^\infty(\mathcal{S} \times \mathcal{A})$ for adversarial robustness. Moreover, $\mathcal{B}'$ cannot be $L^q(\mathcal{S} \times \mathcal{A})$ for all $1 \leq q < \infty$ according to Theorem 5.3.(1).

6. Consistent Adversarial Robust DQN

Our theoretical analysis has revealed the feasibility of training a deep Q-network (DQN) by minimizing the Bellman error in L^∞ space to achieve the ORP. However, the exact computation of the L^∞-norm is intractable because of the unknown environment and continuous state space. Therefore, we introduce a surrogate objective based on the L^∞-norm and present the Consistent Adversarial Robust deep Q-network (CAR-DQN), enhancing both the natural and robust performance of agents.

6.1. Stability of Deep Q-learning

Define the state-action visitation distribution as

$$d^\pi_{\mu_0}(s,a) = \mathbb{E}_{s_0\sim\mu_0}(1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr^\pi(s_t = s, a_t = a|s_0).$$

Deep Q-learning algorithms, *e.g.*, DQN, leverage the following objective due to interactions with the environment:

$$\mathcal{L}(Q_\theta; \pi_\theta) = \mathbb{E}_{(s,a)\sim d^{\pi_\theta}_{\mu_0}} \left|\mathcal{T}_B Q_\theta(s,a) - Q_\theta(s,a)\right|.$$

The former theoretical analysis of functional equations stability can be extended to $\mathcal{L}(Q_\theta; \pi_\theta)$ by integrating sampling probability into a seminorm.

Definition 6.1 ($(p, d^\pi_{\mu_0})$-seminorm)**.** Given a policy π, $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $1 \leq p \leq \infty$, if $d^\pi_{\mu_0}$ is a probability density function, we define the $(p, d^\pi_{\mu_0})$-seminorm as the following, which satisfies the absolute homogeneity and triangle inequality:

$$\|f\|_{p,d^\pi_{\mu_0}} := \left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|d^\pi_{\mu_0}(s,a)f(s,a)\right|^p d\mu(s,a)\right)^{\frac{1}{p}}.$$

We note that $(p, d^\pi_{\mu_0})$-seminorm will be upgraded to a norm, if $d^\pi_{\mu_0}(s,a) > 0$ almost everywhere for $(s,a) \in \mathcal{S} \times \mathcal{A}$. The deep Q-learning objective can be streamlined as $\mathcal{L}(Q_\theta; \pi_\theta) = \|\mathcal{T}_B Q_\theta - Q_\theta\|_{1,d^{\pi_\theta}_{\mu_0}}$ based on the definition. Similar to Theorem 5.3, we prove this objective cannot ensure robustness, while $(\infty, d^\pi_{\mu_0})$-norm is necessary for adversarial robustness (see Appendix C.2).

Theorem 6.2. *In the practical DQN procedure, the Bellman optimality equations $\mathcal{T}_B Q = Q$ is $(L^{\infty,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A}), L^p(\mathcal{S}\times\mathcal{A}))$-stable for all $1 \leq p \leq \infty$, while it is not $(L^{q,d^\pi_{\mu_0}}(\mathcal{S} \times \mathcal{A}), L^p(\mathcal{S} \times \mathcal{A}))$-stable for all $1 \leq q < p \leq \infty$.*

We also investigate the stability when $d^\pi_{\mu_0}$ is a probability mass function in Appendix C.1 and C.3.

6.2. Consistent Adversarial Robust Objective

Inspired by the theoretical analysis, we propose to train robust DQNs with $\mathcal{L}_{car}(\theta) = \|\mathcal{T}_B Q_\theta - Q_\theta\|_{\infty, d_{\mu_0}^{\pi_\theta}}$, which is equal to (see Appendix D)

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a) \max_{s_\nu \in B_\epsilon(s)} \left|\mathcal{T}_B Q_\theta(s_\nu, a) - Q_\theta(s_\nu, a)\right|,$$

where π_θ is a behavior policy related to Q_θ and it is usually ϵ-greedy policy derived from Q_θ. Since interactions with the environment in an SA-MDP happen based on the true state s rather than s_ν, $\mathcal{T}_B Q_\theta(s_\nu, a)$ cannot be directly estimated. Thus, we exploit $\mathcal{T}_B Q_\theta(s, a)$ to substituted it, attaining a surrogate objective $\mathcal{L}_{car}^{train}(\theta)$:

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a) \max_{s_\nu \in B_\epsilon(s)} \left|\mathcal{T}_B Q_\theta(s, a) - Q_\theta(s_\nu, a)\right|,$$

which can bound $\mathcal{L}_{car}$, especially in smooth environments. Denote $\mathcal{L}_{car}^{diff}(\theta)$ as the following:

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a) \max_{s_\nu \in B_\epsilon(s)} \left|\mathcal{T}_B Q_\theta(s_\nu, a) - \mathcal{T}_B Q_\theta(s, a)\right|.$$

Theorem 6.3 (Bounding $\mathcal{L}_{car}$ with $\mathcal{L}_{car}^{train}$)**.** *We have that*

$$\left|\mathcal{L}_{car}^{train}(\theta) - \mathcal{L}_{car}^{diff}(\theta)\right| \le \mathcal{L}_{car}(\theta) \le \mathcal{L}_{car}^{train}(\theta) + \mathcal{L}_{car}^{diff}(\theta).$$

Further, suppose the environment is $(L_r, L_\mathbb{P})$-smooth and suppose Q_θ and r are uniformly bounded, i.e. $\exists\, M_Q, M_r > 0$ such that $|Q_\theta(s,a)| \le M_Q$, $|r(s,a)| \le M_r\ \forall s \in \mathcal{S}, a \in \mathcal{A}$. If $M := \sup_{\theta,(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a) < \infty$, then we have

$$\mathcal{L}_{car}^{diff}(\theta) \le C_{\mathcal{T}_B}\epsilon,$$

where $C_{\mathcal{T}_B} = L_{\mathcal{T}_B} M$, $L_{\mathcal{T}_B} = L_r + \gamma C_Q L_\mathbb{P}$ and $C_Q = \max\left\{M_Q, \frac{M_r}{1-\gamma}\right\}$. The definition of $(L_r, L_\mathbb{P})$-smooth environment is shown in Appendix A.2.3.

Theorem 6.3 suggests that $\mathcal{L}_{car}^{train}(\theta)$ is a proper surrogate objective from the optimization perspective. It also provides an insight into potential instability during robust training: If $\mathcal{L}_{car}^{train}(\theta)$ is minimized to a small value yet less than $\mathcal{L}_{car}^{diff}(\theta)$, $\mathcal{L}_{car}(\theta)$ may tend to increase or overfit.

In implementation, to fully utilize each sample in the batch, we derive the soft version $\mathcal{L}_{car}^{soft}(\theta)$ of the CAR objective (derivation seen in Appendix D):

$$\sum_{i\in|\mathcal{B}|} \alpha_i \max_{s_\nu \in B_\epsilon(s_i)} \left|r_i + \gamma \max_{a'} Q_{\bar{\theta}}(s_i', a') - Q_\theta(s_\nu, a_i)\right|,$$

where $\alpha_i = \dfrac{e^{\frac{1}{\lambda}\max_{s_\nu}\left|r_i+\gamma\max_{a'} Q_{\bar{\theta}}(s_i',a')-Q_\theta(s_\nu,a_i)\right|}}{\sum_{i\in|\mathcal{B}|} e^{\frac{1}{\lambda}\max_{s_\nu}\left|r_i+\gamma\max_{a'} Q_{\bar{\theta}}(s_i',a')-Q_\theta(s_\nu,a_i)\right|}}$.

$\mathcal{B}$ represents a batch of transition pairs sampled from the replay buffer. $\bar{\theta}$ is the parameter of target network and λ is the coefficient to control the level of softness.

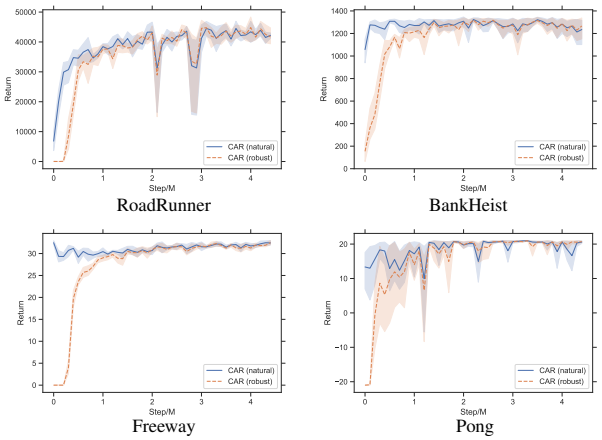

Figure 3. Episode rewards of CAR-DQN agents with and without 10-step PGD attacks on 4 Atari games and 5 random seeds. As evidenced by the overlap of the two curves, CAR-DQN achieves the consistency between Bellman optimal policy and ORP.

7. Experiments

In this section, we conduct extensive comparison and ablation experiments to validate the rationality of our theoretical analysis and the effectiveness of CAR-DQN.

7.1. Implementation details

Environments. Following recent works (Zhang et al., 2020; Oikarinen et al., 2021; Liang et al., 2022), we conduct experiments on four Atari video games (Brockman et al., 2016), including Pong, Freeway, BankHeist, and RoadRunner. These environments are characterized by high-dimensional pixel inputs and discrete action spaces. We pre-process the input images into 84×84 grayscale images and normalize the pixel values to the range $[0, 1]$. In each environment, agents execute an action every 4 frames, skipping the other frames without frame stacking. All rewards are clipped to the range $[-1, 1]$.

Baselines. We compare CAR-DQN with several state-of-the-art robust training methods. SA-DQN (Zhang et al., 2020) incorporates a KL-based regularization and solves the inner maximization problem using PGD (Madry et al., 2017) and CROWN-IBP (Zhang et al., 2019), respectively. RADIAL-DQN (Oikarinen et al., 2021) applies adversarial regularizations based on robustness verification bounds computed by IBP (Gowal et al., 2018). We utilize the official code of SA-DQN and RADIAL-DQN and replicate WocaR-DQN, as its official implementation uses a different environment wrapper compared to SA-DQN and RADIAL.

Evaluations. We evaluate the robustness of agents using three metrics on Atari games: (1) episode return under a 10-steps untargeted PGD attack (Madry et al., 2017), (2) episode return under MinBest (Huang et al., 2017), both of which minimize the probability of selecting the learned

Table 1. Average episode rewards ± standard error of the mean over 50 episodes on baselines and CAR-DQN. The best results of the algorithm with the same type of solver are highlighted in bold. CAR-DQN (PGD) outperforms SA-DQN (PGD) in all metrics and achieves remarkably better robustness (110% higher reward) on RoadRunner. CAR-DQN (cov) outperforms baselines in a majority of cases.

Model		Pong				BankHeist			
		Natural Reward	PGD	MinBest	ACR	Natural Reward	PGD	MinBest	ACR
			$\epsilon = 1/255$				$\epsilon = 1/255$		
Standard	DQN	21.0 ± 0.0	-21.0 ± 0.0	-21.0 ± 0.0	0	1317.2 ± 4.2	22.2 ± 1.9	0.0 ± 0.0	0
PGD	SA-DQN	21.0 ± 0.0	21.0 ± 0.0	21.0 ± 0.0	0	1248.8 ± 1.4	965.8 ± 35.9	1118.0 ± 6.3	0
	CAR-DQN (Ours)	21.0 ± 0.0	21.0 ± 0.0	21.0 ± 0.0	0	$\mathbf{1307.0 \pm 6.1}$	$\mathbf{1243.2 \pm 7.4}$	$\mathbf{1242.6 \pm 8.4}$	0
Convex Relaxation	SA-DQN	21.0 ± 0.0	21.0 ± 0.0	21.0 ± 0.0	1.000	1236.0 ± 1.4	1232.2 ± 2.5	1232.2 ± 2.5	0.991
	RADIAL-DQN	21.0 ± 0.0	21.0 ± 0.0	21.0 ± 0.0	0.898	1341.8 ± 3.8	1341.8 ± 3.8	1341.8 ± 3.8	0.982
	WocaR-DQN	21.0 ± 0.0	21.0 ± 0.0	21.0 ± 0.0	0.979	1315.0 ± 6.1	1312.0 ± 6.1	1312.0 ± 6.1	0.987
	CAR-DQN (Ours)	21.0 ± 0.0	21.0 ± 0.0	21.0 ± 0.0	0.986	$\mathbf{1349.6 \pm 3.0}$	$\mathbf{1347.6 \pm 3.6}$	$\mathbf{1347.4 \pm 3.6}$	0.974
Model		Freeway				RoadRunner			
		Natural Reward	PGD	MinBest	ACR	Natural Reward	PGD	MinBest	ACR
			$\epsilon = 1/255$				$\epsilon = 1/255$		
Standard	DQN	33.9 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0	41492 ± 903	0 ± 0	0 ± 0	0
PGD	SA-DQN	33.6 ± 0.1	23.4 ± 0.2	21.1 ± 0.2	0.250	33380 ± 611	20482 ± 1087	24632 ± 812	0
	CAR-DQN (Ours)	$\mathbf{34.0 \pm 0.0}$	$\mathbf{33.7 \pm 0.1}$	$\mathbf{33.7 \pm 0.1}$	0	$\mathbf{49700 \pm 1015}$	$\mathbf{43286 \pm 801}$	$\mathbf{48908 \pm 1107}$	0
Convex Relaxation	SA-DQN	30.0 ± 0.0	30.0 ± 0.0	30.0 ± 0.0	1.000	46372 ± 882	44960 ± 1152	45226 ± 1102	0.819
	RADIAL-DQN	33.1 ± 0.1	$\mathbf{33.3 \pm 0.1}$	$\mathbf{33.3 \pm 0.1}$	0.998	46224 ± 1133	45990 ± 1112	46082 ± 1128	0.994
	WocaR-DQN	30.8 ± 0.1	31.0 ± 0.0	31.0 ± 0.0	0.992	43686 ± 1608	45636 ± 706	45636 ± 706	0.956
	CAR-DQN (Ours)	$\mathbf{33.2 \pm 0.1}$	33.2 ± 0.1	33.2 ± 0.1	0.981	$\mathbf{49398 \pm 1106}$	$\mathbf{49456 \pm 992}$	$\mathbf{47526 \pm 1132}$	0.760

best action, and (3) Action Certification Rate (ACR) (Zhang et al., 2020). ACR uses relaxation bounds to estimate the percentage of frames where the learned best action is guaranteed to remain unchanged during rollouts under attacks.

CAR-DQN. We implement CAR-DQN based on Double Dueling DQN (Van Hasselt et al., 2016; Wang et al., 2016) and train all baselines and CAR-DQN for 4.5 million steps, based on the same standard model released by Zhang et al. (2020), which has been trained for 6 million steps. We increase the attack ϵ from 0 to $1/255$ in the first 4 million steps using the same smoothed schedule as in Zhang et al. (2020); Oikarinen et al. (2021); Liang et al. (2022), and then continue training with a fixed ϵ for the remaining 0.5 million steps. We use Huber loss to replace the absolute value function and separately apply classic gradient-based methods (PGD) and cheap convex relaxation (IBP) to resolve the inner optimization in $\mathcal{L}_{car}^{soft}(\theta)$. For CAR-DQN with PGD solver, hyper-parameters are the same as SA-DQN (Zhang et al., 2020). For CAR-DQN with IBP solver, we update the target network every 2000 steps, and set learning rate as 1.25×10^{-4}, batch size as 32, exploration ϵ_{exp}-end as 0.01, soft coefficient $\lambda = 1.0$ and discount factor as 0.99. We use a replay buffer with a capacity of 2×10^5 and Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

7.2. Comparison Results

Evaluation on benchmarks. Table 1 presents the natural and robust performance, with all agents trained and attacked using a perturbation radius of $\epsilon = 1/255$. More results and discussion are provided in Appendix H and G. Notably, CAR-DQN agents exhibit superior performance compared to baselines in the most challenging RoadRunner environment, achieving significant improvements in both natural and robust rewards. In the other three games, CAR-DQN can well match the performance of the baselines. Our proposed loss function coupled with the PGD solver, achieves a remarkable return of around 45000 on the RoadRunner environment, outperforming the SA-DQN with the PGD approach. It also attains 60% higher robust rewards under the MinBest attack on the Freeway game. In these two solvers, we observe that PGD exhibits relatively weaker robust performance compared to the convex relaxation, especially failing to ensure the ACR computed with relaxation bounds. This discrepancy can be attributed to that the PGD solver offers a lower bound surrogate function of the loss, while the IBP solver gives an upper bound.

Consistency in natural and PGD attack returns. Figure 3 records the natural and PGD attack returns of CAR-DQN agents during training, showcasing a strong alignment between natural performance and robustness across all environments. This alignment validates our theory that the ORP is consistent with the Bellman optimal policy, and confirms the rationality of the proposed CAP. In addition, Figure 4 illustrates the natural episode return and robustness during training for different algorithms. It is worth noting that CAR-DQN agents can fast and stably converge to peak robustness and natural performance across all environments, while other algorithms exhibit unstable trends. For instance, the natural reward curves of SA-DQN and WocaR-DQN on BankHeist and RADIAL-DQN on RoadRunner distinctly tend to decrease, and the robust curves of SA-DQN and

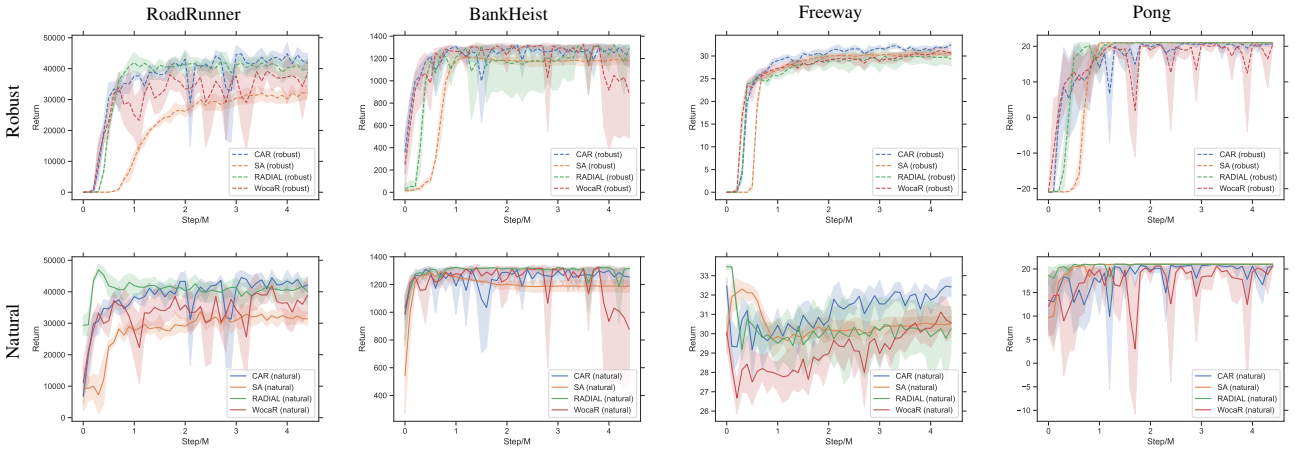

Figure 4. Episode rewards of baselines and CAR-DQN with and without PGD attacks on 4 Atari games. Shaded regions are computed over 5 random seeds. CAR-DQN demonstrates superior natural and robust performance in all environments.

Table 2. Performance of DQN with different Bellman p-error.

Environment	Norm	Natural	PGD	MinBest	ACR
Pong	L^1	$\mathbf{21.0 \pm 0.0}$	-21.0 ± 0.0	-21.0 ± 0.0	0
	L^2	$\mathbf{21.0 \pm 0.0}$	-21.0 ± 0.0	-20.8 ± 0.1	0
	L^∞	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	0.985
Freeway	L^1	$\mathbf{33.9 \pm 0.1}$	0.0 ± 0.0	0.0 ± 0.0	0
	L^2	21.8 ± 0.3	21.7 ± 0.3	22.1 ± 0.3	0
	L^∞	33.3 ± 0.1	$\mathbf{33.2 \pm 0.1}$	$\mathbf{33.2 \pm 0.1}$	0.981
BankHeist	L^1	1325.5 ± 5.7	27.0 ± 2.0	0.0 ± 0.0	0
	L^2	1314.5 ± 4.0	18.5 ± 1.5	22.5 ± 2.6	0
	L^∞	$\mathbf{1356.0 \pm 1.7}$	$\mathbf{1356.5 \pm 1.1}$	$\mathbf{1356.5 \pm 1.1}$	0.969
RoadRunner	L^1	43795 ± 1066	0 ± 0	0 ± 0	0
	L^2	30620 ± 990	0 ± 0	0 ± 0	0
	L^∞	$\mathbf{49500 \pm 2106}$	$\mathbf{48230 \pm 1648}$	$\mathbf{48050 \pm 1642}$	0.947

WocaR-DQN on BankHeist tend to decline. This discrepancy primarily stems from their robustness objectives, which diverge from the standard training loss and consequently result in learning sub-optimal actions. In contrast, the proposed consistent objective ensures that CAR-DQN always learns optimal actions in both natural and robust directions.

Training efficiency. Training SA-DQN, RADIAL-DQN, WocaR-DQN, and CAR-DQN costs approximately 27, 12, 20, and 14 hours, respectively. All these models are trained for 4.5 million frames on identical hardware. Additionally, our proposed loss does not incur additional memory consumption compared to vanilla training.

7.3. Ablation Studies

Necessity of infinity-norm. To verify the necessity of the $(\infty, d^\pi_{\mu_0})$-norm for adversarial robustness, we train DQN agents using the Bellman error under $(1, d^\pi_{\mu_0})$-norm and $(2, d^\pi_{\mu_0})$-norm, respectively. We then compare their performance with our CAR-DQN, which approximates the Bellman error under $(\infty, d^\pi_{\mu_0})$-norm. As shown in Table 2, all agents perform well without attacks in the four games. However, the performance of $(1, d^\pi_{\mu_0})$-norm and $(2, d^\pi_{\mu_0})$-norm agents highly degrades under strong attacks, receiving

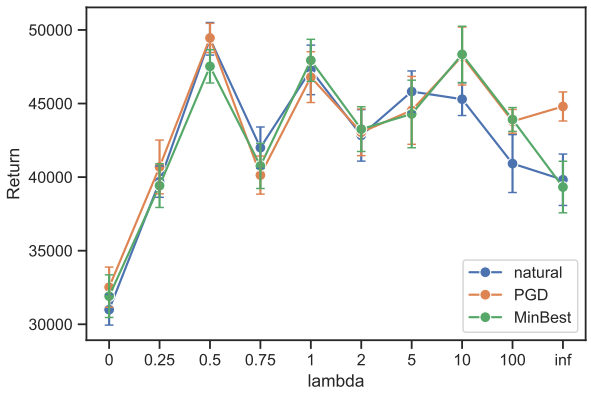

Figure 5. Natural, PGD attack, and MinBest attack rewards of CAR-DQN with different soft coefficients on RoadRunner game.

episode rewards close to the lowest in each game. These empirical results are highly consistent with Theorem 5.1.

Effects of soft coefficient. We validate the effectiveness of the soft CAR-DQN loss by adjusting the soft coefficient λ. We train CAR-DQN agents on the RoadRunner environment with λ values ranging from 0 to ∞. When $\lambda = 0$ we utilize the sample with the largest adversarial TD-error from a batch, while $\lambda = \infty$ corresponds to averaging over all samples in a batch. It is worth noting that a small λ may lead to numerical instability. As depicted in Figure 5, the agents exhibit similar capabilities when $0.5 \leq \lambda \leq 10$, indicating that the learned policies are not sensitive to the soft coefficient within this range. Table 3 displays the performance of CAR-DQN agents with $\lambda = 0, 1, \infty$ across four Atari environments. In the RoadRunner, the case $\lambda = 0$ yields poor performance, achieving returns around 25000 due to inadequate utilization of the samples. Interestingly, utilizing only the sample with the largest adversarial TD error from a batch achieves good robustness on the other three simpler games. The case $\lambda = \infty$ results in worse robustness

Table 3. Ablation studies for soft coefficients on 4 Atari games.

Environment	λ	Natural	PGD	MinBest	ACR
Pong	0	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	0.972
	1	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	0.985
	∞	20.6 ± 0.1	20.7 ± 0.1	20.7 ± 0.1	0.980
Freeway	0	31.6 ± 0.2	31.5 ± 0.1	31.5 ± 0.1	0.966
	1	$\mathbf{33.3 \pm 0.1}$	$\mathbf{33.2 \pm 0.1}$	$\mathbf{33.2 \pm 0.1}$	0.981
	∞	31.5 ± 0.1	30.9 ± 0.3	31.2 ± 0.2	0.967
BankHeist	0	1307.5 ± 11.0	1288.5 ± 14.0	1284.0 ± 13.8	0.980
	1	$\mathbf{1356.0 \pm 1.7}$	$\mathbf{1356.5 \pm 1.1}$	$\mathbf{1356.5 \pm 1.1}$	0.969
	∞	1326.0 ± 4.8	1316.0 ± 6.8	1314.0 ± 6.6	0.979
RoadRunner	0	25160 ± 802	24540 ± 760	26785 ± 617	0.007
	1	$\mathbf{49500 \pm 2106}$	$\mathbf{48230 \pm 1648}$	$\mathbf{48050 \pm 1642}$	0.947
	∞	40890 ± 2075	36760 ± 1874	36740 ± 2098	0.940

compared to other cases with differentiated weights. This suggests that each sample in a batch plays a distinct role in robust training, and we can enhance robust performance by specifying weightings. These results further validate the efficacy of our CAR-DQN loss.

8. Conclusion

In this paper, we prove the alignment of the optimal robust policy with the Bellman optimal policy under the consistency assumption of policy. We show that measuring Bellman error in differed L^p spaces yields varied performance, underscoring the necessity of Bellman infinity-error for robustness. We validate these findings through experiments with CAR-DQN, which optimizes a surrogate objective of Bellman infinity-error. We believe this work contributes significantly to unveiling the nature of robustness in Q-learning. Since our work focuses on value-based DRL with discrete action space, we will extend future research into the policy-based DRL and continuous action space setting.

Acknowledgements

This paper is supported by the National Key R&D (research and development) Program of China (2022YFA1004001) and the National Natural Science Foundation of China (Nos. 11991022, U23B2012). We thank Anqi Li and Wenzhao Liu for valuable feedback on earlier drafts of the paper.

Impact Statement

Reinforcing the robustness of Deep Reinforcement Learning agents is crucial for commercial and industrial applications. Unlike previous approaches that focus on imposing smoothness or stability on neural networks through elaborated regularization techniques, this study uncovers the inherent robustness of the Bellman optimal policy under adversarial attack scenarios. We believe this discovery can offer fresh insights to the research community, potentially driving advancements in related applications.

References

Behzadan, V. and Munir, A. Vulnerability of deep reinforcement learning to policy induction attacks. In *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings 13*, pp. 262–275. Springer, 2017a.

Behzadan, V. and Munir, A. Whatever does not kill deep reinforcement learning, makes it stronger. *arXiv preprint arXiv:1712.09344*, 2017b.

Bharti, S., Zhang, X., Singla, A., and Zhu, J. Provable defense against backdoor policies in reinforcement learning. *Advances in Neural Information Processing Systems*, 35: 14704–14714, 2022.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Bukharin, A., Li, Y., Yu, Y., Zhang, Q., Chen, Z., Zuo, S., Zhang, C., Zhang, S., and Zhao, T. Robust multi-agent reinforcement learning via adversarial regularization: Theoretical foundation and stable algorithms. *arXiv preprint arXiv:2310.10810*, 2023.

Fischer, M., Mirman, M., Stalder, S., and Vechev, M. Online robustness training for deep reinforcement learning. *arXiv preprint arXiv:1911.00887*, 2019.

Gleave, A., Dennis, M., Wild, C., Kant, N., Levine, S., and Russell, S. Adversarial policies: Attacking deep reinforcement learning. *arXiv preprint arXiv:1905.10615*, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Guo, J., Li, A., Wang, L., and Liu, C. Policycleanse: Backdoor detection and mitigation for competitive reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4699–4708, 2023.

He, S., Han, S., Su, S., Han, S., Zou, S., and Miao, F. Robust multi-agent reinforcement learning with state uncertainty. *arXiv preprint arXiv:2307.16212*, 2023.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284*, 2017.

Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., and Levine, S. How to train your robot with deep reinforcement learning: lessons we have learned. *The International Journal of Robotics Research*, 40(4-5):698–721, 2021.

Ilahi, I., Usama, M., Qadir, J., Janjua, M. U., Al-Fuqaha, A., Hoang, D. T., and Niyato, D. Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *IEEE Transactions on Artificial Intelligence*, 3 (2):90–109, 2021.

Inkawhich, M., Chen, Y., and Li, H. Snooping attacks on deep reinforcement learning. *arXiv preprint arXiv:1905.11832*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kiourti, P., Wardega, K., Jha, S., and Li, W. Trojdrl: evaluation of backdoor attacks on deep reinforcement learning. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6. IEEE, 2020.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23 (6):4909–4926, 2021.

Korkmaz, E. Adversarial robust deep reinforcement learning requires redefining robustness. *arXiv preprint arXiv:2301.07487*, 2023.

Kos, J. and Song, D. Delving into adversarial attacks on deep policies. *arXiv preprint arXiv:1705.06452*, 2017.

Liang, Y., Sun, Y., Zheng, R., and Huang, F. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. *Advances in Neural Information Processing Systems*, 35:22547–22561, 2022.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Lin, Y.-C., Hong, Z.-W., Liao, Y.-H., Shih, M.-L., Liu, M.-Y., and Sun, M. Tactics of adversarial attack on deep reinforcement learning agents. *arXiv preprint arXiv:1703.06748*, 2017.

Lu, C., Willi, T., Letcher, A., and Foerster, J. N. Adversarial cheap talk. In *International Conference on Machine Learning*, pp. 22917–22941. PMLR, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pp. 1928–1937. PMLR, 2016.

Nie, B., Ji, J., Fu, Y., and Gao, Y. Improve robustness of reinforcement learning against observation perturbations via l_∞ lipschitz policy networks. *arXiv preprint arXiv:2312.08751*, 2023.

Oikarinen, T., Zhang, W., Megretski, A., Daniel, L., and Weng, T.-W. Robust deep reinforcement learning through adversarial loss. *Advances in Neural Information Processing Systems*, 34:26156–26167, 2021.

Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., and Chowdhary, G. Robust deep reinforcement learning with adversarial attacks. *arXiv preprint arXiv:1712.03632*, 2017.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shen, Q., Li, Y., Jiang, H., Wang, Z., and Zhao, T. Deep reinforcement learning with robust and smooth policy. In *International Conference on Machine Learning*, pp. 8707–8718. PMLR, 2020.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., and Riedmiller, M. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. Pmlr, 2014.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sun, Y., Zheng, R., Liang, Y., and Huang, F. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep rl. *arXiv preprint arXiv:2106.05087*, 2021.

Van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

Wang, C., Li, S., He, D., and Wang, L. Is l^2 physics informed loss always suitable for training physics informed neural network? *Advances in Neural Information Processing Systems*, 35:8278–8290, 2022.

Wang, L., Javed, Z., Wu, X., Guo, W., Xing, X., and Song, D. Backdoorl: Backdoor attack against competitive reinforcement learning. *arXiv preprint arXiv:2105.00579*, 2021.

Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., and Freitas, N. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pp. 1995–2003. PMLR, 2016.

Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Zhang, B., Jiang, D., He, D., and Wang, L. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. *Advances in Neural Information Processing Systems*, 35:19398–19413, 2022.

Zhang, H., Chen, H., Xiao, C., Gowal, S., Stanforth, R., Li, B., Boning, D., and Hsieh, C.-J. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*, 2019.

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems*, 33: 21024–21037, 2020.

Zhang, H., Chen, H., Boning, D., and Hsieh, C.-J. Robust reinforcement learning on state observations with learned optimal adversary. *arXiv preprint arXiv:2101.08452*, 2021.

Zheng, G., Zhang, F., Zheng, Z., Xiang, Y., Yuan, N. J., Xie, X., and Li, Z. Drn: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 world wide web conference*, pp. 167–176, 2018.

A. Theorems and Proofs of Optimal Adversarial Robustness

A.1. Reasonablity of the Consistency Assumption

Theorem A.1. *For any MDP $\mathcal{M}$, let $\mathcal{S}_{nu}$ denote the state set where the optimal action is not unique, i.e. $\mathcal{S}_{nu} = \{s \in \mathcal{S} | \arg\max_a Q^*(s,a)$ is not a singleton$\}$. If $Q^*(\cdot, a)$ is continuous almost everywhere in $\mathcal{S}$ for all $a \in \mathcal{A}$, we have the following conclusions:*

- *For almost everywhere $s \in \mathcal{S} \setminus \mathcal{S}_{nu}$, there exists $\epsilon > 0$ such that $B_\epsilon(s) = B^*_\epsilon(s)$.*
- *Given $\epsilon > 0$, let $\mathcal{S}_{nin}$ denote the set of states where the intrinsic state ϵ-neighourhood is not the same as the ϵ-neighourhood, i.e. $\mathcal{S}_{nin} = \{s \in \mathcal{S} | B_\epsilon(s) \neq B^*_\epsilon(s)\}$. Then, we have $\mathcal{S}_{nin} \subseteq \mathcal{S}_{nu} \cup \mathcal{S}_0 + B_\epsilon = \{s_1 + s_2 | s_1 \in \mathcal{S}_{nu} \cup \mathcal{S}_0,\ \|s_2\| \leq \epsilon\}$, where $\mathcal{S}_0$ is a zero measure set.*

Proof. (1) Let $\mathcal{S}' = \{s \in \mathcal{S} | \exists a \in \mathcal{A}$, s.t. $Q^*(s,a)$ is not continuous at $s\}$. Then $\mu(\mathcal{S}') = 0$ because $Q^*(\cdot, a)$ is continuous almost everywhere in $\mathcal{S}$ for all $a \in \mathcal{A}$ and $\mathcal{A}$ is a finite discrete set. And $Q^*(s,a)$ is continuous in $(\mathcal{S} \setminus \mathcal{S}_{nu}) \setminus \mathcal{S}'$. Because $\arg\max_a Q^*(s,a)$ is a singleton for $s \in (\mathcal{S} \setminus \mathcal{S}_{nu}) \setminus \mathcal{S}'$, define $\arg\max_a Q^*(s,a) = \{a^*_s\}$ for any $s \in (\mathcal{S} \setminus \mathcal{S}_{nu}) \setminus \mathcal{S}'$. Then $Q^*(s, a^*_s) > Q^*(s,a)$ for a fixed $a \in \mathcal{A} \setminus \{a^*_s\}$. According to continuity of $Q^*(\cdot, a)$ for all $a \in \mathcal{A}$, there exists $\epsilon_a > 0$, such that $Q^*(s', a^*_s) > Q^*(s', a)$ for all $s' \in B_{\epsilon_a}(s)$. Because $\mathcal{A}$ is a finite discrete set, let $\epsilon = \min_{a \in \mathcal{A} \setminus a^*_s}\{\epsilon_a\}$, then $Q^*(s', a^*_{s'}) > Q^*(s', a)$ for all $s' \in B_\epsilon(s)$ and for all $a \in \mathcal{A} \setminus \{a^*_{s'}\}$, i.e. $B_\epsilon(s) = B^*_\epsilon(s)$.

(2)Let $\mathcal{S}_n = \{s \in \mathcal{S} | \forall \epsilon_1 > 0, \exists s' \in B_{\epsilon_1}(s),$ s.t. $\arg\max_a Q^*(s', a) \neq \arg\max_a Q^*(s,a)\}$ and $\mathcal{S}_0 = \mathcal{S}_n \cap \mathcal{S}'$. Then S_0 is the set of discontinuous points that cause the optimal action to change. And $\mu(\mathcal{S}_0) = \mu(\mathcal{S}_n \cap \mathcal{S}') = 0$ because $\mu(\mathcal{S}') = 0$.

For any $s \in \mathcal{S}_{nin} = \{s \in \mathcal{S} | B_\epsilon(s) \neq B^*_\epsilon(s)\}$, we have the following two cases.

Case 1. $\exists s' \in B_\epsilon(s)$ s.t. $s' \in \mathcal{S}_{nu}$, then $s \in \mathcal{S}_{nu} + B_\epsilon$, i.e.

$$s \in \mathcal{S}_{nu} \cup \mathcal{S}_0 + B_\epsilon. \tag{1}$$

Case 2. $\forall s' \in B_\epsilon(s), s' \notin \mathcal{S}_{nu}$, which means that $\arg\max_a Q^*(s', a)$ is a singleton for all $s' \in B_\epsilon(s)$. Define $\arg\max_a Q^*(s', a) = \{a^*_{s'}\}$ for any $s' \in B_\epsilon(s)$.

Because $s \in \mathcal{S}_{nin}$, there exist a $s' \in B_\epsilon(s)$ such that $a^*_{s'} \neq a^*_s$. Let s_1 be the point that closest to s satisfing $a^*_{s_1} \neq a^*_s$, then $s_1 \in B_\epsilon(s)$. We have

$$s_1 \in \mathcal{S}_n. \tag{2}$$

Otherwise $s_1 \notin \mathcal{S}_n$ means that $\exists \epsilon_1 > 0, \forall s' \in B_{\epsilon_1}(s_1)$, $a^*_{s'} = a^*_{s_1}$, then s_1 is not the point that closest to s satisfing $a^*_{s_1} \neq a^*_s$, which is a contradiction. We also have

$$s_1 \in \mathcal{S}'. \tag{3}$$

Otherwise $s_1 \notin \mathcal{S}'$ means that $\forall a \in \mathcal{A}$, $Q^*(\cdot, a)$ is continuous in s_1. First, we have

$$Q^*(s_1, a^*_{s_1}) > Q^*(s_1, a), \forall a \in \mathcal{A} \setminus \{a^*_{s_1}\}. \tag{4}$$

Then $\exists \epsilon_2 > 0, \forall s \in B_{\epsilon_2}(s_1)$, s.t.

$$Q^*(s, a^*_{s_1}) > Q^*(s, a), \forall a \in \mathcal{A} \setminus \{a^*_{s_1}\}. \tag{5}$$

because of the continuity of point s_1. This contradicts the definition of s_1.

According to (2) and (3), we have $s_1 \in \mathcal{S}' \cap \mathcal{S}_n$ i.e. $s_1 \in \mathcal{S}_0$. Then $s \in \mathcal{S}_0 + B_\epsilon(s)$, i.e.

$$s \in \mathcal{S}_{nu} \cup \mathcal{S}_0 + B_\epsilon. \tag{6}$$

Thus

$$\mathcal{S}_{nin} \subseteq \mathcal{S}_{nu} \cup \mathcal{S}_0 + B_\epsilon. \tag{7}$$

□

Remark A.2. In practical and complex tasks, we can view $\mathcal{S}_{nu}$ as an empty set.

Remark A.3. Except for the smooth environment, many tasks can be modeled as environments with sparse rewards. Further, the value function and action-value function in these environments are almost everywhere continuous.

Remark A.4. According to the construction in the above proof, we know that $\mathcal{S}_0$ is a set of special discontinuous points and its elements are rare in practical complex environments.

Further, we can get the following corollary in the setting of continuous functions and there are better conclusions.

Corollary A.5. *For any MDP $\mathcal{M}$, let $\mathcal{S}_{nu}$ denote the state set where the optimal action is not unique, i.e. $\mathcal{S}_{nu} = \{s \in \mathcal{S} | \arg\max_a Q^*(s,a)$ is not a singleton$\}$. If $Q^*(\cdot, a)$ is continuous in $\mathcal{S}$ for all $a \in \mathcal{A}$, we have the following conclusions:*

- *For $s \in \mathcal{S} \setminus \mathcal{S}_{nu}$, there exists $\epsilon > 0$ such that $B_\epsilon(s) = B^*_\epsilon(s)$.*
- *Given $\epsilon > 0$, let $\mathcal{S}_{nin}$ denote the set of states where the intrinsic state ϵ-neighourhood is not the same as the ϵ-neighourhood, i.e. $\mathcal{S}_{nin} = \{s \in \mathcal{S} | B_\epsilon(s) \neq B^*_\epsilon(s)\}$. Then, we have $\mathcal{S}_{nin} \subseteq \mathcal{S}_{nu} + B_\epsilon = \{s_1 + s_2 | s_1 \in \mathcal{S}_{nu},\ \|s_2\| \leq \epsilon\}$. Especially, when $\mathcal{S}_{nu}$ is a finite set, we have $\mu(\mathcal{S}_{nin}) \leq |\mathcal{S}_{nu}| \mu(B_\epsilon) = C_d |\mathcal{S}_{nu}| \epsilon^d$, where C_d is a constant with respect to dimension d and norm.*

Proof. Corollary A.5 can be derived from Theorem A.1 because we have the following conclusion in continuous case.

$$\mathcal{S}_0 \subseteq \{s \in \mathcal{S} | \exists a \in \mathcal{A}, \text{s.t. } Q^*(s,a) \text{ is not continuous at } s\} = \emptyset \tag{8}$$

□

Remark A.6. Certain natural environments show smooth reward function and transition dynamics, especially in continuous control tasks where the transition dynamics come from some physical laws. Further, the value function and action-value function in these environments is continuous.

A.2. ORP and CAR Operator

Define the consistent adversarial robust operator for adversarial action-value function:

$$(\mathcal{T}_{car} Q)(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\min_{s'_\nu \in B_\epsilon(s')} Q\left(s', \arg\max_{a_{s'_\nu}} Q\left(s'_\nu, a_{s'_\nu}\right)\right)\right]. \tag{9}$$

A.2.1. Equivalence with Optimal Adversarial Value Function

Lemma A.7 (Bellman equations for fixed π and ν in SA-MDP, Zhang et al. (2020))**.** *Given $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and $\nu : \mathcal{S} \to \mathcal{S}$, we have*

$$V^{\pi \circ \nu}(s) = \mathbb{E}_{a \sim \pi(\cdot|\nu(s))} Q^{\pi \circ \nu}(s,a) \tag{10}$$

$$= \mathbb{E}_{a \sim \pi(\cdot|\nu(s))} \left[r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} V^{\pi \circ \nu}(s') \right], \tag{11}$$

$$Q^{\pi \circ \nu}(s,a) = r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} V^{\pi \circ \nu}(s') \tag{12}$$

$$= r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a), a' \sim \pi(\cdot|\nu(s'))} Q^{\pi \circ \nu}(s', a'). \tag{13}$$

Lemma A.8 (Bellman equation for strongest adversary ν^* in SA-MDP, Zhang et al. (2020))**.**

$$V^{\pi \circ \nu^*(\pi)}(s) = \min_{\nu(s) \in B_\epsilon(s)} \mathbb{E}_{a \sim \pi(\cdot|\nu(s))} Q^{\pi \circ \nu^*(\pi)}(s,a). \tag{14}$$

Definition A.9. Define the linear functional $\mathcal{L}^{\pi \circ \nu} : L^p(\mathcal{S} \times \mathcal{A}) \to L^p(\mathcal{S} \times \mathcal{A})$ for fixed π and ν:

$$(\mathcal{L}^{\pi \circ \nu} Q)(s,a) := \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a), a' \sim \pi(\cdot|\nu(s'))} Q(s', a'). \tag{15}$$

Then, by lemma A.7, we have that

$$Q^{\pi \circ \nu} = r + \gamma \mathcal{L}^{\pi \circ \nu} Q^{\pi \circ \nu}. \tag{16}$$

Lemma A.10. *$\mathcal{T} : \mathcal{X} \to \mathcal{X}$ is a linear functional where $\mathcal{X}$ are normed vector space. If there exists $m > 0$ such that*

$$\|\mathcal{T}x\| \geq m\|x\| \quad \forall x \in \mathcal{X}, \tag{17}$$

then $\mathcal{T}$ has a bounded inverse operator $\mathcal{T}^{-1}$.

Proof. If $\mathcal{T}x_1 = \mathcal{T}x_2$, then $\mathcal{T}(x_1 - x_2) = 0$. While $0 = \|\mathcal{T}(x_1 - x_2)\| \geq m\|x_1 - x_2\|$, thus $x_1 = x_2$. Then $\mathcal{T}$ is a bijection and thus the inverse operator of $\mathcal{T}$ exists.

For any $y \in \mathcal{X}$, $\mathcal{T}^{-1}y \in \mathcal{X}$. We have that

$$\|y\| = \|\mathcal{T}\left(\mathcal{T}^{-1}y\right)\| \geq m\|\mathcal{T}^{-1}y\|. \tag{18}$$

Thus, we attain that

$$\|\mathcal{T}^{-1}y\| \leq \frac{1}{m}\|y\|, \quad \forall y \in \mathcal{X}, \tag{19}$$

which shows that $\mathcal{T}^{-1}$ is bounded. □

Lemma A.11. *$I - \gamma\mathcal{L}^{\pi\circ\nu}$ is invertible and thus we have that*

$$Q^{\pi\circ\nu} = \left(I - \gamma\mathcal{L}^{\pi\circ\nu}\right)^{-1} r. \tag{20}$$

Proof. Firstly, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\begin{aligned}
\left(\mathcal{L}^{\pi\circ\nu}Q\right)(s,a) &= \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a),a'\sim\pi(\cdot|\nu(s'))}Q(s',a') && (21)\\
&\leq \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a),a'\sim\pi(\cdot|\nu(s'))}\|Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})} && (22)\\
&= \|Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})} && (23)
\end{aligned}$$

Thus, we have that

$$\|\mathcal{L}^{\pi\circ\nu}Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})} \leq \|Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})}. \tag{24}$$

For any $Q \in L^p(\mathcal{S} \times \mathcal{A})$, we have

$$\begin{aligned}
\|\left(I - \gamma\mathcal{L}^{\pi\circ\nu}\right)Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})} &= \|Q - \gamma\mathcal{L}^{\pi\circ\nu}Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})} && (25)\\
&\geq \|Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})} - \gamma\|\mathcal{L}^{\pi\circ\nu}Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})} && (26)\\
&\geq \|Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})} - \gamma\|Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})} && (27)\\
&= (1-\gamma)\|Q\|_{L^\infty(\mathcal{S}\times\mathcal{A})}, && (28)
\end{aligned}$$

where the first inequality comes from the triangle inequality and the second inequality comes from (24). Then, according to lemma A.10, we attain that $I - \gamma\mathcal{L}^{\pi\circ\nu}$ is invertible. □

Lemma A.12. *If $Q > 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, then we have that $\left(I - \gamma\mathcal{L}^{\pi\circ\nu}\right)^{-1}Q > 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.*

Proof. At first, we have

$$\begin{aligned}
&\left(I - \gamma\mathcal{L}^{\pi\circ\nu}\right)\left(\sum_{t=0}^{\infty}\gamma^t\left(\mathcal{L}^{\pi\circ\nu}\right)^t\right) && (29)\\
&= \sum_{t=0}^{\infty}\gamma^t\left(\mathcal{L}^{\pi\circ\nu}\right)^t - \sum_{t=1}^{\infty}\gamma^t\left(\mathcal{L}^{\pi\circ\nu}\right)^t && (30)\\
&= I. && (31)
\end{aligned}$$

Thus, we get that

$$\left(I - \gamma\mathcal{L}^{\pi\circ\nu}\right)^{-1} = \sum_{t=0}^{\infty}\gamma^t\left(\mathcal{L}^{\pi\circ\nu}\right)^t. \tag{32}$$

If $Q(s,a) > 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, then for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have

$$\left(\mathcal{L}^{\pi \circ \nu} Q\right)(s,a) = \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a), a' \sim \pi(\cdot|\nu(s'))} Q(s',a') \geq 0. \tag{33}$$

Further, we have that $\left(\left(\mathcal{L}^{\pi \circ \nu}\right)^k Q\right)(s,a) > 0$ for all $k \in \mathbb{N}$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$. Thus, we have

$$\begin{aligned}
&\left(I - \gamma \mathcal{L}^{\pi \circ \nu}\right)^{-1} Q(s,a) && (34)\\
&= \sum_{t=0}^{\infty} \gamma^t \left(\left(\mathcal{L}^{\pi \circ \nu}\right)^t Q\right)(s,a) && (35)\\
&> 0. && (36)
\end{aligned}$$

□

Theorem A.13. *If the optimal adversarial action-value function under the strongest adversary* $Q_0(s,a) := \max_\pi \min_\nu Q^{\pi \circ \nu}(s,a)$ *exists for all* $s \in \mathcal{S}$ *and* $a \in \mathcal{A}$*, then it is the fixed point of CAR operator.*

Proof. Denote $V_0(s) = \max_\pi \min_\nu V^{\pi \circ \nu}(s)$. For all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have

$$\begin{aligned}
Q_0(s,a) &= \max_\pi \min_\nu Q^{\pi \circ \nu}(s,a) && (37)\\
&= r(s,a) + \gamma \max_\pi \min_\nu \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} V^{\pi \circ \nu}(s') && (38)\\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} V_0(s') && (39)\\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \min_{\nu(s) \in B_\epsilon(s)} \max_\pi \mathbb{E}_{a \sim \pi(\cdot|\nu(s))} Q_0(s,a) && (40)\\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\min_{s'_\nu \in B_\epsilon(s')} Q_0\left(s', \arg\max_{a_{s'_\nu}} Q_0\left(s'_\nu, a_{s'_\nu}\right)\right)\right] && (41)\\
&= \left(\mathcal{T}_{car} Q\right)(s,a), && (42)
\end{aligned}$$

where the fourth equation comes from lemma (A.8). This completes the proof. □

Theorem A.14. *If the consistency assumption holds, then* Q^* *is the fixed point of the CAR operator. Further,* Q^* *is the optimal adversarial action-value function under the strongest adversary, i.e.* $Q^*(s,a) = \max_\pi \min_\nu Q^{\pi \circ \nu}(s,a)$*, for all* $s \in \mathcal{S}$ *and* $a \in \mathcal{A}$*.*

Proof.

$$\begin{aligned}
\left(\mathcal{T}_{car} Q^*\right)(s,a) &= r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\min_{s'_\nu \in B^*_\epsilon(s')} Q^*\left(s', \arg\max_{a_{s'_\nu}} Q^*\left(s'_\nu, a_{s'_\nu}\right)\right)\right] && (43)\\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\min_{s'_\nu \in B^*_\epsilon(s')} \max_{a'} Q^*\left(s', a'\right)\right] && (44)\\
&= r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\max_{a'} Q^*\left(s', a'\right)\right] && (45)\\
&= Q^*(s,a), && (46)
\end{aligned}$$

where the second equality utilizes the definition of $B^*_\epsilon(s')$. Thus, Q^* is a fixed point of the CAR operator.

Define π and ν as the following:

$$\begin{aligned}
\pi(s) &:= \arg\max_a Q^*(s,a), && (47)\\
\nu(s) &:= \arg\min_{s_\nu \in B_\epsilon(s)} Q^*\left(s, \arg\max_{a_{s_\nu}} Q^*\left(s_\nu, a_{s_\nu}\right)\right). && (48)
\end{aligned}$$

Then, we have

$$\left(\mathcal{T}_{car}Q^*\right)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\min_{s'_\nu \in B_\epsilon(s')} Q^*\left(s', \arg\max_{a_{s'_\nu}} Q^*\left(s'_\nu, a_{s'_\nu}\right)\right)\right] \tag{49}$$

$$= r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q^*\left(s', \arg\max_{a_{\nu(s')}} Q^*\left(\nu(s'), a_{\nu(s')}\right)\right)\right] \tag{50}$$

$$= r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q^*\left(s', \pi\left(\nu(s')\right)\right)\right] \tag{51}$$

$$= r(s,a) + \gamma\left(\mathcal{L}^{\pi\circ\nu}Q^*\right)(s,a). \tag{52}$$

Thus, we have

$$Q^* = \left(I - \gamma\mathcal{L}^{\pi\circ\nu}\right)^{-1} r = Q^{\pi\circ\nu}, \tag{53}$$

where equations comes from lemma A.11. Further, according to the consistency assumption, we attain $Q^{\pi\circ\nu}(s,a) = Q^{\pi\circ\nu^*(\pi)}$. This shows that Q^* is the action-value adversarial function of policy π under the strongest adversary $\nu = \nu^*(\pi)$.

According to the consistency assumption and the definition of B^*_ϵ, we have that

$$\pi(\nu(s)) = \pi(s), \quad \forall s \in \mathcal{S}. \tag{54}$$

Then, for any stationary policy π', we have that

$$\left[\left(\mathcal{L}^{\pi\circ\nu} - \mathcal{L}^{\pi'\circ\nu^*(\pi')}\right)Q^{\pi\circ\nu}\right](s,a) \tag{55}$$

$$= \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q^{\pi\circ\nu}\left(s', \pi\left(\nu(s')\right)\right) - \mathbb{E}_{a'\sim\pi'(\cdot|\nu^*(s';\pi'))}Q^{\pi\circ\nu}(s',a')\right] \tag{56}$$

$$= \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q^{\pi\circ\nu}\left(s', \pi\left(s'\right)\right) - \mathbb{E}_{a'\sim\pi'(\cdot|\nu^*(s';\pi'))}Q^{\pi\circ\nu}(s',a')\right] \tag{57}$$

$$= \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a),a'\sim\pi'(\cdot|\nu^*(s';\pi'))}\left[Q^{\pi\circ\nu}\left(s', \pi\left(s'\right)\right) - Q^{\pi\circ\nu}(s',a')\right] \tag{58}$$

$$\geq 0, \tag{59}$$

where the second equality comes from (54) and the last inequality comes from (47).

Further, we have that

$$Q^* - Q^{\pi'\circ\nu^*(\pi')} = Q^{\pi\circ\nu} - Q^{\pi'\circ\nu^*(\pi')} \tag{60}$$

$$= Q^{\pi\circ\nu} - \left(I - \gamma\mathcal{L}^{\pi'\circ\nu^*(\pi')}\right)^{-1} r \tag{61}$$

$$= Q^{\pi\circ\nu} - \left(I - \gamma\mathcal{L}^{\pi'\circ\nu^*(\pi')}\right)^{-1}\left(I - \gamma\mathcal{L}^{\pi\circ\nu}\right)Q^{\pi\circ\nu} \tag{62}$$

$$= \left(I - \gamma\mathcal{L}^{\pi'\circ\nu^*(\pi')}\right)^{-1}\left(\left(I - \gamma\mathcal{L}^{\pi'\circ\nu^*(\pi')}\right) - \left(I - \gamma\mathcal{L}^{\pi\circ\nu}\right)\right)Q^{\pi\circ\nu} \tag{63}$$

$$= \gamma\left(I - \gamma\mathcal{L}^{\pi'\circ\nu^*(\pi')}\right)^{-1}\left(\mathcal{L}^{\pi\circ\nu} - \mathcal{L}^{\pi'\circ\nu^*(\pi')}\right)Q^{\pi\circ\nu} \tag{64}$$

$$\geq 0, \tag{65}$$

where the last inequality comes from (59) and lemma A.12. Thus, we have that $Q^{\pi\circ\nu} = Q^* \geq Q^{\pi'\circ\nu^*(\pi')}$ for all policy π' which shows that π is the optimal robust policy under strongest adversary. □

Corollary A.15. *If the consistency assumption holds, there exists a deterministic and stationary policy π^* which satisfies $V^{\pi^*\circ\nu^*(\pi^*)}(s) \geq V^{\pi\circ\nu^*(\pi)}(s)$ and $Q^{\pi^*\circ\nu^*(\pi^*)}(s,a) \geq Q^{\pi\circ\nu^*(\pi)}(s,a)$ for all $\pi \in \Pi$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

Proof. According to theorem A.14, we have that $Q^*(s,a) = \max_\pi \min_\nu Q^{\pi\circ\nu}(s,a)$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Define π^* and ν^* as the following:

$$\pi^*(s) := \arg\max_a Q^*(s,a), \tag{66}$$

$$\nu^*(s) := \arg\min_{s_\nu \in B_\epsilon(s)} Q^*\left(s, \arg\max_{a_{s_\nu}} Q^*\left(s_\nu, a_{s_\nu}\right)\right). \tag{67}$$

Then, we have that

$$\left(\mathcal{T}_{car}Q^*\right)(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\min_{s'_\nu \in B_\epsilon(s')} Q^*\left(s', \arg\max_{a_{s'_\nu}} Q^*\left(s'_\nu, a_{s'_\nu}\right)\right)\right] \tag{68}$$

$$= r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q^*\left(s', \arg\max_{a_{\nu(s')}} Q^*\left(\nu^*(s'), a_{\nu(s')}\right)\right)\right] \tag{69}$$

$$= r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q^*\left(s', \pi^*\left(\nu^*(s')\right)\right)\right] \tag{70}$$

$$= r(s,a) + \gamma\left(\mathcal{L}^{\pi^*\circ\nu^*}Q^*\right)(s,a). \tag{71}$$

Thus, we have

$$Q^* = \left(I - \gamma\mathcal{L}^{\pi\circ\nu}\right)^{-1} r = Q^{\pi^*\circ\nu^*}, \tag{72}$$

where equations comes from lemma A.11. Further, according to the consistency assumption, we attain $Q^{\pi^*\circ\nu^*}(s,a) = Q^{\pi^*\circ\nu^*(\pi^*)}$. This shows that Q^* is the action-value adversarial function of policy π^* under the strongest adversary $\nu* = \nu^*(\pi^*)$. Thus, we have that

$$Q^{\pi^*\circ\nu^*(\pi^*)}(s,a) \geq Q^{\pi\circ\nu^*(\pi)}(s,a), \quad \forall s\in\mathcal{S}, a\in\mathcal{A}. \tag{73}$$

For any policy π and $s\in\mathcal{S}$, we have that

$$V^{\pi^*\circ\nu^*(\pi^*)}(s) = \mathbb{E}_{a\sim\pi^*(\cdot|\nu^*(s;\pi^*))}Q^{\pi^*\circ\nu^*(\pi^*)}(s,a) \tag{74}$$

$$= \max_a Q^{\pi^*\circ\nu^*(\pi^*)}(s,a) \tag{75}$$

$$\geq \mathbb{E}_{a\sim\pi(\cdot|\nu^*(s;\pi))}Q^{\pi^*\circ\nu^*}(s,a) \tag{76}$$

$$\geq \mathbb{E}_{a\sim\pi(\cdot|\nu^*(s;\pi))}Q^{\pi\circ\nu^*(\pi)}(s,a) \tag{77}$$

$$= V^{\pi\circ\nu^*(\pi)}(s), \tag{78}$$

where the first and last equations come from lemma A.7 and the last inequality comes from (73). □

A.2.2. Not a Contraction

Theorem A.16. *$\mathcal{T}_{car}$ is not a contraction.*

Proof. Let $\mathcal{S} = [-1,1]$, $\mathcal{A} = \{a_1, a_2\}$, $0 < \epsilon \ll 1$ and dynamic transition $\mathbb{P}(\cdot|s,a)$ be a determined function. Let $n > \max\{\frac{\delta}{\gamma}, 2\delta\}$, $\delta > 0$ and

$$\begin{aligned} Q_1(s,a_1) = {} & 2n\cdot\mathbb{1}_{\{s\in[-1,0)\}} + \left[2n - \frac{2n-2\delta}{\frac{1}{8}\epsilon}s\right]\cdot\mathbb{1}_{\left\{s\in\left[0,\frac{1}{8}\epsilon\right)\right\}} + 2\delta\cdot\mathbb{1}_{\left\{s\in\left[\frac{1}{8}\epsilon,\frac{3}{8}\epsilon\right)\right\}} \\ & + \left[2\delta + \frac{n-2\delta}{\frac{1}{8}\epsilon}\left(s - \frac{3\epsilon}{8}\right)\right]\cdot\mathbb{1}_{\left\{s\in\left[\frac{3}{8}\epsilon,\frac{1}{2}\epsilon\right)\right\}} + n\cdot\mathbb{1}_{\left\{s\in\left[\frac{1}{2}\epsilon,1\right]\right\}}, \end{aligned} \tag{79}$$

$$\begin{aligned} Q_1(s,a_2) = {} & n\cdot\mathbb{1}_{\{s\in[-1,0)\}} + \left[n - \frac{n-\delta}{\frac{1}{8}\epsilon}s\right]\cdot\mathbb{1}_{\left\{s\in\left[0,\frac{1}{8}\epsilon\right)\right\}} + \delta\cdot\mathbb{1}_{\left\{s\in\left[\frac{1}{8}\epsilon,\frac{3}{8}\epsilon\right)\right\}} \\ & + \left[\delta + \frac{2n-\delta}{\frac{1}{8}\epsilon}\left(s - \frac{3\epsilon}{8}\right)\right]\cdot\mathbb{1}_{\left\{s\in\left[\frac{3}{8}\epsilon,\frac{1}{2}\epsilon\right)\right\}} + 2n\cdot\mathbb{1}_{\left\{s\in\left[\frac{1}{2}\epsilon,1\right]\right\}}, \end{aligned} \tag{80}$$

$$\begin{aligned} Q_2(s,a_1) = {} & 2n\cdot\mathbb{1}_{\{s\in[-1,0)\}} + \left[2n - \frac{2n-\delta}{\frac{1}{8}\epsilon}s\right]\cdot\mathbb{1}_{\left\{s\in\left[0,\frac{1}{8}\epsilon\right)\right\}} + \delta\cdot\mathbb{1}_{\left\{s\in\left[\frac{1}{8}\epsilon,\frac{3}{8}\epsilon\right)\right\}} \\ & + \left[\delta + \frac{n-\delta}{\frac{1}{8}\epsilon}\left(s - \frac{3\epsilon}{8}\right)\right]\cdot\mathbb{1}_{\left\{s\in\left[\frac{3}{8}\epsilon,\frac{1}{2}\epsilon\right)\right\}} + n\cdot\mathbb{1}_{\left\{s\in\left[\frac{1}{2}\epsilon,1\right]\right\}}, \end{aligned} \tag{81}$$

$$
\begin{aligned}
Q_2(s,a_2) = n \cdot \mathbb{1}_{\{s\in[-1,0)\}} + \left[n - \frac{n-2\delta}{\frac{1}{8}\epsilon}s\right] \cdot \mathbb{1}_{\left\{s\in\left[0,\frac{1}{8}\epsilon\right)\right\}} + 2\delta \cdot \mathbb{1}_{\left\{s\in\left[\frac{1}{8}\epsilon,\frac{3}{8}\epsilon\right)\right\}} \\
+ \left[2\delta + \frac{2n-2\delta}{\frac{1}{8}\epsilon}\left(s - \frac{3\epsilon}{8}\right)\right] \cdot \mathbb{1}_{\left\{s\in\left[\frac{3}{8}\epsilon,\frac{1}{2}\epsilon\right)\right\}} + 2n \cdot \mathbb{1}_{\left\{s\in\left[\frac{1}{2}\epsilon,1\right]\right\}}.
\end{aligned} \tag{82}
$$

Then

$$
\|Q_1 - Q_2\|_{L^\infty(\mathcal{S}\times\mathcal{A})} = \delta. \tag{83}
$$

We have

$$
\begin{aligned}
\mathcal{T}_{car}Q_1(s,a) - \mathcal{T}_{car}Q_2(s,a) = \gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\min_{s_\nu^1\in B_\epsilon(s')} Q_1\left(s', \arg\max_{a_{s_\nu^1}} Q_1\left(s_\nu^1, a_{s_\nu^1}\right)\right) - \right. \\
\left. \min_{s_\nu^2\in B_\epsilon(s')} Q_2\left(s', \arg\max_{a_{s_\nu^2}} Q_2\left(s_\nu^2, a_{s_\nu^2}\right)\right)\right].
\end{aligned} \tag{84}
$$

Let $\mathbb{P}(s' = -\frac{\epsilon}{2}|s,a) = 1$ and $s' = -\frac{\epsilon}{2}$, then

$$
\min_{s_\nu^1\in B_\epsilon(s')} Q_1\left(s', \arg\max_{a_{s_\nu^1}} Q_1\left(s_\nu^1, a_{s_\nu^1}\right)\right) = Q_1(s', a_1), \tag{85}
$$

$$
\min_{s_\nu^1\in B_\epsilon(s')} Q_2\left(s', \arg\max_{a_{s_\nu^2}} Q_2\left(s_\nu^2, a_{s_\nu^2}\right)\right) = Q_2(s', a_2). \tag{86}
$$

Thus

$$
\mathcal{T}_{car}Q_1(s,a) - \mathcal{T}_{car}Q_2(s,a) = \gamma\left[Q_1(s',a_1) - Q_2(s',a_2)\right] = \gamma n > \delta, \tag{87}
$$

which means that

$$
\|\mathcal{T}_{car}Q_1 - \mathcal{T}_{car}Q_2\|_{L^\infty(\mathcal{S}\times\mathcal{A})} > \|Q_1 - Q_2\|_{L^\infty(\mathcal{S}\times\mathcal{A})}. \tag{88}
$$

Therefore, $\mathcal{T}_{car}$ is not a contraction. □

A.2.3. CONVERGENCE

In this section, we prove a conclusion for convergence of the fixed point iterations of the CAR operator under the $(L_r, L_\mathbb{P})$-smooth environment assumption.

Definition A.17 (Bukharin et al. (2023))**.** Let $\mathcal{S} \subseteq \mathbb{R}^d$. We say the environment is $(L_r, L_\mathbb{P})$-smooth, if the reward function $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$, and the transition dynamics $\mathbb{P} : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$ satisfy

$$
|r(s,a) - r(s',a)| \le L_r\|s - s'\| \text{ and } \|\mathbb{P}(\cdot \mid s,a) - \mathbb{P}(\cdot \mid s',a)\|_{L^1(\mathcal{S})} \le L_\mathbb{P}\|s - s'\|,
$$

for $(s,s',a) \in \mathcal{S}\times\mathcal{S}\times\mathcal{A}$. $\|\cdot\|$ denotes a metric on $\mathbb{R}^d$.

The definition is motivated by observations that certain natural environments show smooth reward function and transition dynamics, especially in continuous control tasks where the transition dynamics come from some physical laws.

The following lemma shows that $\mathcal{T}_{car}^k Q$ is uniformly bounded.

Lemma A.18. *Suppose Q and r are uniformly bounded, i.e. $\exists\, M_Q, M_r > 0$ such that $|Q(s,a)| \le M_Q$, $|r(s,a)| \le M_r\ \forall s\in\mathcal{S}, a\in\mathcal{A}$. Then $\mathcal{T}_{car}Q(\cdot,a)$ is uniformly bounded, i.e.*

$$
|\mathcal{T}_{car}Q(s,a)| \le C_Q,\ \forall s\in\mathcal{S}, a\in\mathcal{A}, \tag{89}
$$

where $C_Q = \max\left\{M_Q, \frac{M_r}{1-\gamma}\right\}$. Further, for any $k\in\mathbb{N}$, $\mathcal{T}_{car}^k Q(\cdot,a)$ has the same uniform bound as $\mathcal{T}_{car}Q(\cdot,a)$, i.e.

$$
\left|\mathcal{T}_{car}^k Q(s,a)\right| \le C_Q,\ \forall s\in\mathcal{S}, a\in\mathcal{A}. \tag{90}
$$

Proof.

$$|\mathcal{T}_{car}Q(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[\min_{s'_\nu \in B_\epsilon(s')} Q\left(s', \arg\max_{a_{s'_\nu}} Q\left(s'_\nu, a_{s'_\nu}\right)\right)\right]\right| \tag{91}$$

$$\leq |r(s,a)| + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \left| \min_{s'_\nu \in B_\epsilon(s')} Q\left(s', \arg\max_{a_{s'_\nu}} Q\left(s'_\nu, a_{s'_\nu}\right)\right)\right| \tag{92}$$

$$\leq M_r + \gamma M_Q \tag{93}$$

$$\leq \max\left\{M_Q, \frac{M_r}{1-\gamma}\right\}, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{94}$$

Let $C_Q = \max\left\{M_Q, \frac{M_r}{1-\gamma}\right\}$. Suppose the inequality (90) holds for $k = n$. Then, for $k = n+1$, we have

$$\left|\mathcal{T}^{n+1}_{car}Q(s,a)\right| = \left| r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[\min_{s'_\nu \in B_\epsilon(s')} \mathcal{T}^n_{car}Q\left(s', \arg\max_{a_{s'_\nu}} \mathcal{T}^n_{car}Q\left(s'_\nu, a_{s'_\nu}\right)\right)\right]\right| \tag{95}$$

$$\leq |r(s,a)| + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \left| \min_{s'_\nu \in B_\epsilon(s')} \mathcal{T}^n_{car}Q\left(s', \arg\max_{a_{s'_\nu}} \mathcal{T}^n_{car}Q\left(s'_\nu, a_{s'_\nu}\right)\right)\right| \tag{96}$$

$$\leq M_r + \gamma C_Q \tag{97}$$

$$\leq (1-\gamma)C_Q + \gamma C_Q \tag{98}$$

$$= C_Q. \tag{99}$$

By induction, we have $\left|\mathcal{T}^k_{car}Q(s,a)\right| \leq C_Q,\ \forall s \in \mathcal{S}, a \in \mathcal{A}, k \in \mathbb{N}$. □

The following lemma shows that $\mathcal{T}^k_{car}Q$ is uniformly Lipschitz continuous in the $(L_r, L_\mathbb{P})$-smooth environment.

Lemma A.19. *Suppose the environment is $(L_r, L_\mathbb{P})$-smooth and suppose Q and r are uniformly bounded, i.e. $\exists\, M_Q, M_r > 0$ such that $|Q(s,a)| \leq M_Q,\ |r(s,a)| \leq M_r\ \forall s \in \mathcal{S}, a \in \mathcal{A}$. Then $\mathcal{T}_{car}Q(\cdot,a)$ is Lipschitz continuous, i.e.*

$$|\mathcal{T}_{car}Q(s,a) - \mathcal{T}_{car}Q(s',a)| \leq L_{\mathcal{T}_{car}}\|s - s'\|, \tag{100}$$

where $L_{\mathcal{T}_{car}} = L_r + \gamma C_Q L_\mathbb{P}$ and $C_Q = \max\left\{M_Q, \frac{M_r}{1-\gamma}\right\}$. Further, for any $k \in \mathbb{N}$, $\mathcal{T}^k_{car}Q(\cdot,a)$ is Lipschitz continuous and has the same Lipschitz constant as $\mathcal{T}_{car}Q(\cdot,a)$, i.e.

$$\left|\mathcal{T}^k_{car}Q(s,a) - \mathcal{T}^k_{car}Q(s',a)\right| \leq L_{\mathcal{T}_{car}}\|s - s'\|. \tag{101}$$

Proof. For all $s_1, s_2 \in \mathcal{S}$, we have

$$\mathcal{T}_{car}Q(s_1,a) - \mathcal{T}_{car}Q(s_2,a) \tag{102}$$

$$= r(s_1,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_1,a)} \left[\min_{s'_\nu \in B_\epsilon(s')} Q\left(s', \arg\max_{a_{s'_\nu}} Q\left(s'_\nu, a_{s'_\nu}\right)\right)\right] \tag{103}$$

$$- r(s_2,a) - \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_2,a)} \left[\min_{s'_\nu \in B_\epsilon(s')} Q\left(s', \arg\max_{a_{s'_\nu}} Q\left(s'_\nu, a_{s'_\nu}\right)\right)\right] \tag{104}$$

$$= (r(s_1,a) - r(s_2,a)) \tag{105}$$

$$+ \gamma \int_{s'} (\mathbb{P}(s'|s_1,a) - \mathbb{P}(s'|s_2,a)) \min_{s'_\nu \in B_\epsilon(s')} Q\left(s', \arg\max_{a_{s'_\nu}} Q\left(s'_\nu, a_{s'_\nu}\right)\right) ds'. \tag{106}$$

Then, we have

$$
\begin{aligned}
&|\mathcal{T}_{car}Q(s_1,a) - \mathcal{T}_{car}Q(s_2,a)| && (107)\\
&\leq |(r(s_1,a) - r(s_2,a))| && (108)\\
&\quad + \left|\gamma \int_{s'} \left(\mathbb{P}(s'|s_1,a) - \mathbb{P}(s'|s_2,a)\right) \min_{s'_\nu \in B_\epsilon(s')} Q\left(s', \arg\max_{a_{s'_\nu}} Q\left(s'_\nu, a_{s'_\nu}\right)\right) ds'\right| && (109)\\
&\leq L_r\|s_1 - s_2\| && (110)\\
&\quad + \gamma \int_{s'} |\mathbb{P}(s'|s_1,a) - \mathbb{P}(s'|s_2,a)| \left|\min_{s'_\nu \in B_\epsilon(s')} Q\left(s', \arg\max_{a_{s'_\nu}} Q\left(s'_\nu, a_{s'_\nu}\right)\right)\right| ds' && (111)\\
&\leq L_r\|s_1 - s_2\| + \gamma C_Q \int_{s'} |\mathbb{P}(s'|s_1,a) - \mathbb{P}(s'|s_2,a)|\, ds' && (112)\\
&\leq L_r\|s_1 - s_2\| + \gamma C_Q L_{\mathbb{P}}\|s_1 - s_2\| && (113)\\
&= \left(L_r + \gamma C_Q L_{\mathbb{P}}\right)\|s_1 - s_2\|. && (114)
\end{aligned}
$$

The second inequality comes from the Lipschitz property of r. The third inequality comes from the uniform boundedness of Q and the last inequality utilizes the Lipschitz property of $\mathbb{P}$.

Note that $\mathcal{T}_{car}^k$ and $\mathcal{T}_{car}$ have the same uniform boundedness C_Q. Then, due to lemma A.18, we can extend the above proof to $\mathcal{T}_{car}^k$. □

Remark A.20. Note that if replace the operator $\mathcal{T}_{car}$ in the Lemma A.18 and Lemma A.19 with Bellman optimality operator $\mathcal{T}_B$, these lemmas still hold.

The following lemma shows that the fixed point iteration has a property close to contraction.

Lemma A.21. *Suppose Q and r are uniformly bounded, i.e. $\exists\ M_Q, M_r > 0$ such that $|Q(s,a)| \leq M_Q,\ |r(s,a)| \leq M_r\ \forall s \in \mathcal{S}, a \in \mathcal{A}$. Let Q^* denote the Bellman optimality Q-function. If the consistency assumption holds, we have*

$$
\|\mathcal{T}_{car}Q - \mathcal{T}_{car}Q^*\|_\infty \leq \gamma\left(\|Q - Q^*\|_\infty + 2\max_s \max_{s_\nu \in B_\epsilon^*(s)} \max_a |Q(s,a) - Q(s_\nu,a)|\right). \tag{115}
$$

Further, if $Q(\cdot,a)$ is L-Lipschitz continuous with respect to $s \in \mathcal{S}$, i.e

$$
|Q(s,a) - Q(s',a)| \leq L\|s - s'\|, \quad \forall s, s' \in \mathcal{S},\ a \in \mathcal{A}, \tag{116}
$$

we have

$$
\|\mathcal{T}_{car}Q - \mathcal{T}_{car}Q^*\|_\infty \leq \gamma\|Q - Q^*\|_\infty + 2\gamma L\epsilon. \tag{117}
$$

Proof. Denote $a^*_{s'_\nu,Q} = \arg\max_a Q(s'_\nu, a)$ and $s'^*_\nu = \arg\min_{s'_\nu \in B_\epsilon^*(s')} Q\left(s', a^*_{s'_\nu,Q}\right)$. If $\mathcal{T}_{car}Q > \mathcal{T}_{car}Q^*$, we have

$$
\begin{aligned}
&\left(\mathcal{T}_{car}Q\right)(s,a) - \left(\mathcal{T}_{car}Q^*\right)(s,a) && (118)\\
&= \gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\min_{s'_\nu \in B_\epsilon^*(s')} Q\left(s', a^*_{s'_\nu,Q}\right) - \min_{s'_\nu \in B_\epsilon^*(s')} Q^*\left(s', a^*_{s'_\nu,Q^*}\right)\right] && (119)\\
&= \gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q\left(s', a^*_{s'^*_\nu,Q}\right) - Q^*\left(s', a^*_{s',Q^*}\right)\right] && (120)\\
&= \gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q\left(s', a^*_{s'^*_\nu,Q}\right) - Q^*\left(s', a^*_{s'^*_\nu,Q}\right) + Q^*\left(s', a^*_{s'^*_\nu,Q}\right) - Q^*\left(s', a^*_{s',Q^*}\right)\right] && (121)\\
&\leq \gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q\left(s', a^*_{s'^*_\nu,Q}\right) - Q^*\left(s', a^*_{s'^*_\nu,Q}\right)\right] && (122)\\
&\leq \gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'}\left(Q(s',a') - Q^*(s',a')\right)\right] && (123)\\
&\leq \gamma\|Q - Q^*\|_\infty, && (124)
\end{aligned}
$$

where the second equality utilize the definition of $B_\epsilon^*(s')$ and the first inequality comes from the optimality of a^*_{s',Q^*}. If $\mathcal{T}_{car}Q < \mathcal{T}_{car}Q^*$, we have

$$
\begin{aligned}
&\left(\mathcal{T}_{car}Q^*\right)(s,a) - \left(\mathcal{T}_{car}Q\right)(s,a) && (125)\\
&= \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\min_{s'_\nu \in B_\epsilon^*(s')} Q^*\left(s', a^*_{s'_\nu,Q^*}\right) - \min_{s'_\nu \in B_\epsilon^*(s')} Q\left(s', a^*_{s'_\nu,Q}\right)\right] && (126)\\
&= \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q^*\left(s', a^*_{s',Q^*}\right) - Q\left(s', a^*_{s'^*_\nu,Q}\right)\right] && (127)\\
&= \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q^*\left(s', a^*_{s',Q^*}\right) - Q\left(s', a^*_{s',Q^*}\right)\right] && (128)\\
&\quad + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q\left(s', a^*_{s',Q^*}\right) - Q\left(s'^*_\nu, a^*_{s',Q^*}\right)\right] && (129)\\
&\quad + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q\left(s'^*_\nu, a^*_{s',Q^*}\right) - Q\left(s', a^*_{s'^*_\nu,Q}\right)\right]. && (130)
\end{aligned}
$$

We will separately analyze the items 128, 129 and 130. Firstly, we can bound the item 128 with $\|Q - Q^*\|_\infty$.

$$
\begin{aligned}
&\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q^*\left(s', a^*_{s',Q^*}\right) - Q\left(s', a^*_{s',Q^*}\right)\right] && (131)\\
&\le \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'}\left(Q\left(s',a'\right) - Q^*\left(s',a'\right)\right)\right] && (132)\\
&\le \|Q - Q^*\|_\infty. && (133)
\end{aligned}
$$

For the item 129, we have

$$
\begin{aligned}
&\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q\left(s', a^*_{s',Q^*}\right) - Q\left(s'^*_\nu, a^*_{s',Q^*}\right)\right] && (134)\\
&\le \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'}\left(Q\left(s',a'\right) - Q\left(s'^*_\nu,a'\right)\right)\right] && (135)\\
&\le \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{s'_\nu \in B_\epsilon^*(s')}\max_{a'}\left|Q\left(s',a'\right) - Q\left(s'_\nu,a'\right)\right|\right] && (136)\\
&\le \max_{s}\max_{s_\nu \in B_\epsilon^*(s)}\max_{a}\left|Q\left(s,a\right) - Q\left(s_\nu,a\right)\right|. && (137)
\end{aligned}
$$

Due to $a^*_{s'^*_\nu,Q} = \arg\max_a Q\left(s'^*_\nu, a\right)$, we have $Q\left(s'^*_\nu, a\right) \le Q\left(s'^*_\nu, a^*_{s'^*_\nu,Q}\right),\ \forall a$. Then, for the item 130, we have

$$
\begin{aligned}
&\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q\left(s'^*_\nu, a^*_{s',Q^*}\right) - Q\left(s', a^*_{s'^*_\nu,Q}\right)\right] && (138)\\
&\le \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[Q\left(s'^*_\nu, a^*_{s'^*_\nu,Q}\right) - Q\left(s', a^*_{s'^*_\nu,Q}\right)\right] && (139)\\
&\le \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'}\left|Q\left(s',a'\right) - Q\left(s'^*_\nu,a'\right)\right|\right] && (140)\\
&\le \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{s'_\nu \in B_\epsilon^*(s')}\max_{a'}\left|Q\left(s',a'\right) - Q\left(s'_\nu,a'\right)\right|\right] && (141)\\
&\le \max_{s}\max_{s_\nu \in B_\epsilon^*(s)}\max_{a}\left|Q\left(s,a\right) - Q\left(s_\nu,a\right)\right|. && (142)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
&\left(\mathcal{T}_{car}Q^*\right)(s,a) - \left(\mathcal{T}_{car}Q\right)(s,a) && (143)\\
&\le \gamma\left(\|Q - Q^*\|_\infty + 2\max_{s}\max_{s_\nu \in B_\epsilon^*(s)}\max_{a}\left|Q\left(s,a\right) - Q\left(s_\nu,a\right)\right|\right). && (144)
\end{aligned}
$$

In a summary, we get

$$
\|\mathcal{T}_{car}Q - \mathcal{T}_{car}Q^*\|_\infty \le \gamma\left(\|Q - Q^*\|_\infty + 2\max_{s}\max_{s_\nu \in B_\epsilon^*(s)}\max_{a}\left|Q\left(s,a\right) - Q\left(s_\nu,a\right)\right|\right). \quad (145)
$$

Further, when $Q(\cdot, a)$ is L-Lipschitz continuous, i.e

$$|Q(s,a) - Q(s',a)| \leq L\|s - s'\|, \quad \forall s, s' \in \mathcal{S},\ a \in \mathcal{A}, \tag{146}$$

we have

$$\max_s \max_{s_\nu \in B_\epsilon^*(s)} \max_a |Q(s,a) - Q(s_\nu, a)| \tag{147}$$

$$\leq \max_s \max_{s_\nu \in B_\epsilon^*(s)} L\|s - s_\nu\| \tag{148}$$

$$\leq L\epsilon. \tag{149}$$

Then, we have

$$\|\mathcal{T}_{car}Q - \mathcal{T}_{car}Q^*\|_\infty \leq \gamma\left(\|Q - Q^*\|_\infty + 2L\epsilon\right). \tag{150}$$

□

Remark A.22. We can relax the Lipschitz condition to local lipschitz continuous in the $B_\epsilon^*(s)$.

We prove that the fixed point iterations of $\mathcal{T}_{car}$ at least converge to a sub-optimal solution close to Q^* in the $(L_r, L_\mathbb{P})$-smooth environment.

Theorem A.23. *Suppose the environment is $(L_r, L_\mathbb{P})$-smooth and suppose Q_0 and r are uniformly bounded, i.e. $\exists\ M_{Q_0}, M_r > 0$ such that $|Q_0(s,a)| \leq M_{Q_0},\ |r(s,a)| \leq M_r\ \forall s \in \mathcal{S}, a \in \mathcal{A}$. Let Q^* denote the Bellman optimality Q-function and $Q_{k+1} = \mathcal{T}_{car}Q_k = \mathcal{T}_{car}^{k+1}Q_0$ for all $k \in \mathbb{N}$. If the consistency assumption holds, we have*

$$\|Q_{k+1} - Q^*\|_\infty \leq \gamma^{k+1}\|Q_0 - Q^*\|_\infty + \gamma^{k+1}D_{Q_0} + \frac{2\gamma\epsilon}{1-\gamma}L_{\mathcal{T}_{car}}, \tag{151}$$

where $D_{Q_0} = 2\max_s \max_{s_\nu \in B_\epsilon^*(s)} \max_a |Q_0(s,a) - Q_0(s_\nu, a)|$, $L_{\mathcal{T}_{car}} = L_r + \gamma C_{Q_0} L_\mathbb{P}$ and $C_{Q_0} = \max\left\{M_{Q_0}, \frac{M_r}{1-\gamma}\right\}$.

Proof. For any $k \in \mathbb{N}$, we have

$$\|Q_{k+1} - Q^*\|_\infty \tag{152}$$

$$= \|\mathcal{T}_{car}^{k+1}Q_0 - \mathcal{T}_{car}^{k+1}Q^*\|_\infty \tag{153}$$

$$\leq \gamma\|\mathcal{T}_{car}^{k}Q_0 - \mathcal{T}_{car}^{k}Q^*\|_\infty + 2\gamma L_{\mathcal{T}_{car}}\epsilon \tag{154}$$

$$\leq \gamma\left(\gamma\|\mathcal{T}_{car}^{k-1}Q_0 - \mathcal{T}_{car}^{k-1}Q^*\|_\infty + 2\gamma L_{\mathcal{T}_{car}}\epsilon\right) + 2\gamma L_{\mathcal{T}_{car}}\epsilon \tag{155}$$

$$= \gamma^2\|\mathcal{T}_{car}^{k-1}Q_0 - \mathcal{T}_{car}^{k-1}Q^*\|_\infty + 2\epsilon L_{\mathcal{T}_{car}}\sum_{l=1}^{2}\gamma^l \tag{156}$$

$$\leq \cdots \tag{157}$$

$$\leq \gamma^k\|\mathcal{T}_{car}Q_0 - \mathcal{T}_{car}Q^*\|_\infty + 2\epsilon L_{\mathcal{T}_{car}}\sum_{l=1}^{k}\gamma^l \tag{158}$$

$$\leq \gamma^{k+1}\|Q_0 - Q^*\|_\infty + 2\gamma^{k+1}\max_s \max_{s_\nu \in B_\epsilon^*(s)} \max_a |Q_0(s,a) - Q_0(s_\nu, a)| + 2\epsilon L_{\mathcal{T}_{car}}\sum_{l=1}^{k}\gamma^l \tag{159}$$

$$\leq \gamma^{k+1}\|Q_0 - Q^*\|_\infty + 2\gamma^{k+1}\max_s \max_{s_\nu \in B_\epsilon^*(s)} \max_a |Q_0(s,a) - Q_0(s_\nu, a)| + \frac{2\gamma\epsilon}{1-\gamma}L_{\mathcal{T}_{car}}. \tag{160}$$

The first and second inequalities come from Lemma A.19 and Lemma A.21. The penultimate inequality comes from Lemma A.21. □

B. Theorems and Proofs of Policy Robustness under Bellman p-error

Banach Space is a complete normed space $(X, \|\cdot\|)$, consisting of a vector space X together with a norm $\|\cdot\| : X \to \mathbb{R}^+$. In this paper, we consider the setting where the continuous state space $\mathcal{S} \subset \mathbb{R}^d$ is a compact set and the action space $\mathcal{A}$ is a finite set. We discuss in the Banach space $\left(L^p\left(\mathcal{S}\times\mathcal{A}\right), \|\cdot\|_p\right),\ 1 \le p \le \infty$. Define $L^p\left(\mathcal{S}\times\mathcal{A}\right) = \{f \| \|f\|_p < \infty\}$, where $\|f\|_p = \left(\int_{\mathcal{S}} \sum_{a\in\mathcal{A}} |f(s,a)|^p \, d\mu(s)\right)^{\frac{1}{p}}$ for $1 \le p < \infty$, μ is the measure over $\mathcal{S}$ and $\|f\|_\infty = \inf\{M \in \mathbb{R}_{\ge 0} | \, |f(s,a)| \le M \text{ for almost every } (s,a)\}$. For simplicity, we refer to this Banach space as $L^p\left(\mathcal{S}\times\mathcal{A}\right)$.

B.1. L^∞ is Necessary for Adversarial Robustness

Theorem B.1. *There exists an MDP instance $\mathcal{M}$ such that the following statements hold. Given a function Q and adversary perturbation size ϵ, let $\mathcal{S}^Q_{sub}$ denote the set of states where the greedy policy according to Q is suboptimal, i.e. $\mathcal{S}^Q_{sub} = \{s | Q^*(s, \arg\max_a Q(s,a)) < \max_a Q^*(s,a)\}$ and let $\mathcal{S}^Q_{adv}$ denote the set of states in whose ϵ-neighbourhood there exists the adversarial state, i.e. $\mathcal{S}^Q_{adv} = \{s | \exists s_\nu \in B_\epsilon(s),\ s.t.\ Q^*(s, \arg\max_a Q(s_\nu, a)) < \max_a Q^*(s,a)\}$, where Q^* is the Bellman optimal Q-function.*

- *For any $1 \le p < \infty$ and $\delta > 0$, there exists a function $Q \in L^p\left(\mathcal{S}\times\mathcal{A}\right)$ satisfying $\|Q - Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})} \le \delta$ such that $\mu\left(\mathcal{S}^Q_{sub}\right) = O(\delta)$ yet $\mu\left(\mathcal{S}^Q_{adv}\right) = \mu\left(\mathcal{S}\right)$.*
- *There exists a $\bar{\delta} > 0$ such that for any $0 < \delta \le \bar{\delta}$, for any function $Q \in L^\infty\left(\mathcal{S}\times\mathcal{A}\right)$ satisfying $\|Q - Q^*\|_{L^\infty(\mathcal{S}\times\mathcal{A})} \le \delta$, we have that $\mu\left(\mathcal{S}^Q_{sub}\right) = O(\delta)$ and $\mu\left(\mathcal{S}^Q_{adv}\right) = 2\epsilon + O\left(\delta\right)$.*

Proof. Given a MDP instance $\mathcal{M}$ such that $\mathcal{S} = [-1, 1]$, $\mathcal{A} = \{a_1, a_2\}$ and

$$\mathbb{P}(s'|s,a_1) = \begin{cases} \mathbb{1}_{\{s'=s-\epsilon_1\}}, & s \in [-1+\epsilon_1, 1] \\ \mathbb{1}_{\{s'=-1\}}, & s \in [-1, -1+\epsilon_1) \end{cases} \tag{161}$$

$$\mathbb{P}(s'|s,a_2) = \begin{cases} \mathbb{1}_{\{s'=s+\epsilon_1\}}, & s \in [-1, 1-\epsilon_1] \\ \mathbb{1}_{\{s'=1\}}, & s \in (1-\epsilon_1, 1] \end{cases} \tag{162}$$

$$\begin{aligned} r(s,a_1) &= -ks, \\ r(s,a_2) &= ks. \end{aligned} \tag{163}$$

where $\mathbb{P}$ is the transition dynamic, r is the reward function, $k > 0$, $0 < \epsilon_1 \ll 1$ and $\mathbb{1}_{\{\cdot\}}$ is the indicator function. Let γ be the discount factor.

First, we prove that equation (164) is the optimal policy.

$$\pi^*(s) = \arg\max_a Q^*(s,a) = \begin{cases} \{a_2\}, s > 0 \\ \{a_1\}, s < 0 \\ \{a_1, a_2\}, s = 0 \end{cases} \tag{164}$$

Define s^π_t is state rollouted by policy π in time step t.

Let $s_0 > 0$, then

- If $s^{\pi^*}_t = 1$, while $s^\pi_t \in [-1, 1]$, then $s^{\pi^*}_t \ge |s^\pi_t|$ hold for any policy π.
- If $s^{\pi^*}_t < 1$. First we have

$$0 < s^{\pi^*}_t = s_0 + t\epsilon_1 < 1, \tag{165}$$

$$-1 < s_0 - t\epsilon_1. \tag{166}$$

Then for any policy π, we have the following equation by definition of transition,

$$s^\pi_t = s_0 + \sum_{i=1}^{t} x_i \epsilon_1, \tag{167}$$

where $x_i \in \{-1, 1\}, i = 1, ..., t$. Then

$$s_t^{\pi^*} \geq |s_t^\pi| \tag{168}$$

Then for any policy π ,$s_0 > 0$ and $t \geq 0$, we have

$$s_t^{\pi^*} \geq |s_t^\pi| \tag{169}$$

and

$$\begin{aligned}
&\pi(a_2|s_t^\pi) r(s_t^\pi, a_2) + \pi(a_1|s_t^\pi) r(s_t^\pi, a_1) \\
&\leq \max\{\pi(a_2|s_t^\pi)(ks_t^\pi) + \pi(a_1|s_t^\pi)(-ks_t^\pi), \pi(a_2|s_t^\pi)(-ks_t^\pi) + \pi(a_1|s_t^\pi)(ks_t^\pi)\} \\
&= |ks_t^\pi \left(\pi(a_2|s_t^\pi) - \pi(a_1|s_t^\pi)\right)| \\
&\leq |ks_t^\pi| \\
&\leq ks_t^{\pi^*} \\
&= r(s_t^{\pi^*}, a_2)
\end{aligned} \tag{170}$$

Let s_0 be the initial state, $\tau = (s_0, ...)$ be the trajectory of policy π. Define $J(\pi, s_0) = \mathbb{E}_\tau \sum_t \gamma^t r(s_t, a_t)$ is expected reward in initial state s_0 about policy π. Then

$$J(\pi^*, s_0) - J(\pi, s_0) = \mathbb{E}_{s_t^{\pi^*}, a_t \sim \pi^*(\cdot|s_t^{\pi^*})} \sum_t \gamma^t r(s_t^{\pi^*}, a_t) - \mathbb{E}_{s_t^\pi, a_t \sim \pi(\cdot|s_t^\pi)} \sum_t \gamma^t r(s_t^\pi, a_t) \tag{171}$$

$$= \sum_t \gamma^t r(s_t^{\pi^*}, a_2) - \sum_t \gamma^t \mathbb{E}_{s_t^\pi, a_t \sim \pi(\cdot|s_t^\pi)} r(s_t^\pi, a_t) \tag{172}$$

$$= \sum_t \gamma^t r(s_t^{\pi^*}, a_2) - \sum_t \gamma^t \mathbb{E}_{s_t^\pi} \left[\pi(a_2|s_t^\pi) r(s_t^\pi, a_2) + \pi(a_1|s_t^\pi) r(s_t^\pi, a_1)\right] \tag{173}$$

$$= \sum_t \gamma^t \mathbb{E}_{s_t^\pi} \left[r(s_t^{\pi^*}, a_2) - \left[\pi(a_2|s_t^\pi) r(s_t^\pi, a_2) + \pi(a_1|s_t^\pi) r(s_t^\pi, a_1)\right]\right] \tag{174}$$

$$\geq 0. \tag{175}$$

For (172), the policy π^* and dynamic transition $\mathbb{P}$ are deterministic. For (175), We use property (170).

Then for $s > 0$, we get that the optimal policy is $\pi(\cdot|s) = a_2$. By symmetry, we can also get that the optimal policy is $\pi(\cdot|s) = a_1$ for $s < 0$ and a_1,a_2 are also optimal action for $s = 0$. Thus we have proved equation (164) is the optimal policy.

First,we have the following equation according to (164)

$$Q^*(0, a_2) = Q^*(0, a_1). \tag{176}$$

For $s > 0$, we have

$$\begin{aligned}
Q^*(s, a_2) &= ks + \gamma k(s + \epsilon_1) + \gamma^2 k(s + 2\epsilon_1) + ... + \gamma^{t_s} k(s + t_s \epsilon_1) + \sum_{n=1}^{\infty} \gamma^{t_s + n} k \times 1 \\
&= ks + k \left[\sum_{t=1}^{t_s} \gamma^t \left(s + t\epsilon_1\right) + \sum_{t=t_s+1}^{\infty} \gamma^t\right].
\end{aligned} \tag{177}$$

where $s + t_s \epsilon_1 \in (1 - \epsilon_1, 1]$, i.e. $t_s = \lfloor \frac{1-s}{\epsilon_1} \rfloor$.

For $s \geq \epsilon_1$, we have

$$\begin{aligned}
Q^*(s, a_1) &= -ks + \gamma k(s - \epsilon_1) + \gamma^2 ks + \cdots + \gamma^{t_s+2} k(s + t_s \epsilon_1) + \sum_{n=1}^{\infty} \gamma^{t_s + 2 + n} k \times 1 \\
&= -ks + k \left[\sum_{t=1}^{t_s+2} \gamma^t \left(s + (t-2)\epsilon_1\right) + \sum_{t=t_s+3}^{\infty} \gamma^t\right].
\end{aligned} \tag{178}$$

For $0 < s < \epsilon_1$, we have

$$\begin{aligned} Q^*(s,a_1) &= -ks + \gamma(-k)(s-\epsilon_1) + \gamma^2(-k)(s-2\epsilon_1) + \cdots + \gamma^{q_s}(-k)(s-q_s\epsilon_1) + \sum_{n=1}^{\infty}\gamma^{q_s+n}(-k)(-1) \\ &= -ks + k\left[\sum_{t=1}^{q_s}\gamma^t\left(t\epsilon_1 - s\right) + \sum_{t=q_s+1}^{\infty}\gamma^t\right]. \end{aligned} \tag{179}$$

where $s - q_s\epsilon_1 \in [-1, -1+\epsilon_1)$,i.e. $q_s = \lfloor \frac{1+s}{\epsilon_1} \rfloor > t_s$.

According to (177), (178), (179) and $q_s > t_s$, we have

$$Q^*(s,a_2) - Q^*(s,a_1) > 2ks, s > 0 \tag{180}$$

By symmetry, we can also get

$$Q^*(s,a_1) - Q^*(s,a_2) > -2ks, s < 0 \tag{181}$$

(1)First, we have

$$0 < Q^*(s,a) < \sum_{t=0}^{\infty}\gamma^t = \frac{1}{1-\gamma} \tag{182}$$

For any $1 \le p < \infty$, let $n > \max\left\{\frac{1}{\epsilon}, \left(\frac{1}{1-\gamma}\right)^p, \delta^p, \delta^{p-1}\right\}$, $n \in \mathbb{N}$ and

$$Q(s,a_2) = \begin{cases} Q^*(s,a_2) - n^{\frac{1}{p}}, s \in [\frac{k}{n}, \frac{k}{n} + \frac{\delta^p}{n^2}], k = 0, 1, \ldots, n-1 \\ Q^*(s,a_2), others \end{cases} \tag{183}$$

$$Q(s,a_1) = \begin{cases} Q^*(s,a_1) - n^{\frac{1}{p}}, s \in [-\frac{k+1}{n}, -\frac{k+1}{n} + \frac{\delta^p}{n^2}], k = 0, 1, \ldots, n-1 \\ Q^*(s,a_1), others \end{cases} \tag{184}$$

Then

$$\|Q(s,a_1) - Q^*(s,a_1)\|_{L^p(\mathcal{S})} = \|Q(s,a_2) - Q^*(s,a_2)\|_{L^p(\mathcal{S})} = \left[n * \frac{\delta^p}{n^2} * \left(n^{\frac{1}{p}}\right)^p\right]^{\frac{1}{p}} \le \delta. \tag{185}$$

And

$$\begin{aligned} \|Q(s,a)\|_{L^p(\mathcal{S})} &= \|Q(s,a) - Q^*(s,a) + Q^*(s,a)\|_{L^p(\mathcal{S})} && (186) \\ &\le \|Q(s,a) - Q^*(s,a)\|_{L^p(\mathcal{S})} + \|Q^*(s,a)\|_{L^p(\mathcal{S})} && (187) \\ &< \infty. && (188) \end{aligned}$$

which means $Q \in L^p\left(\mathcal{S} \times \mathcal{A}\right)$.

We have the following two inequalities because $n > \left(\frac{1}{1-\gamma}\right)^p$ and (182),

$$Q^*(s,a_2) - n^{\frac{1}{p}} < Q^*(s,a_1), \tag{189}$$

$$Q^*(s,a_1) - n^{\frac{1}{p}} < Q^*(s,a_2). \tag{190}$$

Then

$$\mathcal{S}_{sub}^Q = \bigcup_{k=-n}^{n-1}\left[\frac{k}{n}, \frac{k}{n} + \frac{\delta^p}{n^2}\right] \tag{191}$$

and

$$\mu\left(\mathcal{S}_{sub}^{Q}\right)=2n*\frac{\delta^{p}}{n^{2}}<2\delta=O\left(\delta\right) \tag{192}$$

because $n>\delta^{p-1}$.

According to (191), the distance between any two adjacent intervals of $\mathcal{S}_{sub}^{Q}$ is less than ϵ. For any $s\in\mathcal{S}$, $\exists k\in\{-n,-n+1,...,n-1\}$ s.t. $s\in[\frac{k}{n},\frac{k+1}{n}]$. Because $n>\frac{1}{\epsilon}$ (i.e. $\frac{1}{n}<\epsilon$), then $d(s,\frac{k}{n})<\epsilon$ i.e. $d(s,\mathcal{S}_{sub}^{Q})<\epsilon$, where $d(\cdot,\cdot)$ is Euclid distance. According to the definition of $\mathcal{S}_{adv}^{Q}$, we have $\mathcal{S}_{adv}^{Q}=\mathcal{S}$ and

$$\mu\left(\mathcal{S}_{adv}^{Q}\right)=\mu\left(\mathcal{S}\right). \tag{193}$$

(2)Let $\bar{\delta}\in(0,k]$, for any $0<\delta\leq\bar{\delta}$, for any state-action value function $Q\in L^{\infty}\left(\mathcal{S}\times\mathcal{A}\right)$ satisfying $\|Q-Q^{*}\|_{L^{\infty}(\mathcal{S}\times\mathcal{A})}\leq\delta$, we can get the following two inequalities by (180) and (181).

$$Q(s,a_{2})\geq Q^{*}(s,a_{2})-\delta>Q^{*}(s,a_{1})+\delta\geq Q(s,a_{1}),s\in(\frac{\delta}{k},1], \tag{194}$$

$$Q(s,a_{1})\geq Q^{*}(s,a_{1})-\delta>Q^{*}(s,a_{2})+\delta\geq Q(s,a_{2}),s\in[-1,-\frac{\delta}{k}). \tag{195}$$

Then

$$\mu\left(\mathcal{S}_{sub}^{Q}\right)\leq\frac{2\delta}{k}=O\left(\delta\right), \tag{196}$$

$$\mu\left(\mathcal{S}_{adv}^{Q}\right)\leq\frac{2\delta}{k}+2\epsilon=2\epsilon+O\left(\delta\right). \tag{197}$$

□

B.2. Stability of Bellman Optimality Equations

We propose the following concept of stability drawing on relevant research in the field of partial differential equations (Wang et al., 2022).

Definition B.2. Given two Banach spaces $\mathcal{B}_1$ and $\mathcal{B}_2$, if there exist $\delta>0$ and $C>0$ such that for all $Q\in\mathcal{B}_1\cap\mathcal{B}_2$ satisfying $\|\mathcal{T}Q-Q\|_{\mathcal{B}_1}<\delta$, we have that $\|Q-Q^{*}\|_{\mathcal{B}_2}<C\|\mathcal{T}Q-Q\|_{\mathcal{B}_1}$, where Q^{*} is the exact solution of this functional equation. Then, we say that a nonlinear functional equation $\mathcal{T}Q=Q$ is $(\mathcal{B}_1,\mathcal{B}_2)$-stable.

Remark B.3. This definition indicates that if $\mathcal{T}Q=Q$ is $(\mathcal{B}_1,\mathcal{B}_2)$-stable, then $\|Q-Q^{*}\|_{\mathcal{B}_2}=O\left(\|\mathcal{T}Q-Q\|_{\mathcal{B}_1}\right)$, as $\|\mathcal{T}Q-Q\|_{\mathcal{B}_1}\longrightarrow 0,\forall Q\in\mathcal{B}_1\cap\mathcal{B}_2$.

Lemma B.4. *For any functions $f,g:\mathcal{X}\to\mathbb{R}$, we have*

$$\max_{x\in\mathcal{X}}f(x)-\max_{x\in\mathcal{X}}g(x)\leq\max_{x\in\mathcal{X}}\left(f(x)-g(x)\right). \tag{198}$$

Proof.

$$\max_{x\in\mathcal{X}}f(x)-\max_{x\in\mathcal{X}}g(x)=f(x_{f}^{*})-\max_{x\in\mathcal{X}}g(x)\leq f(x_{f}^{*})-g(x_{f}^{*})\leq\max_{x\in\mathcal{X}}\left(f(x)-g(x)\right), \tag{199}$$

where x_{f}^{*} is the maximizer of function f, i.e. $x_{f}^{*}=\arg\max_{x\in\mathcal{X}}f(x)$. □

Lemma B.5. *For any functions $f,g:\mathcal{X}\to\mathbb{R}$, we have*

$$\left|\max_{x\in\mathcal{X}}\left(f+g\right)(x)-\max_{x\in\mathcal{X}}f(x)\right|\leq\max_{x\in\mathcal{X}}\left|g(x)\right|. \tag{200}$$

Proof. If $\max_{x\in\mathcal{X}}(f+g)(x) \geq \max_{x\in\mathcal{X}} f(x)$, we have

$$\begin{aligned}
&\max_{x\in\mathcal{X}}(f+g)(x) - \max_{x\in\mathcal{X}} f(x) && (201)\\
\leq &\max_{x\in\mathcal{X}} f(x) + \max_{x\in\mathcal{X}} g(x) - \max_{x\in\mathcal{X}} f(x) && (202)\\
= &\max_{x\in\mathcal{X}} g(x) && (203)\\
\leq &\max_{x\in\mathcal{X}} |g(x)|. && (204)
\end{aligned}$$

If $\max_{x\in\mathcal{X}}(f+g)(x) < \max_{x\in\mathcal{X}} f(x)$, we have

$$\max_{x\in\mathcal{X}} f(x) - \max_{x\in\mathcal{X}}(f+g)(x) \leq \max_{x\in\mathcal{X}}(-g(x)) \leq \max_{x\in\mathcal{X}}|g(x)|, \tag{205}$$

where the first inequality comes from Lemma B.4. □

Theorem B.6. *For any MDP $\mathcal{M}$, let $C_{\mathbb{P},p} := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\mathbb{P}(\cdot \mid s,a)\|_{L^{\frac{p}{p-1}}(\mathcal{S})}$. Assume p and q satisfy the following conditions:*

$$C_{\mathbb{P},p} < \frac{1}{\gamma}; \quad p \geq \max\left\{1, \frac{\log(|\mathcal{A}|) + \log(\mu(\mathcal{S}))}{\log\frac{1}{\gamma C_{\mathbb{P},p}}}\right\}; \quad p \leq q \leq \infty. \tag{206}$$

Then, Bellman optimality equation $\mathcal{T}_B Q = Q$ is $(L^q(\mathcal{S}\times\mathcal{A}), L^p(\mathcal{S}\times\mathcal{A}))$-stable.

Proof. For any $1 \leq p \leq q \leq \infty$ and $Q \in L^p(\mathcal{S}\times\mathcal{A}) \cap L^q(\mathcal{S}\times\mathcal{A})$, denote that

$$\begin{aligned}
\mathcal{L}_0 Q(s,a) &:= \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} Q(s',a')\right], && (207)\\
\mathcal{L}Q &:= \mathcal{T}_B Q - Q = r + \mathcal{L}_0 Q - Q. && (208)
\end{aligned}$$

Let Q^* denote the Bellman optimality Q-function. Note that $\mathcal{T}_B Q^* = Q^*$ and $\mathcal{L}Q^* = 0$. Define

$$\begin{aligned}
w = w_Q &:= Q - Q^*, && (209)\\
f = f_Q &:= \mathcal{L}Q = \mathcal{L}Q - \mathcal{L}Q^*. && (210)
\end{aligned}$$

Based on the above notations, we have

$$\begin{aligned}
f &= \mathcal{L}Q - \mathcal{L}Q^* && (211)\\
&= \mathcal{L}_0 Q - Q - \mathcal{L}_0 Q^* + Q^* && (212)\\
&= (\mathcal{L}_0 Q - \mathcal{L}_0 Q^*) - (Q - Q^*) && (213)\\
&= -w + \mathcal{L}_0(Q^* + w) - \mathcal{L}_0 Q^*. && (214)
\end{aligned}$$

Then, we have

$$\begin{aligned}
|w(s,a)| &= |-f + \mathcal{L}_0(Q^*+w) - \mathcal{L}_0 Q^*|\Big|_{(s,a)} && (215)\\
&\leq |f| + |\mathcal{L}_0(Q^*+w) - \mathcal{L}_0 Q^*|\Big|_{(s,a)}. && (216)
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\|w\|_{L^p(\mathcal{S}\times\mathcal{A})} &\leq \||f| + |\mathcal{L}_0(Q^*+w) - \mathcal{L}_0 Q^*|\|_{L^p(\mathcal{S}\times\mathcal{A})} && (217)\\
&\leq \|f\|_{L^p(\mathcal{S}\times\mathcal{A})} + \|\mathcal{L}_0(Q^*+w) - \mathcal{L}_0 Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})}, && (218)
\end{aligned}$$

where the last inequality comes from the Minkowski's inequality. In the following, we analyze the relation between $\|\mathcal{L}_0(Q^*+w)-\mathcal{L}_0Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})}$ and $\|w\|_{L^p(\mathcal{S}\times\mathcal{A})}$.

$$
\begin{aligned}
&\left|\mathcal{L}_0(Q^*+w)-\mathcal{L}_0Q^*\right|\Big|_{(s,a)} && (219)\\
&=\left|\gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}}(Q^*(s',a')+w(s',a'))-\max_{a'\in\mathcal{A}}Q^*(s',a')\right]\right| && (220)\\
&\leq\gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left|\max_{a'\in\mathcal{A}}(Q^*(s',a')+w(s',a'))-\max_{a'\in\mathcal{A}}Q^*(s',a')\right| && (221)\\
&\leq\gamma\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}}|w(s',a')|\right] && (222)\\
&=\gamma\int_{s'}\max_{a'\in\mathcal{A}}|w(s',a')|\,\mathbb{P}(s'\mid s,a)ds' && (223)\\
&\leq\gamma\left\|\max_{a\in\mathcal{A}}|w(s,a)|\right\|_{L^p(\mathcal{S})}\left(\int_{s'}\left(\mathbb{P}(s'\mid s,a)\right)^{\frac{p}{p-1}}ds'\right)^{1-\frac{1}{p}} && (224)\\
&=\gamma\left\|\mathbb{P}(\cdot\mid s,a)\right\|_{L^{\frac{p}{p-1}}(\mathcal{S})}\left\|\max_{a\in\mathcal{A}}|w(s,a)|\right\|_{L^p(\mathcal{S})}. && (225)
\end{aligned}
$$

where the second inequality comes from Lemma B.5 and the last inequality comes from the Holder's inequality. Let $C_{\mathbb{P},p}:=\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left\|\mathbb{P}(\cdot\mid s,a)\right\|_{L^{\frac{p}{p-1}}(\mathcal{S})}$. Then, we have

$$
\begin{aligned}
&\|\mathcal{L}_0(Q^*+w)-\mathcal{L}_0Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})} && (226)\\
&\leq\left(\int_{\mathcal{S}\times\mathcal{A}}\left(\gamma\left\|\mathbb{P}(\cdot\mid s,a)\right\|_{L^{\frac{p}{p-1}}(\mathcal{S})}\left\|\max_{a\in\mathcal{A}}|w(s,a)|\right\|_{L^p(\mathcal{S})}\right)^p d\mu(s,a)\right)^{\frac{1}{p}} && (227)\\
&\leq\left(\int_{\mathcal{S}\times\mathcal{A}}1d\mu(s,a)\right)^{\frac{1}{p}}\gamma C_{\mathbb{P},p}\left\|\max_{a\in\mathcal{A}}|w(s,a)|\right\|_{L^p(\mathcal{S})} && (228)\\
&=\left(\mu(\mathcal{S}\times\mathcal{A})\right)^{\frac{1}{p}}\gamma C_{\mathbb{P},p}\left\|\max_{a\in\mathcal{A}}|w(s,a)|\right\|_{L^p(\mathcal{S})} && (229)\\
&=\gamma C_{\mathbb{P},p}\left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{1}{p}}\left\|\max_{a\in\mathcal{A}}|w(s,a)|\right\|_{L^p(\mathcal{S})} && (230)\\
&\leq\gamma C_{\mathbb{P},p}\left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{1}{p}}\|w\|_{L^p(\mathcal{S}\times\mathcal{A})}, && (231)
\end{aligned}
$$

where the last inequality comes from $\|w\|_{l^\infty(\mathcal{A})}\leq\|w\|_{l^p(\mathcal{A})}$. Thus, when $C_{\mathbb{P},p}<\frac{1}{\gamma}$ and $p\geq\frac{\log(|\mathcal{A}|)+\log(\mu(\mathcal{S}))}{\log\frac{1}{\gamma C_{\mathbb{P},p}}}$ and $q\geq p$, we have

$$
\|w\|_{L^p(\mathcal{S}\times\mathcal{A})}\leq\frac{1}{1-\gamma C_{\mathbb{P},p}\left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{1}{p}}}\|f\|_{L^p(\mathcal{S}\times\mathcal{A})}\leq\frac{\left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{1}{p}-\frac{1}{q}}}{1-\gamma C_{\mathbb{P},p}\left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{1}{p}}}\|f\|_{L^q(\mathcal{S}\times\mathcal{A})}, \tag{232}
$$

where the last inequality comes from $\|f\|_{L^p(\mathcal{S}\times\mathcal{A})}\leq\mu(\mathcal{S}\times\mathcal{A})^{\frac{1}{p}-\frac{1}{q}}\|f\|_{L^q(\mathcal{S}\times\mathcal{A})}$. □

Remark B.7. Note that we have proved a stronger conclusion than stability because the equation (232) holds for all Q rather than for Q satisfying $\|\mathcal{T}Q-Q\|_{\mathcal{B}_1}\longrightarrow 0$.

Remark B.8. When $\mathbb{P}(\cdot\mid s,a)$ is a probability mass function, then we have that $C_{\mathbb{P},p}\leq 1<\frac{1}{\gamma}$ holds for all $1<p\leq\infty$. Generally, note that $\lim_{p\to\infty}C_{\mathbb{P},p}=1$ and as a consequence, when p is large enough, $C_{\mathbb{P},p}<\frac{1}{\gamma}$ holds.

B.3. Instability of Bellman Optimality Equations

Theorem B.9. *There exists a MDP $\mathcal{M}$ such that Bellman optimality equation $\mathcal{T}_BQ=Q$ is not $(L^p(\mathcal{S}\times\mathcal{A}),L^\infty(\mathcal{S}\times\mathcal{A}))$-stable, for $1\leq p<\infty$.*

Generally, we have the following theorem.

Theorem B.10. *There exists an MDP $\mathcal{M}$ such that Bellman optimality equation $\mathcal{T}_B Q = Q$ is not $(L^q(\mathcal{S}\times\mathcal{A}), L^p(\mathcal{S}\times\mathcal{A}))$-stable, for all $1 \le q < p \le \infty$.*

Proof. In order to show a Bellman optimality equation $\mathcal{T}_B Q = Q$ is not $(L^q(\mathcal{S}\times\mathcal{A}), L^p(\mathcal{S}\times\mathcal{A}))$-stable, it is sufficient and necessary to prove $\forall n \in \mathbb{N}, \forall \delta > 0, \exists Q(s,a),$ such that $\|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})} < \delta,$ but $\|Q - Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})} \ge n\|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})}$.

Define an MDP $\mathcal{M}$ where $\mathcal{S} = [-1,1], \mathcal{A} = \{a_1, a_2\}$,

$$\mathbb{P}(s'|s,a_1) = \begin{cases} \mathbb{1}_{\{s'=s-0.1\}}, & s \in [-0.9, 1] \\ \mathbb{1}_{\{s'=s\}}, & else \end{cases}, \quad \mathbb{P}(s'|s,a_2) = \begin{cases} \mathbb{1}_{\{s'=s+0.1\}}, & s \in [-1, 0.9] \\ \mathbb{1}_{\{s'=s\}}, & else \end{cases},$$

$r(s,a_i) = k_i s,\ k_2 \ge k_1 > 0$. The transition function is essentially a deterministic transition dynamic and for convenience, we denote that

$$p(s,a_1) = \begin{cases} s - 0.1, & s \in [-0.9, 1] \\ s, & else \end{cases}, \quad p(s,a_2) = \begin{cases} s + 0.1, & s \in [-1, 0.9] \\ s, & else \end{cases}.$$

Let $Q^*(s,a) = Q^{\pi^*}(s,a)$ be the optimal Q-function, where π^* is the optimal policy.

We have $Q^*(s,a_2) \ge Q^*(s,a_1),\ \forall s \ge 0$. To prove this, we define $\bar{\pi}(s) \equiv a_2, \forall s \ge 0$, and thus

$$Q^{\bar{\pi}}(s,a_2) = \sum_{i=0}^{\infty} \gamma^i r(s_i, a_2), \tag{233}$$

where $s_0 = s \ge 0,\ s_i = p(s_{i-1}, a_2) \ge 0,\ i \ge 1$. Consider Q-function of any policy π

$$Q^{\pi}(s,\alpha_0) = \sum_{i=0}^{\infty} \gamma^i r(\tilde{s}_i, \pi(\tilde{s}_i)), \tag{234}$$

where $\pi(\tilde{s}_0) = \alpha_0 \in \mathcal{A},\ \tilde{s}_0 = s \ge 0,\ \tilde{s}_{i+1} = p(\tilde{s}_i, \pi(\tilde{s}_i))$ and $r(\tilde{s}_i, \pi(\tilde{s}_i)) = \pi(a_1|\tilde{s}_i) r(\tilde{s}_i, a_1) + \pi(a_2|\tilde{s}_i) r(\tilde{s}_i, a_2)$.

We first notice that all s_i and $\tilde{s}_i$ lie on the grid points $\mathcal{S} \cap \{s + 0.1z :\ z \in \mathbb{Z}\}$, actually, $-\lfloor \frac{1+s}{0.1} \rfloor \le z \le \lfloor \frac{1-s}{0.1} \rfloor$. In the following, we prove $s_i \ge \tilde{s}_i,\ \forall i$. We consider the recursion method and suppose $s_i \ge \tilde{s}_i$. Then, we have the following two cases. If $s_i \le 0.9$, we obtain

$$s_{i+1} = s_i + 0.1 \ge \tilde{s}_i + 0.1 \ge \tilde{s}_{i+1}.$$

If $s_i > 0.9$, it follows from $z \le \lfloor \frac{1-s}{0.1} \rfloor$ that

$$s_{i+1} = s_i = s + 0.1 \lfloor \frac{1-s}{0.1} \rfloor \ge \tilde{s}_{i+1}.$$

Thus, we have $s_{i+1} \ge \tilde{s}_{i+1},\ \forall i$. Note that $s_0 = \tilde{s}_0 = s$, and by recursion, it can be obtained that $s_i \ge \tilde{s}_i$ holds for all i.

Noticing that the reward $r(s,a)$ is an increasing function in terms of s and satisfies $r(s,a_2) \ge r(s,a_1), \forall s \ge 0$, we have

$$r(s_i, a_2) \ge r(s_i, \alpha_i) \ge r(\tilde{s}_i, \alpha_i), \quad \forall \alpha_i \in \mathcal{A}, i = 0, 1, 2, \ldots,$$

where the second inequality is due to $s_i \ge \tilde{s}_i$. As a consequence,

$$r(s_i, a_2) \ge r(\tilde{s}_i, \pi(\tilde{s}_i)), \quad \forall i = 0, 1, 2, \ldots. \tag{235}$$

Combining (233), (234), and (235), we obtain that $Q^{\bar{\pi}}(s,a_2) \ge Q^{\pi}(s,\alpha_0)$. Further, with $\alpha_0 = a_2$, we derive $\bar{\pi}(s) = \pi^*(s)$ on $s > 0$. With $\alpha_0 = a_1$, we derive $Q^*(s,a_2) = Q^{\bar{\pi}}(s,a_2) \ge Q^*(s,a_1),\ \forall s \ge 0$.

We then prove that given $1 \leq q < p$, $\forall n \in \mathbb{N}, \delta > 0$, there exists $Q(s,a)$ with $\|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})} \leq \delta$, such that $\|Q - Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})} \geq n\|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})}$. Let $Q(s,a_1) = Q^*(s,a_1)$,

$$Q(s,a_2) = Q^*(s,a_2) + h \cdot \mathbb{1}_{(\frac{1}{4}\epsilon, \frac{3}{4}\epsilon)} + \frac{4h}{\epsilon} s \cdot \mathbb{1}_{(0,\frac{1}{4}\epsilon]} + \left(-\frac{4h}{\epsilon} s + 4h\right) \cdot \mathbb{1}_{[\frac{3}{4}\epsilon,\epsilon)},$$

where $h > 0$, $\epsilon = \min\left\{ \left(\frac{\delta}{3h}\right)^q, \left(3n \cdot 2^{\frac{1}{p}}\right)^{-\frac{pq}{p-q}} \right\}$ and $\mathbb{1}_A(s) = \begin{cases} 1, & s \in A \\ 0, & else \end{cases}$ denotes the indicator function. It can be seen from the definition that

$$Q^*(s,a_2) \leq Q^*(s,a_2) + h \cdot \mathbb{1}_{(\frac{1}{4}\epsilon, \frac{3}{4}\epsilon)} \leq Q(s,a_2) \leq Q^*(s,a_2) + h \cdot \mathbb{1}_{(0,\epsilon)}. \tag{236}$$

We consider the following cases.

- When $s \in (-0.1, -0.1+\epsilon), a = a_2$,

  $$\mathcal{T}_B Q(s,a_2) = r(s,a_2) + \gamma \max_{a_i} Q(s+0.1, a_i) = r(s,a_2) + \gamma Q(s+0.1, a_2),$$

  Together with (236), we have

  $$\begin{aligned} &Q(s,a_2) = Q^*(s,a_2) = r(s,a_2) + \gamma Q^*(s+0.1,a_2) \\ \leq &\mathcal{T}_B Q(s,a_2) \leq r(s,a_2) + \gamma[Q^*(s+0.1,a_2) + h] \\ = &Q^*(s,a_2) + h\gamma = Q(s,a_2) + h\gamma, \end{aligned}$$

  thus $|\mathcal{T}_B Q(s,a_2) - Q(s,a_2)| \leq h\gamma$.

- When $s \in (0, \epsilon), a = a_2$,

  $$\begin{aligned} \mathcal{T}_B Q(s,a_2) &= r(s,a_2) + \gamma \max_{a_i} Q(s+0.1, a_i) = r(s,a_2) + \gamma Q(s+0.1, a_2) \\ &= r(s,a_2) + \gamma Q^*(s+0.1, a_2) = Q^*(s,a_2), \end{aligned}$$

  Again from (236), there is

  $$|\mathcal{T}_B Q(s,a_2) - Q(s,a_2)| = |Q^*(s,a_2) - Q(s,a_2)| \leq h.$$

- When $s \in (0.1, 0.1+\epsilon), a = a_1$,

  $$\mathcal{T}_B Q(s,a_1) = r(s,a_1) + \gamma \max_{a_i} Q(s-0.1, a_i) = r(s,a_1) + \gamma Q(s-0.1, a_2),$$

  Utilizing (236), we have

  $$\begin{aligned} &Q(s,a_1) = Q^*(s,a_1) = r(s,a_1) + \gamma Q^*(s-0.1,a_2) \\ \leq &\mathcal{T}_B Q(s,a_1) \leq r(s,a_1) + \gamma[Q^*(s-0.1,a_2) + h] \\ = &Q^*(s,a_1) + h\gamma = Q(s,a_1) + h\gamma, \end{aligned}$$

  thus $|\mathcal{T}_B Q(s,a_1) - Q(s,a_1)| \leq h\gamma$.

- Otherwise,

  $$\mathcal{T}_B Q(s,a_i) = r(s,a_i) + \gamma Q\left(p(s,a_i), \pi^*(p(s,a_i))\right) = Q^*(s,a_i),$$

  also note that $Q(s,\cdot) = Q^*(s,\cdot)$ for $s \notin (0,\epsilon)$, thus

  $$|\mathcal{T}_B Q(s,a) - Q(s,a)| = |Q^*(s,a) - Q(s,a)| = 0.$$

From the analysis above, we have

$$\|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})} \leq (2h\gamma + h)\epsilon^{\frac{1}{q}} \leq 3h\epsilon^{\frac{1}{q}} \leq \delta, \tag{237}$$

and

$$\|Q - Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})} \geq \|(Q-Q^*)\mathbb{1}_{(\frac{1}{4}\epsilon,\frac{3}{4}\epsilon)}\|_{L^p(\mathcal{S}\times\mathcal{A})} \geq h\left(\frac{\epsilon}{2}\right)^{\frac{1}{p}} \geq n\|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})}. \tag{238}$$

Inequality (237) and (238) come from $\epsilon = \min\left\{ \left(\frac{\delta}{3h}\right)^q, \left(3n \cdot 2^{\frac{1}{p}}\right)^{-\frac{pq}{p-q}} \right\}$, which prove the desired property. □

C. Theorems and Proofs of Stability Analysis of DQN

In practical DQN training, we use the following loss:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s') \in \mathcal{B}} \left| r + \gamma \max_{a'} Q(s', a'; \bar{\theta}) - Q(s, a; \theta) \right|, \tag{239}$$

where $\mathcal{B}$ represents a batch of transition pairs sampled from the replay buffer and $\bar{\theta}$ is the parameter of target network.

$\mathcal{L}(\theta)$ is a approximation of the following objective:

$$\mathcal{L}(Q; \pi) = \mathbb{E}_{s \sim d^{\pi}_{\mu_0}(\cdot)} \mathbb{E}_{a \sim \pi(\cdot|s)} \left| \mathcal{T}_B Q(s, a) - Q(s, a) \right| \tag{240}$$

$$= \mathbb{E}_{(s,a) \sim d^{\pi}_{\mu_0}(\cdot,\cdot)} \left| \mathcal{T}_B Q(s, a) - Q(s, a) \right|, \tag{241}$$

where $d^{\pi}_{\mu_0}(s) = \mathbb{E}_{s_0 \sim \mu_0} \left[(1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr^{\pi}(s_t = s | s_0) \right]$ is the state visitation distribution and $d^{\pi}_{\mu_0}(s, a) = \mathbb{E}_{s_0 \sim \mu_0} \left[(1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr^{\pi}(s_t = s, a_t = a | s_0) \right]$ is the state-action visitation distribution.

C.1. Definition and Properties of $L^{p, d^{\pi}_{\mu_0}}$

Definition C.1. Given a policy π, for any function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $1 \leq p \leq \infty$, we define the seminorm $L^{p, d^{\pi}_{\mu_0}}$.

- If $d^{\pi}_{\mu_0}$ is a probability density function, we define

$$\begin{aligned} \|f\|_{L^{p, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A})} &:= \left\| d^{\pi}_{\mu_0} f \right\|_{L^p(\mathcal{S} \times \mathcal{A})} \\ &= \left(\int_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| d^{\pi}_{\mu_0}(s, a) f(s, a) \right|^p d\mu(s, a) \right)^{\frac{1}{p}}. \end{aligned} \tag{242}$$

- If $d^{\pi}_{\mu_0}$ is a probability mass function, we define

$$\|f\|_{L^{p, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A})} := \left(\sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left| d^{\pi}_{\mu_0}(s, a) f(s, a) \right|^p \right)^{\frac{1}{p}}. \tag{243}$$

Remark C.2. Note that $\mathcal{L}(Q; \pi) = \|\mathcal{T}_B Q - Q\|_{L^{1, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A})}$.

Theorem C.3. *For any $d^{\pi}_{\mu_0}(s, a)$ and $1 \leq p \leq \infty$, $L^{p, d^{\pi}_{\mu_0}}$ is a seminorm.*

Proof. Firstly, we show that $L^{p, d^{\pi}_{\mu_0}}$ satisfies the absolute homogeneity. For any function f and $\lambda \in \mathbb{R}$, we have

$$\|\lambda f\|_{L^{p, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A})} = \left\| d^{\pi}_{\mu_0} \lambda f \right\|_{L^p(\mathcal{S} \times \mathcal{A})} = |\lambda| \left\| d^{\pi}_{\mu_0} f \right\|_{L^p(\mathcal{S} \times \mathcal{A})} = |\lambda| \|f\|_{L^{p, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A})}. \tag{244}$$

Next, we show that the triangle inequality holds. For any functions f and g, we have

$$\|f + g\|_{L^{p, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A})} = \left\| d^{\pi}_{\mu_0}(f + g) \right\|_{L^p(\mathcal{S} \times \mathcal{A})} \tag{245}$$

$$\leq \left\| d^{\pi}_{\mu_0} f \right\|_{L^p(\mathcal{S} \times \mathcal{A})} + \left\| d^{\pi}_{\mu_0} g \right\|_{L^p(\mathcal{S} \times \mathcal{A})} \tag{246}$$

$$= \|f\|_{L^{p, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A})} + \|g\|_{L^{p, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A})}, \tag{247}$$

where the inequality comes from the triangle inequality of $L^p(\mathcal{S} \times \mathcal{A})$. □

Theorem C.4. *If $d^{\pi}_{\mu_0}(s, a) > 0$ for almost everywhere $(s, a) \in \mathcal{S} \times \mathcal{A}$, then $L^{p, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A}) := \left\{ f \,\middle|\, \|f\|_{L^{p, d^{\pi}_{\mu_0}}(\mathcal{S} \times \mathcal{A})} \leq \infty \right\}$ is a Banach space, for $1 \leq p \leq \infty$.*

Proof. Firstly, we show the $L^{p,d^\pi_{\mu_0}}$ is positive definite. If $\|f\|_{L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}=0$, we have that $d^\pi_{\mu_0}(s,a)f(s,a)=0$, for almost everywhere $(s,a)\in\mathcal{S}\times\mathcal{A}$. Due to the nonnegativity of $d^\pi_{\mu_0}(s,a)$, we have $f(s,a)=0$, for almost everywhere $(s,a)\in\mathcal{S}\times\mathcal{A}$.

We show the completeness of $L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})$ in the following. For any Cauchy sequence $\{f_i\}\subset L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})$, $\left\{d^\pi_{\mu_0}f_i\right\}$ is a Cauchy sequence in $L^p(\mathcal{S}\times\mathcal{A})$. Then, due to the completeness of $L^p(\mathcal{S}\times\mathcal{A})$, there exists $g\in L^p(\mathcal{S}\times\mathcal{A})$ such that

$$\lim_{i\to\infty}\left\|d^\pi_{\mu_0}f_i\right\|_{L^p(\mathcal{S}\times\mathcal{A})}=\|g\|_{L^p(\mathcal{S}\times\mathcal{A})}.$$

Let $f=\frac{g}{d^\pi_{\mu_0}}$ for almost everywhere $(s,a)\in\mathcal{S}\times\mathcal{A}$, we have $\|f\|_{L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}=\|g\|_{L^p(\mathcal{S}\times\mathcal{A})}\le\infty$ and

$$\lim_{i\to\infty}\|f_i\|_{L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}=\|f\|_{L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}.$$

Thus, $L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})$ is a Banach space. □

We analyze the properties of $L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})$ in the following lemma.

Lemma C.5. *Given a policy π, for any function $f:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$, then we have the following properties.*

- *If $M_{d^\pi_{\mu_0}}:=\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}d^\pi_{\mu_0}(s,a)<\infty$, then*

$$\|f\|_{L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}\le M_{d^\pi_{\mu_0}}\|f\|_{L^p(\mathcal{S}\times\mathcal{A})},\ \forall 1\le p\le\infty.$$

- *If $C_{d^\pi_{\mu_0}}:=\inf_{(s,a)\in\mathcal{S}\times\mathcal{A}}d^\pi_{\mu_0}(s,a)>0$, then we have*

$$C_{d^\pi_{\mu_0}}\|f\|_{L^p(\mathcal{S}\times\mathcal{A})}\le\|f\|_{L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})},\ \forall 1\le p\le\infty.$$

- $\|f\|_{L^{1,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}\le\left\|d^\pi_{\mu_0}\right\|_{L^{\frac{p}{p-1}}(\mathcal{S}\times\mathcal{A})}\|f\|_{L^p(\mathcal{S}\times\mathcal{A})},\ \forall 1\le p\le\infty.$

- *If $d^\pi_{\mu_0}(s,a)\ne 0$ for almost everywhere $(s,a)\in\mathcal{S}\times\mathcal{A}$, then we have*

$$\|f\|_{L^1(\mathcal{S}\times\mathcal{A})}\le C_{d^\pi_{\mu_0},p}\|f\|_{L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})},\ \forall 1<p<\infty,$$

  where $C_{d^\pi_{\mu_0},p}=\left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|d^\pi_{\mu_0}(s,a)\right|^{-\frac{p}{p-1}}d\mu(s,a)\right)^{\frac{p-1}{p}}$.

- *If $d^\pi_{\mu_0}(s,a)\ne 0$ for almost everywhere $(s,a)\in\mathcal{S}\times\mathcal{A}$, then we have*

$$\|f\|_{L^p(\mathcal{S}\times\mathcal{A})}\le C_{d^\pi_{\mu_0},p}\|f\|_{L^{p^2,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})},\ \forall 1<p<\infty,$$

  where $C_{d^\pi_{\mu_0},p}=\left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|d^\pi_{\mu_0}(s,a)\right|^{-\frac{p^2}{p-1}}d\mu(s,a)\right)^{\frac{p-1}{p^2}}$.

Proof. (1) If $M_{d^\pi_{\mu_0}}:=\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}d^\pi_{\mu_0}(s,a)<\infty$, we have

$$\|f\|_{L^{p,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}=\left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|d^\pi_{\mu_0}(s,a)f(s,a)\right|^p d\mu(s,a)\right)^{\frac{1}{p}} \tag{248}$$

$$\le M_{d^\pi_{\mu_0}}\left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}}|f(s,a)|^p d\mu(s,a)\right)^{\frac{1}{p}} \tag{249}$$

$$=M_{d^\pi_{\mu_0}}\|f\|_{L^p(\mathcal{S}\times\mathcal{A})}. \tag{250}$$

(2) If $C_{d_{\mu_0}^\pi} := \inf_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^\pi(s,a) > 0$, we have

$$\|f\|_{L^{p,d_{\mu_0}^\pi}(\mathcal{S}\times\mathcal{A})} = \left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|d_{\mu_0}^\pi(s,a)f(s,a)\right|^p d\mu(s,a)\right)^{\frac{1}{p}} \tag{251}$$

$$\geq C_{d_{\mu_0}^\pi} \|f\|_{L^p(\mathcal{S}\times\mathcal{A})}. \tag{252}$$

(3)

$$\|f\|_{L^{1,d_{\mu_0}^\pi}(\mathcal{S}\times\mathcal{A})} = \int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|d_{\mu_0}^\pi(s,a)f(s,a)\right| d\mu(s,a) \tag{253}$$

$$\leq \|f\|_{L^p(\mathcal{S}\times\mathcal{A})} \left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|d_{\mu_0}^\pi(s,a)\right|^{\frac{p}{p-1}} d\mu(s,a)\right)^{1-\frac{1}{p}} \tag{254}$$

$$= \left\|d_{\mu_0}^\pi\right\|_{L^{\frac{p}{p-1}}(\mathcal{S}\times\mathcal{A})} \|f\|_{L^p(\mathcal{S}\times\mathcal{A})}, \tag{255}$$

where the first inequality comes from the Holder's inequality.

(4) For $1 < p < \infty$, we have

$$\|f\|^p_{L^{p,d_{\mu_0}^\pi}(\mathcal{S}\times\mathcal{A})} = \int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|d_{\mu_0}^\pi(s,a)f(s,a)\right|^p d\mu(s,a) \tag{256}$$

$$\geq \|f\|^p_{L^1(\mathcal{S}\times\mathcal{A})} \left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|d_{\mu_0}^\pi(s,a)\right|^{-\frac{p}{p-1}} d\mu(s,a)\right)^{-(p-1)}, \tag{257}$$

where the inequality comes from reverse Holder's inequality.

(5) Further, we have

$$\|f\|^{p^2}_{L^{p^2,d_{\mu_0}^\pi}(\mathcal{S}\times\mathcal{A})} = \int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|d_{\mu_0}^\pi(s,a)f(s,a)\right|^{p^2} d\mu(s,a) \tag{258}$$

$$\geq \|f\|^{p^2}_{L^p(\mathcal{S}\times\mathcal{A})} \left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|d_{\mu_0}^\pi(s,a)\right|^{-\frac{p^2}{p-1}} d\mu(s,a)\right)^{-(p-1)}, \tag{259}$$

where the inequality comes from reverse Holder's inequality. □

Remark C.6. Note that in a practical Q-learning scheme, we take the ϵ-greedy policy for exploration and as a result, for any state-action pair (s,a), we can visit it with positive probability, i.e. $d_{\mu_0}^\pi(s,a) > 0$. Furthermore, the condition, $d_{\mu_0}^\pi(s,a) \neq 0$ for almost everywhere $(s,a) \in \mathcal{S}\times\mathcal{A}$, always holds.

C.2. Stability of DQN: the Good

Theorem C.7. *For any MDP $\mathcal{M}$ and fixed policy π, assume $C_{d_{\mu_0}^\pi} := \inf_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^\pi(s,a) > 0$ and let $C_{\mathbb{P},p} := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\mathbb{P}(\cdot \mid s,a)\|_{L^{\frac{p}{p-1}}(\mathcal{S})}$. Assume p and q satisfy the following conditions:*

$$C_{\mathbb{P},p} < \frac{1}{\gamma}; \quad p \geq \max\left\{1, \frac{\log(|\mathcal{A}|) + \log(\mu(\mathcal{S}))}{\log\frac{1}{\gamma C_{\mathbb{P},p}}}\right\}; \quad p \leq q \leq \infty. \tag{260}$$

Then, Bellman optimality equation $\mathcal{T}_B Q = Q$ is $\left(L^{q,d_{\mu_0}^\pi}(\mathcal{S}\times\mathcal{A}), L^p(\mathcal{S}\times\mathcal{A})\right)$-stable.

Proof. For any $1 \leq p \leq q \leq \infty$ and $Q \in L^p(\mathcal{S}\times\mathcal{A}) \cap L^q(\mathcal{S}\times\mathcal{A})$, denote that

$$\mathcal{L}_0 Q(s,a) := \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} Q(s',a')\right], \tag{261}$$

$$\mathcal{L}Q := \mathcal{T}_B Q - Q = r + \mathcal{L}_0 Q - Q. \tag{262}$$

Let Q^* denote the Bellman optimality Q-function. Note that $\mathcal{T}_B Q^* = Q^*$ and $\mathcal{L}Q^* = 0$. Define

$$w = w_Q := Q - Q^*, \tag{263}$$
$$f = f_Q := \mathcal{L}Q = \mathcal{L}Q - \mathcal{L}Q^*. \tag{264}$$

Based on the above notations, we have

$$f = \mathcal{L}Q - \mathcal{L}Q^* \tag{265}$$
$$= -w + \mathcal{L}_0 \left(Q^* + w\right) - \mathcal{L}_0 Q^*. \tag{266}$$

According to the inequality (232), we have that when $C_{\mathbb{P},p} < \frac{1}{\gamma}$ and $p \geq \frac{\log(|\mathcal{A}|)+\log(\mu(\mathcal{S}))}{\log \frac{1}{\gamma C_{\mathbb{P},p}}}$ and $q \geq p$, we have

$$\|w\|_{L^p(\mathcal{S}\times\mathcal{A})} \leq \frac{\left(|\mathcal{A}|\,\mu\left(\mathcal{S}\right)\right)^{\frac{1}{p}-\frac{1}{q}}}{1-\gamma C_{\mathbb{P},p}\left(|\mathcal{A}|\,\mu\left(\mathcal{S}\right)\right)^{\frac{1}{p}}} \|f\|_{L^q(\mathcal{S}\times\mathcal{A})}. \tag{267}$$

According to Lemma C.5, when $C_{d^\pi_{\mu_0}} := \inf_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^\pi_{\mu_0}(s,a) > 0$, we have

$$\|w\|_{L^p(\mathcal{S}\times\mathcal{A})} \leq \frac{\left(|\mathcal{A}|\,\mu\left(\mathcal{S}\right)\right)^{\frac{1}{p}-\frac{1}{q}}}{C_{d^\pi_{\mu_0}}\left(1-\gamma C_{\mathbb{P},p}\left(|\mathcal{A}|\,\mu\left(\mathcal{S}\right)\right)^{\frac{1}{p}}\right)} \|f\|_{L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}. \tag{268}$$

□

Remark C.8. Note that in a practical Q-learning scheme, we take the ϵ-greedy policy for exploration and as a result, for any state-action pair (s,a), we can visit it with positive probability, and thus the condition $C_{d^\pi_{\mu_0}} > 0$ is fulfilled.

We also demonstrate a theorem with better bound yet stronger condition.

Theorem C.9. *For any MDP $\mathcal{M}$ and fixed policy π, assume $C_{d^\pi_{\mu_0}} := \inf_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^\pi_{\mu_0}(s,a) > 0$. Assume p, q and γ satisfy the following conditions:*

$$C_{d^\pi_{\mu_0},\mathbb{P},p} := \frac{\left\|d^\pi_{\mu_0}\right\|_{L^{p^2}(\mathcal{S}\times\mathcal{A})} C_{\mathbb{P},p}}{C_{d^\pi_{\mu_0}}} < \frac{1}{\gamma}; \quad p \geq \frac{\log\left(|\mathcal{A}|\right) + \log\left(\mu\left(\mathcal{S}\right)\right)}{\log \frac{1}{\gamma C_{d^\pi_{\mu_0},\mathbb{P},p}}} - 1; \quad q \geq p^2. \tag{269}$$

Then, Bellman optimality equation $\mathcal{T}_B Q = Q$ is $\left(L^{q,d^\pi_{\mu_0}}\left(\mathcal{S}\times\mathcal{A}\right), L^p\left(\mathcal{S}\times\mathcal{A}\right)\right)$-stable.

Proof. For any $1 \leq p \leq q \leq \infty$ and $Q \in L^p\left(\mathcal{S}\times\mathcal{A}\right) \cap L^q\left(\mathcal{S}\times\mathcal{A}\right)$, denote that

$$\mathcal{L}_0 Q(s,a) := \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}\left[\max_{a'\in\mathcal{A}} Q\left(s',a'\right)\right], \tag{270}$$
$$\mathcal{L}Q := \mathcal{T}_B Q - Q = r + \mathcal{L}_0 Q - Q. \tag{271}$$

Let Q^* denote the Bellman optimality Q-function. Note that $\mathcal{T}_B Q^* = Q^*$ and $\mathcal{L}Q^* = 0$. Define

$$w = w_Q := Q - Q^*, \tag{272}$$
$$f = f_Q := \mathcal{L}Q = \mathcal{L}Q - \mathcal{L}Q^*. \tag{273}$$

Based on the above notations, we have

$$f = \mathcal{L}Q - \mathcal{L}Q^* \tag{274}$$
$$= -w + \mathcal{L}_0 \left(Q^* + w\right) - \mathcal{L}_0 Q^*. \tag{275}$$

Then, we have

$$|w(s,a)| = |-f + \mathcal{L}_0(Q^* + w) - \mathcal{L}_0 Q^*| \Big|_{(s,a)} \tag{276}$$

$$\leq |f| + |\mathcal{L}_0(Q^* + w) - \mathcal{L}_0 Q^*| \Big|_{(s,a)}. \tag{277}$$

Thus, we obtain

$$\|w\|_{L^{p^2, d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})} \leq \left\| d^\pi_{\mu_0} |f| + d^\pi_{\mu_0} |\mathcal{L}_0(Q^* + w) - \mathcal{L}_0 Q^*| \right\|_{L^{p^2}(\mathcal{S}\times\mathcal{A})} \tag{278}$$

$$\leq \|f\|_{L^{p^2, d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})} + \|\mathcal{L}_0(Q^* + w) - \mathcal{L}_0 Q^*\|_{L^{p^2, d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}, \tag{279}$$

where the last inequality comes from the Minkowski's inequality. Owing to Lemma C.5, we have

$$\|w\|_{L^{p^2, d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})} \geq \frac{1}{C_{d^\pi_{\mu_0}, p}} \|w\|_{L^p(\mathcal{S}\times\mathcal{A})}, \tag{280}$$

where $C_{d^\pi_{\mu_0}, p} = \left(\int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| d^\pi_{\mu_0}(s,a) \right|^{-\frac{p^2}{p-1}} d\mu(s,a) \right)^{\frac{p-1}{p^2}}$. In the following, we analyze the relation between $\|\mathcal{L}_0(Q^* + w) - \mathcal{L}_0 Q^*\|_{L^{p^2, d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}$ and $\|w\|_{L^p(\mathcal{S}\times\mathcal{A})}$.

$$|\mathcal{L}_0(Q^* + w) - \mathcal{L}_0 Q^*| \Big|_{(s,a)} \tag{281}$$

$$= \left| \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[\max_{a'\in\mathcal{A}} (Q^*(s',a') + w(s',a')) - \max_{a'\in\mathcal{A}} Q^*(s',a') \right] \right| \tag{282}$$

$$\leq \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \left| \max_{a'\in\mathcal{A}} (Q^*(s',a') + w(s',a')) - \max_{a'\in\mathcal{A}} Q^*(s',a') \right| \tag{283}$$

$$\leq \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[\max_{a'\in\mathcal{A}} |w(s',a')| \right] \tag{284}$$

$$= \gamma \int_{s'} \max_{a'\in\mathcal{A}} |w(s',a')| \, \mathbb{P}(s' \mid s,a) ds' \tag{285}$$

$$\leq \gamma \left\| \max_{a\in\mathcal{A}} |w(s,a)| \right\|_{L^p(\mathcal{S})} \left(\int_{s'} (\mathbb{P}(s' \mid s,a))^{\frac{p}{p-1}} ds' \right)^{1-\frac{1}{p}} \tag{286}$$

$$= \gamma \|\mathbb{P}(\cdot \mid s,a)\|_{L^{\frac{p}{p-1}}(\mathcal{S})} \left\| \max_{a\in\mathcal{A}} |w(s,a)| \right\|_{L^p(\mathcal{S})}, \tag{287}$$

where the second inequality comes from Lemma B.5 and the third inequality comes from the Holder's inequality. let $C_{\mathbb{P},p} := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\mathbb{P}(\cdot \mid s,a)\|_{L^{\frac{p}{p-1}}(\mathcal{S})}$ and then, we have

$$\|\mathcal{L}_0(Q^* + w) - \mathcal{L}_0 Q^*\|_{L^{p^2, d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})} \tag{288}$$

$$\leq \left(\int_{\mathcal{S}\times\mathcal{A}} \left(\gamma \|\mathbb{P}(\cdot \mid s,a)\|_{L^{\frac{p}{p-1}}(\mathcal{S})} \left\| \max_{a\in\mathcal{A}} |w(s,a)| \right\|_{L^p(\mathcal{S})} d^\pi_{\mu_0}(s,a) \right)^{p^2} d\mu(s,a) \right)^{\frac{1}{p^2}} \tag{289}$$

$$= \left(\int_{\mathcal{S}\times\mathcal{A}} \left(d^\pi_{\mu_0}(s,a) \right)^{p^2} d\mu(s,a) \right)^{\frac{1}{p^2}} \gamma C_{\mathbb{P},p} \left\| \max_{a\in\mathcal{A}} |w(s,a)| \right\|_{L^p(\mathcal{S})} \tag{290}$$

$$\leq \gamma \left\| d^\pi_{\mu_0} \right\|_{L^{p^2}(\mathcal{S}\times\mathcal{A})} C_{\mathbb{P},p} \left\| \max_{a\in\mathcal{A}} |w(s,a)| \right\|_{L^p(\mathcal{S})} \tag{291}$$

$$\leq \gamma \left\| d^\pi_{\mu_0} \right\|_{L^{p^2}(\mathcal{S}\times\mathcal{A})} C_{\mathbb{P},p} \|w\|_{L^p(\mathcal{S}\times\mathcal{A})}, \tag{292}$$

where the last inequality comes from $\|w\|_{l^\infty(\mathcal{A})} \leq \|w\|_{l^p(\mathcal{A})}$. Then, we have that

$$\left(\frac{1}{C_{d^\pi_{\mu_0},p}} - \gamma \left\|d^\pi_{\mu_0}\right\|_{L^{p^2}(\mathcal{S}\times\mathcal{A})} C_{\mathbb{P},p}\right) \|w\|_{L^p(\mathcal{S}\times\mathcal{A})} \leq \|f\|_{L^{p^2,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}. \tag{293}$$

When $C_{d^\pi_{\mu_0}} := \inf_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^\pi_{\mu_0}(s,a) > 0$, we have

$$C_{d^\pi_{\mu_0},p} \leq \frac{\left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{p-1}{p^2}}}{C_{d^\pi_{\mu_0}}} \tag{294}$$

Thus, when the following conditions hold

$$C_{d^\pi_{\mu_0},\mathbb{P},p} := \frac{\left\|d^\pi_{\mu_0}\right\|_{L^{p^2}(\mathcal{S}\times\mathcal{A})} C_{\mathbb{P},p}}{C_{d^\pi_{\mu_0}}} < \frac{1}{\gamma}; \quad p \geq \frac{\log(|\mathcal{A}|) + \log(\mu(\mathcal{S}))}{\log\frac{1}{\gamma C_{d^\pi_{\mu_0},\mathbb{P},p}}} - 1; \quad q \geq p^2,$$

we have

$$\|w\|_{L^p(\mathcal{S}\times\mathcal{A})} \leq \frac{\left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{p-1}{p^2}}}{C_{d^\pi_{\mu_0}} - \gamma \left\|d^\pi_{\mu_0}\right\|_{L^{p^2}(\mathcal{S}\times\mathcal{A})} C_{\mathbb{P},p} \left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{p-1}{p^2}}} \|f\|_{L^{p^2,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})} \tag{295}$$

$$\leq \frac{\left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{1}{p}-\frac{1}{q}}}{C_{d^\pi_{\mu_0}} - \gamma \left\|d^\pi_{\mu_0}\right\|_{L^{p^2}(\mathcal{S}\times\mathcal{A})} C_{\mathbb{P},p} \left(|\mathcal{A}|\,\mu(\mathcal{S})\right)^{\frac{p-1}{p^2}}} \|f\|_{L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}, \tag{296}$$

where the last inequality comes from $\|f\|_{L^{p^2,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})} \leq \mu(\mathcal{S}\times\mathcal{A})^{\frac{1}{p^2}-\frac{1}{q}} \|f\|_{L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})}$. □

Remark C.10. The conditions are not satisfactory. The result implicitly adds the constrain for γ because $\lim_{p\to\infty} C_{d^\pi_{\mu_0},\mathbb{P},p} = \frac{1}{C_{d^\pi_{\mu_0}}}$, which indicates γ may be very small, i.e. $\gamma < C_{d^\pi_{\mu_0}}$.

In the following theorem, we describe the instablility of DQN.

Theorem C.11. *There exists a MDP $\mathcal{M}$ such that for all π satisfying $M_{d^\pi_{\mu_0}} := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^\pi_{\mu_0}(s,a) < \infty$, Bellman optimality equation $\mathcal{T}_B Q = Q$ is not $\left(L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A}), L^p(\mathcal{S}\times\mathcal{A})\right)$-stable, for all $1 \leq q < p \leq \infty$.*

Proof. According to the proof of Theorem B.10, for $1 \leq q < p \leq \infty$, there exists a MDP $\mathcal{M}$ satisfying the following statement. For all $\delta > 0$ and $n \in \mathbb{N}$, there exists a $Q \in L^p(\mathcal{S}\times\mathcal{A}) \cap L^q(\mathcal{S}\times\mathcal{A})$ satisfying $\|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})} \leq \frac{\delta}{M_{d^\pi_{\mu_0}}}$ such that $\|Q - Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})} > n\|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})}$.

According to Lemma C.5, if $M_{d^\pi_{\mu_0}} := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^\pi_{\mu_0}(s,a) < \infty$, we have

$$\|Q\|_{L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})} \leq M_{d^\pi_{\mu_0}} \|Q\|_{L^q(\mathcal{S}\times\mathcal{A})} < \infty. \tag{297}$$

Thus, we have $Q \in L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A}) \cap L^p(\mathcal{S}\times\mathcal{A})$. For the same reason, we have

$$\|\mathcal{T}_B Q - Q\|_{L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})} \leq M_{d^\pi_{\mu_0}} \|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})} \leq \delta. \tag{298}$$

Hence, we get that For all $\delta > 0$ and $n \in \mathbb{N}$, there exists a $Q \in L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A}) \cap L^p(\mathcal{S}\times\mathcal{A})$ satisfying $\|\mathcal{T}_B Q - Q\|_{L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A})} \leq \delta$ such that $\|Q - Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})} > n\|\mathcal{T}_B Q - Q\|_{L^q(\mathcal{S}\times\mathcal{A})}$. □

Remark C.12. If $M_{d^\pi_{\mu_0}} = \infty$, $d^\pi_{\mu_0}$ degenerates to the discrete probability distribution. Then, $L^{p,d^\pi_{\mu_0}}$ can be considered as a norm defined on a finite dimension space. In this setting, we also have $C_{d^\pi_{\mu_0}} = 0$ and Bellman optimality equation $\mathcal{T}_B Q = Q$ is not $\left(L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A}), L^p(\mathcal{S}\times\mathcal{A})\right)$-stable, for any p and q.

According to the above theorems and remarks, we have the following corollary in the DQN procedure.

Corollary C.13. *In practical DQN procedure, the Bellman optimality equations $\mathcal{T}_B Q = Q$ is $(L^{\infty,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A}), L^p(\mathcal{S}\times\mathcal{A}))$-stable for all $1 \leq p \leq \infty$, while it is not $(L^{q,d^\pi_{\mu_0}}(\mathcal{S}\times\mathcal{A}), L^p(\mathcal{S}\times\mathcal{A}))$-stable for all $1 \leq q < p \leq \infty$.*

C.3. Stability of DQN: the Bad

Theorem C.14. *There exists an MDP $\mathcal{M}$ such that for all π satisfying $d^{\pi}_{\mu_0}$ is a discrete probability distribution, Bellman optimality equation $\mathcal{T}_B Q = Q$ is not $\left(L^{q,d^{\pi}_{\mu_0}}(\mathcal{S}\times\mathcal{A}), L^{p}(\mathcal{S}\times\mathcal{A})\right)$-stable, for any p and q.*

Proof. We only need to show there exists an MDP such that $\forall n \in \mathbb{N}, \forall \delta > 0, \exists Q(s,a)$, such that $\|\mathcal{T}_B Q - Q\|_{L^{q,d^{\pi}_{\mu_0}}(\mathcal{S}\times\mathcal{A})} < \delta$, but $\|Q - Q^*\|_{L^p(\mathcal{S}\times\mathcal{A})} \geq n$.

Consider an MDP $\mathcal{M}$ where $\mathcal{S} = [-1,1], \mathcal{A} = \{a_1, a_2\}$,

$$\mathbb{P}(s'|s,a_1) = \begin{cases} \mathbb{1}_{\{s'=s-0.1\}}, & s \in [-0.9, 1] \\ \mathbb{1}_{\{s'=s\}}, & else \end{cases}, \quad \mathbb{P}(s'|s,a_2) = \begin{cases} \mathbb{1}_{\{s'=s+0.1\}}, & s \in [-1, 0.9] \\ \mathbb{1}_{\{s'=s\}}, & else \end{cases},$$

$r(s,a_i) = k_i s,\ k_2 \geq k_1 > 0$. The transition function is essentially a deterministic transition dynamic and for convenience, we denote that

$$p(s,a_1) = \begin{cases} s - 0.1, & s \in [-0.9, 1] \\ s, & else \end{cases}, \quad p(s,a_2) = \begin{cases} s + 0.1, & s \in [-1, 0.9] \\ s, & else \end{cases}.$$

Let $Q^*(s,a) = Q^{\pi^*}(s,a)$ be the optimal Q-function, where π^* is the optimal policy.

Define $B_0 = \{s \in \mathcal{S} : \exists a \in \mathcal{A}, s.t.\, d^{\pi}_{\mu_0}(s,a) \neq 0\}$, which contains all the states that can be explored. Let $B = \{B_0 \cup \{B_0 + 0.1\} \cup \{B_0 - 0.1\}\} \cap \mathcal{S}$, then $\forall s \in B,\ p(s,a) \in B$. Since $d^{\pi}_{\mu_0}$ is a discrete probability distribution, $\mu(B) = \mu(B_0) = 0$.

Let $D = [-1,1] \setminus B$, and $Q(s,a) = Q^*(s,a) + h \cdot \mathbb{1}_D$, where $h = \frac{n}{2^{\frac{1}{p}}}$. We have that $Q(s,a) = Q^*(s,a), \forall s \notin D$. We then have for any $s \in B$ that

$$\begin{aligned} \mathcal{T}_B Q(s,a) &= r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} Q(p(s,a), a') \\ &= r(s,a) + \gamma \cdot \max_{a' \in \mathcal{A}} Q^*(p(s,a), a') \\ &= Q^*(s,a) \\ &= Q(s,a). \end{aligned}$$

For any $s \notin B$, $d^{\pi}_{\mu_0}(s,a) = 0$ for all $a \in \mathcal{A}$. Hence, $\|\mathcal{T}_B Q(s,a) - Q(s,a)\|_{L^{q,d^{\pi}_{\mu_0}}(\mathcal{S}\times\mathcal{A})} = 0 < \delta$.

However, we find that

$$\|Q(s,a) - Q^*(s,a)\|_{L^p(\mathcal{S}\times\mathcal{A})} = h \cdot \mu(D)^{\frac{1}{p}} = h \cdot 2^{\frac{1}{p}} \geq n,$$

which completes the proof. □

Remark C.15. If $d^{\pi}_{\mu_0}$ is a discrete probability distribution, $L^{\infty,d^{\pi}_{\mu_0}}(\mathcal{S}\times\mathcal{A})$ is not a good choice. However, the sample process should be considered in practical reinforcement learning algorithms and as a consequence, we have to apply the space $L^{\infty,d^{\pi}_{\mu_0}}(\mathcal{S}\times\mathcal{A})$ rather than $L^{\infty}(\mathcal{S}\times\mathcal{A})$.

D. Derivation of CAR-DQN

Our theory motivates us to use the following objective:

$$\begin{aligned} \mathcal{L}_{car}(\theta) &:= \|\mathcal{T}_B Q_\theta - Q_\theta\|_{L^{\infty,d^{\pi_\theta}_{\mu_0}}(\mathcal{S}\times\mathcal{A})} && (299) \\ &= \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\pi_\theta}_{\mu_0}(s,a) \left|\mathcal{T}_B Q_\theta(s,a) - Q_\theta(s,a)\right| && (300) \\ &= \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d^{\pi_\theta}_{\mu_0}(s,a) \left| r(s,a) + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \left[\max_{a'\in\mathcal{A}} Q_\theta(s',a')\right] - Q_\theta(s,a)\right|. && (301) \end{aligned}$$

However, the objective is intractable in a model-free setting, due to the unknown environment, i.e. unknown reward function and unknown transition function.

Remark D.1. We apply the space $L^{\infty,d_{\mu_0}^{\pi_\theta}}(\mathcal{S}\times\mathcal{A})$ rather than $L^{\infty}(\mathcal{S}\times\mathcal{A})$ because the sampling process should be considered in practical reinforcement learning algorithms.

D.1. Surrogate Objective

We can derive that

$$\mathcal{L}_{car}(\theta) = \sup_{s\in\mathcal{S}}\max_{a\in\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\left|\mathcal{T}_B Q_\theta(s,a) - Q_\theta(s,a)\right| \tag{302}$$

$$= \sup_{s\in\mathcal{S}}\max_{s_\nu\in B_\epsilon(s)}\max_{a\in\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\left|\mathcal{T}_B Q_\theta(s_\nu,a) - Q_\theta(s_\nu,a)\right| \tag{303}$$

$$= \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\mathcal{T}_B Q_\theta(s_\nu,a) - Q_\theta(s_\nu,a)\right|. \tag{304}$$

However, in a practical reinforcement learning setting, we cannot directly get the estimation of $\mathcal{T}_B Q_\theta(s_\nu,a)$.

Theorem D.2. *Let* $\mathcal{L}_{car}^{train}(\theta) := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\mathcal{T}_B Q_\theta(s,a) - Q_\theta(s_\nu,a)\right|$ *and* $\mathcal{L}_{car}^{diff}(\theta) := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\mathcal{T}_B Q_\theta(s_\nu,a) - \mathcal{T}_B Q_\theta(s,a)\right|$. *We have that*

$$\left|\mathcal{L}_{car}^{train}(\theta) - \mathcal{L}_{car}^{diff}(\theta)\right| \le \mathcal{L}_{car}(\theta) \le \mathcal{L}_{car}^{train}(\theta) + \mathcal{L}_{car}^{diff}(\theta). \tag{305}$$

Proof. On one hand, we have

$$\mathcal{L}_{car}(\theta) = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\mathcal{T}_B Q_\theta(s_\nu,a) - Q_\theta(s_\nu,a)\right| \tag{306}$$

$$\le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left(\left|\mathcal{T}_B Q_\theta(s,a) - Q_\theta(s_\nu,a)\right| + \left|\mathcal{T}_B Q_\theta(s_\nu,a) - \mathcal{T}_B Q_\theta(s,a)\right|\right) \tag{307}$$

$$\le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\mathcal{T}_B Q_\theta(s_\nu,a) - \mathcal{T}_B Q_\theta(s,a)\right| \tag{308}$$

$$+ \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\mathcal{T}_B Q_\theta(s,a) - Q_\theta(s_\nu,a)\right|. \tag{309}$$

On the other hand, we have

$$\mathcal{L}_{car}(\theta) = \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\mathcal{T}_B Q_\theta(s_\nu,a) - Q_\theta(s_\nu,a)\right| \tag{310}$$

$$\ge \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\left|\mathcal{T}_B Q_\theta(s_\nu,a) - \mathcal{T}_B Q_\theta(s,a)\right| - \left|\mathcal{T}_B Q_\theta(s,a) - Q_\theta(s_\nu,a)\right|\right| \tag{311}$$

$$\ge \Bigg|\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\mathcal{T}_B Q_\theta(s_\nu,a) - \mathcal{T}_B Q_\theta(s,a)\right| \tag{312}$$

$$- \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a)\max_{s_\nu\in B_\epsilon(s)}\left|\mathcal{T}_B Q_\theta(s,a) - Q_\theta(s_\nu,a)\right|\Bigg|, \tag{313}$$

where the second inequality comes from Lemma B.5. □

It is hard to calculate or estimate $\mathcal{L}_{car}^{diff}(\theta)$ in practice. Fortunately, we think $\mathcal{L}_{car}^{diff}(\theta)$ should be small in practice and we give a constant upper bound of $\mathcal{L}_{car}^{diff}(\theta)$ in the smooth environment.

Lemma D.3. *Suppose Q and r are uniformly bounded, i.e. $\exists\, M_Q, M_r > 0$ such that $|Q(s,a)| \le M_Q$, $|r(s,a)| \le M_r\ \forall s\in\mathcal{S}, a\in\mathcal{A}$. Then $\mathcal{T}_B Q(\cdot,a)$ is uniformly bounded, i.e.*

$$\left|\mathcal{T}_B Q(s,a)\right| \le C_Q,\ \forall s\in\mathcal{S}, a\in\mathcal{A}, \tag{314}$$

where $C_Q = \max\left\{M_Q, \frac{M_r}{1-\gamma}\right\}$. Further, for any $k\in\mathbb{N}$, $\mathcal{T}_B^k Q(\cdot,a)$ has the same uniform bound as $\mathcal{T}_B Q(\cdot,a)$, i.e.

$$\left|\mathcal{T}_B^k Q(s,a)\right| \le C_Q,\ \forall s\in\mathcal{S}, a\in\mathcal{A}. \tag{315}$$

Proof.

$$|\mathcal{T}_B Q(s,a)| = \left| r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} Q(s',a') \right] \right| \tag{316}$$

$$\leq |r(s,a)| + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left| \max_{a' \in \mathcal{A}} Q(s',a') \right| \tag{317}$$

$$\leq M_r + \gamma M_Q \tag{318}$$

$$\leq \max\left\{ M_Q, \frac{M_r}{1-\gamma} \right\}, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{319}$$

Let $C_Q = \max\left\{ M_Q, \frac{M_r}{1-\gamma} \right\}$. Suppose the inequality (315) holds for $k = n$. Then, for $k = n+1$, we have

$$\left|\mathcal{T}_B^{n+1} Q(s,a)\right| = \left| r(s,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\max_{a' \in \mathcal{A}} \mathcal{T}_B^n Q(s',a') \right] \right| \tag{320}$$

$$\leq |r(s,a)| + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left| \max_{a' \in \mathcal{A}} \mathcal{T}_B^n Q(s',a') \right| \tag{321}$$

$$\leq M_r + \gamma C_Q \tag{322}$$

$$\leq (1-\gamma) C_Q + \gamma C_Q \tag{323}$$

$$= C_Q. \tag{324}$$

By induction, we have $\left|\mathcal{T}_B^k Q(s,a)\right| \leq C_Q,\ \forall s \in \mathcal{S}, a \in \mathcal{A}, k \in \mathbb{N}$. □

Lemma D.4. *Suppose the environment is $(L_r, L_\mathbb{P})$-smooth and suppose Q and r are uniformly bounded, i.e. $\exists\, M_Q, M_r > 0$ such that $|Q(s,a)| \leq M_Q,\ |r(s,a)| \leq M_r\ \forall s \in \mathcal{S}, a \in \mathcal{A}$. Then $\mathcal{T}_B Q(\cdot, a)$ is Lipschitz continuous, i.e.*

$$|\mathcal{T}_B Q(s,a) - \mathcal{T}_B Q(s',a)| \leq L_{\mathcal{T}_B} \|s - s'\|, \tag{325}$$

where $L_{\mathcal{T}_B} = L_r + \gamma C_Q L_\mathbb{P}$ and $C_Q = \max\left\{ M_Q, \frac{M_r}{1-\gamma} \right\}$. Further, for any $k \in \mathbb{N}$, $\mathcal{T}_B^k Q(\cdot, a)$ is Lipschitz continuous and has the same Lipschitz constant as $\mathcal{T}_B Q(\cdot, a)$, i.e.

$$\left|\mathcal{T}_B^k Q(s,a) - \mathcal{T}_B^k Q(s',a)\right| \leq L_{\mathcal{T}_B} \|s - s'\|. \tag{326}$$

Proof. For all $s_1, s_2 \in \mathcal{S}$, we have

$$\mathcal{T}_B Q(s_1,a) - \mathcal{T}_B Q(s_2,a) \tag{327}$$

$$= r(s_1,a) + \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_1,a)} \left[\max_{a' \in \mathcal{A}} Q(s',a') \right] - r(s_2,a) - \gamma \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_2,a)} \left[\max_{a' \in \mathcal{A}} Q(s',a') \right] \tag{328}$$

$$= (r(s_1,a) - r(s_2,a)) + \gamma \int_{s'} \left(\mathbb{P}(s'|s_1,a) - \mathbb{P}(s'|s_2,a)\right) \max_{a' \in \mathcal{A}} Q(s',a')\, ds'. \tag{329}$$

Then, we have

$$|\mathcal{T}_B Q(s_1,a) - \mathcal{T}_B Q(s_2,a)| \tag{330}$$

$$\leq |(r(s_1,a) - r(s_2,a))| + \left| \gamma \int_{s'} \left(\mathbb{P}(s'|s_1,a) - \mathbb{P}(s'|s_2,a)\right) \max_{a' \in \mathcal{A}} Q(s',a')\, ds' \right| \tag{331}$$

$$\leq L_r \|s_1 - s_2\| + \gamma \int_{s'} \left|\mathbb{P}(s'|s_1,a) - \mathbb{P}(s'|s_2,a)\right| \left| \max_{a' \in \mathcal{A}} Q(s',a') \right| ds' \tag{332}$$

$$\leq L_r \|s_1 - s_2\| + \gamma C_Q \int_{s'} \left|\mathbb{P}(s'|s_1,a) - \mathbb{P}(s'|s_2,a)\right| ds' \tag{333}$$

$$\leq L_r \|s_1 - s_2\| + \gamma C_Q L_\mathbb{P} \|s_1 - s_2\| \tag{334}$$

$$= \left(L_r + \gamma C_Q L_\mathbb{P}\right) \|s_1 - s_2\|. \tag{335}$$

The second inequality comes from the Lipschitz property of r. The third inequality comes from the uniform boundedness of Q and the last inequality utilizes the Lipschitz property of $\mathbb{P}$.

Note that the uniform boundedness used in the above proof is C_Q rather than M_Q. Then, due to lemma D.3, we can extend the above proof to $\mathcal{T}_B^k$. □

Theorem D.5. *Suppose the environment is $(L_r, L_{\mathbb{P}})$-smooth and suppose Q_θ and r are uniformly bounded, i.e. $\exists\, M_Q, M_r > 0$ such that $|Q_\theta(s,a)| \le M_Q,\ |r(s,a)| \le M_r\ \forall s \in \mathcal{S}, a \in \mathcal{A}$. If $M := \sup_{\theta,(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a) < \infty$, then we have*

$$\mathcal{L}_{car}^{diff}(\theta) \le C_{\mathcal{T}_B}\epsilon, \tag{336}$$

where $C_{\mathcal{T}_B} = L_{\mathcal{T}_B} M$, $L_{\mathcal{T}_B} = L_r + \gamma C_Q L_{\mathbb{P}}$ and $C_Q = \max\left\{M_Q, \frac{M_r}{1-\gamma}\right\}$.

Proof.

$$\begin{aligned}
\mathcal{L}_{car}^{diff}(\theta) &= \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a) \max_{s_\nu\in B_\epsilon(s)} |\mathcal{T}_B Q_\theta(s_\nu,a) - \mathcal{T}_B Q_\theta(s,a)| && (337)\\
&\le \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a) \max_{s_\nu\in B_\epsilon(s)} (L_r + \gamma C_Q L_{\mathbb{P}}) \|s_\nu - s\| && (338)\\
&\le (L_r + \gamma C_Q L_{\mathbb{P}})\,\epsilon \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\mu_0}^{\pi_\theta}(s,a) && (339)\\
&\le M\,(L_r + \gamma C_Q L_{\mathbb{P}})\,\epsilon, && (340)
\end{aligned}$$

where the first inequality comes from Lemma D.4 and the last inequality comes from the uniform boundedness of $d_{\mu_0}^{\pi_\theta}$. □

D.2. Approximate Objective

Lemma D.6. *For any function $f : \Omega \to \mathbb{R}$ and $\lambda > 0$, we have*

$$\max_{p\in\Delta(\Omega)} [\mathbb{E}_{\omega\sim p} f(\omega) - \lambda\,\mathrm{KL}(p\|p_0)] = \lambda \ln \mathbb{E}_{\omega\sim p_0}\left[e^{\frac{f(\omega)}{\lambda}}\right], \tag{341}$$

where $\Delta(\Omega)$ denotes the set of probability distributions on Ω. And the solution is achieved by the following distribution q:

$$q(\omega) = \frac{p_0(\omega)e^{\frac{f(\omega)}{\lambda}}}{\int_{\omega\in\Omega} p_0(\omega)e^{\frac{f(\omega)}{\lambda}} d\mu(\omega)} = \frac{1}{C} p_0(\omega) e^{\frac{f(\omega)}{\lambda}}. \tag{342}$$

Proof. Let

$$C := \mathbb{E}_{\omega\sim p_0}\left[e^{\frac{f(\omega)}{\lambda}}\right] = \int_{\omega\in\Omega} p_0(\omega) e^{\frac{f(\omega)}{\lambda}} d\mu(\omega),$$

$$q(\omega) = \frac{p_0(\omega)e^{\frac{f(\omega)}{\lambda}}}{\int_{\omega\in\Omega} p_0(\omega)e^{\frac{f(\omega)}{\lambda}} d\mu(\omega)} = \frac{1}{C} p_0(\omega) e^{\frac{f(\omega)}{\lambda}}.$$

Then, we have

$$\begin{aligned}
&\mathbb{E}_{\omega\sim p} f(\omega) - \lambda\,\mathrm{KL}(p\|p_0) && (343)\\
=&\mathbb{E}_{\omega\sim p}\left[\lambda \ln e^{\frac{f(\omega)}{\lambda}} - \lambda \ln \frac{p(\omega)}{p_0(\omega)}\right] && (344)\\
=&\lambda \mathbb{E}_{\omega\sim p}\left[\ln \frac{e^{\frac{f(\omega)}{\lambda}} p_0(\omega)}{p(\omega)}\right] && (345)\\
=&\lambda \mathbb{E}_{\omega\sim p}\left[\ln \frac{C q(\omega)}{p(\omega)}\right] && (346)\\
=&\lambda\,[\ln C - \mathrm{KL}(p\|q)] && (347)\\
\le&\lambda \ln C && (348)\\
=&\lambda \ln \mathbb{E}_{\omega\sim p_0}\left[e^{\frac{f(\omega)}{\lambda}}\right]. && (349)
\end{aligned}$$

Note that the equal sign holds if and only if $p = q$. Thus, we get

$$q \in \arg\max_{p \in \Delta(\Omega)} \left[\mathbb{E}_{\omega \sim p} f(\omega) - \lambda \, \mathrm{KL}(p \| p_0)\right].$$

□

We get the following approximate objective of $\mathcal{L}_{car}^{train}(\theta)$:

$$\mathcal{L}_{car}^{app}(\theta) = \max_{(s,a,r,s') \in \mathcal{B}} \frac{1}{|\mathcal{B}|} \max_{s_\nu \in B_\epsilon(s)} \left| r + \gamma \max_{a'} Q_{\bar{\theta}}(s', a') - Q_\theta(s_\nu, a) \right|. \tag{350}$$

Denote

$$f_i = f(s_i, a_i, r_i, s_i') := \max_{s_\nu \in B_\epsilon(s_i)} \left| r_i + \gamma \max_{a'} Q_{\bar{\theta}}(s_i', a') - Q_\theta(s_\nu, a_i) \right|. \tag{351}$$

To fully utilize each sample in the batch, we derive the soft version of the above objective:

$$\begin{aligned}
\mathcal{L}_{car}^{train}(\theta) &= \max_{(s,a,r,s') \in \mathcal{B}} \frac{1}{|\mathcal{B}|} f(s, a, r, s') && (352)\\
&= \frac{1}{|\mathcal{B}|} \max_{p \in \Delta(\mathcal{B})} \sum_{(s_i, a_i, r_i, s_i') \in \mathcal{B}} p_i f(s_i, a_i, r_i, s_i') && (353)\\
&\geq \frac{1}{|\mathcal{B}|} \max_{p \in \Delta(\mathcal{B})} \left(\sum_{(s_i, a_i, r_i, s_i') \in \mathcal{B}} p_i f(s_i, a_i, r_i, s_i') - \lambda \, \mathrm{KL}\left(p \| U\left(\mathcal{B}\right)\right) \right), && (354)
\end{aligned}$$

where $U\left(\mathcal{B}\right)$ represents the uniform distribution over $\mathcal{B}$. According to Lemma D.6, the optimal solution of the maximization problem (354) is p^*:

$$p_i^* = \frac{e^{\frac{1}{\lambda} f_i}}{\sum_{i \in |\mathcal{B}|} e^{\frac{1}{\lambda} f_i}}. \tag{355}$$

The maximization problem (354) is the lower bound of the maximization problem (353) so p^* is a proper approximation of the optimal solution of the maximization problem (353). Thus, we get the soft version of the CAR objective:

$$\mathcal{L}_{car}^{soft}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{(s_i, a_i, r_i, s_i') \in \mathcal{B}} \frac{e^{\frac{1}{\lambda} f_i}}{\sum_{i \in |\mathcal{B}|} e^{\frac{1}{\lambda} f_i}} \max_{s_\nu \in B_\epsilon(s_i)} \left| r_i + \gamma \max_{a'} Q_{\bar{\theta}}(s_i', a') - Q_\theta(s_\nu, a_i) \right|. \tag{356}$$

E. Examples of Intrinsic States

In Figure 6, 7, 8, we show some examples in 3 Atari games (Pong, Freeway, and RoadRunner) that the state observation s and adversarial observation s_ν have the same intrinsic state.

F. Additional Algorithm Details

Algorithm. We present the CAR-DQN training algorithm in Algorithm 1.

DQN architecture. We implement Dueling network architectures (Wang et al., 2016) and the same architecture following (Zhang et al., 2020; Oikarinen et al., 2021) which has 3 convolutional layers and a two-head fully connected layers. The first convolutional layer has 8×8 kernel, stride 4, and 32 channels. The second convolutional layer has 4×4 kernel, stride 2, and 64 channels. The third convolutional layer has 3×3 kernel, stride 1, and 64 channels and is then flattened. The fully connected layers have 512 units for both heads wherein one head outputs a state value and the other outputs advantages of each action. Every middle layer is applied by the ReLU activation function.

Algorithm 1 Consistent Adversarial Robust Deep Q-Learning (CAR-DQN).

Input: Number of iterations T, target network update frequency M, a schedule β_t for the exploration probability β, a schedule ϵ_t for the perturbation radius ϵ.
Initialize current Q network $Q(s,a)$ with parameters θ and target Q network $Q'(s,a)$ with parameters $\theta' \leftarrow \theta$.
Initialize replay buffer $\mathcal{B}$.
for $t = 1$ **to** T **do**
With probability β_t select random action a_t, otherwise select $a_t = \arg\max_a Q(s_t, a; \theta)$.
Execute action a_t in environment and observe reward r_t and the next state s_{t+1}.
Store transition pair $\{s_t, a_t, r_t, s_{t+1}\}$ in $\mathcal{B}$.
Randomly sample a minibatch of N transition pairs $\{s_i, a_i, r_i, s_{i+1}\}$ from $\mathcal{B}$.
Set $y_i = r_i + \gamma Q'(s_{i+1}, \arg\max_{a'} Q(s_i, a'; \theta); \theta')$ for non-terminal s_i, and $y_i = r_i$ for terminal s_i.
Define $\mathcal{L}_{car}^{soft}(\theta)$:

$$\mathcal{L}_{car}^{soft}(\theta) := \sum_{i \in |\mathcal{B}|} \alpha_i \max_{s_\nu \in B_{\epsilon_t}(s_i)} |y_i - Q(s_\nu, a_i; \theta)| ,$$

where $\alpha_i = \frac{e^{\frac{1}{\lambda} \max_{s_\nu} |y_i - Q(s_\nu, a_i; \theta)|}}{\sum_{i \in |\mathcal{B}|} e^{\frac{1}{\lambda} \max_{s_\nu} |y_i - Q(s_\nu, a_i; \theta)|}}$.
Option 1: Use projected gradient descent (PGD) to solve $\mathcal{L}_{car}^{soft}(\theta)$.
For every $i \in |\mathcal{B}|$, run PGD to solve:

$$s_{i,\nu} = \arg\max_{s_\nu \in B_{\epsilon_t}(s_i)} |y_i - Q(s_\nu, a_i; \theta)| .$$

Compute the approximation of $\mathcal{L}_{car}^{soft}(\theta)$:

$$\mathcal{L}_{car}(\theta) = \sum_{i \in |\mathcal{B}|} \alpha_i |y_i - Q(s_{i,\nu}, a_i; \theta)| ,$$

where $\alpha_i = \frac{e^{\frac{1}{\lambda} |y_i - Q(s_{i,\nu}, a_i; \theta)|}}{\sum_{i \in |\mathcal{B}|} e^{\frac{1}{\lambda} |y_i - Q(s_{i,\nu}, a_i; \theta)|}}$.
Option 2: Use convex relaxations of neural networks to solve a surrogate loss for $\mathcal{L}_{car}^{soft}(\theta)$.
For every $i \in |\mathcal{B}|$, obtain upper and lower bounds on $Q(s, a_i; \theta)$ for all $s \in B_{\epsilon_t}(s_i)$:

$$u_i(\theta) = \text{ConvexRelaxUB}\left(Q(s, a_i; \theta), \theta, s \in B_{\epsilon_t}(s_i)\right) ,$$

$$l_i(\theta) = \text{ConvexRelaxLB}\left(Q(s, a_i; \theta), \theta, s \in B_{\epsilon_t}(s_i)\right) .$$

Compute the surrogate loss for $\mathcal{L}_{car}^{soft}(\theta)$:

$$\mathcal{L}_{car}(\theta) = \sum_{i \in |\mathcal{B}|} \alpha_i \max\left\{|y_i - u_i(\theta)|, |y_i - l_i(\theta)|\right\} ,$$

where $\alpha_i = \frac{e^{\frac{1}{\lambda} \max\{|y_i - u_i(\theta)|, |y_i - l_i(\theta)|\}}}{\sum_{i \in |\mathcal{B}|} e^{\frac{1}{\lambda} \max\{|y_i - u_i(\theta)|, |y_i - l_i(\theta)|\}}}$.
Update the Q network by performing a gradient descent step to minimize $\mathcal{L}_{car}(\theta)$.
Update target Q network every M steps: $\theta' \leftarrow \theta$.
end for

G. Comparation between RADIAL-DQN and CAR-DQN with Increasing Perturbation Radius

The core at RADIAL-DQN is a heuristic robust regularization that minimizes the overlap between bounds of perturbed Q values of the current action and others:

$$\mathcal{L}_{radial}(\theta) = \mathbb{E}_{(s,a,s',r)} \left[\sum_y Q_{\text{diff}}(s, y) \cdot Ovl(s, y, \epsilon) \right] ,$$

where $Q_{\text{diff}}(s,y) = \max(0, Q(s,a) - Q(s,y))$, $Ovl(s,y,\epsilon) = \max(0, \bar{Q}(s,y,\epsilon) - \underline{Q}(s,a,\epsilon) + \eta)$ and $\eta = c \cdot Q_{\text{diff}}(s,y), c = 0.5$. Q_{diff} is treated as a constant for the optimization. We consider RADIAL-DQN could perform better than CAR-DQN with increasing perturbation radius since $\mathcal{L}_{radial}(\theta)$ is a stronger regularization to enhance robustness while compromising natural rewards. The stronger robust constraint is mainly reflected in two aspects:

- The loose bounds. RADIAL-DQN uses the cheap but loose convex relaxation method (IBP) to estimate $\bar{Q}(s,y,\epsilon)$ and $\underline{Q}(s,a,\epsilon)$.
- The positive margin η.

They both result in $Ovl(s,y,\epsilon)$ a stronger constraint for representing the overlap of perturbed Q values. However, RADIAL-DQN has the following weaknesses:

- $\mathcal{L}_{radial}(\theta)$ will harm natural rewards. As shown in Figure 4, the natural rewards curve of RADIAL-DQN on RoadRunner distinctly tends to decrease, especially around 0.5 million steps. In contrast, the natural curves of our CAR-DQN showcase more stable upward trends in all environments. Besides, as shown in Table 5, RADIAL-DQN training with a larger radius attains lower natural rewards which also restricts robustness according to our theory, while CAR-DQN keeps a better and consistent natural and robust performance.
- Heuristic implementation lacks theoretical guarantees and introduces sensitive hyperparameter c. We conduct additional experiments on RoadRunner with different c and observe the sensitivity of RADIAL-DQN to the choice of c. Larger c could cause poor performance because the robustness constraint is too strict and thus the policy degrades to some simple policy with lower rewards. Smaller c may result in much weaker robustness. By contrast, our CAR-DQN is developed based on the theory of optimal robust policy and stability. Although we also introduce a hyperparameter λ, our ablation studies in Figure 5 show that our algorithm is insensitive to it.

Table 4. Performance of RADIAL-DQN sensitive to different positive margins $c \cdot Q_{\text{diff}}(s,y)$.

Model	Natural Return	PGD			MinBest		
		$\epsilon = 1/255$	$\epsilon = 3/255$	$\epsilon = 5/255$	$\epsilon = 1/255$	$\epsilon = 3/255$	$\epsilon = 5/255$
RADIAL-DQN ($c = 0.25$)	14678 ± 329	14836 ± 314	13670 ± 466	13512 ± 617	14712 ± 309	14804 ± 457	13226 ± 351
RADIAL-DQN ($c = 0.5$)	$\mathbf{46224 \pm 1133}$	$\mathbf{45990 \pm 1112}$	$\mathbf{42162 \pm 1147}$	$\mathbf{23248 \pm 499}$	$\mathbf{46082 \pm 1128}$	$\mathbf{42036 \pm 1048}$	$\mathbf{25434 \pm 756}$
RADIAL-DQN ($c = 0.75$)	3992 ± 482	3992 ± 482	3992 ± 482	3992 ± 482	3992 ± 482	3992 ± 482	3992 ± 482

- Depending on the currently learned optimal action. $\mathcal{L}_{radial}(\theta)$ essentially takes the currently learned action as a robust label which may produce a wrong direction for robustness training if the learned action is not optimal. In contrast, our CAR-DQN seeks optimal robust policies with theoretical guarantees and does not utilize the learned action for robustness training, simultaneously improving natural and robust performance.

The main motivation of CAR-DQN based on our theory is to improve natural and robust performance concurrently which makes sense in real-world scenarios where strong adversarial attacks are relatively rare. Our training loss can guarantee robustness under the attack with the same perturbation radius as the training. We also think it is a very significant problem whether and how we can design an algorithm training with little epsilon and theoretically guarantee robustness for larger epsilon. However, this is beyond the scope of our paper and we will consider this problem in subsequent work.

Moreover, as shown in Table 6, CAR-DQN also achieves the top performance in larger perturbation radiuses on Pong and BankHeist and matches the RADIAL-DQN on Freeway. To show the superiority of CAR-DQN further, we also train CAR-DQN with a perturbation radius of 3/255 and 5/255 on RoadRunner for 4.5 million steps (see Table 5). We can see that CAR-DQN still attains superior natural and robust performance training with larger attack radiuses while RADIAL-DQN markedly degrades its natural performance due to the too-strong robustness constraint. CAR-DQN always has a higher robust return on the training radius than RADIAL-DQN.

H. Additional Experiment Results

Training stability. We also observe that there are some instability phenomena in the training of CAR, RADIAL, and WocaR in Figure 4. We conjecture that the occasional instability in CAR training comes from the unified loss combining natural and

Table 5. Performance of CAR-DQN and RADIAL-DQN trained with different perturbation radiuses on the RoadRunner environment. The best results of the algorithm with the same training radius are highlighted in bold.

Model	Natural Return	PGD			MinBest		
		$\epsilon = 1/255$	$\epsilon = 3/255$	$\epsilon = 5/255$	$\epsilon = 1/255$	$\epsilon = 3/255$	$\epsilon = 5/255$
RADIAL-DQN ($\epsilon = 1/255$)	46224 ± 1133	45990 ± 1112	$\mathbf{42162 \pm 1147}$	$\mathbf{23248 \pm 499}$	46082 ± 1128	$\mathbf{42036 \pm 1048}$	$\mathbf{25434 \pm 756}$
CAR-DQN ($\epsilon = 1/255$)	$\mathbf{49398 \pm 1106}$	$\mathbf{49456 \pm 992}$	28588 ± 1575	15592 ± 885	$\mathbf{47526 \pm 1132}$	32878 ± 1898	21102 ± 1427
RADIAL-DQN ($\epsilon = 3/255$)	34656 ± 1104	35094 ± 1277	35082 ± 948	$\mathbf{32770 \pm 1062}$	35096 ± 1277	34374 ± 996	$\mathbf{27926 \pm 881}$
CAR-DQN ($\epsilon = 3/255$)	$\mathbf{47348 \pm 1305}$	$\mathbf{46284 \pm 1114}$	$\mathbf{43578 \pm 1315}$	27060 ± 1117	$\mathbf{46286 \pm 1122}$	$\mathbf{42602 \pm 1336}$	24862 ± 1195
RADIAL-DQN ($\epsilon = 5/255$)	35160 ± 1157	36158 ± 1104	36732 ± 1076	34826 ± 913	36158 ± 1104	36732 ± 1076	34592 ± 913
CAR-DQN ($\epsilon = 5/255$)	$\mathbf{42545 \pm 2028}$	$\mathbf{43230 \pm 1468}$	$\mathbf{37845 \pm 2344}$	$\mathbf{39235 \pm 1519}$	$\mathbf{43645 \pm 1531}$	$\mathbf{37535 \pm 2112}$	$\mathbf{38150 \pm 1316}$

robustness objectives which may cause undesirable optimization direction under a batch of special samples. The instability of RADIAL is particularly evident in the robustness curve on the BankHeist environment and natural curve on the Freeway environment and it may be from the larger batch size (=128) setting during the RADIAL training while CAR, SA-DQN and WocaR set the batch size as 32 or 16. The worst-case estimation of WocaR may be inaccurate in some states and WocaR also uses a small batch of size 16. The combination of these two can lead to instability, especially in complex environments such as RoadRunner and BankHeist. Another possible reason is that CAR, RADIAL, and WocaR all use the cheap relaxation method leading to a loose bound while SA-DQN utilizes a tighter relaxation.

Table 6. Average episode rewards $\pm$ standard error of the mean over 50 episodes on baselines and CAR-DQN. The best results of the algorithm with the same type of solver are highlighted in bold.

Environment	Model		Natural Return	PGD			MinBest			ACR		
				$\epsilon = 1/255$	$\epsilon = 3/255$	$\epsilon = 5/255$	$\epsilon = 1/255$	$\epsilon = 3/255$	$\epsilon = 5/255$	$\epsilon = 1/255$	$\epsilon = 3/255$	$\epsilon = 5/255$
Pong	Standard	DQN	21.0 ± 0.0	-21.0 ± 0.0	-21.0 ± 0.0	-20.8 ± 0.1	-21.0 ± 0.0	-21.0 ± 0.0	-21.0 ± 0.0	0	0	0
	PGD	SA-DQN	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	-19.4 ± 0.3	-21.0 ± 0.0	$\mathbf{21.0 \pm 0.0}$	-19.4 ± 0.2	-21.0 ± 0.0	0	0	0
		CAR-DQN (Ours)	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{16.8 \pm 0.7}$	-21.0 ± 0.0	$\mathbf{21.0 \pm 0.0}$	$\mathbf{20.7 \pm 0.1}$	$\mathbf{-0.8 \pm 2.8}$	0	0	0
	Convex Relaxation	SA-DQN	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	-19.6 ± 0.1	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	-9.5 ± 1.3	1.000	0	0
		RADIAL-DQN	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	4.9 ± 0.6	0.898	0	0
		WocaR-DQN	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	20.5 ± 0.1	20.6 ± 0.1	$\mathbf{21.0 \pm 0.0}$	20.7 ± 0.1	20.9 ± 0.1	0.979	0	0
		CAR-DQN (Ours)	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	$\mathbf{21.0 \pm 0.0}$	0.986	0	0
Freeway	Standard	DQN	33.9 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0	0	0
	PGD	SA-DQN	33.6 ± 0.1	23.4 ± 0.2	20.6 ± 0.3	$\mathbf{7.6 \pm 0.3}$	21.1 ± 0.2	21.3 ± 0.2	21.8 ± 0.3	0.250	0.275	0.275
		CAR-DQN (Ours)	$\mathbf{34.0 \pm 0.0}$	$\mathbf{33.7 \pm 0.1}$	$\mathbf{25.8 \pm 0.2}$	3.8 ± 0.2	$\mathbf{33.7 \pm 0.1}$	$\mathbf{30.0 \pm 0.3}$	$\mathbf{26.2 \pm 0.2}$	0	0	0
	Convex Relaxation	SA-DQN	30.0 ± 0.0	30.0 ± 0.0	30.2 ± 0.1	27.7 ± 0.1	30.0 ± 0.0	30.0 ± 0.0	29.2 ± 0.1	1.000	0.912	0
		RADIAL-DQN	33.1 ± 0.1	$\mathbf{33.3 \pm 0.1}$	$\mathbf{33.3 \pm 0.1}$	$\mathbf{29.0 \pm 0.1}$	$\mathbf{33.3 \pm 0.1}$	$\mathbf{33.3 \pm 0.1}$	$\mathbf{31.2 \pm 0.2}$	0.998	0	0
		WocaR-DQN	30.8 ± 0.1	31.0 ± 0.0	30.6 ± 0.1	29.0 ± 0.2	31.0 ± 0.0	31.1 ± 0.1	29.0 ± 0.2	0.992	0.150	0
		CAR-DQN (Ours)	$\mathbf{33.2 \pm 0.1}$	33.2 ± 0.1	32.3 ± 0.2	27.6 ± 0.3	33.2 ± 0.1	32.8 ± 0.2	31.0 ± 0.1	0.981	0	0
BankHeist	Standard	DQN	1317.2 ± 4.2	22.2 ± 1.9	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0	0	0
	PGD	SA-DQN	1248.8 ± 1.4	965.8 ± 35.9	35.6 ± 3.4	0.6 ± 0.3	1118.0 ± 6.3	50.8 ± 2.5	4.8 ± 0.7	0	0	0
		CAR-DQN (Ours)	$\mathbf{1307.0 \pm 6.1}$	$\mathbf{1243.2 \pm 7.4}$	$\mathbf{908.2 \pm 17.0}$	$\mathbf{83.0 \pm 2.2}$	$\mathbf{1242.6 \pm 8.4}$	$\mathbf{970.8 \pm 9.6}$	$\mathbf{819.4 \pm 9.0}$	0	0	0
	Convex Relaxation	SA-DQN	1236.0 ± 1.4	1232.2 ± 2.5	1208.8 ± 1.7	1029.8 ± 34.6	1232.2 ± 2.5	1214.8 ± 2.6	1051.0 ± 35.5	0.991	0.409	0
		RADIAL-DQN	1341.8 ± 3.8	1341.8 ± 3.8	$\mathbf{1346.4 \pm 3.2}$	1092.6 ± 37.8	1341.8 ± 3.8	1328.6 ± 5.4	732.6 ± 11.5	0.982	0	0
		WocaR-DQN	1315.0 ± 6.1	1312.0 ± 6.1	1323.4 ± 2.2	1094.0 ± 10.2	1312.0 ± 6.1	1301.6 ± 3.9	1041.4 ± 17.4	0.987	0.093	0
		CAR-DQN (Ours)	$\mathbf{1349.6 \pm 3.0}$	$\mathbf{1347.6 \pm 3.6}$	1332.0 ± 7.3	$\mathbf{1191.0 \pm 9.0}$	$\mathbf{1347.4 \pm 3.6}$	$\mathbf{1338.0 \pm 2.9}$	$\mathbf{1233.6 \pm 5.0}$	0.974	0	0
RoadRunner	Standard	DQN	41492 ± 903	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	0	0	0
	PGD	SA-DQN	33380 ± 611	20482 ± 1087	0 ± 0	0 ± 0	24632 ± 812	614 ± 72	214 ± 26	0	0	0
		CAR-DQN (Ours)	$\mathbf{49700 \pm 1015}$	$\mathbf{43286 \pm 801}$	$\mathbf{25740 \pm 1468}$	$\mathbf{2574 \pm 261}$	$\mathbf{48908 \pm 1107}$	$\mathbf{35882 \pm 904}$	$\mathbf{23218 \pm 698}$	0	0	0
	Convex Relaxation	SA-DQN	46372 ± 882	44960 ± 1152	20910 ± 827	3074 ± 179	45226 ± 1102	25548 ± 737	12324 ± 529	0.819	0	0
		RADIAL-DQN	46224 ± 1133	45990 ± 1112	$\mathbf{42162 \pm 1147}$	$\mathbf{23248 \pm 499}$	46082 ± 1128	$\mathbf{42036 \pm 1048}$	$\mathbf{25434 \pm 756}$	0.994	0	0
		WocaR-DQN	43686 ± 1608	45636 ± 706	19386 ± 721	6538 ± 464	45636 ± 706	21068 ± 1026	15050 ± 683	0.956	0	0
		CAR-DQN (Ours)	$\mathbf{49398 \pm 1106}$	$\mathbf{49456 \pm 992}$	28588 ± 1575	15592 ± 885	$\mathbf{47526 \pm 1132}$	32878 ± 1898	21102 ± 1427	0.760	0	0

Insensitivity of learning rate and batch size. We compare the performance of CAR-DQN with different small batch size $(16, 32)$ and learning rate $(1.25 \times 10^{-4}, 6.25 \times 10^{-5})$ which are respectively used by Zhang et al. (2020); Liang et al. (2022). As shown in Figure 13, we can see CAR-DQN is insensitive to these parameters.

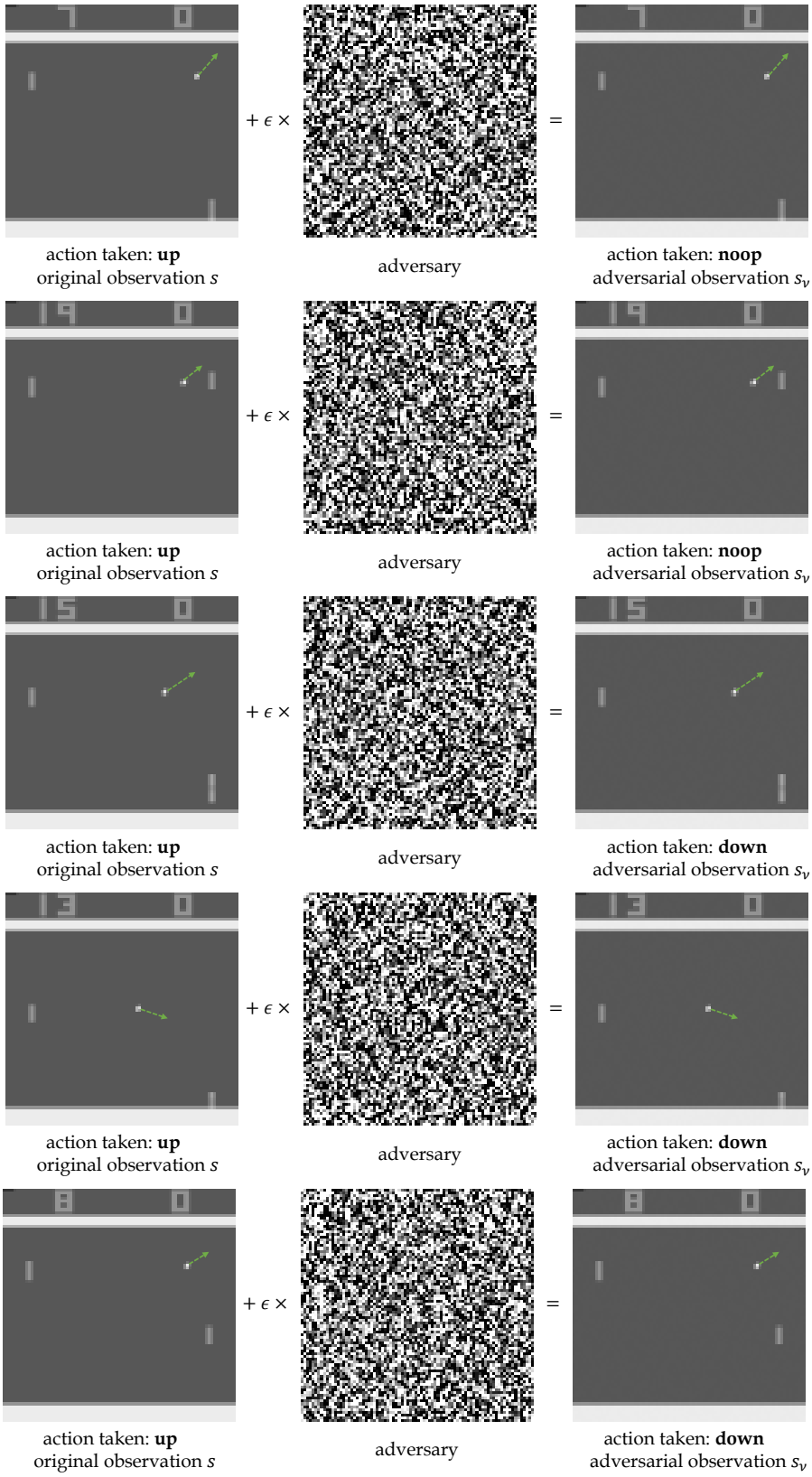

Figure 6. Examples of intrinsic states in Pong games. The direction of movement of the ball is marked.

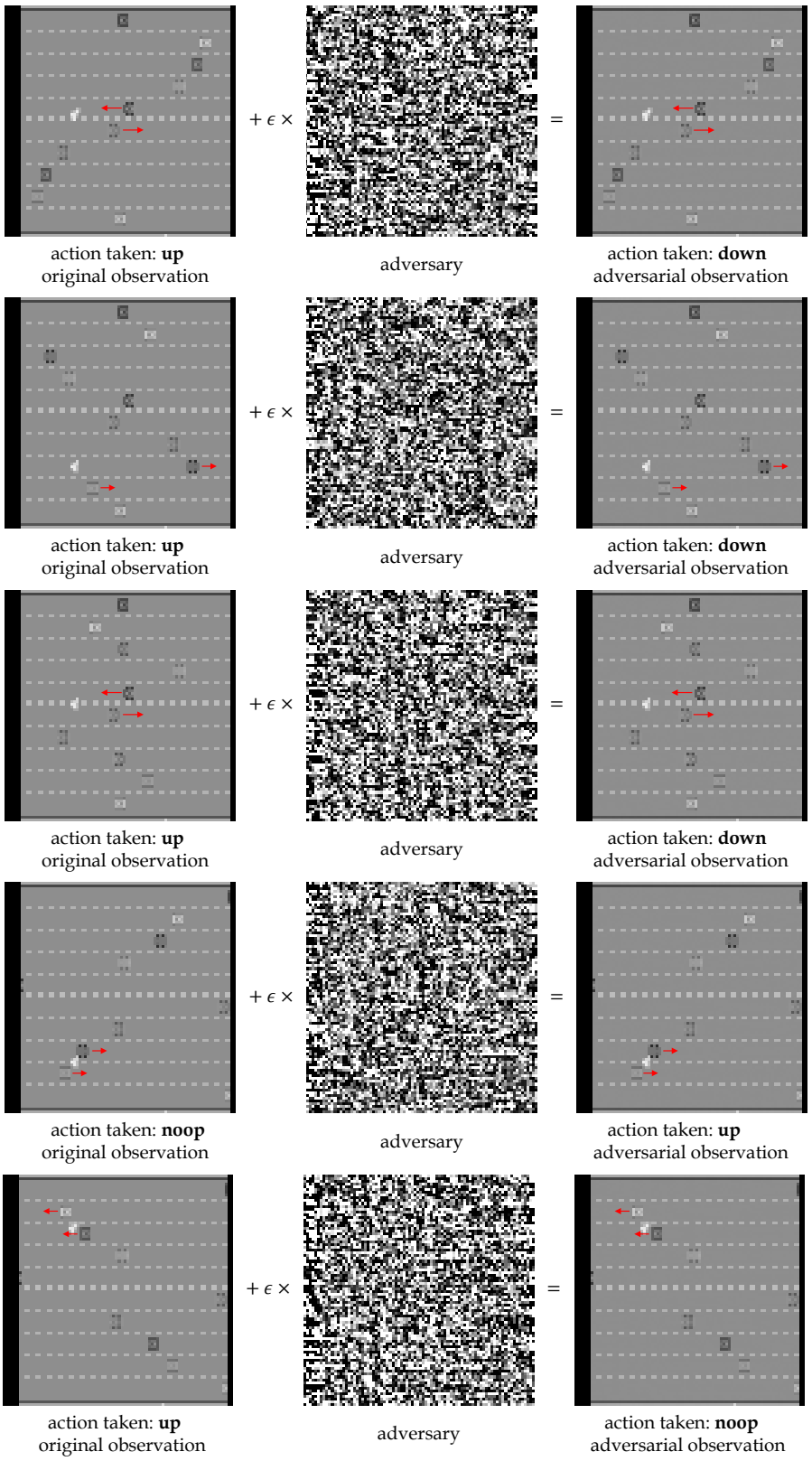

Figure 7. Examples of intrinsic states in Freeway games. The directions of movement of cars around the chicken are marked.

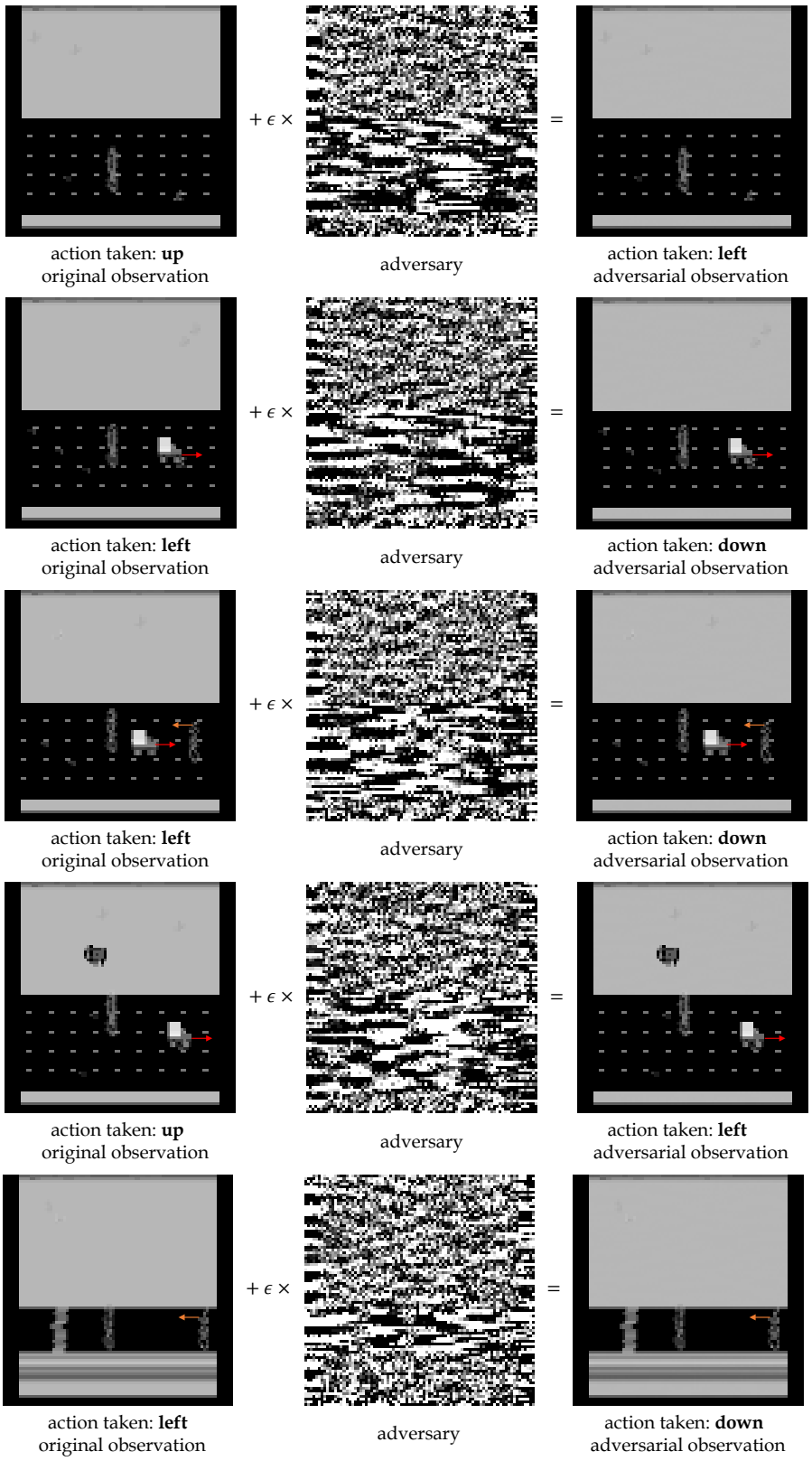

Figure 8. Examples of intrinsic states in Road Runner games. The directions of movement of trucks and the competitors are marked.

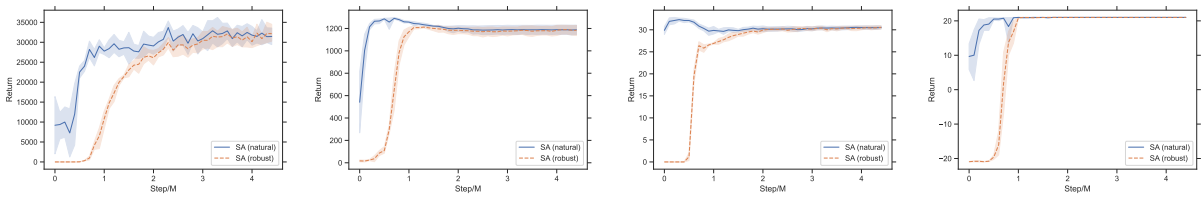

Figure 9. Natural and robustness performance exhibited by SA-DQN agents during the training process on 4 Atari games.

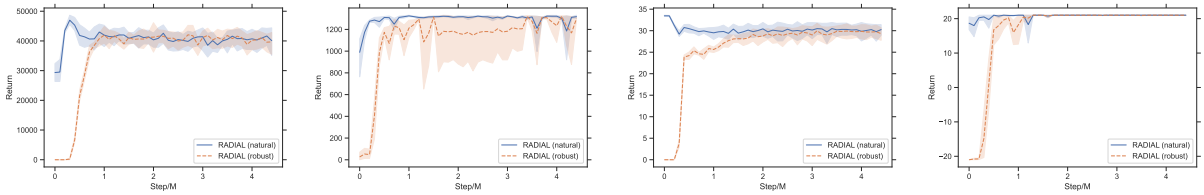

Figure 10. Natural and robustness performance exhibited by RADIAL-DQN agents during the training process on 4 Atari games.

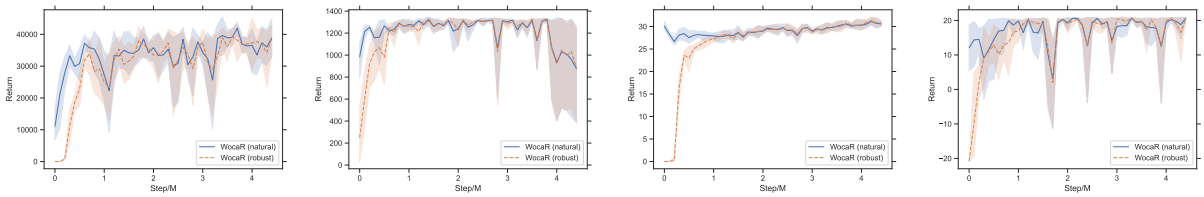

Figure 11. Natural and robustness performance exhibited by WocaR-DQN agents during the training process on 4 Atari games.

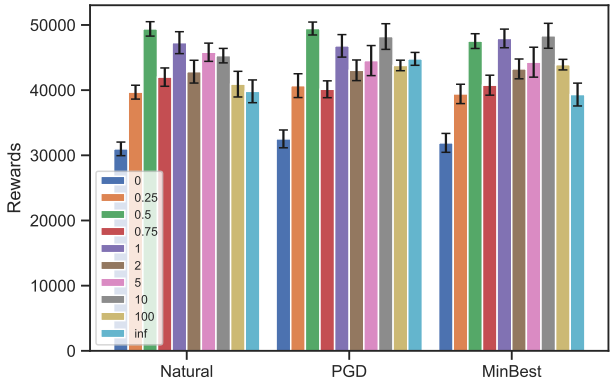

Figure 12. Natural, PGD attack and MinBest attack rewards of CAR-DQN with different soft coefficients on RoadRunner.

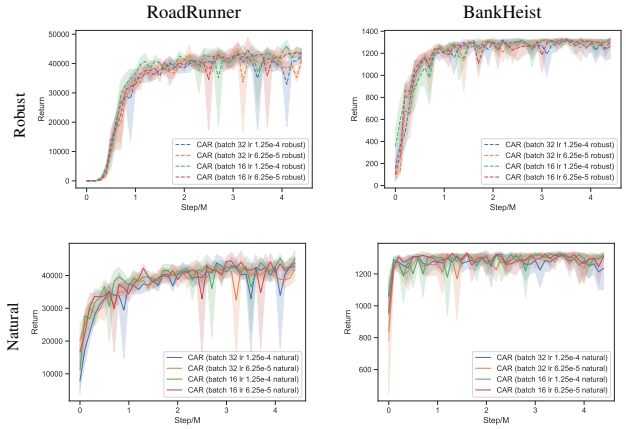

Figure 13. Episode rewards of CAR-DQN with different batch sizes and learning rates during training on RoadRunner and BankHeist with and without PGD attack.