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

Haoran Li

lihaoran21@mails.ucas.ac.cn

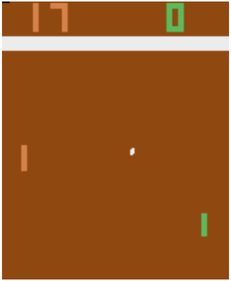
Joint work with Zicheng Zhang, Wang Luo, Congying Han, Yudong Hu, Tiande Guo, Shichen Liao

School of Mathematical Sciences, University of Chinese Academy of Sciences

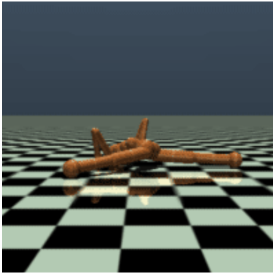
ICML 2024, Vienna

Vulnerability of Deep Reinforcement Learning

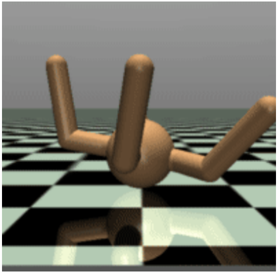
- Deep reinforcement learning agents are quite vulnerable to minor perturbations in their state observations.



DQN Pong
PGD attack
Reward: -21
(lowest)



PPO Humanoid
Robust Sarsa Attack
Reward: 719
(original 4386)



DDPG Ant
Robust Sarsa Attack
Reward: 258
(original 2462)

Image Source: Zhang et al. 2020.

This poses a major challenge for deploying DRL in the real world.

Markov Decision Process: Formulation of RL

- $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, \mathbb{P}, \gamma, \mu_0)$
  - State space $\mathcal{S} \subset \mathbb{R}^d$ is a compact set.
  - Action space $\mathcal{A}$ is a finite set.
  - Reward function $r : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$.
  - Transition dynamics $\mathbb{P} : \mathcal{S} \times \mathcal{A} \longrightarrow \Delta(\mathcal{S})$, where $\Delta(\mathcal{S})$ is the probability space over $\mathcal{S}$.
  - Discount factor $\gamma \in [0, 1)$.
  - Initial state distribution $\mu_0 \in \Delta(\mathcal{S})$.

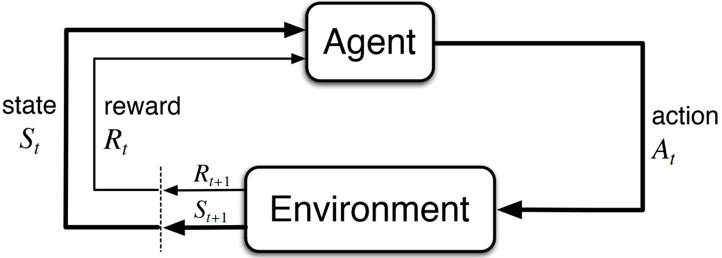


Image Source: Sutton and Barto 2018

Bellman Optimal Policy: Objective of RL

- Given a MDP $\mathcal{M}$, for any policy π, define
  - value function: $V^{\pi}_{\mathcal{M}}(s) = \mathbb{E}_{\tau\sim\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)|s_0 = s\right]$,
  - Q function: $Q^{\pi}_{\mathcal{M}}(s, a) = \mathbb{E}_{\tau\sim\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)|s_0 = s, a_0 = a\right]$,

  where trajectory $\tau = (s_0, a_0, r_0, s_1, a_1, r_1 \dots)$.
- Objective:

$$\max_{\pi} V^{\pi}_{\mathcal{M}}(s), \quad \text{for a given state } s.$$

Bellman Optimal Policy: Objective of RL

- Given a MDP $\mathcal{M}$, for any policy π, define
  - value function: $V^{\pi}_{\mathcal{M}}(s) = \mathbb{E}_{\tau\sim\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)|s_0 = s\right],$
  - Q function: $Q^{\pi}_{\mathcal{M}}(s, a) = \mathbb{E}_{\tau\sim\pi,\mathbb{P}}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t, a_t)|s_0 = s, a_0 = a\right],$

  where trajectory $\tau = (s_0, a_0, r_0, s_1, a_1, r_1 \dots)$.
- Objective:

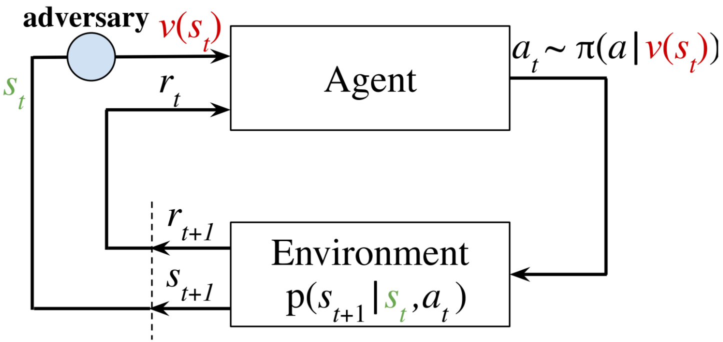
$$\max_{\pi} V^{\pi}_{\mathcal{M}}(s), \quad \text{for a given state } s.$$

- There exists a *stationary* and *deterministic* policy π^* that simultaneously maximizes $V^{\pi}(s)$ for *all* $s \in \mathcal{S}$, and $Q^* := Q^{\pi^*}$ satisfies the Bellman optimality equations, i.e.

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

State-Adversarial MDP (Zhang et al. 2020): Elegant Formulation of RL against Perturbations on Observations

- $\mathcal{M}_\nu = (\mathcal{S}, \mathcal{A}, r, \mathbb{P}, \nu, \gamma, \mu)$
  - Adversary $\nu : \mathcal{S} \longrightarrow \mathcal{S}, \quad s \longmapsto s_\nu \in B(s)$.
- value function: $V^{\pi\circ\nu}(s) = \mathbb{E}_{\pi\circ\nu,\mathbb{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s\right]$,
- Q function: $Q^{\pi\circ\nu}(s, a) = \mathbb{E}_{\pi\circ\nu,\mathbb{P}}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\right]$.



Challenges of SA-MDP

Objective of SA-MDP: **Optimal Robust Policy (ORP)**

- Strongest Adversary: Given a policy π, the strongest adversary $\nu^*(\pi) = \arg\min_\nu V^{\pi\circ\nu}$ exists.
- An ORP π^* should maximize the value function against this strongest adversary for all states, i.e. $V^{\pi^*\circ\nu^*(\pi^*)}(s) = \max_\pi V^{\pi\circ\nu^*(\pi)}(s), \ \forall s.$

Challenges of SA-MDP

Objective of SA-MDP: **Optimal Robust Policy (ORP)**

- Strongest Adversary: Given a policy π, the strongest adversary $\nu^*(\pi) = \arg\min_\nu V^{\pi\circ\nu}$ exists.
- An ORP π^* should maximize the value function against this strongest adversary for all states, i.e. $V^{\pi^*\circ\nu^*(\pi^*)}(s) = \max_\pi V^{\pi\circ\nu^*(\pi)}(s),\ \forall s$.

However, unlike standard MDPs, ORP of SA-MDPs may not exist.

- Deterministic policies are *not* sufficient to achieve ORP.
- Even stochastic ORP may *not* always exist.

This reveals a potential conflict between robustness and policy optimality, making it challenging to enforce strict robustness constraints.

When does the ORP exist?

When does the ORP exist?

Under the Consistency Assumption of Policy, ORP exists, and aligns with the Bellman optimal policy!

Consistency Assumption of Policy (CAP)

Define the *intrinsic state ϵ-neighbourhood* for any state s as

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

Assumption (Consistency Assumption of Policy)

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

Consistency Assumption of Policy (CAP)

Define the *intrinsic state ϵ-neighbourhood* for any state s as

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

Assumption (Consistency Assumption of Policy)

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

- The set of states violating CAP is nearly empty.

Theorem (Rationality of the CAP)

Let $\mathcal{S}_{nin}$ denote the set of states violating the CAP. Then, we have that $\mathcal{S}_{nin} \subseteq \mathcal{S}_{nu} \cup \mathcal{S}_0 + B_\epsilon$, where $\mathcal{S}_{nu}$ is the state set where the optimal action is not unique, and $\mathcal{S}_0$ is the set of discontinuous points that cause the optimal action to change. These sets are nearly empty in practical tasks.

Existence of the Optimal Robust Policy under CAP

Define the consistent adversarial robust (CAR) operator as

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

- $\mathcal{T}_{car}$ is not contractive.

Existence of the Optimal Robust Policy under CAP

Define the consistent adversarial robust (CAR) operator as

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

- $\mathcal{T}_{car}$ is not contractive.

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

- *If the optimal adversarial action-value function $Q^{\pi^* \circ \nu^*(\pi^*)}$ under the strongest adversary exists for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then it is the fixed point of CAR operator.*
- *If the CAP holds, then Q^* is the fixed point of CAR operator $\mathcal{T}_{car}$. Furthermore, Q^* is the optimal adversarial action-value function under the strongest adversary, i.e., $Q^*(s,a) = Q^{\pi^* \circ \nu^*(\pi^*)}(s,a)$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

Bellman optimal policy doubles as ORP!
— Improving adversarial robustness does **not require sacrificing natural performance**.

Bellman optimal policy doubles as ORP!
— Improving adversarial robustness does **not require sacrificing natural performance**.

Why do conventional DRL algorithms, which aim for the Bellman optimal policy, fail to ensure adversarial robustness?

Bellman optimal policy doubles as ORP!
— Improving adversarial robustness does **not require sacrificing natural performance**.

Why do conventional DRL algorithms, which aim for the Bellman optimal policy, fail to ensure adversarial robustness?

Infinity-error is necessary!

Previously 1-error ✗

L^∞ is Necessary for Adversarial Robustness

- For any Banach space $\mathcal{B}$, if $\|Q_\theta - Q^*\|_{\mathcal{B}} = 0$, then $Q_\theta = Q^*$.
- However, in practice, $0 < \|Q_\theta - Q^*\|_{\mathcal{B}} = \delta \ll 1 \quad \Longrightarrow \quad Q_\theta = ?$

L^∞ is Necessary for Adversarial Robustness

- For any Banach space $\mathcal{B}$, if $\|Q_\theta - Q^*\|_{\mathcal{B}} = 0$, then $Q_\theta = Q^*$.
- However, in practice, $0 < \|Q_\theta - Q^*\|_{\mathcal{B}} = \delta \ll 1 \quad \Longrightarrow \quad Q_\theta = ?$

Theorem (Necessity of L^∞-norm)

Let $\mathcal{S}_{sub}^{Q}$ denote the set of states where the greedy policy according to Q is suboptimal and $\mathcal{S}_{adv}^{Q}$ denote the set of states within whose ϵ-neighborhood there exists the adversarial state. 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 satisfying $\|Q - Q^*\|_p \leq \delta$ such that $\mu\left(\mathcal{S}_{sub}^{Q}\right) = O(\delta)$ yet $\mu\left(\mathcal{S}_{adv}^{Q}\right) = \mu\left(\mathcal{S}\right)$.*

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

Stability of Nonlinear Functional Equations

- $\|Q_\theta - Q^*\|_{\mathcal{B}}$ cannot be directly measured.
- Instead, minimize the Bellman error $\|\mathcal{T}_B Q_\theta - Q_\theta\|_{\mathcal{B}'}$ to train Q_θ, where $\mathcal{T}_B$ is the Bellman optimality operator.
- We have shown that $\mathcal{B}$ **should be** $L^\infty(\mathcal{S} \times \mathcal{A})$**.**

$$\mathcal{B}' = ?$$

Stability of Nonlinear Functional Equations

- $\|Q_\theta - Q^*\|_{\mathcal{B}}$ cannot be directly measured.
- Instead, minimize the Bellman error $\|\mathcal{T}_B Q_\theta - Q_\theta\|_{\mathcal{B}'}$ to train Q_θ, where $\mathcal{T}_B$ is the Bellman optimality operator.
- We have shown that $\mathcal{B}$ **should be** $L^\infty(\mathcal{S}\times\mathcal{A})$**.**

$$\mathcal{B}' = \,?$$

- $\|\mathcal{T}_B Q_\theta - Q_\theta\|_{\mathcal{B}'} = 0 \quad \Longrightarrow \quad Q_\theta = Q^*.$
- $0 < \|\mathcal{T}_B Q_\theta - Q_\theta\|_{\mathcal{B}'} = \delta \ll 1 \quad \Longrightarrow \quad \|Q_\theta - Q^*\|_{\mathcal{B}} < \,?$

Definition (Stability of Functional Equations)

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 $\|Q - Q^*\|_{\mathcal{B}_2} < C\|\mathcal{T}Q - Q\|_{\mathcal{B}_1}$, then we say a nonlinear functional equation $\mathcal{T}Q = Q$ is $(\mathcal{B}_1, \mathcal{B}_2)$-stable.

- 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 $\|\mathcal{T}Q - Q\|_{\mathcal{B}_1} \longrightarrow 0$, $\forall Q \in \mathcal{B}_1 \cap \mathcal{B}_2$.

Stability of Bellman Optimality Equations

Theorem (Stable and Unstable Properties of $\mathcal{T}_B$ in L^p Spaces)

- *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:*

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

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

$\mathcal{B}'$ should also be $L^\infty(\mathcal{S}\times\mathcal{A})$.

Stability of Deep Q-network (DQN) in Practice

Theorem (Stable and Unstable Properties of $\mathcal{T}_B$ in $(p, d^{\pi}_{\mu_0})$ Spaces)

- *For any MDP $\mathcal{M}$ and policy π, let* $C_{\mathbb{P},p} := \sup_{(s,a)} \left\| \mathbb{P}(\cdot \mid s,a) \right\|_{L^{\frac{p}{p-1}}}$. *Assume* $C_{d^{\pi}_{\mu_0}} := \inf_{(s,a)} d^{\pi}_{\mu_0}(s,a) > 0$ *and p and q satisfy:*

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

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

$\left\| \mathcal{T}_B Q_\theta - Q_\theta \right\|_{\infty, d^{\pi_\theta}_{\mu_0}}$ is crucial for ensuring both the natural performance and robustness of DQN.

Consistent Adversarial Robust DQN (CAR-DQN)

- Theoretical Objective: Bellman Infinity-error, i.e.,

$$\mathcal{L}_{car}(\theta) = \|\mathcal{T}_B Q_\theta - Q_\theta\|_{\infty, d_{\mu_0}^{\pi_\theta}}.$$

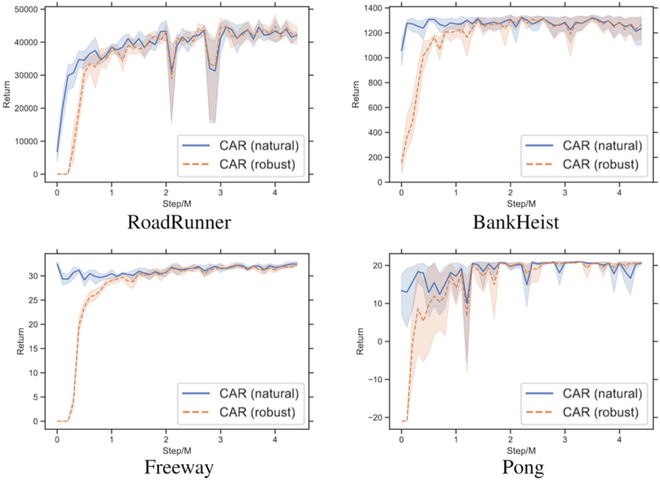
- Surrogate Objective:

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

where

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

Natural and Robust Returns Show Consistency



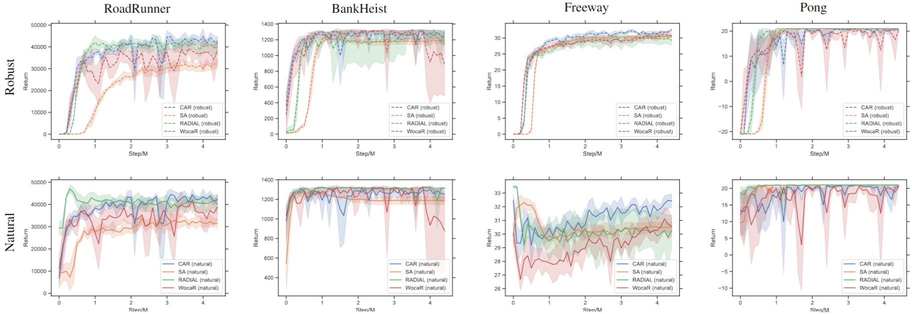
CAR-DQN exhibits consistent natural and robust performance.

Bellman Infinity-error is Necessary

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

Ablation studies across different L^p spaces
confirm our theoretical findings on
the necessity of the Bellman infinity-error for robustness.

CAR-DQN Shows Superior Natural and Robust Returns



Model		Pong Natural Reward	Pong PGD $\epsilon = 1/255$	Pong MinBest $\epsilon = 1/255$	Pong ACR	BankHeist Natural Reward	BankHeist PGD $\epsilon = 1/255$	BankHeist MinBest $\epsilon = 1/255$	BankHeist ACR
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 Natural Reward	Freeway PGD $\epsilon = 1/255$	Freeway MinBest $\epsilon = 1/255$	Freeway ACR	RoadRunner Natural Reward	RoadRunner PGD $\epsilon = 1/255$	RoadRunner MinBest $\epsilon = 1/255$	RoadRunner ACR
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

Summary

- Under the mild consistency assumption of policy, the optimal robust policy exists and aligns with the Bellman optimal policy.
- This theoretically highlights that improving the adversarial robustness does not require sacrificing natural performance.
- The Bellman infinity-error is necessary for achieving ORP, while prior DRL algorithms lack robustness due to their use of 1-error.
- CAR-DQN employs a surrogate objective of the Bellman infinity-error to learn both natural return and robustness.

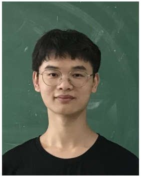
Q & A

Feel free to contact Haoran Li! Welcome collaboration!
Contact: @leolmia or `lihaoran21@mails.ucas.ac.cn`

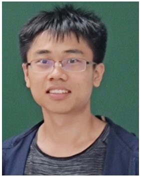

Tiande Guo


Congying Han


Zicheng Zhang


Wang Luo


Yudong Hu


Shichen Liao

Paper: arxiv.org/abs/2402.02165
Code: github.com/leoranlmia/CAR-DQN

Selected Reference

[1] Huan Zhang et al. "Robust deep reinforcement learning against adversarial perturbations on state observations". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21024–21037.

[2] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[3] Tuomas Oikarinen et al. "Robust deep reinforcement learning through adversarial loss". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 26156–26167.

[4] Yongyuan Liang et al. "Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 22547–22561.