EXPLORING THE ROBUSTNESS OF DISTRIBUTIONAL REINFORCEMENT LEARNING AGAINST NOISY STATE OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

In real scenarios, state observations that an agent observes may contain measurement errors or adversarial noises, misleading the agent to take suboptimal actions or even collapse while training. In this paper, we study the training robustness of distributional Reinforcement Learning (RL), a class of state-of-the-art methods that estimate the whole distribution, as opposed to only the expectation, of the total return. Firstly, we propose State-Noisy Markov Decision Process (SN-MDP) in the tabular case to incorporate both random and adversarial state observation noises, in which the contraction of both expectation-based and distributional Bellman operators is derived. Beyond SN-MDP with the function approximation, we theoretically characterize the bounded gradient norm of histogram-based distributional loss, accounting for the better training robustness of distribution RL. We also provide stricter convergence conditions of the Temporal-Difference (TD) learning under more flexible state noises, as well as the sensitivity analysis by the leverage of influence function. Finally, extensive experiments on the suite of games show that distributional RL enjoys better training robustness compared with its expectation-based counterpart across various state observation noises.

1 INTRODUCTION

Learning robust and high-performance policies for continuous state-action reinforcement learning (RL) domains is crucial to enable the successful adoption of deep RL in robotics, autonomy, and control problems. However, recent works have demonstrated that deep RL algorithms are vulnerable either to model uncertainties or external disturbances (Huang et al., 2017; Pattanaik et al., 2017; Ilahi et al., 2020; Chen et al., 2019; Zhang et al., 2020; Shen et al., 2020; Singh et al., 2020; Guan et al., 2020). Particularly, model uncertainties normally occur in a noisy reinforcement learning environment where the agent often encounters systematic or stochastic measurement errors on state observations, such as the inexact locations and velocity obtained from the equipped sensors of a robot. On the other hand, external disturbances are normally adversarial in nature. For instance, the adversary can construct adversarial perturbations on state observations to degrade the performance of deep RL algorithms. These two factors lead to noisy state observations that influence the performance of algorithms, precluding the success of reinforcement learning in real-world applications.

Existing works mainly focus on improving the robustness of algorithms in the *test environment* with noisy state observations. Smooth Regularized Reinforcement Learning (Shen et al., 2020) introduced a regularization to enforce smoothness in the learned policy, and thus improved its robustness against measurement errors in the test environment. Similarly, the State-Adversarial Markov decision process (SA-MDP) (Zhang et al., 2020) was proposed and the resulting principled policy regularization enhances the adversarial robustness of various kinds of RL algorithms against adversarial noisy state observations. However, both of these works assumed that the agent can access *clean* state observations during training, which is normally not feasible when the environment is inherently noisy, such as unavoidable measurement errors. Thus, the maintenance and formal analysis of policies robust to noisy state observations during *training* is a worthwhile area of research.

On the other hand, recent distributional reinforcement learning algorithms, including C51 (Bellemare et al., 2017), Quantile-Regression DQN (Dabney et al., 2018b), Implicit Quantile Net-

works (Dabney et al., 2018a) and Moment-Matching DQN (Nguyen et al., 2020), constantly set new records in Atari games, gaining huge attention in the research community. However, existing literature mainly focuses on the performance of algorithms, other benefits, including the robustness in the noisy environment, of distributional RL algorithms are less studied. As distributional RL can leverage additional information about distribution that captures the uncertainty of the environment more accurately, it is natural to expect that distributional RL with this better representation capability can be less vulnerable to the noisy environment while training, which motivates our research.

In this paper, we investigate the robustness of distributional RL against various kinds of state observation noises encountered during training. Firstly, we propose a general State-Noisy MDP in the tabular setting, in which we prove the convergence of distributional Bellman operator. We further extend SN-MDP to the function approximation case by considering more complex noisy state observations. Notably, we characterize the Lipschitz continuity *blessing* resulting from the Histogram distributional loss in distributional RL, which leads to a bounded gradient norm. This better behaved gradient mitigates the impact of noisy states on the objective function, accounting for the less vulnerability of distributional RL while training. Moreover, we also provide the convergence conditions of TD learning under noisy state observations as well as a sensitivity analysis of state noises on the learning of the function approximator via the influence function. Finally, extensive experiments demonstrate that distributional RL algorithms tend to achieve better robust performance in the presence of more complex state observation noises compared with its expectation-based counterpart that may even diverge in some cases. These empirical results in Section 5 echo our previous theoretical results in both Section 3 and 4. Overall, the training robustness advantage of distributional RL algorithms we revealed facilitates their deployment especially in the noisy environment.

2 BACKGROUND

2.1 DISTRIBUTIONAL REINFORCEMENT LEARNING

In the tabular setting without noisy states, the agent's interaction with its environment can be naturally modeled as a standard Markov Decision Process (MDP), a 5-tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$. $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the environment transition dynamics, $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function and $\gamma \in (0, 1)$ is the discount factor.

Value Function vs Value Distribution. Firstly, we denote the *return* where $s_t = s$ as $Z^\pi(s) = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$, representing the cumulative rewards following a policy π, and r_{t+k+1} is reward scalar obtained in the step $t + k + 1$. In the algorithm design, traditional expectation-based RL normally focuses on *value function* $V^\pi(s)$, the expectation of the random variable $Z^\pi(s)$:

$$V^\pi(s) := \mathbb{E}\left[Z^\pi(s)\right] = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \mid s_t = s\right]. \tag{1}$$

In contrast, in the distributional RL setting, we focus on the *value distribution*, the full distribution of $Z^\pi(s)$, and the *state-action value distribution* $Z^\pi(s, a)$ in the control problem where $s_t = s, a_t = a$. Both of these distributions can better capture the uncertainty of returns in the MDP beyond just its expectation (Dabney et al., 2018a; Mavrin et al., 2019).

Distributional Bellman Operator. In expectation-based RL, we update the value function via the Bellman operator $\mathcal{T}^\pi$, while in distributional RL, the updating is on the value distribution via the *distributional Bellman operator* $\mathfrak{T}^\pi$. To derive $\mathfrak{T}^\pi$, we firstly define the transition operator $P^\pi : \mathcal{Z} \to \mathcal{Z}$:

$$\mathcal{P}^\pi Z(s, a) :\overset{D}{=} Z\left(S', A'\right), S' \sim P(\cdot|s, a), A' \sim \pi\left(\cdot|S'\right), \tag{2}$$

where we use capital letters S' and A' to emphasize the random nature of both, and $:\overset{D}{=}$ indicates convergence in distribution. For simplicity, we denote $Z^\pi(s, a)$ by $Z(s, a)$. Thus, the distributional Bellman operator $\mathfrak{T}^\pi$ is defined as:

$$\mathfrak{T}^\pi Z(s, a) :\overset{D}{=} R(s, a, S') + \gamma \mathcal{P}^\pi Z(s, a). \tag{3}$$

More importantly, $\mathfrak{T}^\pi$ is still a contraction for policy evaluation under the maximal form of the Wasserstein metric d_p (more details are given in Appendix A) over the true and parametric value distributions (Bellemare et al., 2017; Dabney et al., 2018b).

2.2 TWO KINDS OF NOISY STATE OBSERVATIONS

We investigate both random and adversarial training robustness, *i.e., the performance of RL algorithms under these two types of noisy state observations*, between the expectation-based and distributional RL algorithms. We consider continuous state observations with continuous noises. In the random noisy state case, we apply Gaussian noises with mean 0 and different standard deviations to state features to simulate the measurement error stemming from various sources.

In the adversarial state perturbation setting, we construct white-box adversarial perturbations on state observations for the current policy during training, following the strategy proposed in (Huang et al., 2017; Pattanaik et al., 2017) that leveraged the gradient information of an engineered loss function. In particular, we denote a_w^t as the "worst" action, with the lowest probability from the current policy $\pi_t(a|s)$ in the training step t. Thus, the optimal adversarial perturbation η_t, constrained in an ϵ-ball, can be derived by minimizing the objective function J:

$$\min_{\eta} J(s_t + \eta, \pi_t) = -\sum_{i=1}^{n} p_i^t \log \pi_t(a_i|s_t + \eta), s.t. \|\eta\| \leq \epsilon, \tag{4}$$

where $p_i^t = 1$ if i corresponds to the index of the least-chosen action, i.e. the w-th index in the vector a, otherwise $p_i^t = 0$. In other words, we construct a target one-hot action p^t with 1 assigned to the index of the least-chosen action. Through this minimization in the form of the cross entropy loss, we can construct the state perturbations η_t that can force the policy to choose the least-chosen action a_w^t in each t step.

3 TABULAR CASE: STATE-NOISY MARKOV DECISION PROCESS

In this section, we extend State-Adversarial MDP (Zhang et al., 2020) to a more general State-Noisy Markov Decision Process (SN-MDP), and particularly provide a proof of the convergence and contraction of distributional Bellman operator in this setting.

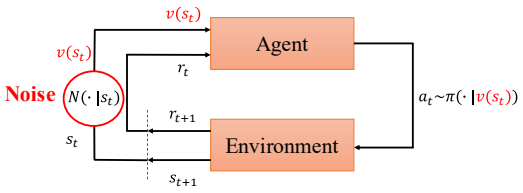

Figure 1: State-Noisy Markov Decision Process. $v(s_t)$ is perturbed by the noise mechanism N.

3.1 DEFINITIONS

As shown in Figure 1, SN-MDP is a 6-tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma, N)$, where the noise generating mechanism $N(\cdot|s)$ maps the state from s to $v(s)$ using either random or adversarial noise with the Markovian and stationary probability $N(v(s)|s)$. It is worthwhile to note that the explicit definition of the noise mechanism N here is based on discrete state transitions, but the analysis can be naturally extended to the continuous case if we let the state space go to infinity. Moreover, let $\mathcal{B}(s)$ be the set that contains the allowed noise space for the noise generating mechanism N, i.e., $v(s) \in \mathcal{B}(s)$.

Following the setting in (Zhang et al., 2020), we only manipulate state observations but do not change the underlying environment transition dynamics based on s or the agent's actions directly. As such, our SN-MDP is more suitable to model the random measurement error, e.g., sensor errors and equipment inaccuracies, and adversarial state observation perturbations in safety-critical scenarios.

3.2 ANALYSIS OF SN-MDP FOR EXPECTATION-BASED RL

We define the value function $\tilde{V}_{\pi \circ N}$ given π in SN-MDP. The Bellman Equations are given by:

$$\tilde{V}_{\pi \circ N}(s) = \sum_{a} \sum_{v(s)} N(v(s)|s)\pi(a|v(s)) \sum_{s'} p(s'|s,a) \cdot \left[R(s,a,s') + \gamma \tilde{V}_{\pi \circ N}(s') \right]. \tag{5}$$

The random noise transits s into $v(s)$ with a certain probability and the adversarial noise is the special case of $N(v(s)|s)$ where $N(v^*(s)|s) = 1$ if $v^*(s)$ is the optimal adversarial noisy state given s, and $N(v(s)|s) = 0$ otherwise. We denote Bellman operators under random noise mechanism $N^r(\cdot|s)$ and adversarial noise mechanism $N^*(\cdot|s)$ as $\mathcal{T}_r^\pi$ and $\mathcal{T}_a^\pi$, respectively. This implies that

$\mathcal{T}_r^\pi \tilde{V}_{\pi \circ N} = \tilde{V}_{\pi \circ N^r}$ and $\mathcal{T}_a^\pi \tilde{V}_{\pi \circ N} = \tilde{V}_{\pi \circ N^*}$. We extend Theorem 1 in Zhang et al. (2020) to both random and adversarial noise scenario, and immediately obtain that both $\mathcal{T}_r^\pi$ and $\mathcal{T}_a^\pi$ are contraction operators in SN-MDP. We explain this in Theorem 5 of Appendix B.

The pivotal conclusion from Theorem 5 is $\mathcal{T}_a^\pi \tilde{V}_{\pi \circ N} = \min_N \tilde{V}_{\pi \circ N}$. This implies that the adversary attempts to minimize the value function, forcing the agent to select the worse-case action among the allowed transition probability space $N(\cdot|s)$ for each state s. The main proof idea is that Bellman updates in SN-MDP result in the convergence to the value function for another "merged" policy π' where $\pi'(a|s) = \sum_{v(s)} N(v(s)|s)\pi(a|v(s))$. The value function for the merged policy might ba far away from that for the original policy π, which tends to worsen the performance of RL algorithms.

3.3 Analysis of SN-MDP in distributional RL

In the SN-MDP setting for distributional RL, the new distributional Bellman equations use new transition operators in place of $\mathcal{P}^\pi$ in Eq. 2. The new transition operators $\mathcal{P}_r^\pi$ and $\mathcal{P}_a^\pi$, for the random and adversarial settings, are defined as:

$$\mathcal{P}_r^\pi Z_N(s,a) :\overset{D}{=} Z_{N^r}(S', A'), A' \sim \pi(\cdot|V(S')), \text{and } \mathcal{P}_a^\pi Z_N(s,a) :\overset{D}{=} Z_{N^*}(S', A'), A' \sim \pi(\cdot|V^*(S')), \tag{6}$$

where $V(S') \sim N^r(\cdot|S')$ is the state random variable after the transition, and $V^*(S')$ is attained from $N^*(\cdot|S')$ under the optimal adversary. Besides, $S' \sim P(\cdot|s,a)$. Thus, the corresponding new distributional Bellman operators $\mathfrak{T}_r^\pi$ and $\mathfrak{T}_a^\pi$ are:

$$\mathfrak{T}_r^\pi Z_N(s,a) :\overset{D}{=} R(s,a,S') + \gamma \mathcal{P}_r^\pi Z_N(s,a), \text{and } \mathfrak{T}_a^\pi Z_N(s,a) :\overset{D}{=} R(s,a,S') + \gamma \mathcal{P}_a^\pi Z_N(s,a). \tag{7}$$

In this sense, four sources of randomness define the new compound distribution in the SN-MDP: (1) randomness of reward, (2) randomness in the new environment transition dynamics $\mathcal{P}_r^\pi$ or $\mathcal{P}_a^\pi$ that additionally includes (3) the stochasticity of the noisy transition N, and (4) the random next-state value distribution $Z(S', A')$. Besides, the premise of the robustness of distributional RL against noisy state observations lies in the convergence of the new derived distribution Bellman Operators in SN-MDP setting. We proved this convergence and contraction for policy evaluation in Theorem 1.

Theorem 1. *(Convergence and Contraction of Distributional Bellman operators in the SN-MDP) Given a policy π, we define the distributional Bellman operators $\mathfrak{T}_r^\pi$ and $\mathfrak{T}_a^\pi$ in Eq. 7, and consider the Wasserstein metric d_p, the following results hold.*

(1) $\mathfrak{T}_r^\pi$ is a contraction under the maximal form of d_p.

(2) $\mathfrak{T}_a^\pi$ is also a contraction under the maximal form of d_p, following the greedy adversarial rule, i.e., $N^*(\cdot|s') = \arg\min_{N(\cdot|s')} \mathbb{E}\left[Z(s', a')\right]$ where $a' \sim \pi(\cdot|V(s'))$ and $V(s') \sim N(\cdot|s')$.

We provide the proof in Appendix C. The convergence of distributional Bellman operators in the SN-MDP is one of our main contributions. This result allows us to deploy distributional reinforcement learning algorithms comfortably even in settings with noisy state observations.

4 Function Approximation Case: Noisy Settings Beyond SN-MDP

In real scenarios, especially safety-critical cases, perturbations on state observations can be more complicated. For instance, the adversary might perform attacks at certain intervals, yielding unbalanced state observation pairs with a perturbed current state and a benign next state and vice versa. This type of unbalanced perturbations is outside the scope of State-Noisy MDP we analyzed in the last section and can have different impacts on the convergence of expectation-based and distributional RL algorithms. In this section, we firstly characterize the robustness *blessing* of distributional RL based on Histogram distributional loss (Imani & White, 2018), and then analyze the impact of more complex state observations on TD convergence and further conduct a sensitivity analysis by the influence function.

Notation. We derive theoretical results with the linear function approximator. For the expectation-based RL, the value estimate $\hat{v} : \mathcal{S} \times \mathbb{R}^d \to \mathbb{R}$ is formed simply as the inner product between state features $\mathbf{x}(s)$ and weights $\mathbf{w} \in \mathbb{R}^d$, given by $\hat{v}(s, \mathbf{w}) \overset{\text{def}}{=} \mathbf{w}^\top \mathbf{x}(s)$. At each step, the state feature can be rewritten as $\mathbf{x}_t \overset{\text{def}}{=} \mathbf{x}(S_t) \in \mathbb{R}^d$. The distributional RL setting is given in Section 4.1.

4.1 ROBUSTNESS BLESSING FOR DISTRIBUTIONAL RL

We show that in the function approximation setting, the distributional loss in distributional RL can additionally yield Lipschitz continuity regarding state features, thus leading to more stable gradients relatively to expectation-based RL. Simply, in distributional RL our goal is to minimize $\mathcal{L}\left(Z_\theta, \mathfrak{T}Z_\theta\right)$, where $\mathfrak{T}$ is the distributional Bellman operator. Here we leverage histogram to parameterize the distribution Z_θ based on KL divergence as $\mathcal{L}$, yielding the *histogram distributional loss* (Imani & White, 2018). The histogram distributional loss $\mathcal{L}(Z_\theta, \mathfrak{T}Z_\theta)$ between Z_θ and $\mathfrak{T}Z_\theta$ can be derived as $\mathcal{L}_\theta = -\sum_{i=1}^{k} p_i \log f_i^\theta(\mathbf{x}(s))$, where the support of $\mathbf{x}(s)$ is uniformly partitioned into k bins. We let function $f : \mathcal{X} \to [0,1]^k$ provide k-dimensional vector $f(\mathbf{x}(s))$ of the coefficients indicating the probability the target is in that bin given $\mathbf{x}(s)$, and use *softmax* based on the linear approximation $\mathbf{x}(s)^\top \theta_i$ to express f, i.e., $f_i(\mathbf{x}(s)) = \exp\left(\mathbf{x}(s)^\top \theta_i\right) / \sum_{j=1}^{k} \exp\left(\mathbf{x}(s)^\top \theta_j\right)$. Moreover, $\theta = \{\theta_1, ..., \theta_k\}$ and the target probability p_i is the cumulative probability increment of target distribution $\mathfrak{T}Z_\theta$ within the i-th bin. Details of the histogram distributional loss are given in Appendix D.

Based on this histogram distributional loss in distribution RL, we obtain Theorem 2 (proof in Appendix D), which reveals that the distribution loss can result in additional Lipschitz continuity property that bounds the norm of gradient over state features $\mathbf{x}(s)$:

Theorem 2. *(Lipschitz Continuity of distributional RL) Consider the histogram distributional loss* $\mathcal{L}_\theta = -\sum_{i=1}^{k} p_i \log f_i^\theta(\mathbf{x}(s))$*, where* $f_i(\mathbf{x}(s)) = \exp\left(\mathbf{x}(s)^\top \theta_i\right) / \sum_{j=1}^{k} \exp\left(\mathbf{x}(s)^\top \theta_j\right)$ *parameterized by* $\theta = \{\theta_1, ..., \theta_k\}$*. Assume* $\|\theta_i\| \leq l$ *for* $\forall i = 1, .., k$*, then* $\mathcal{L}_\theta$ *is* kl*-Lipschitz continuous w.r.t.* $\mathbf{x}(s)$*, yielding a bounded norm of its gradient, i.e.,* $\left\| \frac{\partial}{\partial \mathbf{x}(s)} \sum_{j=1}^{k} p_j \log f_j^\theta(\mathbf{x}(s)) \right\| \leq kl$*.*

Note that the norm of gradient in expectation-based RL with the linear function approximation can be written as $|U_t - \mathbf{w}_t^\top \mathbf{x}_t| \|\mathbf{w}_t\|$, where the target U_t can be evaluated by either Monte Carlo method or TD learning (Sutton & Barto, 2018). However, this upper bound can be arbitrary large as there is no restriction on $|U_t - \mathbf{w}_t^\top \mathbf{x}_t|$. In conclusion, Theorem 2 shows that distributional loss in distributional RL can additionally enjoy kl-Lipschitz continuity compared with the expectation-based RL. The bounded norm of gradient regarding state features mitigates the impact of noisy state observations on the objective function while training, therefore yielding better training robustness.

4.2 TD CONVERGENCE UNDER NOISY STATE OBSERVATIONS

Let $\mu(s)$ be the stationary distribution under the policy π and $p(s'|s)$ be the transition probability from s to s' satisfying $p(s'|s) = \sum_a \pi(a|s) p(s'|s,a)$. We analyze conditions of TD convergence when exposing state observation noises. Firstly, we recall the classical TD update at step t:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \alpha_t (R_{t+1} + \gamma \mathbf{w}_t^\top \mathbf{x}_{t+1} - \mathbf{w}_t^\top \mathbf{x}_t) \mathbf{x}_t \tag{8}$$

where α_t is the step size at time t. Once the system has reached the steady state for any $\mathbf{w}_t$, then the expected next weight vector can be written as $\mathbb{E}[\mathbf{w}_{t+1}|\mathbf{w}_t] = \mathbf{w}_t + \alpha_t(\mathbf{b} - \mathbf{A}\mathbf{w}_t)$, where $\mathbf{b} = \mathbb{E}(R_{t+1}\mathbf{x}_t) \in \mathbb{R}^d$ and $\mathbf{A} \doteq \mathbb{E}\left[\mathbf{x}_t d_t^\top\right] \in \mathbb{R}^{d \times d}$. The TD fixed point $\mathbf{w}_{\text{TD}}$ to the system satisfies $\mathbf{A}\mathbf{w}_{\text{TD}} = \mathbf{b}$. From (Sutton & Barto, 2018), we know that the matrix $\mathbf{A}$ determines the convergence in the linear TD setting. In particular, $\mathbf{w}_t$ converges with probability one to the TD fixed point if $\mathbf{A}$ is positive definite. However, if we add state noises η on either $\mathbf{x}_t$ or $\mathbf{x}_{t+1}$ in Eq. 8, the convergence condition will be different. Theorem 3 (proof in Appendix E) provides conditions for TD convergence in three different noisy state observation settings.

Theorem 3. *(Conditions for TD Convergence under Noisy State Observations) Define* $\mathbf{P}$ *as the* $|\mathcal{S}| \times |\mathcal{S}|$ *matrix forming from* $p(s'|s)$*,* $\mathbf{D}$ *as the* $|\mathcal{S}| \times |\mathcal{S}|$ *diagonal matrix with* $\mu(s)$ *on its diagonal, and* $\mathbf{X}$ *as the* $|\mathcal{S}| \times d$ *matrix with* $\mathbf{x}(s)$ *as its rows, and* $\mathbf{E}$ *is the* $|\mathcal{S}| \times d$ *perturbation matrix with each perturbation vector* $\mathbf{e}(s)$ *as its rows. The stepsizes* $\alpha_t \in (0, 1]$ *satisfy* $\sum_{t=0}^{\infty} \alpha_t = \infty$ *and* $\sum_{t=0}^{\infty} \alpha_t^2 = 0$*. For noisy states, we consider the following three cases:* ***(i)*** $\mathbf{e}(s)$ *on current state features, i.e.,* $\mathbf{x}_t \leftarrow \mathbf{x}_t + \mathbf{e}_t$*,* ***(ii)*** $\mathbf{e}(s')$ *on next state features, i.e.,* $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_{t+1} + \mathbf{e}_{t+1}$*,* ***(iii)*** *the same* $\mathbf{e}$ *on both state features. We can attain that* $\mathbf{w}_t$ *converges to TD fixed point if the following conditions are satisfied, respectively.*

Case (i): *both* $\mathbf{A}$ *and* $(\mathbf{X} + \mathbf{E})^\top \mathbf{D}\mathbf{P}\mathbf{E}$ *are positive definite.* ***Case (ii):*** *both* $\mathbf{A}$ *and* $-\mathbf{X}^\top \mathbf{D}\mathbf{P}\mathbf{E}$ *are positive definite.* ***Case (iii):*** $\mathbf{A}$ *is positive definite.*

From the convergence conditions for the three cases in Theorem 3, it is clear that (iii) is the mildest. This is the same condition as that in the normal TD learning without noisy state observations. Note that the case (iii) can be viewed as the SN-MDP setting, whose convergence has been already rigorously analyzed in Section 3. In Section 5, our experiments demonstrate that both expectation-based and distribution RL are more likely to converge in case (iii) compared with case (i) and (ii).

In cases (i) and (ii), the positive definiteness of $\mathbf{X}^\top\mathbf{DPE}+\mathbf{E}^\top\mathbf{DPE}$ and $-\mathbf{X}^\top\mathbf{DPE}$ is crucial. We partition $(\mathbf{X}+\mathbf{E})^\top\mathbf{DPE}$ into $\mathbf{X}^\top\mathbf{DPE}+\mathbf{E}^\top\mathbf{DPE}$, where the first term has the opposite positive definiteness to $-\mathbf{X}^\top\mathbf{DPE}$, and the second term is positive definite (Sutton & Barto, 2018). Based on these observations, we discuss the subtle convergence relationship in cases (i) and (ii):

(1) If $-\mathbf{X}^\top\mathbf{DPE}$ is positive definite, which indicates that TD is convergent in case (ii), TD can still converge in case (i) **unless** the positive definiteness of $\mathbf{E}^\top\mathbf{DPE}$ dominates in $\mathbf{X}^\top\mathbf{DPE}+\mathbf{E}^\top\mathbf{DPE}$.

(2) If $-\mathbf{X}^\top\mathbf{DPE}$ is negative definite, TD is likely to diverge in case (ii). By contrast, TD will converge in case (i).

In summary, there exists a subtle trade-off of TD convergence in case (i) and (ii) if we approximately ignore the term $\mathbf{E}^\top\mathbf{DPE}$ in case (i). The key of it lies in the positive definiteness of the matrix $\mathbf{X}^\top\mathbf{DPE}$, which heavily depends on the task. In Section 5, we empirically verify that the convergence situations for current and next state observations are normally different. Which situation is superior is heavily dependent on the task.

4.3 Sensitivity Analysis by Influence Function

Next, we conduct an outlier analysis by the *influence function*, a key facet in the robust statistics (Huber, 2004). The influence function characterizes the effect that the noise in particular observation has on an estimator, and can be utilized to investigate the impact of one particular state observation noise on the training of reinforcement learning algorithms. Specifically, suppose that F_ϵ is the contaminated distribution function that combines the clear data distribution F and an outlier x. The distribution F_ϵ can be defined as

$$F_\epsilon = (1-\epsilon)F + \epsilon\delta_x, \tag{9}$$

where δ_x is a probability measure assigning probability 1 to x. Let $\hat{\theta}$ be a regression estimator. The influence function of θ at F, $\psi : \mathcal{X} \to \Gamma$ is defined as

$$\psi_{\hat{\theta},F}(x) = \lim_{\epsilon\to 0} \frac{\hat{\theta}\left(F_\epsilon(x)\right) - \hat{\theta}(F)}{\epsilon}. \tag{10}$$

Mathematically, the influence function is the Gateaux derivative of θ at F in the direction δ_x. Owing to the fact that traditional value-based RL algorithms, e.g., DQN (Mnih et al., 2015), can be viewed as a regression problem (Fan et al., 2020), the linear TD approximator also has a strong connection with regression problems. Based on this correlation, in the following Theorem 4, we quantitatively evaluate the influence function of TD learning in the case of linear function approximation.

Theorem 4. *(Influence Function Analysis in TD Learning with linear function approximation) Denote $d_t = \mathbf{x}_t - \gamma\mathbf{x}_{t+1} \in \mathbb{R}^d$, and $\mathbf{A} \doteq \mathbb{E}\left[\mathbf{x}_t d_t^\top\right] \in \mathbb{R}^{d\times d}$. Let F_π be the data distribution generated from the environment dynamics given a policy π. Consider an outlier pair $(\mathbf{x}_t, \mathbf{x}_{t+1})$ with the reward R_{t+1}, the influence function ψ of this pair on the estimator $\mathbf{w}$ is derived as*

$$\psi_{\mathbf{w},F_\pi}(\mathbf{x}_t, \mathbf{x}_{t+1}) = \mathbb{E}(\mathbf{A}^\top\mathbf{A})^{-1} d_t \mathbf{x}_t^\top \mathbf{x}_t (R_{t+1} - d_t^\top \mathbf{w}). \tag{11}$$

Please refer to Appendix F for the proof. Theorem 4 shows the quantitative impact of an outlier pair $(\mathbf{x}_t, \mathbf{x}_{t+1})$ on the learned parameter $\mathbf{w}$. Moreover, a corollary can be immediately obtained to make a precise comparison of the impacts of perturbations on current and next state features.

Corollary 1. *Given the same perturbation η on either current or next state features, i.e., $\mathbf{x}_t$, and $\mathbf{x}_{t+1}$, at the step t, if we approximate $\eta\eta^\top\mathbf{x}_t$ and $\eta\eta^\top\mathbf{w}$ as $\mathbf{0}$ as η is small enough, the following relationship between the resulting variations of influence function, $\Delta_{\mathbf{x}_t}\psi$ and $\Delta_{\mathbf{x}_{t+1}}\psi$, holds:*

$$\gamma\Delta_{\mathbf{x}_t}\psi + \Delta_{\mathbf{x}_{t+1}}\psi = 2\gamma d_t \eta \mathbf{x}_t^\top (R_{t+1} - d_t^\top \mathbf{w}). \tag{12}$$

We provide the proof of Corollary 1 in Appendix F. Under this equation, the sensitivity of noises on $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$, measured by $\Delta_{\mathbf{x}_t}\psi$ and $\Delta_{\mathbf{x}_{t+1}}\psi$, present a trade-off relationship as their weighted sum

is definite. However, there is not an ordered relationship between $\Delta_{\mathbf{x}_t}\psi$ and $\Delta_{\mathbf{x}_{t+1}}\psi$. In summary, we conclude that the sensitivity of current and next state features against perturbations is normally divergent, and the degree of sensitivity is heavily determined by the task. These conclusions are similar to those we derived in the TD convergence part.

Remark. The Lipschitz continuity blessing derived in Section 4.1 explains the less vulnerability of distributional RL than expectation-based RL, while in Section 4.2 and 4.3 we characterize the convergence conditions and sensitivity of different noisy state observations albeit being in expectation-based case. Our following experiment observation coincide with our theoretical results.

5 EXPERIMENTS

We make a comparison between expectation-based and distributional RL algorithms against various noisy state observations. We select DQN (Mnih et al., 2015) as the baseline, and QR-DQN (Dabney et al., 2018b) as its distributional counterpart. The previous analysis is either for policy evaluation or linear function approximation, but there are natural—though in some cases heuristic—extensions to the control setting and to non-linear function approximation.

Experimental Setup. We perform our algorithms on Cart Pole, Mountain Car, Breakout and Qbert games. We followed the procedure in (Ghiassian et al., 2020; Zhang & Yao, 2019). All the experimental settings, including parameters, are identical to the distributional RL baselines implemented by Zhang (2018); Dabney et al. (2018b). Please refer to Appendix G for more details.

Noisy State Observations. For the random noise, we use Gaussian noise with different standard deviations. For the adversarial noise, we followed (Zhang et al., 2020), where the set of noises $B(s)$ is defined as an ℓ_∞ norm ball around s with a radius ϵ, given by $\ell_\infty B(s) := \{\hat{s} : \|s - \hat{s}\|_\infty \leq \epsilon\}$. We apply Projected Gradient Descent (PGD) version in (Pattanaik et al., 2017), with 3 fixed iterations while adjusting ϵ to control the perturbation strength.

5.1 PERFORMANCE ON CART POLE

We select the standard deviations as 0.05 and 0.1 in the random noisy state setting, and the perturbation sizes ϵ as 0.05 and 0.1 in the adversarial noisy state case. Figure 2 shows the tendency of average return with standard deviation over 200 runs on Cart Pole during the whole training process for both DQN and QRDQN under the adversarial state observation noises. A similar result in random setting is provided in Appendix H with more experimental details in Appendix G.

Firstly, Figure 2 reveals that QRDQN (solid lines) consistently outperforms DQN (dashed lines) in the same color under different state noise strengths, although the performance of QRDQN can degenerate to that of DQN when exposed to strong perturbations shown in the left plot. This conclusion agrees with Theorem 2 in Section 4.1. Secondly, under the same perturbations, next state observations (in the middle plot) are less vulnerable than current states (in the left plot). Both DQN

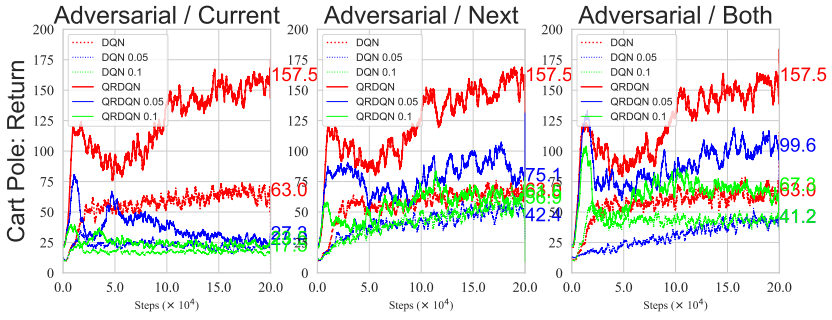

Figure 2: Average returns of DQN and QRDQN against adversarial state observation noises on Cart Pole over 200 runs with smooth size 20. QRDQN (solid lines) almost consistently outperforms DQN (dashed lines) in the same color, demonstrating better robustness.

and QRDQN converge more easily and achieve better performance in the SN-MDP setting (in the right plot). These observations are consistent of the subtle trade-off relationship of current and next states, and the mildest TD convergence condition analyzed in Section 4.2 and 4.3.

5.2 Performance on Mountain Car

From the experimental results in Cart Pole, some may contend that the robust performance of QRDQN can be largely attributed to its superiority in this task. To more rigorously explore the robustness of distributional RL, we select Mountain Car task where QRDQN can only achieve comparable performance relative to DQN and even worse in the early stage of training. We select the standard deviations as 0.01 and 0.0125 in the random setting, and the perturbation sizes ϵ as 0.01 and 0.1 in the adversarial case. We provide the result in the random setting in Figure 3, and a similar result in the adversarial setting is provided in Appendix H.

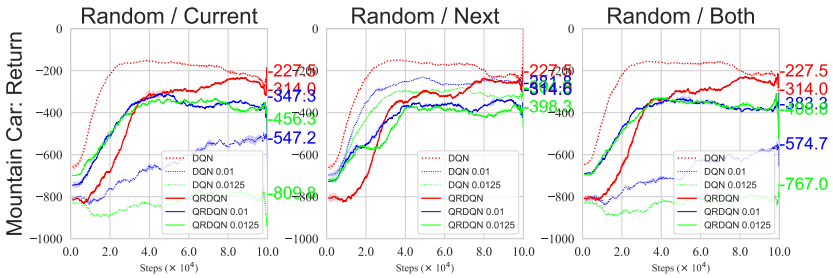

Figure 3: Average returns of DQN and QRDQN against random state observation noises on Mountain Car over 200 runs with smooth size 100. QRDQN (solid lines) almost consistently outperforms DQN (dashed lines) in the same color, demonstrating better robustness.

Similar to Cart Pole, Figure 3 shows that QRDQN is capable of achieving significantly better performance under most noisy scenarios relative to DQN, except the comparable results when imposing random noises on the next state observations (in the middle plot). The next state is still less vulnerable than the current state. These results demonstrate that QRDQN enjoys better robustness than DQN, regardless of whether QRDQN outperforms DQN under the noise-free environment.

5.3 Performance on Breakout

To further verify the superior robustness of QRDQN over DQN, we conduct more realistic experiments on the Atari game: Breakout. In this environment, QRDQN eventually achieves similar performance as DQN, although QRDQN significantly reduces the sample efficiency. Thus, it is a fair comparison to investigate the robust performance on the Breakout environment. We set the number of quantiles in QRDQN to 200 and report the average return over 3 runs for each noisy setting. We choose the standard deviations as 0.01 and 0.05 in the random setting, and the perturbation sizes

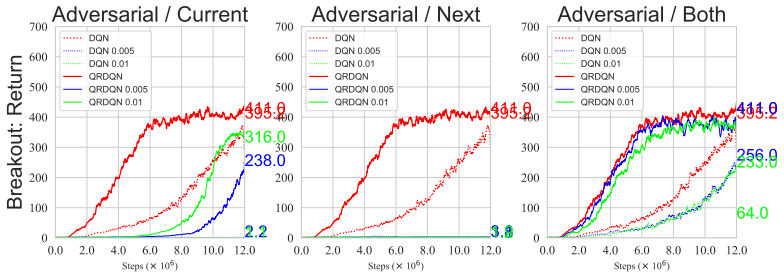

Figure 4: Average returns of DQN and QRDQN against adversarial state observation noises on Breakout over 3 runs with smooth size 1000. QRDQN (solid lines) almost consistently outperforms DQN (dashed lines) in the same color, demonstrating better robustness.

ϵ as 0.005 and 0.01 in the adversarial case. We present results in the adversarial setting as shown in Figure 4 and provide the similar results under random noises in Appendix H.

A key observation from Figure 4 is that both QRDQN and DQN converge around 430 average returns, though the former outperforms the latter across the vast majority of the training process. Consistently, the relationships we revealed in the previous two tasks still exist in Breakout. Firstly, the solid lines (QRDQN) are always above the dashed (DQN) counterpart across various noisy scenarios. The phenomenon in the left figure is of vital importance as it shows the case when distributional RL converges to a satisfactory point while expectation-based RL algorithm even diverges. Moreover, in the middle part both DQN and QRDQN are overly sensitive to noises imposed on next state observations, ultimately failing to converge. This result further demonstrates the divergent sensitivity of current and next states. It is worthwhile to mention that the sensitivity ordering of current and next state observations in Breakout is opposite to those on both Cart Pole and Mountain Car. This result is still consistent with our analysis in Section 4.2 and 4.3, where the sensitivity ordering of both states heavily depends on the task. The milder convergence of both DQN and QRDQN in the SN-MDP (in the right plot) is also exhibited in Breakout, matching our analysis in Theorem 3.

5.4 Performance on Qbert

Furthermore, we conduct another realistic experiment on the Atari game Qbert. For convenience, we choose the one standard deviation 0.05 in the random setting and one perturbation size ϵ as 0.005 in the adversarial case. Similarly, QRDQN also achieves similar performance as DQN in the end in this environment. We elaborate results in the adversarial setting as shown in Figure 5 and provide the similar results in the random setting in Appendix H.

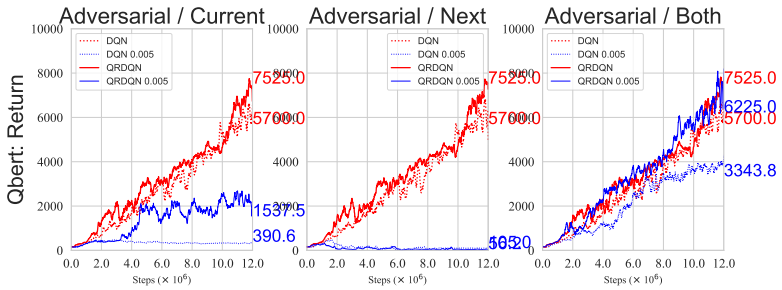

Figure 5: Average returns of DQN and QRDQN against adversarial state observation noises on Qbert environment over 3 runs with smooth size 1000. QRDQN (solid lines) almost consistently outperforms DQN (dashed lines) in the same color, demonstrating better robustness.

The results in Qbert are closely to those in Breakout. Briefly speaking, QRDQN (solid lines) achieves better or at least comparable robustness relative to DQN (dashed lines). In addition, the next state is more sensitive to noises where both DQN and QRDQN diverge, compared with the current state, and both DQN and QRDQN converge more easily in the SN-MDP setting (in the right plot). Most importantly, Figure 5 also illustrates the case when distributional RL algorithm attains a relatively desirable return while the expectation-based counterpart is divergent. This result further demonstrates the robustness advantage of distributional RL over the expectation-based RL.

6 Discussion and Conclusion

The Lipschitz continuity blessing is based on the histogram distributional loss, but it is more expected that similar conclusions can be made under Wasserstein or Crammer distance as these distances are more approachable in real distributional RL algorithms. We leave it as future works.

In this paper, we explored the training robustness of distributional RL against noisy state observations. After the convergence analysis of distributional RL in the SN-MDP, we proved the Lipschitz continuity property of distributional RL, accounting for its less vulnerability. We also provided the TD convergence conditions and a sensitivity analysis on more complex noisy settings. Experimental observations coincides with our theoretical results.

Ethics Statement. Our works reveals that distributional RL can enjoy the training robustness against noisy state observations. The advantage is useful to defend against the poisoning attacks, thus contributing to the privacy of algorithms. Based on our experience, there is no other ethic concerns of our work.

Reproducibility Statement. For the theoretical part, we clearly state the related assumption and detailed proof process in the appendix. In terms of the algorithm, our implementation is directly adapted from the public RL algorithms, including DQN and QR-DQN.

REFERENCES

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. *International Conference on Machine Learning (ICML)*, 2017.

Vivek S Borkar and Sean P Meyn. The ode method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000.

Tong Chen, Jiqiang Liu, Yingxiao Xiang, Wenjia Niu, Endong Tong, and Zhen Han. Adversarial attack and defense in reinforcement learning-from ai security view. *Cybersecurity*, 2(1):11, 2019.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. *International Conference on Machine Learning (ICML)*, 2018a.

Will Dabney, Mark Rowland, Marc G Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018b.

Jianqing Fan, Zhaoran Wang, Yuchen Xie, and Zhuoran Yang. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

Sina Ghiassian, Andrew Patterson, Shivam Garg, Dhawal Gupta, Adam White, and Martha White. Gradient temporal-difference learning with regularized corrections. In *International Conference on Machine Learning*, pp. 3524–3534. PMLR, 2020.

Ziwei Guan, Kaiyi Ji, Donald J Bucci Jr, Timothy Y Hu, Joseph Palombo, Michael Liston, and Yingbin Liang. Robust stochastic bandit algorithms under probabilistic unbounded adversarial attack. In *AAAI*, pp. 4036–4043, 2020.

Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *Advances in Neural Information Processing Systems*, 2017.

Peter J Huber. *Robust Statistics*, volume 523. John Wiley & Sons, 2004.

Inaam Ilahi, Muhammad Usama, Junaid Qadir, Muhammad Umar Janjua, Ala Al-Fuqaha, Dinh Thai Hoang, and Dusit Niyato. Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *arXiv preprint arXiv:2001.09684*, 2020.

Ehsan Imani and Martha White. Improving regression performance with distributional losses. In *International Conference on Machine Learning*, pp. 2157–2166. PMLR, 2018.

Borislav Mavrin, Shangtong Zhang, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. *International Conference on Machine Learning (ICML)*, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Thanh Tang Nguyen, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning with maximum mean discrepancy. *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary. Robust deep reinforcement learning with adversarial attacks. *Advances in Neural Information Processing Systems*, 2017.

Qianli Shen, Yan Li, Haoming Jiang, Zhaoran Wang, and Tuo Zhao. Deep reinforcement learning with robust and smooth policy. In *International Conference on Machine Learning*, pp. 8707–8718. PMLR, 2020.

Rahul Singh, Qinsheng Zhang, and Yongxin Chen. Improving robustness via risk averse distributional reinforcement learning. *arXiv preprint arXiv:2005.00585*, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An Introduction*. MIT press, 2018.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Duane Boning, and Cho-Jui Hsieh. Robust deep reinforcement learning against adversarial perturbations on observations. *Advances in Neural Information Processing Systems*, 2020.

Shangtong Zhang. Modularized implementation of deep rl algorithms in pytorch. `https://github.com/ShangtongZhang/DeepRL`, 2018.

Shangtong Zhang and Hengshuai Yao. Quota: The quantile option architecture for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5797–5804, 2019.

A CONVERGENCE UNDER p-WASSERSTEIN METRIC

We provide more detailed introduction of the convergence of distributional Bellman operator. Firstly, the p-Wasserstein metric W_p is defined as

$$d_p = W_p(Z^*, Z_\theta) = \left(\int_0^1 \left| F_{Z^*}^{-1}(\omega) - F_{Z_\theta}^{-1}(\omega) \right|^p d\omega \right)^{1/p}, \tag{13}$$

which minimizes the distance between the true value distribution Z^* and the parametric distribution Z_θ. F^{-1} is the inverse cumulative distribution function of a random variable with the cumulative distribution function as F. In the control setting, the distributional analogue of the Bellman optimality operator converges to the set of optimal value distributions, although it is in a weak sense and requires more involved arguments (Dabney et al., 2018b).

B THEOREM 5 WITH PROOF

Theorem 5. *(Convergence and Contraction of Bellman operators in the SN-MDP) Given a policy π, define the Bellman operator $\mathcal{T} : \mathbb{R}^{|S|} \to \mathbb{R}^{|S|}$ under random and adversarial states noises by $\mathcal{T}_r^\pi$ and $\mathcal{T}_a^\pi$, respectively. Denote a "merged" policy π' where $\pi'(a|s) = \sum_{v(s)} N(v(s)|s)\pi(a|v(s))$ and $\mathbf{S}(\pi)$ is a policy set given π. Then we have:*

(1) $\mathcal{T}_r^\pi$ is a contraction operator and can converge to $V_{\pi'}$, i.e., $\mathcal{T}_r^\pi \tilde{V}_{\pi \circ N} = \tilde{V}_{\pi \circ N} = V_{\pi'}$, where multiple policies $\pi_r \in \mathbf{S}(\pi)$ might exist with $\sum_{v(s)} N(v(s)|s)\pi_r(a|v(s)) = \pi'(a|s)$.

(2) $\mathcal{T}_a^\pi$ is a contraction with the convergence satisfying $\mathcal{T}_a^\pi \tilde{V}_{\pi \circ N^*} = \min_N \tilde{V}_{\pi \circ N} = V_{\pi \circ N^*}$, where N^* is the optimal adversarial noise strategy. If the optimal policy π_a exists, it satisfies $\pi_a(a|v^*(s)) = \pi(a|s)$ for each s and a, where $v^*(s)$ is the adversarial noisy state manipulated by $N^*(\cdot|s)$.

Proof. Our proof is partly based on Theorem 1 and 2 in (Zhang et al., 2020), but adds more analysis on the converged policy especially under the random noisy states setting. The most important insight in the following proof is that the noise transition can be merged into the agent's policy, resulting in a new "merged" policy π'.

Proof of (1) Firstly, as the Bellman Equation under the random noisy states is right the general form in Eq. 5, it automatically satisfies that $\mathcal{T}_r^\pi \tilde{V}_{\pi \circ N} = \tilde{V}_{\pi \circ N}$ when it converges. As for the proof of contraction, based on our insight about the new "merged" policy π' where $\pi'(a|s) = \sum_{v(s)} N(v(s)|s)\pi(a|v(s))$, we can rewrite our Bellman Operator as:

$$\begin{aligned} &\mathcal{T}_r^\pi \tilde{V}_{\pi \circ N}(s) \\ &= \sum_a \pi'(a|s) \sum_{s'} p(s'|s,a) \left[R(s,a,s') + \gamma \tilde{V}_{\pi \circ N}(s') \right] \\ &= \boldsymbol{R}(s) + \gamma \sum_{s'} \boldsymbol{P}'_{s,s'} \tilde{V}_{\pi \circ N}(s') \end{aligned} \tag{14}$$

where $\boldsymbol{R}(s) = \sum_a \pi'(a|s) \sum_{s'} p(s'|s,a) R(s,a,s')$, and $\boldsymbol{P}'_{s,s'} = \sum_a \pi'(a|s) p(s'|s,a)$ determined by the "merged" policy π'. Then for two different value function $\tilde{V}^1_{\pi \circ N}$ and $\tilde{V}^2_{\pi \circ N}$ we have:

$$\begin{aligned} &\|\mathcal{T}_r^\pi \tilde{V}^1_{\pi \circ N} - \mathcal{T}_r^\pi \tilde{V}^2_{\pi \circ N}\|_\infty \\ &= \max_s |\gamma \sum_{s'} \boldsymbol{P}'_{s,s'} \tilde{V}^1_{\pi \circ N}(s') - \gamma \sum_{s'} \boldsymbol{P}'_{s,s'} \tilde{V}^2_{\pi \circ N}(s')| \\ &\leq \gamma \max_s \sum_{s'} \boldsymbol{P}'_{s,s'} |\tilde{V}^1_{\pi \circ N}(s') - \tilde{V}^2_{\pi \circ N}(s')| \\ &\leq \gamma \max_s \sum_{s'} \boldsymbol{P}'_{s,s'} \max_{s'} |\tilde{V}^1_{\pi \circ N}(s') - \tilde{V}^2_{\pi \circ N}(s')| \\ &= \gamma \max_s \sum_{s'} \boldsymbol{P}'_{s,s'} \|\tilde{V}^1_{\pi \circ N} - \tilde{V}^2_{\pi \circ N}\|_\infty \\ &= \gamma \|\tilde{V}^1_{\pi \circ N} - \tilde{V}^2_{\pi \circ N}\|_\infty \end{aligned} \tag{15}$$

Then according to the Banach fixed-point theorem, since $\gamma \in (0,1)$, $\tilde{V}_{\pi \circ N}$ converges to a unique fixed-point $V_{\pi'}$. However, even though the obtained policy π' satisfies that $\pi'(a|s) = \sum_{v(s)} N(v(s)|s)\pi(a|v(s))$ for each s, a, these equations can not necessarily guarantee a unique π especially when these equations behind this condition are underdetermined. In such scenario, multiple policies π_r will exist as long as they satisfy the equations above.

Proof of (2) Firstly, based on Theorem 1 (Zhang et al., 2020) that shows an optimal policy does not always exist, we assume that an optimal policy exists in the adversarial noisy state setting for the convenience of following analysis. Based on this assumption, we need to derive the explicit value function under the adversary. Inspired by (Zhang et al., 2020), the proof insight is that the behavior of optimal adversary can be also viewed as finding another optimal policy, yielding a zero-sum two player game. Specifically, in the SN-MDP setting, the adversary selects an action $\hat{a} \in \mathcal{S}$ satisfying $\hat{a} = v(s)$, attempting to maximize its state-action value function $\tilde{Q}_{\pi_a}(s, \hat{a})$. Then the adversary's value function $\hat{V}_{\pi_a}(s)$ can be formulated as:

$$\begin{aligned}\hat{V}_{\pi_a}(s) &= \max_{\hat{a}} \hat{Q}_{\pi_a}(s, \hat{a}) \\ &= \max_{\hat{a}} \sum_{s'} \hat{p}(s'|s, \hat{a})(\hat{R}(s, \hat{a}, s') + \gamma \hat{V}_{\pi_a}(s')) \\ &= \max_{v(s)} \sum_{s'} \sum_{a} \pi(a|v(s)) p(s'|s, a)(-R(s, a, s') \\ &+ \gamma \hat{V}_{\pi_a}(s'))\end{aligned} \tag{16}$$

where $\hat{p}(s'|s, \hat{a})$ is the transition dynamics of the adversary, satisfying $\hat{p}(s'|s, \hat{a}) = \sum_a \pi(a|v(s))p(s'|s, a)$ from the perspective of the agent. $\hat{R}(s, \hat{a}, s')$ is the adversary's reward function while taking action $\hat{a}$, which is the opposite number of $R(s, a, s')$ given the action a. In addition, since both the adversary and agent can serve as a zero-sum two-player game, it indicates that $\tilde{V}_{\pi_a}(s) = -\hat{V}_{\pi_a}(s)$ for the agent's value function $\tilde{V}_{\pi_a}$ in the adversary setting. Then we rearrange the equation above as follows:

$$\begin{aligned}\tilde{V}_{\pi_a}(s) &= -\hat{V}_{\pi_a}(s) \\ &= -\min_{N(\cdot|s)} \sum_{s'} \sum_{a} \pi'(a|s) p(s'|s, a)(-R(s, a, s') \\ &- \gamma \tilde{V}_{\pi_a}(s')) \\ &= \min_{v(s)} \sum_{s'} \sum_{a} \pi'(a|s) p(s'|s, a)(R(s, a, s') \\ &+ \gamma \tilde{V}_{\pi_a}(s')) \\ &= \min_{N(\cdot|s)} \sum_{s'} \sum_{a} \pi'(a|s) p(s'|s, a)(r_{t+1} + \gamma \min_{N} \mathbb{E}_{\pi \circ N} \left[\sum_{k=0}^{\infty} r_{t+k+2} | s_{t+1} = s' \right]) \\ &= \min_{N} \tilde{V}_{\pi \circ N}(s)\end{aligned} \tag{17}$$

Note that we optimize over N, which means we consider $N(\cdot|s)$ for each state s. Further, we derive the contraction of the Bellman operator $\mathcal{T}_a^\pi$. We rewrite our Bellman Operator $\mathcal{T}_a^\pi$ as:

$$\begin{aligned}\mathcal{T}_a^\pi \tilde{V}_{\pi \circ N}(s) &= \min_{N} \tilde{V}_{\pi \circ N}(s) \\ &= \min_{N} \boldsymbol{R}(s) + \gamma \sum_{s'} \boldsymbol{P}'_{s,s'} \tilde{V}_{\pi \circ N}(s')\end{aligned} \tag{18}$$

We firstly assume $\mathcal{T}_a^\pi \tilde{V}_{\pi_a}^1(s) \geq \mathcal{T}_a^\pi \tilde{V}_{\pi_a}^2(s)$, then we have:

$$\begin{aligned}&\mathcal{T}_a^\pi \tilde{V}_{\pi \circ N}^1(s) - \mathcal{T}_a^\pi \tilde{V}_{\pi \circ N}^2(s) \\ &\leq \max_{N(\cdot|s)} \{\gamma \sum_{s'} \boldsymbol{P}'_{s,s'} \tilde{V}_{\pi \circ N}^1(s') - \gamma \sum_{s'} \boldsymbol{P}'_{s,s'} \tilde{V}_{\pi \circ N}^2(s')\} \\ &\leq \gamma \max_{N(\cdot|s)} \sum_{s'} \boldsymbol{P}'_{s,s'} |\tilde{V}_{\pi \circ N}^1(s') - \tilde{V}_{\pi \circ N}^2(s')| \\ &\leq \gamma \max_{N(\cdot|s)} \sum_{s'} \boldsymbol{P}'_{s,s'} \max_{s} |\tilde{V}_{\pi \circ N}^1(s') - \tilde{V}_{\pi \circ N}^2(s')| \\ &= \gamma \max_{N(\cdot|s)} \sum_{s'} \boldsymbol{P}'_{s,s'} \|\tilde{V}_{\pi \circ N}^1 - \tilde{V}_{\pi \circ N}^2\|_\infty \\ &\leq \gamma \|\tilde{V}_{\pi \circ N}^1 - \tilde{V}_{\pi \circ N}^2\|_\infty\end{aligned} \tag{19}$$

where the first inequality holds as $\min_{x_1} f(x_1) - \min_{x_2} g(x_2) \leq \max_x (f(x) - g(x))$ and we extends this inequality into the Wasserstein distance in the proof of convergence of distributional RL setting in Appendix C. The last inequality holds since only $\boldsymbol{P}'_{s,s'}$ depends on $N(\cdot|s)$ while the infinity norm is a constant, which is independent with the current $N(\cdot|s)$. Similarly, the other scenario can be still proved. Thus, we have:

$$\|\mathcal{T}_a^\pi \tilde{V}_{\pi \circ N}^1 - \mathcal{T}_a^\pi \tilde{V}_{\pi \circ N}^2\|_\infty \leq \gamma \|\tilde{V}_{\pi \circ N}^1 - \tilde{V}_{\pi \circ N}^2\|_\infty \tag{20}$$

Thus, we proved that $\mathcal{T}_a^\pi$ is still a contraction and converge to $\min_N \tilde{V}_{\pi \circ N}$. We denote it as $\tilde{V}_{\pi \circ N^*}$ In addition, based on the insight of the "merged" policy π_a', we have $\pi_a' = \sum_{v(s)} N^*(v(s)|s)\pi(a|v(s)) = \pi(a|v^*(s))$ where the deterministic state $v^*(s)$ is the adversarial noisy state from the state s.

□

C PROOF OF THEOREM 1

Proof. Firstly, we will provide the properties of Wassertein distance d_p in Lemma 1 that we leverage in our following convergence proof.

Lemma 1. *(Properties of Wasserstein Metric) We consider the distribution distance between the random variable U and V. Denote d_p as the Wasserstein distance between two distribution defined in Eq. 13. For any scalar a and random variable A independent of U and V, the following relationships hold:*

$$\begin{aligned} d_p(aU, aV) &\leq |a| d_p(U, V) \\ d_p(A + U, A + V) &\leq d_p(U, V) \\ d_p(AU, AV) &\leq \|A\|_p d_p(U, V) \end{aligned} \tag{21}$$

Further, let $A_1, A_2, ...$ be a set of random variables describing the a partition of ω, when the partition lemma holds:

$$d_p(U, V) \leq \sum_i d_p(A_i U, A_i V). \tag{22}$$

Then, the following contraction proof is in the maximal form of d_p and we denote it as $\bar{d}_p$.

Proof of (1) This contraction proof is similar to the original one (Bellemare et al., 2017) in the distributional RL without state observation noises. The only difference lies in the new transition operator $\mathcal{P}_r^\pi$, but it dose not change the main proof process. For two different random variables Z_N^1 and Z_N^2 about returns, we have:

$$\begin{aligned} &\bar{d}_p(\mathfrak{T}_r^\pi Z_N^1, \mathfrak{T}_r^\pi Z_N^2) \\ &= \sup_{s,a} d_p(\mathfrak{T}_r^\pi Z_N^1(s,a), \mathfrak{T}_r^\pi Z_N^2(s,a)) \\ &= \sup_{s,a} d_p(R(s,a,S') + \gamma \mathcal{P}_r^\pi Z_N^1(s,a), R(s,a,S') + \gamma \mathcal{P}_r^\pi Z_N^2(s,a)) \\ &\leq \gamma \sup_{s,a} d_p(\mathcal{P}_r^\pi Z_N^1(s,a), \mathcal{P}_r^\pi Z_N^2(s,a)) \\ &\leq \gamma \sup_{s,a} \sup_{s',a'} d_p(Z_N^1(s',a'), Z_N^2(s',a')) \\ &= \gamma \sup_{s',a'} d_p(Z_1(s',a'), Z_2(s',a')) \\ &= \gamma \sup_{s,a} d_p(Z_N^1(s,a), Z_N^2(s,a)) \\ &= \gamma \bar{d}_p(Z_N^1, Z_N^2). \end{aligned} \tag{23}$$

Thus, we conclude that $\mathfrak{T}_r^\pi : \mathcal{Z} \to \mathcal{Z}$ is a γ-contraction in $\bar{d}_p$.

Proof of (2) Firstly, we define the distributional Bellman optimality operator $\mathfrak{T}$ in MDP as

$$\mathfrak{T} Z(s,a) :\stackrel{D}{=} R\left(s, a, S'\right) + \gamma Z(S', \pi_Z(s')) \tag{24}$$

where $S' \sim P(\cdot|s,a)$ and $\pi_Z(S') = \arg\max_{a'} \mathbb{E}\left[Z(S', a')\right]$. By contrast, in SN-MDP, Our greedy adversarial rule $N^*(\cdot|s')$ is based on the greedy policy rule in distributional Bellman optimality operator, which attempts to find adversarial $N^*(\cdot|s')$ in order to minimize $\mathbb{E}\left[Z_N(s', a')\right]$, where $a' \sim \pi(\cdot|V(s'))$ and $V(s') \sim N(\cdot|s')$. We assume $N^*(\cdot|s')$ yields a deterministic state s^*, and thus the agent always takes action based on s^*, which we denote as $A^* \sim \pi(\cdot|s^*)$. Therefore, we can obtain the state-action function $Q_{N^*}^\pi(s,a)$ under the adversary as

$$\begin{aligned} Q_{N^*}^\pi(s,a) &= \min_N \mathbb{E}\left[Z_N^\pi(s,a)\right] \\ &= \mathbb{E}\left[Z^{\pi^*}(s,a)\right] \end{aligned} \tag{25}$$

where $\pi^*(\cdot|s) = \pi(\cdot|s^*)$ for $\forall s$ that follows the adversarial policy A^*.

Next, to derive the contractive property of $\mathfrak{T}_a^\pi$, we denote two state-action valued distributions as $Z_N^1(s,a)$ and $Z_N^2(s,a)$. Then we have:

$$
\begin{aligned}
&\bar{d}_p(\mathfrak{T}_a^\pi Z_N^1, \mathfrak{T}_a^\pi Z_N^2) \\
&= \sup_{s,a} d_p(\mathfrak{T}_a^\pi Z_N^1(s,a), \mathfrak{T}_a^\pi Z_N^2(s,a)) \\
&= \sup_{s,a} d_p(R(s,a,S') + \gamma \mathcal{P}_a^\pi Z_N^1(s,a), R(s,a,S') + \gamma \mathcal{P}_a^\pi Z_N^2(s,a)) \\
&\leq \gamma \sup_{s,a} \sum_{s'} P(s'|s,a) d_p(Z_N^1(s', A^*), Z_N^2(s', A^*)) \\
&= \gamma \sum_{s'} P(s'|s,a) d_p(Z_N^1(s', A^*), Z_N^2(s', A^*)) \\
&\leq \gamma \sup_{s'} d_p(Z_N^1(s', A^*), Z_N^2(s', A^*)) \\
&= \gamma \sup_{s'} d_p(\sum_{a'_*} \pi(a'_*|s^*) Z_N^1(s', a'_*), \sum_{a'_*} \pi(a'_*|s^*) Z_N^2(s', a'_*)) \\
&\leq \gamma \sup_{s'} \sum_{a'_*} \pi(a'_*|s^*) d_p(Z_N^1(s', a'_*), Z_N^2(s', a'_*)) \\
&\leq \gamma \sup_{s', a'_*} d_p(Z_N^1(s', a'_*), Z_N^2(s', a'_*)) \\
&= \gamma \sup_{s,a} d_p(Z_N^1(s,a), Z_N^2(s,a)) \\
&= \bar{d}_p(Z_N^1, Z_N^2)
\end{aligned}
\tag{26}
$$

Thus, we conclude that $\mathfrak{T}_a^\pi$ is still a γ-contraction in $\bar{d}_p$. □

D PROOF OF THEOREM 2

Proof. Firstly, we show the derivation details of the Histogram distribution loss starting from KL divergence between p and q_θ. p_i is the cumulative probability increment of target distribution $\mathfrak{T}Z_\theta$ within the i-th bin, and q_θ corresponds to a (normalized) histogram, and has density values $\frac{f_i^\theta(\mathbf{x}(s))}{w_i}$ per bin. Thus, we have:

$$
\begin{aligned}
\mathcal{L}(Z_\theta, \mathfrak{T}Z_\theta) &= -\int_a^b p(y) \log q_\theta(y) dy \\
&= -\sum_{i=1}^k \int_{l_i}^{l_i + w_i} p(y) \log \frac{f_i^\theta(\mathbf{x}(s))}{w_i} dy \\
&= -\sum_{i=1}^k \log \frac{f_i^\theta(\mathbf{x}(s))}{w_i} \underbrace{\left(F_{\mathfrak{T}Z_\theta}(l_i + w_i) - F_{\mathfrak{T}Z_\theta}(l_i)\right)}_{p_i} \\
&\doteq -\sum_{i=1}^k p_i \log f_i^\theta(\mathbf{x}(s))
\end{aligned}
\tag{27}
$$

where the last equality holds because the width parameter w_i can be ignored for this minimization problem.

Next, we compute the gradient of the Histogram distributional loss.

$$
\begin{aligned}
&\frac{\partial}{\partial \mathbf{x}(s)} \sum_{j=1}^{k} p_j \log f_j^{\theta}(\mathbf{x}(s)) \\
&= \sum_{j=1}^{k} p_j \frac{1}{f_j^{\theta}(\mathbf{x}(s))} \nabla f_j^{\theta}(\mathbf{x}(s)) \\
&= \sum_{j=1}^{k} p_j \frac{1}{f_j^{\theta}(\mathbf{x}(s))} f_j^{\theta}(\mathbf{x}(s)) \sum_{i=1}^{k} \frac{\exp(\mathbf{x}(s)^{\top}\theta_i)}{\sum_{p=1}^{k} \exp(\mathbf{x}(s)^{\top}\theta_p)} (\theta_j - \theta_i) \\
&= \sum_{j=1}^{k} p_j \sum_{i=1}^{k} f_i^{\theta}(\mathbf{x}(s))(\theta_j - \theta_i) \\
&= \sum_{j=1}^{k} p_j \theta_j - \sum_{i=1}^{k} f_i^{\theta}(\mathbf{x}(s))\theta_i \\
&= \sum_{i=1}^{k} (p_i - f_i^{\theta}(\mathbf{x}(s)))\theta_i
\end{aligned}
\tag{28}
$$

Then, as we have $\|\theta_i\| \leq l$ for $\forall i$, we bound the norm of its gradient

$$
\begin{aligned}
&\| \frac{\partial}{\partial \mathbf{x}(s)} \sum_{j=1}^{k} p_j \log f_j^{\theta}(\mathbf{x}(s)) \| \\
&\leq \sum_{i=1}^{k} \| (p_i - f_i^{\theta}(\mathbf{x}(s)))\theta_i \| \\
&= \sum_{i=1}^{k} |p_i - f_i^{\theta}(\mathbf{x}(s))| \|\theta_i\| \\
&\leq kl
\end{aligned}
\tag{29}
$$

The last equality satisfies because $|p_i - f_i^{\theta}(\mathbf{x}(s))|$ is less than 1 and even smaller. By contrast, in the expectation-based RL, our objective function can be viewed as a least squared optimization, and the updating rule regarding parameter $\mathbf{w}$ is

$$
\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t + \alpha \left[v_{\pi}(S_t) - \hat{v}(S_t, \mathbf{w}_t) \right] \nabla \hat{v}(S_t, \mathbf{w}_t) \\
&= \mathbf{w}_t + \alpha \left(U_t - \mathbf{w}_t^{\top} \mathbf{x}_t \right) \mathbf{x}_t
\end{aligned}
\tag{30}
$$

where U_t can be evaluated by either Monte Carlo method or TD learning. Based on the updating rule, we can immediately obtain the gradient of loss, i.e., $\left(U_t - \mathbf{w}_t^{\top} \mathbf{x}_t\right) \mathbf{w}_t$. Thus, its norm is $|U_t - \mathbf{w}_t^{\top} \mathbf{x}_t| \|\mathbf{w}_t\| \leq |U_t - \mathbf{w}_t^{\top} \mathbf{x}_t| l$. However, this upper bound can be arbitrary large as there is no restriction on $|U_t - \mathbf{w}_t^{\top} \mathbf{x}_t|$. In summary, compared with the least squared loss in expectation-based RL, the histogram distributional loss in distributional RL can additionally enjoy kl-Lipschitz continuity with bounded gradient norm regarding the state features $\mathbf{x}(s)$. This upper bound of gradient norm can mitigate the impact of the noises on state observations on the loss function, therefore yielding training robustness for distributional RL.

□

E PROOF OF THEOREM 3

Proof. To prove the convergence of TD under the noisy states, we use the results from (Borkar & Meyn, 2000) that require the condition about stepsizes α_t holds: $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 = 0$. Based on (Sutton & Barto, 2018), the positive definiteness of $\mathbf{A}$ will determine the TD convergence. For linear TD(0), in the

continuing case with $\gamma < 1$, $\mathbf{A}$ can be re-written as:

$$\begin{aligned}\mathbf{A} &= \sum_s \mu(s) \sum_a \pi(a|s) \sum_{r,s'} p(r, s'|s, a)\mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right)^\top \\ &= \sum_s \mu(s) \sum_{s'} p(s'|s)\mathbf{x}_t \left(\mathbf{x}_t - \gamma \mathbf{x}_{t+1}\right)^\top \\ &= \sum_s \mu(s)\mathbf{x}_t (\mathbf{x}_t - \gamma \sum_{s'} p(s'|s)\mathbf{x}_{t+1})^\top \\ &= \mathbf{X}^\top \mathbf{D}\mathbf{X} - \mathbf{X}^\top \mathbf{D}\gamma \mathbf{P}\mathbf{X} \\ &= \mathbf{X}^\top \mathbf{D}(\mathbf{I} - \gamma \mathbf{P})\mathbf{X}\end{aligned} \tag{31}$$

Then we use $\mathbf{A}_t$ to present the convergence matrix in the case (i) where the perturbation vector $\mathbf{e}_t$ is added onto the current state features, i.e., $\mathbf{x}_t \leftarrow \mathbf{x}_t + \mathbf{e}_t$, while we use $\mathbf{A}_{t+1}$ and $\mathbf{A}_{t,t+1}$to present the counterparts in the case (ii) and (iii), respectively. Based on Eq. 31, in the case (iii), we have:

$$\begin{aligned}\mathbf{A}_{t,t+1} &= (\mathbf{X} + \mathbf{E})^\top \mathbf{D}(\mathbf{X} + \mathbf{E}) - (\mathbf{X} + \mathbf{E})^\top \mathbf{D}\gamma \mathbf{P}(\mathbf{X} + \mathbf{E}) \\ &= (\mathbf{X} + \mathbf{E})^\top \mathbf{D}(\mathbf{I} - \gamma \mathbf{P})(\mathbf{X} + \mathbf{E})\end{aligned} \tag{32}$$

From (Sutton & Barto, 2018), we know that the inner matrix $\mathbf{D}(\mathbf{I} - \gamma \mathbf{P})$ is the key to determine the positive definiteness of $\mathbf{A}$. If we assume that $\mathbf{A}$ is positive definite, which also indicates that $\mathbf{D}(\mathbf{I} - \gamma \mathbf{P})$ is positive definite equivalently. As such, $\mathbf{A}_{t,t+1}$ is positive definite automatically, and thus the liner TD would converge to the TD fixed point. Next, in the case (i) we have:

$$\begin{aligned}\mathbf{A}_t &= (\mathbf{X} + \mathbf{E})^\top \mathbf{D}(\mathbf{X} + \mathbf{E}) - (\mathbf{X} + \mathbf{E})^\top \mathbf{D}\gamma \mathbf{P}\mathbf{X} \\ &= \mathbf{A} + \mathbf{X}^\top \mathbf{D}\mathbf{E} + \mathbf{E}^\top \mathbf{D}\mathbf{X} + \mathbf{E}^\top \mathbf{D}\mathbf{E} - \mathbf{E}^\top \mathbf{D}\gamma \mathbf{P}\mathbf{X} \\ &= (\mathbf{X} + \mathbf{E})^\top \mathbf{D}(\mathbf{I} - \gamma \mathbf{P})(\mathbf{X} + \mathbf{E}) + (\mathbf{X} + \mathbf{E})^\top \mathbf{D}\gamma \mathbf{P}\mathbf{E} \\ &= \mathbf{A}_{t,t+1} + \gamma (\mathbf{X} + \mathbf{E})^\top \mathbf{D}\mathbf{P}\mathbf{E} \\ &= \mathbf{A}_{t,t+1} + \gamma (\mathbf{X}^\top \mathbf{D}\gamma \mathbf{P}\mathbf{E} + \mathbf{E}^\top \mathbf{D}\gamma \mathbf{P}\mathbf{E})\end{aligned} \tag{33}$$

Similarly, in the case (ii), we can also attain:

$$\begin{aligned}\mathbf{A}_{t+1} &= \mathbf{X}^\top \mathbf{D}\mathbf{X} - \mathbf{X}^\top \mathbf{D}\gamma \mathbf{P}(\mathbf{X} + \mathbf{E}) \\ &= \mathbf{A} - \gamma \mathbf{X}^\top \mathbf{D}\mathbf{P}\mathbf{E}\end{aligned} \tag{34}$$

We know that the positive definiteness of $\mathbf{A}$ and $\mathbf{A}_{t,t+1}$ is only determined by the positive definiteness of the inner matrix $\mathbf{D}(\mathbf{I} - \gamma \mathbf{P})$. If we assume the positive definiteness of $\mathbf{A}$, i.e., the positive definiteness of $\mathbf{A}_{t,t+1}$ and $\mathbf{D}(\mathbf{I} - \gamma \mathbf{P})$, as $\gamma > 0$, what we only need to focus on are the positive definiteness of $\mathbf{X}^\top \mathbf{D}\mathbf{P}\mathbf{E} + \mathbf{E}^\top \mathbf{D}\mathbf{P}\mathbf{E}$ and $-\mathbf{X}^\top \mathbf{D}\mathbf{P}\mathbf{E}$. If they are positive definite, TD learning will converge under their cases, respectively. □

F Proof of Theorem 4 and Corollary 1

Proof. We combine the proof of Theorem 4 and Corollary 1 together. The TD fixed point $\mathbf{w}_{\text{TD}}$ to the system satisfies $\mathbf{A}\mathbf{w}_{\text{TD}} = \mathbf{b}$. Thus, the TD convergence point, i.e., TD fixed point, can be attained by solving the following regression problem:

$$\min_{\mathbf{w}} \|\mathbf{b} - \mathbf{A}\mathbf{w}\|^2 \tag{35}$$

To derive the influence function, consider the contaminated distribution which puts a little more weight on the outlier pair $(\mathbf{x}_t, \mathbf{x}_{t+1})$:

$$\begin{aligned}\hat{\mathbf{w}} = \arg\min_{\mathbf{w}} (1 - \epsilon)\mathbb{E}[(\mathbf{b} - \mathbf{A}\mathbf{w})^\top (\mathbf{b} - \mathbf{A}\mathbf{w})] + \\ \epsilon (y_b - x_A^\top \mathbf{w})^\top (y_b - x_A^\top \mathbf{w}),\end{aligned} \tag{36}$$

where $y_b = R_{t+1}\mathbf{x}_t$ and $x_b = d_t \mathbf{x}_t^\top$. We take the first condition:

$$(1 - \epsilon)\mathbb{E}(2\mathbf{A}^\top \mathbf{A}\mathbf{w} - 2\mathbf{A}^\top \mathbf{b}) - 2\epsilon x_A (y_b - x_A^\top \mathbf{w}) = 0. \tag{37}$$

Then we arrange this equality and obtain:

$$(1 - \epsilon)\mathbb{E}(\mathbf{A}^\top \mathbf{A} + x_A x_A^\top)\mathbf{w}_\epsilon = (1 - \epsilon)\mathbb{E}(\mathbf{A}^\top \mathbf{b}) + \epsilon x_A y_b. \tag{38}$$

Then we take the gradient on ϵ and let $\epsilon = 0$, then we have:

$$\begin{aligned}(-\mathbb{E}(\mathbf{A}^\top \mathbf{A}) + x_A x_A^\top)\mathbf{w}_\epsilon + \mathbb{E}(\mathbf{A}^\top \mathbf{A})\psi_{\mathbf{w}, F_\pi} = -\mathbb{E}(\mathbf{A}^\top \mathbf{b}) \\ + x_A y_b.\end{aligned} \tag{39}$$

We know that under the least square estimation, the closed-form solution of $\mathbf{w}_\epsilon$ is $\mathbb{E}(\mathbf{A}^\top\mathbf{A})^{-1}\mathbb{E}(\mathbf{A}^\top\mathbf{b})$. Thus, after the simplicity, we finally attain:

$$\begin{aligned}\psi_{\mathbf{w},F_\pi}(\mathbf{x}_t,\mathbf{x}_{t+1}) &= \mathbb{E}(\mathbf{A}^\top\mathbf{A})^{-1}x_A(y_b - x_A^\top\mathbf{w}) \\ &= \mathbb{E}(\mathbf{A}^\top\mathbf{A})^{-1}d_t\mathbf{x}_t^\top\mathbf{x}_t(R_{t+1} - d_t^\top\mathbf{w}).\end{aligned} \tag{40}$$

Next, we prove the Corollary. We only need to focus on the item $d_t\mathbf{x}_t^\top\mathbf{x}_t(R_{t+1} - d_t^\top\mathbf{w})$, which we denote as ψ_0. Then we use $\Delta_{x_t}\psi$ and $\Delta_{x_{t+1}}\psi$ to represent the change of ψ after adding perturbations η on $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$, respectively. In particular, since we approximate $\eta\eta^\top\mathbf{x}_t$ and $\eta\eta^\top\mathbf{w}$ as $\mathbf{0}$, then we have that the change of influence function for the perturbation η on the current state feature $\mathbf{x}_t$:

$$\begin{aligned}\Delta_{x_t}\psi &\approx (d_t+\eta)(\mathbf{x}_t^\top\mathbf{x}_t + 2\eta^\top\mathbf{x}_t)(R_{t+1} - d_t^\top\mathbf{w} - \eta^\top\mathbf{w}) - \psi_0 \\ &\approx -d_t\mathbf{x}_t^\top\mathbf{x}_t\eta^\top\mathbf{w} + 2d_t\eta^\top\mathbf{x}_t(R_{t+1} - d_t^\top\mathbf{w}) + \eta\cdot\mathbf{x}_t^\top\mathbf{x}_t(R_{t+1} - d_t^\top\mathbf{w}) \\ &= 2d_t\eta^\top\mathbf{x}_t(R_{t+1} - d_t^\top\mathbf{w}) - \frac{1}{\gamma}(\gamma d_t\mathbf{x}_t^\top\mathbf{x}_t\eta^\top\mathbf{w} - \gamma\eta\mathbf{x}_t^\top\mathbf{x}_t(R_{t+1} - d_t^\top\mathbf{w})).\end{aligned} \tag{41}$$

Then the influence function for the perturbation η on the next state feature $\mathbf{x}_{t+1}$ is:

$$\begin{aligned}\Delta_{x_{t+1}}\psi &= (d_t - \gamma\eta)\mathbf{x}_t^\top\mathbf{x}_t(R_{t+1} - d_t^\top\mathbf{w} + \gamma\eta^\top\mathbf{w}) - \psi_0 \\ &\approx \gamma d_t\mathbf{x}_t^\top\mathbf{x}_t\eta^\top\mathbf{w} - \gamma\eta\mathbf{x}_t^\top\mathbf{x}_t(R_{t+1} - d_t^\top\mathbf{w}).\end{aligned} \tag{42}$$

Finally, it is easy to observe that the following relationship holds:

$$\gamma\Delta_{x_t}\psi = 2\gamma d_t\eta x_t^\top(R_{t+1} - d_t^\top\mathbf{w}) - \Delta_{x_{t+1}}\psi. \tag{43}$$

□

G Experimental Setup

After a linear search, in the QR-DQN, We set $\kappa = 1$ for the Huber quantile loss across all tasks due to its smoothness.

Cart Pole After a linear search, in the QR-DQN, we set the number of quantiles N to be 20, and evaluate both DQN and QR-DQN on 200,000 training iterations.

Mountain Car After a linear search, in the QR-DQN, we set the number of quantiles N to be 2, and evaluate both DQN and QR-DQN on 100,000 training iterations.

Breakout and Qbert After a linear search, in the QR-DQN, we set the number of quantiles N to be 200, and evaluate both DQN and QR-DQN on 12,000,000 training iterations.

H More Experiment Results

We provide the results of robust performance under random noisy state observations in Figure 6 in Cart Pole.

We provide the results of robust performance under adversarial noisy state observations in Figure 7 in Mountain Car.

We provide the results of robust performance under random noisy state observations in Figure 8 in Breakout.

We provide the results of robust performance under random noisy state observations in Figure 9 in Qbert.

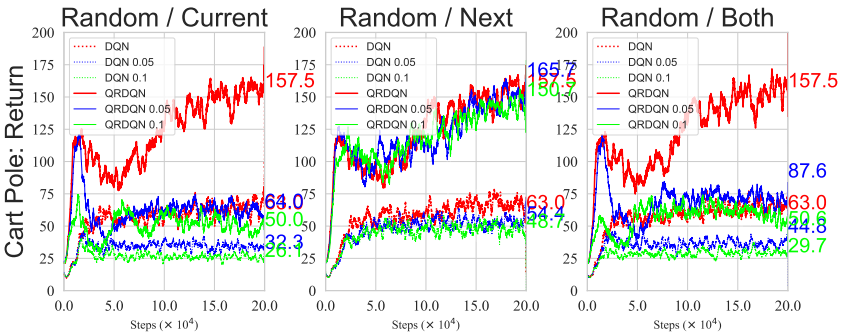

Figure 6: Average returns of DQN and QRDQN against random state observation noises on Cart Pole environment over 200 runs with smooth size 20. QRDQN (solid lines) almost consistently outperforms DQN (dashed lines) in the same color, demonstrating better robustness.

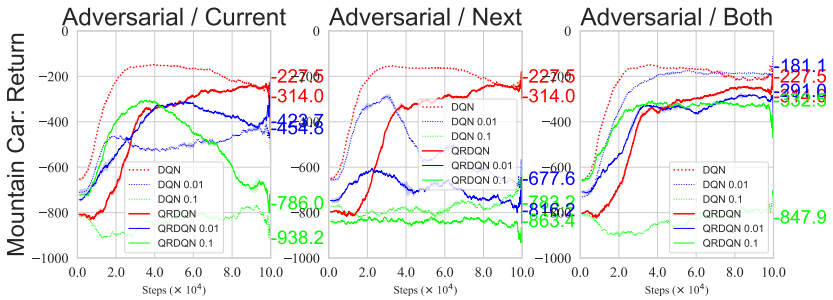

Figure 7: Average returns of DQN and QRDQN against adversarial state observation noises on Mountain Car environment over 200 runs with smooth size 100. QRDQN (solid lines) almost consistently outperforms DQN (dashed lines) in the same color, demonstrating better robustness.

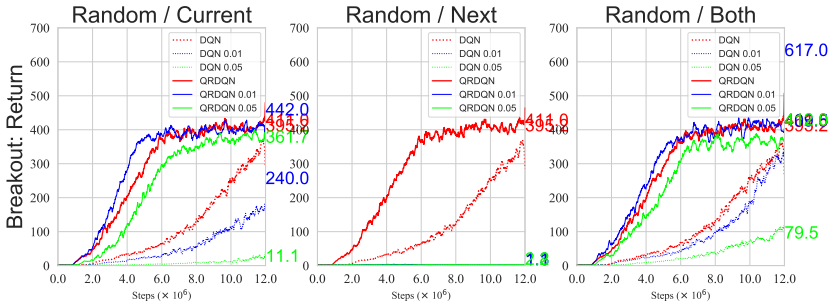

Figure 8: Average returns of DQN and QRDQN against random state observation noises on Breakout environment over 200 runs with smooth size 100. QRDQN (solid lines) almost consistently outperforms DQN (dashed lines) in the same color, demonstrating better robustness.

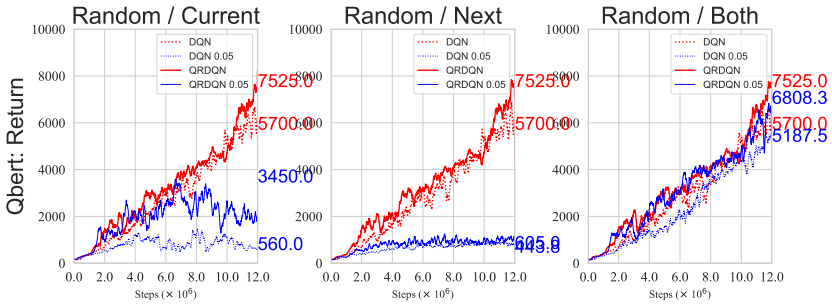

Figure 9: Average returns of DQN and QRDQN against random state observation noises on Qbert environment over 3 runs with smooth size 1000. QRDQN (solid lines) almost consistently outperforms DQN (dashed lines) in the same color, demonstrating better robustness.