Reward Certification for Policy Smoothed Reinforcement Learning

Ronghui Mu[1], **Leandro Soriano Marcolino**[2], **Yanghao Zhang**[1], **Tianle Zhang**[1], **Xiaowei Huang**[1], **Wenjie Ruan**[1*]

[1]Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK
[2]School of Computing & Communication, Lancaster University, Lancaster, LA1 4YW, UK
{r.mu, x.huang, t.zhang,y.zhang}@liverpool.ac.uk, l.marcolino@lancaster.ac.uk, w.ruan@trustai.uk

Abstract

Reinforcement Learning (RL) has achieved remarkable success in safety-critical areas, but it can be weakened by adversarial attacks. Recent studies have introduced "smoothed policies" to enhance its robustness. Yet, it is still challenging to establish a provable guarantee to certify the bound of its total reward. Prior methods relied primarily on computing bounds using Lipschitz continuity or calculating the probability of cumulative reward being above specific thresholds. However, these techniques are only suited for continuous perturbations on the RL agent's observations and are restricted to perturbations bounded by the l_2-norm. To address these limitations, this paper proposes a general *black-box* certification method, called **ReCePS**, which is capable of directly certifying the cumulative reward of the smoothed policy under various l_p-norm bounded perturbations. Furthermore, we extend our methodology to certify perturbations on action spaces. Our approach leverages f-divergence to measure the distinction between the original distribution and the perturbed distribution, subsequently determining the certification bound by solving a convex optimisation problem. We provide a comprehensive theoretical analysis and run experiments in multiple environments. Our results show that our method not only improves the tightness of certified lower bound of the mean cumulative reward but also demonstrates better efficiency than state-of-the-art methods.

Introduction

The utilisation of neural networks in Reinforcement Learning (RL) has achieved remarkable success in safety-critical domains, such as controlling robots and autonomous driving (Sallab et al. 2017; Pan et al. 2017; Johannink et al. 2019; Cai et al. 2023; Wu et al. 2023). Nevertheless, recent research has revealed their vulnerability to the presence of adversarial perturbations (Szegedy et al. 2014; Madry et al. 2017; Goodfellow, Shlens, and Szegedy 2015; Mu et al. 2021; Jin et al. 2022; Zhang et al. 2021; Wang et al. 2022). For example, numerous studies have demonstrated that even well-trained RL policies can suffer significant failures when directly perturbing the observations of the RL agent (Pattanaik et al. 2018) or in action space (Lin et al. 2017). In this regard, it is vital to analyse their robustness before their deployment in safety-critical systems (Christiano et al. 2016; Cheng et al. 2019).

Various empirical defences have been proposed to defend adversarial attacks in RL systems (Manikandan et al. 2011; Pattanaik et al. 2017; Yin et al. 2023), while it has been demonstrated that even robust models can still be compromised by more advanced attack methods (Russo and Proutiere 2019). Hence, there is a need for computing provable guarantees for the trained policy to break the ongoing cycle of attacks and defenses, which is referred to as *robustness certification* (Huang et al. 2020; Ruan, Yi, and Huang 2021). The majority of studies aiming to provide robustness certification primarily focus on classification tasks, while the certification for RL remains largely unexplored.

Compared to classification tasks, certifying RL engages more challenges due to its sequential decision-making nature (Huang, Jin, and Ruan 2023). To address this obstacle, recent efforts have focused on certifying a "smoothed policy" based randomised smoothing strategies. These methods are distinguished due to they do not require access to the internal architecture and parameters of the DNNs, which is referred to as *black-box certification*. For instance, Wu et al. (2021) employed Lipschitz continuity to approximate the final output, but this approach resulted in a loose bound. Instead of directly certifying the lower bound, Kumar, Levine, and Feizi (2021) proposed certifying the reward by breaking it down into several thresholds and then assessing the probability of the cumulative reward staying above a specific threshold. However, this method proved too restrictive, leading to a loss of essential information regarding the output reward and resulting in weaker certificates for smoothed policies. Additionally, both approaches can only handle observation perturbations under the l_2-norm bound. However, in the real world, adversaries can perturb both observations and actions within the bounds of various perturbation constraints of the l_p-norm.

Thus, this paper focuses on effectively certifying the lower bound of mean utility for a policy under diverse l_p-norm perturbations. We present a novel approach based on the generalisation theorem between distributions. By leveraging this theorem, we demonstrate that determining the lower bound of expected utility can be achieved by solving a convex optimisation problem. Doing this enables us to di-

*Corresponding Author

rectly compute the lower bound, resulting in enhanced certification outcomes. Additionally, by employing f-divergence to quantify the distance between distributions, our approach can be expanded to provide certification for a range of l_p-norm bounded perturbations, which includes certifying observation perturbations constrained by the l_1-norm and action space perturbations bounded by the l_0-norm. Our contributions can be summarised as:

(i) We propose a novel methodology to directly certify the cumulative reward of the smoothed policy. This approach uses f-divergence to gauge the separation between the original distribution and the perturbed distribution.

(ii) Our method is capable of handling perturbations bounded by both l_0 and l_1-norm. This work is the first of its kind to consider the certification of the l_0-norm bounded perturbation in the action space.

(iii) By comparing our method with the previous approach, we theoretically validate that our method can enhance certified robustness by taking into account the distribution of cumulative rewards during sampling.

(iv) We empirically demonstrate that our method outperforms the state-of-the-art methods, producing tighter bounds for l_2 perturbations. Our intensive experiments in various environments also validate the effectiveness of our certification method for l_1 perturbations in observation and l_0 perturbations in the action space.

Related Work

Various empirical defences have been proposed to defend adversarial attacks in RL systems (Manikandan et al. 2011; Pattanaik et al. 2017; Eysenbach and Levine 2022), and certain studies have employed adversarial training techniques to enhance policy robustness (Shen et al. 2020; Zhang et al. 2020). The robustness certification mainly focused on classification tasks. They employ various approaches, such as deterministic methods (Ehlers 2017; Ruan, Huang, and Kwiatkowska 2018; Wong and Kolter 2018; Mu et al. 2022; Sun and Ruan 2023; Wang et al. 2023; Zhang, Ruan, and Xu 2023), and probability-based techniques (Lecuyer et al. 2019; Cohen, Rosenfeld, and Kolter 2019; Jin et al. 2023; Zhang, Ruan, and Fieldsend 2022; Mu et al. 2023), to establish lower bounds for the classification accuracy in the presence of specific perturbations. Regarding the certification of robustness in RL, Wu et al. (2021) introduced a method to certify both the reward and action taken at each time step. However, they can only handle non-adaptive adversaries in static scenarios. Another study by Kumar, Levine, and Feizi (2021) aimed to compute a lower bound for the mean cumulative reward in the face of adaptive adversaries, using the Neyman-Pearson Lemma. Yet, their method necessitates dividing the reward into multiple thresholds and subsequently determining whether the reward exceeds these thresholds. In this paper, we present a general technique by solving convex optimization problems, which directly enables the certification of cumulative rewards in the presence of adaptive adversaries, without the need to partition the reward. Broader discussions of related work are given in Appendix H.

Preliminaries

Deep Q-Networks (DQNs)

Markov decision processes (MDP) serve as the foundation of reinforcement learning (RL). It can be formed by a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S} \in \mathbb{R}^D$ represents a set of continuous states, $\mathcal{A}$ is a set of discrete actions, $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the transition function, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ denotes the reward function, and γ is the discount factor $\in [0, 1]$. We denote the stationary policy as $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$. In the state $s_t \in \mathcal{S}$ at the time step t, the agent will take action $a_t \in \mathcal{A} \sim \pi(s_t)$ and then move to the next state $s_{t+1} \sim P(s_t, a_t)$, receiving the reward $R(s_t, a_t)$. The goal is to learn a policy that maximises expected discounted cumulative reward $\mathbb{E}[\sum_t \gamma^t R(s_t, \pi(s_t))]$.

Deep Q networks (Mnih et al. 2013) used a neural network to approximate the maximum expected cumulative reward, which is called the action value function (Q function), after taking action a_t at the state s_t, $Q(s_t, a_t) = R(s_t, a_t) + \gamma \mathbb{E}[\max_a Q(s_{t+1}, a_{t+1})]$. The system consists of two neural networks: one is responsible for updating the network parameters θ, while the other acts as the target network, sharing the same architecture as the first but with fixed parameters. Once the initial network has undergone multiple iterations, its parameters are transmitted to the target network. Additionally, a replay buffer is utilised to store experience tuples generated through interactions with the environment. These tuples are subsequently fed into the neural network to update the parameters.

Policy Smoothed Reinforcement Learning

This paper focuses on evaluating the finite-step Reinforcement Learning with smoothed policy introduced by Kumar, Levine, and Feizi (2021). The smoothing policy training process involves incorporating random smoothing noise from a probability distribution into the observation of the input state s_t of the agents at each time step t. For a detailed demonstration of the training process, please refer to the Appendix A.

Definition 1. *(Smoothed policy) Given policy π, let $\forall s_t \in \mathcal{S}$, the noise vector $\Delta_t \in \mathbb{R}^N$ is $i.i.d$ drawn from the Gaussian distribution $\mathcal{N}(0, \sigma^2 I_N)$, the smoothed policy can be represented as*

$$\tilde{\pi}(s_t) = \pi(s_t + \Delta_t) := \arg\max_{a_t \in \mathcal{A}} \widetilde{Q}^{\pi}(s_t + \Delta_t, a_t) \quad (1)$$

The objective is to establish a lower bound on the expected sum of rewards obtained within a finite time frame when the observation is perturbed under a constraint. Kumar, Levine, and Feizi (2021) proposed to certify the lower bound of the mean reward based on the technique developed by (Kumar et al. 2020) using the empirical cumulative distribution function (CDF) of the probability of surpassing a specific threshold. In this paper, we propose a general framework that can find the lower bound of mean cumulative reward directly via the optimisation approach.

Methodology

Our main objective is to certify the lower bound of the mean cumulative reward during the testing in the presence

of an adversary intentionally perturbing the agent's observations or actions. In particular, perturbations within the action space constitute a distinct instance of l_0 perturbation, which we will explicitly define in following section. In this section, our primary focus is on perturbations in observations.

General Robustness Certification

We consider the general adversarial setting in reinforcement learning (Kumar, Levine, and Feizi 2021; Wu et al. 2021), where the observation of the agent is perturbed by δ_t at each time step t. Suppose that the entire sequence of adversarial perturbation is $\delta = \{\delta_1, \delta_2, ...\}$ and the overall l_p-norm of the perturbation is bounded by ϵ. The adversarial attack aims to reduce the cumulative reward of the smoothed policy by perturbing the observations:

$$\begin{aligned} &\min_{\epsilon} J_\epsilon(\tilde{\pi}) := \sum_{t=0} \gamma^t R(s_t, \pi(s'_t)) = \sum_{t=0} \gamma^t R(s_t, \tilde{\pi}(s_t + \delta_t)), \\ &\text{s.t. } \|\delta_1, \delta_2, \ldots\|_p \leq \epsilon \end{aligned} \tag{2}$$

The robust certification for the provably robust RL with policy π in the finite-step game is to certify the lower bound of the total reward $\underline{J}$ under the norm bound perturbation ϵ.

$$\min_{\epsilon} J_\epsilon(\tilde{\pi}) \geq \underline{J}, \text{ s.t. } \|\delta\|_p \leq \epsilon \tag{3}$$

Nevertheless, it is challenging to obtain the exact solution for this worst-case function. In this paper, instead of investigating the adversarial attack within the observation space $s_t \in \mathcal{S}$, we shift our focus to studying it within the domain of probability measures over observations. The smoothed policy can be interpreted as incorporating random inputs by sampling observations from the distribution $\mu(s_t)$. In other words, we define $\tilde{\pi}$ as $\pi_{o_t \sim \mu(s_t)}(o_t)$, where o_t represents the sampled observation.

In the reinforcement learning (RL) framework, since the perturbation introduced to one step may depend on the current state, previous action, and previous observation. It was shown that incorporating noise into sequential observations within the RL framework does not result in an isometric distribution across the observation space. Therefore, in line with previous research (Kumar, Levine, and Feizi 2021), we expend the complete perturbation budget, denoted as δ, summarised in the initial coordinate of the initial perturbation vector to simplify the analysis.

Definition 2. *Given a state trajectory $(s_0, s_1, ..., s_{T-1})$, we define the **smoothed** state and observation trajectory as $\tau = (s_0, o_0, s_1, o_1, ..., s_{T-1}, o_{T-1})$ where $s_{t+1} \sim P(s_t, \pi(o_t))$ and $o_t \sim \mu(s_t)$. Suppose we perturb the observation at each time step by δ_t, the **perturbed** trajectory can be define as $\tau = (s'_0, o'_0, s'_1, o'_1, ..., s'_{T-1}, o'_{T-1})$ where $s'_t = s_t + \delta_t$, $s'_{t+1} \sim P(s'_t, \pi(o'_t))$ and $o'_t \sim \mu(s'_t)$. In the following analysis in this paper, $\tau \sim \mu(s_t)$ represents the smoothed trajectory and $\tau \sim \mu(s'_t)$ denotes the perturbed trajectory.*

The objective in Eq. 3 can be rewritten as:

$$\min_{q \in \{\mu(s'_t): \|\delta\| \leq \epsilon\}} \underset{\tau \sim q}{J(\tau)} := \sum_{t=0}^{T-1} \gamma^t R(s'_t, \pi(o'_t)) \geq \underline{J} \tag{4}$$

Under the constraint $q \in \mathcal{D}_\epsilon = \{\mu(s'_t) : \|\delta\| \leq \epsilon\}$, this is an infinite-dimension optimisation problem over the space of probability distribution $q \in \mathcal{P}(s_t)$. Suppose that p is the distribution across the smoothed observations, $p = \mu(s_t)$, the divergence constraint can be represented as $\mathcal{D}_\epsilon \subseteq \{q : D(q\|p) \leq \epsilon\}$, where D is the divergence between two distributions.

Divergence Measure

The verification problem can be formulated as an optimisation problem to find the minimum expected utility of the smoothed policy $\tilde{\pi}$:

$$\min_{q \in \mathcal{D}_\epsilon} \underset{\tau \sim q}{\mathrm{E}}[J(\tau)] \tag{5}$$

where $\mathcal{D}_\epsilon \subseteq \{q : D(q\|p) \leq \epsilon\}$. To address this issue, we must establish a particular relaxation of the set $\mathcal{D}_\epsilon$, as it may not be convex and cannot be solved directly. Hence, we will consider the sets $D(q\|p)$ that can be easily optimised over. In this paper, we will explore three broad constraint sets of f-divergences (Csiszár 1967) that can be adapted to evaluate the l_p-norm between two distributions.

Definition 3. *(f-divergence measure (Csiszár 1967)). Given the distribution q, $p \in \mathcal{P}(s)$, f is a convex function with $f(1) = 0$.*

$$D(q||p)_f := \int f\left(\frac{dq}{dp}\right) dp \tag{6}$$

Notably, $D(q||p) \geq 0$, and $D(q||p) = 0$ if $q = p$. Given the reference distribution p and the f-divergence $D(q||p)$, the constraint set under the bound $\epsilon_D \geq 0$ can be defined as

$$\mathcal{D}_f = \{q \in \mathcal{P}(s) : D(q\|p) \leq \epsilon_D\} \tag{7}$$

Optimisation Based Certification Approach

In this section, we demonstrate our approach to tackling the challenges of the solving problem given in Eq.4, which is achieved by converting it into a convex optimisation problem. Our methodology is versatile, capable of accommodating various constraint sets $\mathcal{D}$ defined by diverse f-divergences. The foundational support for constructing the convex optimisation problem is derived from the Generalised Donsker-Varadhan Variational Formula, as presented in Theorem 4.2 of Ben-Tal and Teboulle (2007). This formula plays a pivotal role in establishing a duality connection between the primal and dual forms of the Optimal Control of Expectations (OCE) optimisation problem.

In this paper, we extend the utility of this theorem, adapting it as a fundamental theorem to determine the optimal lower bound for the expected cumulative reward within the framework of RL.

Corollary 1. [1] *(Adaptive Generalized Donsker-Varadhan Variational Formula for RL) Given the f-divergence, Suppose $q = \mu(s'_t), p = \mu(s_t)$, let $z(\tau) = \frac{q(\tau)}{p(\tau)}$, which represents the likelihood ratio over the full trajectory τ. $J(\tau)$ denoting as the cumulative reward of the smoothed policy. Given a convex function f with $f(1) = 0$, we have*

$$\begin{aligned} &\min_z \left\{ D(q||p) + \underset{\tau \sim p}{\mathrm{E}}[z(\tau) J(\tau)] \right\} \\ &= \max_{\eta \in \mathrm{R}} \left\{ \eta - \underset{\tau \sim p}{\mathrm{E}}[f^*(\eta - J(\tau))] \right\}, \end{aligned} \tag{8}$$

[1] All proofs for Theorems are available in the Appendix.

where f^* represents the convex conjugate (the Legendre–Fenchel transform) of f, $f^*(x) = \max_{y>0}(xy - f(y))$.

Building on Corollary 1, we derive the pivotal theorem in this paper, which serves as the objective for the convex optimisation problem.

Theorem 1. *Given a convex function f with $f(1) = 0$, and f^* is its convex conjugate. Under the f-divergence constraint $D_f(q\|p) \leq \epsilon_D$ with $\epsilon_D \geq 0$, let $z(\tau) = \frac{q(\tau)}{p(\tau)}$, we can solve the following convex optimisation problem to find the optimal value for the expected cumulative reward of the smoothed policy $\tilde{\pi}(s)$:*

$$\max_{\nu>0,\eta\in\mathbb{R}} \left\{ \nu \left[\eta - \mathop{\mathrm{E}}_{\tau\sim p} \left(f^* \left(\eta + \epsilon - \frac{J(\tau)}{\nu} \right) \right) \right] \right\} \tag{9}$$

Proof. Based on Corollary 1, we can rewrite the formula on the right side:

$$\begin{aligned}
&\max_{\nu>0,\eta\in\mathbb{R}} \left\{ \nu \left[\eta - \mathop{\mathrm{E}}_{\tau\sim p} \left(f^* \left(\eta + \epsilon - \tfrac{J(\tau)}{\nu} \right) \right) \right] \right\} \\
&= \max_{\nu>0,\eta\in\mathbb{R}} \left\{ \nu \left[\eta + \epsilon - \mathop{\mathrm{E}}_{\tau\sim p} \left(f^* \left(\eta + \epsilon - \tfrac{J(\tau)}{\nu} \right) \right) - \epsilon \right] \right\} \\
&= \max_{\nu>0} \left\{ \nu \left[\max_{t\in\mathbb{R}} \left\{ t - \mathop{\mathrm{E}}_{\tau\sim p} \left(f^* \left(t - \tfrac{J(\tau)}{\nu} \right) \right) \right\} - \epsilon \right] \right\} \\
&= \max_{\nu>0} \left\{ \nu \left[\min_z \left\{ D(q||p) + \mathop{\mathrm{E}}_{\tau\sim p} \left[z(\tau) \tfrac{J(\tau)}{\nu} \right] \right\} - \epsilon \right] \right\} \\
&= \max_{\nu>0} \left\{ \min_z \left\{ \mathop{\mathrm{E}}_{\tau\sim p} \left[z(\tau) J(\tau) \right] + \nu \left(D(q||p) - \epsilon \right) \right\} \right\} \\
&= \min_z \left\{ \mathop{\mathrm{E}}_{\tau\sim p} \left[z(\tau) J(\tau) \right] \right\} \\
&= \min_{\tau\sim q} \left\{ \mathrm{E} \left[J(\tau) \right] \right\}
\end{aligned}$$

As a result, maximising the problem stated in Eq. 9 can give the lower bound for the mean cumulative reward. □

To solve the convex optimisation problem in Eq. 9, we need to estimate the expected value of $\mathop{\mathrm{E}}_{\tau\sim p} \left(f^* \left(\eta + \epsilon - \frac{J(\tau)}{\nu} \right) \right)$. In the RL system, due to its sequential decision-making nature, calculating the precise expectation of cumulative reward in closed form is challenging. To overcome this obstacle, we employ Monte Carlo Sampling as a means to approximate the cumulative reward. Given a random variable with distribution $p(x) = \mu(x)$, where $x^1, x^2, ..., x^M$ are drawn independently and identically from distribution $p(x)$, the expected value of x can be approximated as $\tilde{\mu}(x) = \frac{1}{M} \sum_{i=1}^{M} x^i$.

Our verification procedure is outlined in Algorithm 1. Initially, we generate M episodes by introducing noise Δ to observations, where the noise values Δ can be independently and identically drawn from a normal distribution $\mathcal{N}(0, \sigma^2 I_D)$. In the context of RL, we sample random noise from a fixed distribution and incorporate noise with observations to observe multiple instances of cumulative reward. Subsequently, we observe M cumulative rewards $(J_1, J_2, ..., J_M)$. These samples are then utilised to solve the jointly convex optimisation problem (Eq. 9), which can be efficiently addressed using a powerful tool (Diamond and Boyd 2016). Finally, we obtain the optimal values of ν and η through this optimisation process.

Algorithm 1: Certified Robustness Bound of the Cumulative Reward

Input: Soothed Policy $\tilde{\pi}$; action value function $Q^{\tilde{\pi}}$; perturbation bound ϵ; divergence function f

Parameter: small sampling size M; large sampling size $\tilde{M}$; Gaussian distribution parameter σ; confidence parameter α

function SMOOTHINGREWARDSET($M, Q^{\tilde{\pi}}$)
  for $m \leftarrow 1, M$ **do**
    $J_m \leftarrow 0$
    for $t \leftarrow 1, T$ **do**
      Generate noise $\Delta \sim \mathcal{N}(0, \sigma^2 I_D)$
      $s'_t \leftarrow s_t + \Delta$
      $a_t \leftarrow \arg\max_{a_t \in \mathcal{A}} Q^{\tilde{\pi}}(s'_t, a_t)$
      Execute a_t and obtain R and s_{t+1}
      $J_m \leftarrow J_m + R$
  $\mathbf{J} \leftarrow \{J_1, ..., J_M\}$
return J
function OPTIMISATION($M, Q^{\tilde{\pi}}, \epsilon, f$)
  $\mathbf{J} \leftarrow SmoothingRewardSet(M, Q^{\tilde{\pi}})$
  Solve the convex optimisation problem

$$\max_{\nu>0,\eta\in\mathbb{R}} \left\{ \nu \left[\eta - \sum_{i=1}^{M} \left(f^* \left(\eta + \epsilon - \frac{J_i}{\nu} \right) \right) \right] \right\}$$

  return v^*, η^*
▷ Compute the lower bound of certified reward
$v^*, \eta^* \leftarrow Optimisation(M, Q^{\tilde{\pi}}, \epsilon, f)$
$\tilde{\mathbf{J}} \leftarrow SmoothingRewardSet(\tilde{M}, Q^{\tilde{\pi}})$
Substitute the v^* and η^* to compute $f_i^* \left(\eta^* + \epsilon - \frac{\tilde{J}_i}{\nu^*} \right)$ for $i = 1, 2, ..., \tilde{M}$
Apply Eq. 10 to compute E_U by replacing $h(\tau_i)$ with f_i^* .
Substitute the expectation term by E_U in Eq. 9 to obtain the final lower bound of mean reward.

To establish a confidence guarantee for the lower bound of the objective function, we need to estimate the maximum value of the expectation term denoted as $E_U \geq \mathop{\mathrm{E}}_{\tau\sim p} \left(f^* \left(\eta + \epsilon - \frac{J(\tau)}{\nu} \right) \right)$. To achieve this approximation, we increase the number of i.i.d. samples, denoted as $\tilde{M}$, drawn from the distribution p. These additional samples are used to compute an upper bound with high confidence, centred around the empirical estimation of the expectation term. In this paper, we employ the widely recognised Hoeffding's bound (Shivaswamy and Jebara 2010) to approximate the lower bound of this expectation.

Definition 4. *(Hoeffding's bound evaluation) At each state, we randomly sampling perturbed observation i.i.d from distribution $p = \mu(s_t)$. We run M episodes to obtain cumulative rewards, $(J(\tau_1), J(\tau_2), ..., J(\tau_M))$. Then, for any $\alpha > 0$, we can obtain the provable evaluation bound of the expectation*

value of $h(\tau)$*:*

$$P\left(|\frac{1}{M}\sum_{i}^{M}h(\tau_i) - \underset{\tau\sim p}{E}(h(\tau))| \leq \zeta\right) \leq 1-\alpha$$

where $h(\tau_i) = f^*\left(\eta + \epsilon - \frac{J(\tau_i)}{\nu}\right)$.

The Definition 4 remains valid when $\zeta = \sqrt{\frac{R^2}{2M}\log\frac{2}{\alpha}}$, where R is the range of $h(\tau)$. Therefore, with confidence at least $1-\alpha$, we have

$$E_U = \frac{1}{M}\sum_{i}^{M}h(\tau_i) + \zeta. \tag{10}$$

Comparison with Prior Work

Two existing studies concerning the certification of cumulative rewards in RL are Kumar, Levine, and Feizi (2021) and Wu et al. (2021). In (Kumar, Levine, and Feizi 2021), the authors tackle certification by forming it as a classification problem, determining whether the cumulative reward surpasses a specific threshold. They utilise integrals to calculate the total expected reward through the cumulative distribution function. Conversely, Wu et al. (2021) directly certifies the lower boundary of the anticipated reward by exploiting the Lipschitz continuity of the smoothed reward. Nevertheless, in comparison to both our work and the methodology proposed by Kumar, Levine, and Feizi (2021), this approach demonstrates a relatively weak performance, as indicated by the findings in our experimental results provided in the Appendix I. This divergence in performance could be potentially due to their emphasis on certifying per-step perturbations. Consequently, in this article, we will present a theoretical analysis that elucidates why our method outperforms the approach introduced in (Kumar, Levine, and Feizi 2021). The problem they studied in may be treated as a special case of our methodology. When considering the case of smoothed Gaussian noise i.i.d sampling from distribution $\mathcal{N}(0,\sigma^2 I_D)$, the Hockey-Stick divergence can be analytically calculated, which can be expressed as a function of the l_2-norm, as established in Dong et al. (2021) (detailed proof in Appendix D). The Hockey-Stick divergence, denoted as $D_{HS,\lambda}(q||p)$, is defined by the function $f(x) = max(x-\lambda,0) - max(1-\lambda,0)$.

The certification framework proposed in (Kumar, Levine, and Feizi 2021) involves working with a reward range and separating it into multiple thresholds. We can reformulate our optimisation objective to recover the certification result presented in (Kumar, Levine, and Feizi 2021) as following:

Theorem 2. *(CDF-based optimisation framework) Suppose the reward range is (a, b) and considering* n *thresholds, denoted as* $a < g_1 < g_2 < ... < g_n < b$*. Let* θ_i *represent the lower bound of probability that the total reward obtained by the smoothed policy is above the threshold* g_i*. Let f* be the convex conjugate of Hockey-Stick divergences, we have*

$$\begin{aligned}&\max_{\nu>0,\eta\in\mathbb{R}} \nu\big[\eta - \big((1-\theta_1)f^*\left(\eta+\epsilon-\tfrac{a}{\nu}\right)\\ &+\textstyle\sum_{i=1}^{n}(g_{i+1}-g_i)\theta_i f^*\left(\eta+\epsilon-\tfrac{g_i}{\nu}\right)+\theta_n f^*\left(\eta+\nu\epsilon-\tfrac{g_n}{\nu}\right)\big)\big]\end{aligned}$$

Therefore, to compare our algorithm with the CDF-based algorithm in (Kumar, Levine, and Feizi 2021), let $H(g_i) = P(J(\tau) \geq g_i)$ represent the probability of the cumulative reward being above s_i, we have the following derivation:

$$\begin{aligned}&\eta - ((1-\theta_1)f^*\left(\eta+\epsilon-\tfrac{a}{\nu}\right)+\\ &\textstyle\sum_{i=1}^{n}(g_{i+1}-g_i)\theta_i f^*\left(\eta+\epsilon-\tfrac{g_i}{\nu}\right)+\theta_n f^*\left(\eta+\epsilon-\tfrac{g_n}{\nu}\right))\\ =&\eta - \underset{\tau\sim p}{E}[f^*\left(\eta+\epsilon-\tfrac{a}{\nu}\right)H(a)+\textstyle\sum_{i=1}^{n}f^*\left(\eta+\epsilon-\tfrac{g_i}{\nu}\right)H(g_i)\\ &+f^*\left(\eta+\epsilon-\tfrac{g_n}{\nu}\right)H(g_n)]\\ \leq&\eta - \underset{\tau\sim p}{\mathrm{E}}[f^*(\left(\eta+\epsilon-\tfrac{a}{\nu}\right)H(a)+\textstyle\sum_{i=1}^{n}\left(\eta+\epsilon-\tfrac{g_i}{\nu}\right)H(g_i)\\ &+\left(\eta+\epsilon-\tfrac{g_i}{\nu}\right)H(g_n))]\\ =&\eta - \underset{\tau\sim p}{\mathrm{E}}[f^*((\eta+\epsilon)(H(a)+\\ &\textstyle\sum_{i}^{n}H(g_i)) - \tfrac{a}{\nu}H(a) - \sum_{i=1}^{n}\tfrac{g_i}{\nu}H(g_i))]\\ =&\eta - \underset{\tau\sim p}{\mathrm{E}}[f^*(\eta+\epsilon-\tfrac{a}{\nu}H(a) - \textstyle\sum_{i=1}^{n}\tfrac{g_i}{\nu}H(g_i))]\\ =&\eta - \underset{\tau\sim p}{\mathrm{E}}[f^*(\eta+\epsilon-\tfrac{J(\tau)}{\nu})]\end{aligned}$$

The inequality in the derivation follows the Jensen's inequality. Therefore, by optimising the Eq. 9, it is possible to achieve either equivalent or more stringent bounds as those in Kumar, Levine, and Feizi (2021). The experimental optimisation objective aimed at certifying the l_2-norm perturbation is detailed below:

Theorem 3. *(Optimisation Objective to verify* l_2*-norm bound by Gaussian Noise) Given the parameter* λ *of the Hockey-Stick divergence, we can solve the following convex optimisation problem to find the minimum bound of the mean cumulative reward:*

$$\begin{aligned}\max_{\nu>0,\eta\in\mathbb{R}} \quad &\{\eta - \underset{\tau\sim p}{\mathrm{E}}[\max((\eta+\nu\epsilon-J(\tau))\lambda,0)\\ &+\nu\max(1-\lambda,0)]\} \quad s.t \quad \eta \leq \nu(1-\epsilon)+J(\tau)\end{aligned}$$

Certification on l_1-Norm

To certify the l_1-norm perturbation, the problem in the context of RL can be formulated as the accumulation of perturbation magnitudes applied to the agent's observations, constrained by ϵ_D. The objective is to find ϵ_D that satisfies $\mathcal{D}_{\epsilon,s_t} := \{\mu(s_t+\delta) : ||\delta||_1 \leq \epsilon\} \subseteq \mathcal{D}_{\epsilon_D,s_t} := \{q \in \mathcal{P}(s_t) : D_f(q||p) \leq \epsilon_D\}$. Suppose we draw noise from the Gaussian distribution $\mathcal{N}(0,\sigma^2 I_D)$, then we can adapt Theorem I in (Barsov and Ulyanov 1987). This allows us to employ the total variation distance $TV(q||p) = D_f(q||p)$, where $f(x) = \frac{1}{2}|x-1|$, as a measurement based on the l_1-norm.

Theorem 4. *Given state* $s_t \in \mathcal{S}$*, we consider two Gaussian distributions:* $p = \mathcal{N}(s_t,\sigma^2 I_D)$ *and* $q = \mathcal{N}(s_t',\sigma^2 I_D)$*, where* $s_t' = s_t + \delta_t$*. Under the constraint of* $||(\delta_1,\delta_2,...)||_1 \leq \epsilon$*, the total variation distance for any two distributions* (p,q) *is defined as* $\epsilon_{TV} = 2\Phi(\frac{\epsilon}{2\sigma^2} - 1)$*, where* Φ *is the CDF of standard norm distribution.*

The objective function to certify the lower bound of the expected utility under l_1*-norm perturbation is:*

$$\begin{aligned}\max_{\nu>0,\eta\in\mathbb{R}} &\left\{\eta - \underset{\tau\sim p}{\mathrm{E}}\left[\max(\eta+\nu\epsilon-J(\tau),-\tfrac{\nu}{2})\right]\right\}\\ &s.t \quad \eta \leq \nu(\tfrac{1}{2}-\epsilon)+\min(J(\tau))\end{aligned} \tag{11}$$

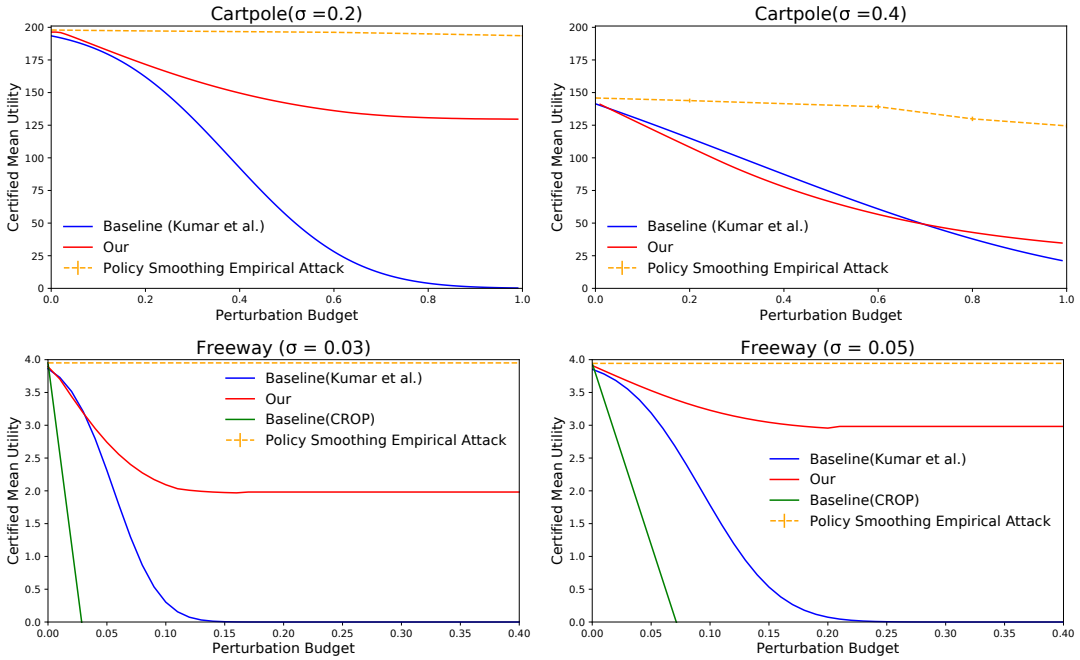

Figure 1: Comparison of the certified lower bound of mean utility under l_2-norm perturbation (where higher certified utility is better). For 'Cartpole', the baseline has a runtime of 0.039s per budget, while ours is 0.016s. For 'Freeway', both our method and the baseline exhibit similar running times, both at 0.03s per budget. The dashed line represents the empirical attack result.

Certification on l_0-Norm

Our approach can also be extended to certify perturbations bounded by l_0-norm. Within the RL domain, the l_0-norm can be employed to quantify the number of actions manipulated when performing attacks in the discrete action space. We define a smoothed policy denoted as $\tilde{\pi}$, which is developed by introducing an element of randomness by not selecting the best action at each step with probability k. For the certification, we choose the worst action with a probability of k at each step. The objective is to establish a lower bound for the mean cumulative reward under an l_0-norm bound, which specifies the maximum number of allowed actions that can be changed. In this paper, we use the Rényi divergence with parameter β to measure the l_0 perturbation, and the optimisation objective can be formulated as follows:

Theorem 5. *Given a trained smoothed policy $\tilde{\pi}$ with action space $\mathcal{A} : \{a_1, ..., a_N\}$. Suppose at each step the policy has a probability of k to select the best action a^*, and a probability of $e = 1 - k$ to select an action different from a^*. The total number of time steps is denoted as T. The action sequences $\mathcal{A}$ that are selected by the smoothed policy can be viewed as on the distribution of $\mu(\mathcal{A}) = \prod_{i=1}^{T}(1-e)^{1[a_i=a_i^*]}\left(\frac{e}{N-1}\right)^{1[a_i\neq a_i^*]}$. The certification objective is: $\max_{\nu>0,\eta\in\mathbb{R}}(\nu\eta - \mathop{\mathrm{E}}_{\tau\sim p}[\mathcal{Z}])$, where*

$$\mathcal{Z} = \begin{cases} \nu + \nu^{\frac{-1}{\beta-1}}(\beta-1)\left(\frac{\max(x,0)}{\beta}\right)^{\frac{\beta}{\beta-1}} & if \quad \beta > 1 \\ -\nu + \nu^{\frac{-1}{\beta-1}}(1-\beta)\left(\frac{x}{-\beta}\right)^{\frac{\beta}{\beta-1}} & if \quad 0 \leq \beta < 1, x \leq 0 \end{cases}$$

where $x = \nu\eta + \nu\epsilon - J(\tau)$.

Experiments

We first demonstrate the experimental results of our method - **ReCePS** [2], compared with the baseline to certify the cumulative reward under l_2-norm bounded perturbation. Then, we show the results of certifying the l_1-norm perturbation on observations and the l_0-norm perturbation on action space.

Environments We followed the baseline (Kumar, Levine, and Feizi 2021) to apply our algorithm on three standard environments: two classic control environments ('Cartpole' and 'Mountain Car') and one high-dimensional Atari game ('Freeway'). Due to the page limit, we present the results for 'Cartpole' and 'Freeway' in the main paper, and extra experiments are provided in Appendix I.

RL algorithm Among the environments, the 'Cartpole' and two Atrari games adapt a discrete action, and 'Mountain Car' utilises a continuous action space. We use the Deep-Q-Network (DQN) (Mnih et al. 2013) to train the policy for the discrete action space and we use deep Deterministic Gradient (DDPG) (Lillicrap et al. 2015) to train the policy based on continuous action space (Details are in Appendix A).

Environments setup In all experiments, we initially select $M = 1000$ to determine the optimal value in the optimisation function. Subsequently, we set $\tilde{M} = 10000$ to derive the lower bound for the mean cumulative reward. The confidence level for all experimental outcomes, including baselines, is set at $1 - \alpha = 0.99$.

[2]Our code is available at https://github.com/TrustAI/ReCePS

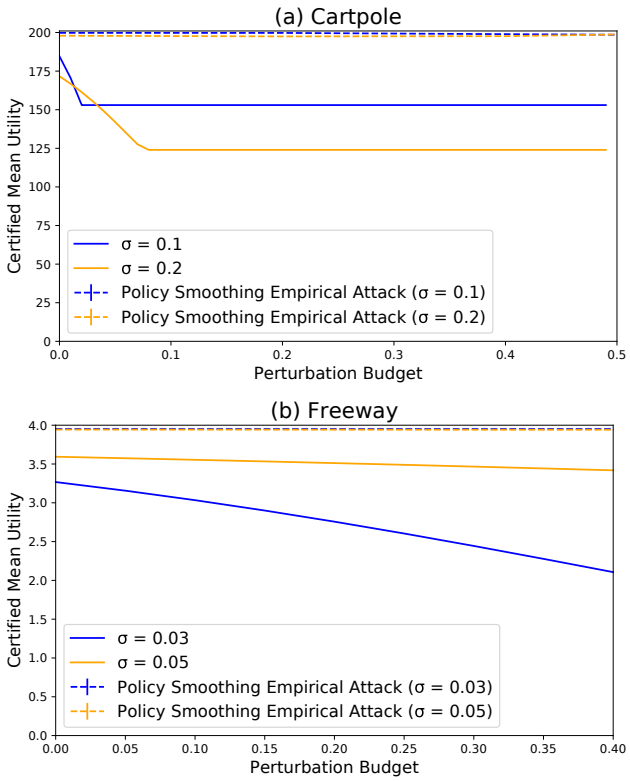

Figure 2: Certified lower bound of mean cumulative reward, under l_1-norm bounded perturbation.

Certified Bound Under l_2 & l_1 Perturbations

In scenarios where the adversary can influence the agent's observations, we adopt the approach from prior research, assuming that the adversary intervenes in each frame only once, right when it is initially observed. For both the training and testing phases, we introduce noise to create a defence policy based on smoothing techniques.

For the certification of l_2-norm perturbation, we compare our method with Kumar, Levine, and Feizi (2021), which is also designed to certify the adaptive RL adversary and follow the same setting to make a fair comparison. Wu et al. (2021) also proposed an algorithm to certify the cumulative reward, but they focused on certifying the policy at each step against a non-adaptive adversary. Their global reward certification method based on Lipchitz continuity is too weak compared with Kumar, Levine, and Feizi (2021) and our method (results presented in Appendix I). We use the Hocky-Stick divergence to measure the norm distance and tune the parameter λ to achieve the best certificate. As shown in Figure 1, our method can obtain a tighter certified bound when we increase the perturbation budget, especially for the 'Freeway' environment. However, we can also observe that when we increase the smoothing parameter σ, the bound of our method gets closer to the bound obtained by the baseline. This could occur because as the σ grows, the variance of the cumulative output reward distribution also increases. Consequently, employing the mean across several simulations to estimate the expectation term might result in inaccuracies when solving the optimisation problem.

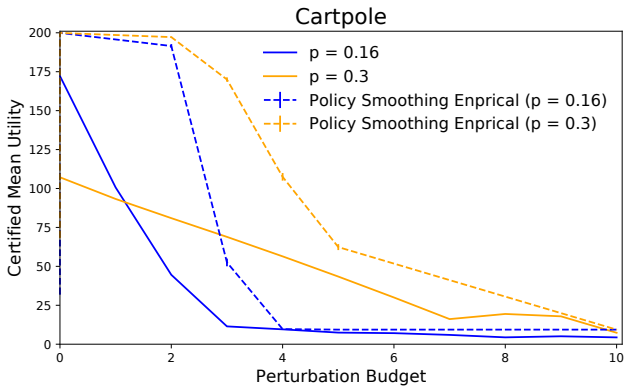

Figure 3: Certified lower bound and empirical attack result of mean cumulative reward, under l_0-norm bounded perturbation on action space in 'Cartpole' environment with two discrete actions. p is the probability of changing the action.

Concerning the certification of the l_1 norm perturbation, we present the results in Figure 2. These perturbations are structured as the sum of perturbations applied to all states, assuming that the adversary introduces perturbations to state observations throughout the entire episode. We follow the same procedure of Kumar, Levine, and Feizi (2021) to perform the attack. The outcomes reveal a clear trend: elevating the σ value enhances policy robustness, but excessive σ values can lead to reduced performance (lower reward).

Certified Bound Under l_0 Perturbations

For the perturbation on action space, we adapted the attack method in Lin et al. (2017), which performs an attack when the gap between the value of the best and worst actions is above a threshold. To defend against such attacks, during training and testing time we will enforce the agent to select the worst action with probability p. By adopting the Rényi divergence, we can compute the lower bound of cumulative reward by finding the optimal value of the convex optimisation problem in Theorem 5. However, as the problem is non-linear, we use the 'SciPy' optimise Python package to solve it, and we tune the parameter of Rényi divergence, β, to obtain the optimal certificate. We present the results of the experiment in 'Cartpole' environment in Figure 3 and experiments on other environments can be found in Appendix I. It's evident that increasing the smoothing parameter p enhances the policy's robustness against attacks.

Conclusion

We introduce **ReCePS** – a general framework to provide a lower bound of mean cumulative reward certification that can handle perturbations bounded by various l_p norms. Our approach employs the f-divergence to quantify the magnitude of adversarial perturbation. By solving a convex optimisation problem, we compute the optimal minimum utility. Our experiments demonstrate that ReCePS yields tighter bounds than state-of-the-art solutions, extending its applicability to perturbations within the action space.

Acknowledgements

The research is supported by the UK EPSRC under project EnnCORE [EP/T026995/1]. We thank the High-End Computing facility at Lancaster University for the computing resources.

References

Barsov, S.; and Ulyanov, V. 1987. Estimates of the proximity of Gaussian measures. *Doklady Mathematics*, 34: 462–.

Ben-Tal, A.; and Teboulle, M. 2007. An old-new concept of convex risk measures: The optimized certainty equivalent. *Mathematical Finance*, 17(3): 449–476.

Cai, Y.; Zhang, C.; Shen, W.; Zhang, X.; Ruan, W.; and Huang, L. 2023. RePreM: Representation Pre-training with Masked Model for Reinforcement Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'23)*.

Cheng, R.; Orosz, G.; Murray, R. M.; and Burdick, J. W. 2019. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3387–3395.

Christiano, P.; Shah, Z.; Mordatch, I.; Schneider, J.; Blackwell, T.; Tobin, J.; Abbeel, P.; and Zaremba, W. 2016. Transfer from simulation to real world through learning deep inverse dynamics model. *arXiv preprint arXiv:1610.03518*.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, 1310–1320. PMLR.

Csiszár, I. 1967. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica*, 2: 299–318.

Diamond, S.; and Boyd, S. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83): 1–5.

Dong, Y.; Yang, X.; Deng, Z.; Pang, T.; Xiao, Z.; Su, H.; and Zhu, J. 2021. Black-box detection of backdoor attacks with limited information and data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16482–16491.

Ehlers, R. 2017. Formal verification of piece-wise linear feed-forward neural networks. In *Automated Technology for Verification and Analysis: 15th International Symposium, ATVA 2017, Pune, India, October 3–6, 2017, Proceedings 15*, 269–286. Springer.

Eysenbach, B.; and Levine, S. 2022. Maximum Entropy RL (Provably) Solves Some Robust RL Problems. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Huang, X.; Jin, G.; and Ruan, W. 2023. Deep reinforcement learning. In *Machine Learning Safety*, 219–235. Springer.

Huang, X.; Kroening, D.; Ruan, W.; Sharp, J.; Sun, Y.; Thamo, E.; Wu, M.; and Yi, X. 2020. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37: 100270.

Jin, G.; Yi, X.; Huang, W.; Schewe, S.; and Huang, X. 2022. Enhancing adversarial training with second-order statistics of weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15273–15283.

Jin, G.; Yi, X.; Wu, D.; Mu, R.; and Huang, X. 2023. Randomized adversarial training via taylor expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16447–16457.

Johannink, T.; Bahl, S.; Nair, A.; Luo, J.; Kumar, A.; Loskyll, M.; Ojea, J. A.; Solowjow, E.; and Levine, S. 2019. Residual reinforcement learning for robot control. In *2019 International Conference on Robotics and Automation (ICRA)*, 6023–6029. IEEE.

Kumar, A.; Levine, A.; and Feizi, S. 2021. Policy smoothing for provably robust reinforcement learning. *arXiv preprint arXiv:2106.11420*.

Kumar, A.; Levine, A.; Feizi, S.; and Goldstein, T. 2020. Certifying confidence via randomized smoothing. *Advances in Neural Information Processing Systems*, 33: 5165–5177.

Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, 656–672. IEEE.

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

Lin, Y.; Hong, Z.; Liao, Y.; Shih, M.; Liu, M.; and Sun, M. 2017. Tactics of Adversarial Attack on Deep Reinforcement Learning Agents. In Sierra, C., ed., *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 3756–3762. ijcai.org.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *CoRR*, abs/1706.06083.

Manikandan, S.; et al. 2011. Measures of central tendency: Median and mode. *J Pharmacol Pharmacother*, 2(3): 214–215.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mu, R.; Ruan, W.; Marcolino, L. S.; Jin, G.; and Ni, Q. 2023. Certified policy smoothing for cooperative multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15046–15054.

Mu, R.; Ruan, W.; Marcolino, L. S.; and Ni, Q. 2022. 3DVerifier: efficient robustness verification for 3D point cloud models. *Machine Learning*, 1–28.

Mu, R.; Ruan, W.; Soriano Marcolino, L.; and Ni, Q. 2021. Sparse Adversarial Video Attacks with Spatial Transformations. In *The 32nd British Machine Vision Conference (BMVC'21)*.

Pan, X.; You, Y.; Wang, Z.; and Lu, C. 2017. Virtual to real reinforcement learning for autonomous driving. *arXiv preprint arXiv:1704.03952*.

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

Pattanaik, A.; Tang, Z.; Liu, S.; Bommannan, G.; and Chowdhary, G. 2018. Robust Deep Reinforcement Learning with Adversarial Attacks. In André, E.; Koenig, S.; Dastani, M.; and Sukthankar, G., eds., *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS 2018, Stockholm, Sweden, July 10-15, 2018*, 2040–2042. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM.

Ruan, W.; Huang, X.; and Kwiatkowska, M. 2018. Reachability analysis of deep neural networks with provable guarantees. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, 2651–2659.

Ruan, W.; Yi, X.; and Huang, X. 2021. Adversarial Robustness of Deep Learning: Theory, Algorithms, and Applications. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM'21)*, 4866–4869.

Russo, A.; and Proutiere, A. 2019. Optimal attacks on reinforcement learning policies. *arXiv preprint arXiv:1907.13548*.

Sallab, A. E.; Abdou, M.; Perot, E.; and Yogamani, S. 2017. Deep reinforcement learning framework for autonomous driving. *arXiv preprint arXiv:1704.02532*.

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

Shivaswamy, P.; and Jebara, T. 2010. Empirical bernstein boosting. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 733–740. JMLR Workshop and Conference Proceedings.

Sun, S.; and Ruan, W. 2023. TextVerifier: Robustness Verification for Textual Classifiers with Certifiable Guarantees. In *Findings of the Association for Computational Linguistics: ACL 2023*, 4362–4380.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Wang, F.; Xu, P.; Ruan, W.; and Huang, X. 2023. Towards Verifying the Geometric Robustness of Large-scale Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'23)*.

Wang, F.; Zhang, C.; Xu, P.; and Ruan, W. 2022. Deep learning and its adversarial robustness: A brief introduction. In *HANDBOOK ON COMPUTER LEARNING AND INTELLIGENCE: Volume 2: Deep Learning, Intelligent Control and Evolutionary Computation*, 547–584.

Wong, E.; and Kolter, Z. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, 5286–5295. PMLR.

Wu, F.; Li, L.; Huang, Z.; Vorobeychik, Y.; Zhao, D.; and Li, B. 2021. Crop: Certifying robust policies for reinforcement learning through functional smoothing. *arXiv preprint arXiv:2106.09292*.

Wu, H.; Yunas, S.; Rowlands, S.; Ruan, W.; and Wahlström, J. 2023. Adversarial driving: Attacking end-to-end autonomous driving. In *2023 IEEE Intelligent Vehicles Symposium (IV)*, 1–7. IEEE.

Yin, X.; Wu, S.; Liu, J.; Fang, M.; Zhao, X.; Huang, X.; and Ruan, W. 2023. ReRoGCRL: Representation-based Robustness in Goal-Conditioned Reinforcement Learning. *arXiv preprint arXiv:2312.07392*.

Zhang, C.; Ruan, W.; and Xu, P. 2023. Reachability Analysis of Neural Network Control Systems. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'23)*.

Zhang, H.; Avrithis, Y.; Furon, T.; and Amsaleg, L. 2021. Walking on the Edge: Fast, Low-Distortion Adversarial Examples. *IEEE Transactions on Information Forensics and Security*, 16: 701–713.

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

Zhang, T.; Ruan, W.; and Fieldsend, J. E. 2022. PRoA: A Probabilistic Robustness Assessment against Functional Perturbations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD'22)*.