Game-Theoretic Robust Reinforcement Learning Handles Temporally-Coupled Perturbations

Yongyuan Liang [1] Yanchao Sun [1] Ruijie Zheng [1] Xiangyu Liu [1] Tuomas Sandholm [2 3] Furong Huang [1] Stephen McAleer [2]

Abstract

Robust reinforcement learning (RL) seeks to train policies that can perform well under environment perturbations or adversarial attacks. Existing approaches typically assume that the space of possible perturbations remains the same across timesteps. However, in many settings, the space of possible perturbations at a given timestep depends on past perturbations. We formally introduce temporally-coupled perturbations, presenting a novel challenge for existing robust RL methods. To tackle this challenge, we propose GRAD, a novel game-theoretic approach that treats the temporally-coupled robust RL problem as a partially-observable two-player zero-sum game. By finding an approximate equilibrium in this game, GRAD ensures the agent's robustness against temporally-coupled perturbations. Empirical experiments on a variety of continuous control tasks demonstrate that our proposed approach exhibits significant robustness advantages compared to baselines against both standard and temporally-coupled attacks, in both state and action spaces.

1. Introduction

In recent years, reinforcement learning (RL) has demonstrated remarkable success in tackling complex decision-making problems in various domains. However, the vulnerability of deep RL algorithms to test-time changes in the environment or adversarial attacks has raised significant concerns for real-world applications. Developing robust RL algorithms that can defend against these adversarial attacks is crucial for the safety, reliability and effectiveness of RL-based systems.

In most existing research on robust RL (Huang et al., 2017; Liang et al., 2022; Sun et al., 2022; Tessler et al., 2019; Zhang et al., 2020), the adversary is able to perturb the observation or action every timestep under a static constraint. Specifically, the adversary's perturbations are constrained within a predefined space, such as an L_p norm, which remains unchanged from one timestep to the next. This *standard* assumption in the robust RL literature can be referred to as a *non-temporally-coupled* assumption. This static constraint, however, can result in much different way of perturbation at every consecutive time steps. For example, the attacker may be able to blow the wind hard southeast at time t but northwest at time $t+1$ within this L_p norm under this static constraint. In contrast, in the realm of real-world settings, the adversary may not have complete flexibility to perturb the environment differently across timesteps. For example, it is unlikely for the wind to move in one direction in one second, then in the opposite direction in the next second. In these *temporally-coupled* settings, employing a robust policy learning technique designed for the static attack strategy would result in an excessively conservative policy. However, by formulating the robust RL problem as a partially-observable two-player game, we introduce a game-theoretic algorithm which lets the agent automatically adapt to the adversary under any attack constraints, either standard or temporally-coupled.

In this paper, we propose a novel approach: Game-theoretic Response approach for Adversarial Defense (GRAD) that leverages Policy Space Response Oracles (PSRO) (Lanctot et al., 2017) for robust training in the *temporally-coupled* setting. Our method aims to enhance the agent's resilience against the most powerful adversary in both state and action spaces. We model the interaction between the agent and the temporally-coupled adversary as a two-player zero-sum game and employ PSRO to ensure the agent's best response against the learned adversary and find an approximate equilibrium. This game-theoretic framework empowers our approach to effectively maximize the agent's worst-case rewards by adapting to the strongest adversarial strategies.

Our contributions are three-fold: First, we propose a novel

[*] Equal contribution [1] University of Maryland [2] Carnegie Mellon University [3] Strategy Robot, Inc., Optimized Markets, Inc., Strategic Machine, Inc.. Correspondence to: Stephen McAleer <smcaleer@cs.cmu.edu>.

2^{nd} AdvML Frontiers workshop at 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

class of temporally-coupled adversarial attacks to identify the realistic pitfalls of prior threat models and propose a challenge for existing robust RL methods which overlook the strength of temporally-coupled adversaries. Secondly, we introduce a game-theoretic response approach, referred to as GRAD, for robust training with a temporally-coupled adversary. We elaborate the theoretical advantages of our approach compared to existing robust RL methods. Lastly, we provide extensive empirical results that demonstrate the effectiveness of our approach in defending against both temporally-coupled attacks and standard (non-temporally coupled) attacks. Our evaluations span across various continuous control tasks, considering perturbations in both state and action spaces. Figure 1 shows interpretable phenomenons of GRAD agent and robust baselines under different types of attacks in Humanoid.

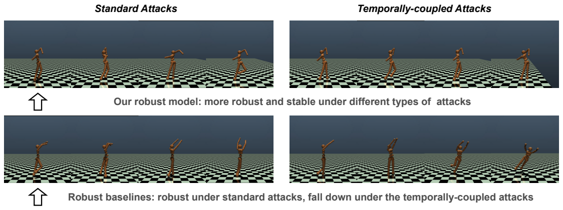

Standard Attacks
Temporally-coupled Attacks
Our robust model: more robust and stable under different types of attacks
Robust baselines: robust under standard attacks, fall down under the temporally-coupled attacks

Figure 1. The robust GRAD agents (top) and the state-of-the-art robust WocaR-RL (Liang et al., 2022)(bottom) show different learned behaviors. Under standard non-temporally-coupled attacks, both agents maintain basic body stability, with the GRAD agent attempting to avoid lateral rotations. However, under temporally-coupled attacks, the baseline agent is prone to falling towards one side, while GRAD maintains a higher level of robustness.

2. Related Work

Robust RL against adversaries perturbations *Regularization-based methods* (Zhang et al., 2020; Shen et al., 2020; Oikarinen et al., 2021) enforce the policy to have similar outputs under similar inputs, which achieves certifiable performance for DQN in some Atari games. But in continuous control tasks, these methods may not reliably improve the worst-case performance. A recent work by Korkmaz (Korkmaz, 2021) points out that these adversarially trained models may still be sensible to new perturbations. *Attack-driven methods* train DRL agents with adversarial examples. Some early works (Kos & Song, 2017; Behzadan & Munir, 2017; Mandlekar et al., 2017; Pattanaik et al., 2018; Franzmeyer et al., 2022) apply weak or strong gradient-based attacks on state observations to train RL agents against adversarial perturbations. Zhang et al. (Zhang et al., 2021) and Sun et al. (Sun et al., 2022) propose to alternately train an RL agent and a strong RL adversary, namely ATLA, which significantly improves the policy robustness against rectangle state perturbations. A recent work by Liang et al. (Liang et al., 2022) introduce a more principled adversarial training framework which does not explicitly learn the adversary, and both the efficiency and robustness of RL agents are boosted. There is also a line of work studying *theoretical guarantees* of adversarial defenses in RL (Lütjens et al., 2020; Oikarinen et al., 2021; Fischer et al., 2019; Kumar et al., 2022; Wu et al., 2022; Sun et al., 2023) in various settings.

Robust RL against action perturbations. Besides observation perturbations, attacks can happen in many other scenarios. For example, the agent's executed actions can be perturbed (Pan et al., 2022; Tan et al., 2020; Tessler et al., 2019; Lee et al., 2021; Lanier et al., 2022). Moreover, in a multi-agent game, an agent's behavior can create adversarial perturbations to a victim agent (Gleave et al., 2020). Pinto et al. (Pinto et al., 2017) model the competition between the agent and the attacker as a zero-sum two-player game, and train the agent under a learned attacker to tolerate both environment shifts and adversarial disturbances.

Two-player zero-sum games. There are a number of related deep reinforcement learning methods for two-player zero-sum games. CFR-based techniques such as Deep CFR (Brown et al., 2019a), DREAM (Steinberger et al., 2020), and ESCHER (McAleer et al., 2023), use deep reinforcement learning to approximate CFR. Policy-gradient techniques such as RPG (Srinivasan et al., 2018), NeuRD (Hennes et al., 2020), Friction-FoReL (Perolat et al., 2021; 2022), and MMD (Sokota et al., 2022), approximate Nash equilibrium via modified actor-critic algorithms. Our robust RL approach takes the double oracle techniques such as PSRO (Lanctot et al., 2017) as the backbone. PSRO-based algorithms have been shown to outperform the previously-mentioned algorithms in certain games (McAleer et al., 2021). More related work on robust MDP, safe RL and game-theoretic RL is discussed in Appendix B.

3. Preliminaries

Notations and Background. A Markov decision process (MDP) can be defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ represent the state space and the action space, $\mathcal{R}$ is the reward function: $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ represents the set of probability distributions over the state space $\mathcal{S}$ and $\gamma \in (0, 1)$ is the discount factor. The agent selects actions based on its policy, $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, which is represented by a function approximator (e.g. a neural network) that is updated during training and fixed during testing. The value function is denoted by $V^\pi(s) := \mathbb{E}_{P,\pi}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \mid s_0 = s]$, which measures the expected cumulative discounted reward that an agent can obtain from state $s \in \mathcal{S}$ by following policy π.

State Adversaries. State adversary is a type of test-time attacker that perturbs the agent's state observation returned by the environment at each time step and aims to reduce the expected episode reward gained by the agent. While the input to the agent's policy is perturbed, the underlying state in the environment remains unchanged. State adversaries, such as those presented in (Zhang et al., 2020; 2021; Sun et al., 2022), typically consider perturbations on a continuous state space under a certain attack budget ϵ. The attacker perturbs a state s into $\tilde{s} \in \mathcal{B}_\epsilon(s)$, where $\mathcal{B}_\epsilon(s)$ is a ℓ_p norm ball centered at s with radius ϵ.

Action Adversaries. Action adversaries' goal is to manipulate the behavior of the agent by directly perturbing the action a executed by the agent to $\tilde{a}$ before the environment receives it (altering the output of the agent's policy), causing it to deviate from the optimal policy. In addition to directly perturbing actions, recent work (Tessler et al., 2019) has also considered the setting where the action adversary selects a different, adversarial action with the probability α as an uncertainty constraint. In this paper, we focus solely on continuous-space perturbations and employ an admissible action perturbation budget as a commonly used ℓ_p threat model, similar to the state perturbation.

Zero-sum Game. We model the game between the agent and the adversary as a two-player zero-sum game that is a tuple $\langle \mathcal{S}, \Pi_a, \Pi_v, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where Π_a and Π_v denote the sets of policies for the agent and the adversary, respectively. In this framework, both the transition kernels $\mathcal{P}$ and the reward function $\mathcal{R}$ of the victim agent depend on not only its own policy $\pi_a \in \Pi_a$, but also the adversary's policy $\pi_v \in \Pi_v$. The adversary's reward $R(s_t, \bar{a}_t)$ is defined as the negative of the victim agent's reward $R(s_t, a_t)$, reflecting the zero-sum nature of the game.

Double Oracle Algorithm (DO) and Policy Space Response Oracles (PSRO). Double oracle (McMahan et al., 2003) is an algorithm for finding a NE in normal-form games. The algorithm operates by keeping a population of strategies Π^t at time t. Each iteration, a NE $\pi^{*,t}$ is computed for the game restricted to strategies in Π^t. Then, a best response $\mathbb{BR}_i(\pi^{*,t}_{-i})$ to this NE is computed for each player i and added to the population, $\Pi^{t+1}_i = \Pi^t_i \cup \{\mathbb{BR}_i(\pi^{*,t}_{-i})\}$ for $i \in \{1, 2\}$. Although in the worst case DO must expand all pure strategies before $\pi^{*,t}$ converges to a NE in the original game, in many games DO terminates early and outperforms alternative methods. An open problem is characterizing games where DO will outperform other methods.

Policy Space Response Oracles (PSRO) (Lanctot et al., 2017; Muller et al., 2019; Feng et al., 2021; McAleer et al., 2022b;a) are a method for approximately solving very large games. PSRO maintains a population of reinforcement learning policies and iteratively trains a best response to a mixture of the opponent's population. PSRO is a fundamentally different method than the previously described methods in that in certain games it can be much faster but in other games it can take exponentially long in the worst case. Neural Extensive Form Double Oracle (NXDO) (McAleer et al., 2021) combines PSRO with extensive-form game solvers and can be used to converge faster that PSRO. The full algorithms of DO and PSRO are in Appendix A.

4. Methodology

In this section, we formally define temporally-coupled attacks and introduce our game-theoretic response approach for adversarial defense against the proposed attacks.

4.1. Temporally-coupled Attack

In adversarial RL, it is common and reasonable to impose restrictions on the power of an adversary. To achieve this, we introduce the concept of *standard* admissible perturbations, as defined in Definition 4.1, which restricts the adversary to perturb a state s_t or an action a_t to a predefined set.

Definition 4.1 (ϵ-Admissible Adversary Perturbations)**.** An adversarial perturbation p_t is considered admissible in the context of a state adversary if, for a given state s_t at timestep t, the perturbed state $\tilde{s}_t$ defined as $\tilde{s}_t = s_t + p_t$ satisfies $\|s_t - \tilde{s}_t\| \leq \epsilon$, where ϵ is the state budget constraint. Similarly, if p_t is generated by an action adversary, the perturbed action $\tilde{a}_t$ defined as $\tilde{a}_t = a_t + p_t$ should be under the action constraint of $\|a_t - \tilde{a}_t\| \leq \epsilon$.

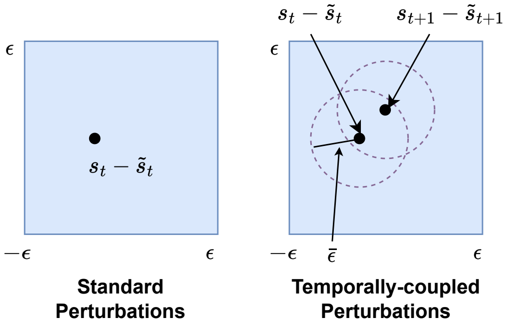


Figure 2. Standard perturbations and *Temporally-coupled* perturbations in a 2d example.

While the budget constraint ϵ is commonly applied in prior adversarial attacks, it may not be applicable in many real-world scenarios where the attacker needs to consider the past perturbations when determining the current perturbations. Specifically, in the temporal dimension, perturbations exhibit a certain degree of correlation. To capture this characteristic, we introduce the concept of temporally-coupled attackers. We propose a temporally-coupled constraint as defined in Definition 4.2, which sets specific limitations on the perturbation at the current timestep based on the previous timestep's perturbation.

Definition 4.2 ($\bar{\epsilon}$-Temporally-coupled Perturbations)**.** A temporally-coupled state perturbation p_t is deemed acceptable if it satisfies the temporally-coupled constraint $\bar{\epsilon}$: $\|s_t - \tilde{s}_t - (s_{t+1} - \tilde{s}_{t+1})\|_a \leq \bar{\epsilon}$ where $\tilde{s}_t$ and $\tilde{s}_{t+1}$ are the perturbed states obtained by adding p_t and p_{t+1} to s_t and s_{t+1}, respectively. For action adversaries, the temporally-coupled constraint $\bar{\epsilon}$ is similarly denoted as

$\|a_t - \tilde{a}_t - (a_{t+1} - \tilde{a}_{t+1})\| \leq \bar{\epsilon}$, where $\tilde{a}_t$ and $\tilde{a}_{t+1}$ are the perturbed actions.

When an adversary is subjected to both of these constraints, it is referred to as a temporally-coupled adversary in this paper. Each timestep's perturbation is restricted within a certain range ϵ, similar to other regular adversarial attacks. However, it is further confined within a smaller range $\bar{\epsilon}$ based on the previous timestep's perturbation. This temporally-coupled design offers two significant benefits.

Considering the temporal coupling between perturbations over time by constraining the next perturbations in a smaller range and discouraging drastic changes in attack direction, the adversary can launch continuous and stronger attacks while preserving a certain degree of stability. Intuitively, if the adversary consistently attacks in one direction, it can be more challenging for the victim to preserve balance and defend effectively compared to when the perturbations alternate between the left and right directions.

Then, the temporally-coupled constraint also enables the adversary to efficiently discover the optimal attack strategy by narrowing down the range of choices for each timestep's perturbation. Reducing the search space does not necessarily weaken the adversary; in fact, it can potentially make the adversary stronger if the optimal attack lies within the temporally-determined search space, which is supported by our empirical results. By constraining the adversary to a more focused exploration of attack strategies, the temporally-coupled constraint facilitates the discovery and exploitation of more effective and targeted adversarial tactics that exhibit less variation at consecutive timesteps. This characteristic enhances the adversary's ability to launch consistent and potent attacks.

Practically, it is crucial to carefully determine $\bar{\epsilon}$ to guarantee that this additional temporally-coupled constraint does not impede the performance of attacks but rather amplifies their effectiveness. The effectiveness of different choices for $\bar{\epsilon}$ was empirically evaluated in our empirical studies, highlighting the benefits it brings to adversarial learning. By leveraging such a temporally-coupled adversary, we propose a novel approach for robust training that enhances the agent's robustness. The detailed advantages of this approach will be elaborated in the following section.

4.2. GRAD: Game-theoretic Response approach for Adversarial Defense

Existing works primarily focus on the non-temporally-coupled assumption and thus may not be suitable in many real-world scenarios, but by treating the game-theoretic framework with a temporally-coupled adversary, our robust RL approach offers a more generalized solution that covers both standard and temporally-coupled settings.

Algorithm 1 Game-theoretic Response approach for Adversarial Defense (GRAD)

```
Input: Initial policy sets for the agent and adversary $\Pi : \{\Pi_a, \Pi_v\}$
Compute expected utilities as empirical payoff matrix $U^\Pi$ for each joint $\pi : \{\pi_a, \pi_v\} \in \Pi$
Compute meta-Nash equilibrium $\sigma_a$ and $\sigma_v$ over policy sets $(\Pi_a, \Pi_v)$
for epoch in $\{1, 2, \ldots\}$ do
    for many iterations $N_{\pi_a}$ do
        Sample the adversary policy $\pi_v \sim \sigma_v$
        Train $\pi'_a$ with trajectories against the fixed adversary $\pi_v$: $\mathcal{D}_{\pi'_a} := \{(\hat{s}^k_t, a^k_t, r^k_t, \hat{s}^k_{t+1})\}\big|^B_{k=1}$
        (when the fixed adversary only attacks the action space, $\hat{s}_t = s_t$.)
    end for
    $\Pi_a = \Pi_a \cup \{\pi'_a\}$
    for many iterations $N_{\pi_v}$ do
        Sample the agent policy $\pi_a \sim \sigma_a$
        Train the adversary policy $\pi'_v$ with trajectories: $\mathcal{D}_{\pi'_v} := \{(s^k_t, \bar{a}^k_t, -r^k_t, s^k_{t+1})\}\big|^B_{k=1}$
        ($\pi'_v$ applies attacks to the fixed victim agent $\pi_a$ based on $\bar{a}_t$ using different methods)
    end for
    $\Pi_v = \Pi_v \cup \{\pi'_v\}$
    Compute missing entries in $U^\Pi$ from $\Pi$
    Compute new meta strategies $\sigma_a$ and $\sigma_v$ from $U^\Pi$
end for
Return: current meta Nash equilibrium on whole population $\sigma_a$ and $\sigma_v$
```

In our Game-theoretic Response approach for Adversarial Defense (GRAD) framework as a modification of PSRO (Lanctot et al., 2017), an agent and a temporally-coupled adversary are trained as part of a two-player game. They play against each other and update their policies in response to each other's policies. The adversary is modeled as a separate agent who attempts to maximize the impact of attacks on the original agent's performance and whose action space is constrained by both ϵ and $\bar{\epsilon}$. Note that existing robust RL approaches such as (Liang et al., 2022) heavily rely on the ϵ-budget assumption, while the temporally-coupled constraints or other types of attack constraints are not considered or addressed. In contrast, GRAD naturally considers both the traditional ϵ-budget constraint and the new temporally-coupled constraint when calculating the best response. Meanwhile, the original agent's objective function is based on the reward obtained from the environment, taking into account the perturbations imposed by the adversary. The process continues until an approximate equilibrium is reached, at which point the original agent is considered to be robust to the attacks learned by the adversary. We show our full algorithm in Algorithm 1.

For different types of attackers, the agent generates different trajectories while training against a fixed attacker. If the attacker only targets the state, then the agent's training data will consist of the altered state $\hat{s}$ after adding the perturbations from the fixed attacker. If the attacker targets the agent's action, the agent's policy output a will be altered as $\hat{a}$ by the attacker, even if the agent receives the correct state s during training. However, this action alteration may not be detectable in the trajectories collected by the agent. As for the adversary's training, after defining the adversary's attack method and policy model, the adversary applies attacks to the fixed agent and collects the original state, along with the negative of the agent's reward $-r$, to train the adversary. Furthermore, the differences between GRAD and ATLA (Zhang et al., 2021) are explained in Appendix C.

5. Experiments

Evaluation Metrics By employing diverse sets of attackers: those specialized in perturbing the state space, those focusing on the action space and those capable of adaptably targeting both spaces, we conduct a comprehensive evaluation of the state-of-the-art robustness of our proposed method, GRAD, in comparison to existing robust baselines. This evaluation sheds light on the effectiveness of GRAD across a wide range of attack scenarios and highlights its robustness against different types of adversaries. In terms of evaluation metrics, we report the average test episodic rewards both under no attack and against the strongest traditional or temporally-coupled adversarial attacks to reflect both the natural performance and robustness of trained agents.

We calculated the average normalized rewards for each evaluation metric and each robust agent in all the environments as in Figure 3a, 3b and 3c. Table 1 presents the detailed comparison of robust moedels under diverse types of best temporally-coupled attacks, while more details of our experiments and full results can be found in Appendix D.

Case I: Against attacks on state space

For state adversaries, among the ATLA (Zhang et al., 2021; Sun et al., 2022) methods, PA-ATLA-PPO is the most robust, which trains with the standard strongest PA-AD attacker. As a modification, we train PA-ATLA-PPO* with a temporally-coupled PA-AD attacker, which is the type of adversary trained with GRAD agent. For a more intuitive and fair comparison, we only present the rewards of the best-performing ATLA agents under the type of attacks they were trained with.

In the absence of any attacks, GRAD maintains a competitive natural reward, which indicates that the agent's performance does not degrade significantly in the environment where is no adversary after approaching an approximate Nash equilibrium with the adversary. Even without training with regular attackers, our method demonstrates significantly better robustness under the non-temporally-coupled type of attack, particularly in the highest-dimensional and challenging environment, Humanoid, where it outperforms other methods by a large margin. Under our proposed temporally-coupled attacks, the average performance of our approach surpasses the state-of-the-art by up to 45%, highlighting the strong robustness of the policies learned by GRAD against all types of state adversarial attacks.

Case II: Against attacks on action space

In addition to state attacks, we assess the robustness of our methods against action adversaries that perturb the actions taken by the agent. We are the first to train an RL-based action adversary using the trajectory outlined in Algorithm 1, which leads to a more significant drop in rewards compared to action noise and showcases the worst-case performance of our robust agents under action perturbations.

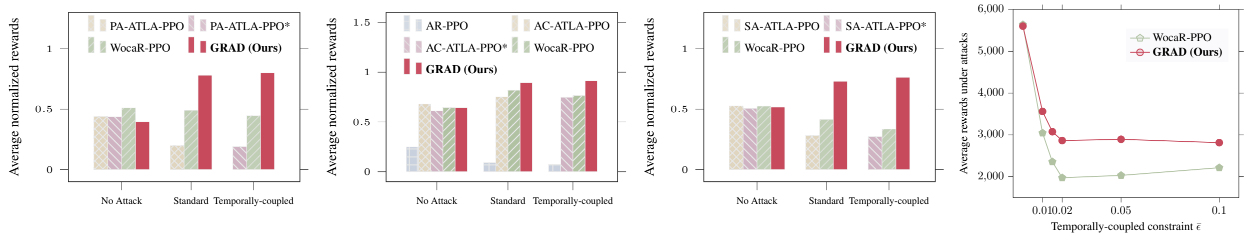
Among our baselines, we include AR-PPO, although it is not robust against strong action adversaries and performs well only under random noise. Another modification we made is AC-ATLA-PPO, where we train the agent alternately with the aforementioned action adversary. Similar to PA-ATLA-PPO*, we also train AC-ATLA-PPO* agents with a temporally-coupled action adversary, which is also utilized to train our GRAD agents.

In general, while action perturbation may not cause as strong of a "damage" as state perturbation, our GRAD method still achieves superior robustness. In terms of natural reward, GRAD performs comparably with other baselines. While the advantage of GRAD may not be apparent or significant under standard action attacks in less challenging environments, it surpasses other methods by more than 10% on Ant and Humanoid. Under temporally-coupled action attacks, GRAD consistently outperforms the most robust baseline by an average of over 20%, particularly exhibiting exceptional robustness on Humanoid. These results demonstrate the effective defense of GRAD against different types of adversarial attacks in the action space.

Case III: Against attacks on either state or action spaces

In prior works, adversarial attacks typically focus on perturbing either the agent's observations or introducing noise to the action space. However, in real-world scenarios, agents may encounter both types of attacks. To address this challenge, we propose an adversary called the State or Action Adversary (SA-AD), which allows the adversary to choose between attacking the agent's state or action at each time step, integrating this choice into the adversary's action space. Similar to the previous experiments, We train SA-ATLA-PPO with SA regular attacker, while SA-ATLA-PPO* and GRAD are trained with temporally-coupled SA attackers.

Our experimental results demonstrate that GRAD obtains



(a) State Attacks

(b) Action Attacks

(c) Adaptable SA Attacks

(d) Ablated studies for $\bar{\epsilon}$

Figure 3. Histograms 3a, 3b and 3c: normalized average rewards for GRAD and baselines across five environments, which mitigate the impact of varying reward ranges across environments. Each bar represents the distribution of rewards obtained by robust agents under state, action, or adaptable attacks. Figure 3d presents ablation results against temporally-coupled adversaries with different $\bar{\epsilon}$.

Best Attack	Model	Hopper	Walker2d	Halfcheetah	Ant	Humanoid
Temporally-Coupled State Attack	PA-ATLA-PPO*	2286 ± 29	2095 ± 34	3572 ± 31	2594 ± 19	1462 ± 27
	WocaR-PPO	2298 ± 13	2562 ± 18	4136 ± 28	3046 ± 10	1974 ± 33
	GRAD (Ours)	**2867 ± 36**	**2945 ± 27**	**4526 ± 29**	**3519 ± 24**	**2864 ± 28**
Temporally-Coupled Action Attack	AC-ATLA-PPO*	2625 ± 15	2780 ± 26	3815 ± 36	3382 ± 27	3092 ± 17
	AR-PPO	974 ± 15	1120 ± 28	1439 ± 21	679 ± 18	585 ± 29
	WocaR-PPO	2673 ± 28	2860 ± 32	4018 ± 37	3260 ± 27	3132 ± 35
	GRAD (Ours)	**3125 ± 26**	**3179 ± 20**	**4320 ± 27**	**3619 ± 34**	**4156 ± 29**
Temporally-Coupled SA Attack	SA-ATLA-PPO*	1994 ± 20	2492 ± 28	3694 ± 23	3145 ± 26	1972 ± 19
	WocaR-PPO	2297 ± 25	2497 ± 22	3935 ± 29	2887 ± 32	2180 ± 23
	GRAD (Ours)	**3051 ± 33**	**2932 ± 24**	**4096 ± 28**	**3336 ± 22**	**3295 ± 34**

Table 1. Average episode rewards ± standard error over 100 episodes for robust baselines and our **GRAD**. **Bold** numbers indicate the best results under different types of temporally-coupled attacks. The gray rows are the most robust agents.

similar natural rewards compared to the ATLA baselines, which is consistent with the findings from previous experiments. To summarize the results under SA attacks, our findings indicate that the combination of two different forms of attacks can effectively target robust agents in most scenarios, providing strong evidence of their robustness. In the case of regular SA attackers, GRAD outperforms other methods in all five environments, with a margin of over 20% in the Humanoid environment. Moreover, when defending against temporally-coupled attacks, GRAD significantly enhances robustness by more than 30% in multiple environments, with a minimum improvement of 10%. These results clearly demonstrate the robustness of GRAD against attackers that can target different domains.

Ablation studies for temporally-coupled constraint $\bar{\epsilon}$. As defined in our framework, the temporally-coupled constraint $\bar{\epsilon}$ limits the perturbations within a range that varies between timesteps. When $\bar{\epsilon}$ is set too large, the constraint becomes ineffective, resembling a standard attacker.

Conversely, setting $\bar{\epsilon}$ close to zero overly restricts perturbations, leading to a decline in attack performance. An appropriate value for $\bar{\epsilon}$ is critical for effective temporally-coupled attacks. Figure 3d illustrates the performance of robust models against temporally-coupled state attackers trained with different maximum $\bar{\epsilon}$. For WocaR-PPO, the temporally-coupled attacker achieves optimal attack performance when the values of $\bar{\epsilon}$ are set to 0.02. As the $\bar{\epsilon}$ values increase and the temporally-coupled constraint weakens, the agent's performance improves, indicating a decrease in the adversary's attack effectiveness. In the case of GRAD agents, they consistently maintain robust performance as the $\bar{\epsilon}$ values become larger. This observation highlights the impact of temporal coupling on the vulnerability of robust baselines to such attacks. In contrast, GRAD agents consistently demonstrate robustness against these attacks.

Conclusions In this paper, we introduce a novel attack model to challenge deep RL models, based on a temporally-coupled constraint that can naturally arise in real life. Since existing robust RL methods usually focus on a traditional threat model that perturbs state observations or actions arbitrarily within an L_p norm ball, they become too conservative and can fail to perform a good defense under the temporally-coupled attacks. In contrast, we propose a game-theoretical response approach GRAD, which finds the best response against attacks with various constraints including temporally-coupled ones. Extensive experiments in continuous control tasks show that GRAD significantly outperforms prior robust RL methods against various adversaries which emphasizes the empirical potential and contributions of our method in improving RL robustness.

References

Bechtle, S., Lin, Y., Rai, A., Righetti, L., and Meier, F. Curious ilqr: Resolving uncertainty in model-based rl. In Kaelbling, L. P., Kragic, D., and Sugiura, K. (eds.), *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 162–171. PMLR, 30 Oct–01 Nov 2020. URL https://proceedings.mlr.press/v100/bechtle20a.html.

Behzadan, V. and Munir, A. Whatever does not kill deep reinforcement learning, makes it stronger. *CoRR*, abs/1712.09344, 2017.

Brown, N. and Sandholm, T. Libratus: The superhuman AI for no-limit poker. In *IJCAI*, pp. 5226–5228, 2017a.

Brown, N. and Sandholm, T. Safe and nested subgame solving for imperfect-information games. *Advances in neural information processing systems*, 30, 2017b.

Brown, N. and Sandholm, T. Superhuman AI for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Brown, N., Sandholm, T., and Amos, B. Depth-limited solving for imperfect-information games. *Advances in neural information processing systems*, 31, 2018.

Brown, N., Lerer, A., Gross, S., and Sandholm, T. Deep counterfactual regret minimization. In *International conference on machine learning*, pp. 793–802. PMLR, 2019a.

Brown, N., Lerer, A., Gross, S., and Sandholm, T. Deep counterfactual regret minimization. In *International Conference on Machine Learning*, pp. 793–802, 2019b.

Brown, N., Bakhtin, A., Lerer, A., and Gong, Q. Combining deep reinforcement learning and search for imperfect-information games. *Advances in Neural Information Processing Systems*, 33:17057–17069, 2020.

Burch, N., Johanson, M., and Bowling, M. Solving imperfect information games using decomposition. In *Twenty-eighth AAAI conference on artificial intelligence*, 2014.

Feng, X., Slumbers, O., Yang, Y., Wan, Z., Liu, B., McAleer, S., Wen, Y., and Wang, J. Discovering multi-agent auto-curricula in two-player zero-sum games. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Fischer, M., Mirman, M., Stalder, S., and Vechev, M. T. Online robustness training for deep reinforcement learning. *CoRR*, abs/1911.00887, 2019.

Franzmeyer, T., McAleer, S., Henriques, J. F., Foerster, J. N., Torr, P. H., Bibi, A., and de Witt, C. S. Illusory attacks: Detectability matters in adversarial attacks on sequential decision-makers. *arXiv preprint arXiv:2207.10170v2*, 2022.

Fu, H., Liu, W., Wu, S., Wang, Y., Yang, T., Li, K., Xing, J., Li, B., Ma, B., Fu, Q., and Wei, Y. Actor-critic policy optimization in a large-scale imperfect-information game. In *Proceedings of the Tenth International Conference on Learning Representations (ICLR)*, 2022.

Garcıa, J. and Fernández, F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

Gaskett, C. Reinforcement learning under circumstances beyond its control. In *International Conference on Computational Intelligence for Modelling Control and Automation*, 2003.

Gleave, A., Dennis, M., Wild, C., Kant, N., Levine, S., and Russell, S. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations*, 2020.

Goyal, V. and Grand-Clement, J. Robust markov decision processes: Beyond rectangularity. *Mathematics of Operations Research*, 48(1):203–226, 2023.

Gray, J., Lerer, A., Bakhtin, A., and Brown, N. Human-level performance in no-press diplomacy via equilibrium search. In *International Conference on Learning Representations*, 2020.

Gruslys, A., Lanctot, M., Munos, R., Timbers, F., Schmid, M., Perolat, J., Morrill, D., Zambaldi, V., Lespiau, J.-B., Schultz, J., et al. The advantage regret-matching actor-critic. *arXiv preprint arXiv:2008.12234*, 2020.

Heger, M. Consideration of risk in reinforcement learning. In *International Conference on Machine Learning*, 1994.

Heinrich, J. and Silver, D. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

Hennes, D., Morrill, D., Omidshafiei, S., Munos, R., Perolat, J., Lanctot, M., Gruslys, A., Lespiau, J.-B., Parmas, P., Duéñez-Guzmán, E., et al. Neural replicator dynamics: Multiagent learning via hedging policy gradients. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 492–501, 2020.

Huang, S. H., Papernot, N., Goodfellow, I. J., Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. In *International Conference on Learning Representations(Workshop)*, 2017.

Korkmaz, E. Investigating vulnerabilities of deep neural policies. In *Uncertainty in Artificial Intelligence*, pp. 1661–1670. PMLR, 2021.

Kos, J. and Song, D. Delving into adversarial attacks on deep policies. In *International Conference on Learning Representations(Workshop)*, 2017.

Kumar, A., Levine, A., and Feizi, S. Policy smoothing for provably robust reinforcement learning. In *International Conference on Learning Representations*, 2022.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D., and Graepel, T. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Lanier, J., McAleer, S., Baldi, P., and Fox, R. Feasible adversarial robust reinforcement learning for underspecified environments. *arXiv preprint arXiv:2207.09597*, 2022.

Lee, X. Y., Esfandiari, Y., Tan, K. L., and Sarkar, S. Query-based targeted action-space adversarial policies on deep reinforcement learning agents. In *Proceedings of the ACM/IEEE 12th International Conference on Cyber-Physical Systems*, ICCPS '21, pp. 87–97, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383530.

Li, J., Koyamada, S., Ye, Q., Liu, G., Wang, C., Yang, R., Zhao, L., Qin, T., Liu, T.-Y., and Hon, H.-W. Suphx: Mastering mahjong with deep reinforcement learning. *arXiv preprint arXiv:2003.13590*, 2020.

Liang, Y., Sun, Y., Zheng, R., and Huang, F. Efficient adversarial training without attacking: Worst-case-aware robust reinforcement learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=y-E1htoQl-n.

Lim, S. H., Xu, H., and Mannor, S. Reinforcement learning in robust markov decision processes. *Advances in Neural Information Processing Systems*, 26:701–709, 2013.

Liu, W., Li, B., and Togelius, J. Model-free neural counterfactual regret minimization with bootstrap learning. *IEEE Transactions on Games*, 2022.

Lütjens, B., Everett, M., and How, J. P. Certified adversarial robustness for deep reinforcement learning. In *Conference on Robot Learning*, pp. 1328–1337. PMLR, 2020.

Mandlekar, A., Zhu, Y., Garg, A., Fei-Fei, L., and Savarese, S. Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3932–3939. IEEE, 2017.

Mankowitz, D. J., Levine, N., Jeong, R., Abdolmaleki, A., Springenberg, J. T., Shi, Y., Kay, J., Hester, T., Mann, T., and Riedmiller, M. Robust reinforcement learning for continuous control with model misspecification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgC60EtwB.

Mannor, S., Mebel, O., and Xu, H. Lightning does not strike twice: Robust mdps with coupled uncertainty. *arXiv preprint arXiv:1206.4643*, 2012.

Mannor, S., Mebel, O., and Xu, H. Robust mdps with k-rectangular uncertainty. *Mathematics of Operations Research*, 41(4):1484–1509, 2016.

McAleer, S., Lanier, J., Fox, R., and Baldi, P. Pipeline PSRO: A scalable approach for finding approximate Nash equilibria in large games. In *Advances in Neural Information Processing Systems*, 2020.

McAleer, S., Lanier, J., Baldi, P., and Fox, R. XDO: A double oracle algorithm for extensive-form games. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

McAleer, S., Lanier, J., Wang, K., Baldi, P., Fox, R., and Sandholm, T. Self-play psro: Toward optimal populations in two-player zero-sum games. *arXiv preprint arXiv:2207.06541*, 2022a.

McAleer, S., Wang, K., Lanctot, M., Lanier, J., Baldi, P., and Fox, R. Anytime optimal psro for two-player zero-sum games. *arXiv preprint arXiv:2201.07700*, 2022b.

McAleer, S., Farina, G., Lanctot, M., and Sandholm, T. Escher: Eschewing importance sampling in games by computing a history value function to estimate regret. *International Conference on Learning Representations*, 2023.

McMahan, H. B., Gordon, G. J., and Blum, A. Planning in the presence of cost functions controlled by an adversary. *Proceedings of the 20th International Conference on Machine Learning (ICML)*, 2003.

Moravcik, M., Schmid, M., Ha, K., Hladik, M., and Gaukrodger, S. Refining subgames in large imperfect information games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Moravčík, M., Schmid, M., Burch, N., Lisỳ, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., and Bowling, M. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

Muller, P., Omidshafiei, S., Rowland, M., Tuyls, K., Perolat, J., Liu, S., Hennes, D., Marris, L., Lanctot, M., Hughes, E., et al. A generalized training approach for multiagent learning. In *International Conference on Learning Representations*, 2019.

Oikarinen, T., Zhang, W., Megretski, A., Daniel, L., and Weng, T.-W. Robust deep reinforcement learning through adversarial loss. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=eaAM_bdW0Q.

Pan, X., Xiao, C., He, W., Yang, S., Peng, J., Sun, M., Liu, M., Li, B., and Song, D. Characterizing attacks on deep reinforcement learning. In *AAMAS*, pp. 1010–1018. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022.

Pattanaik, A., Tang, Z., Liu, S., Bommannan, G., and Chowdhary, G. Robust deep reinforcement learning with adversarial attacks. In *AAMAS*, pp. 2040–2042. International Foundation for Autonomous Agents and Multiagent Systems Richland, SC, USA / ACM, 2018.

Perolat, J., Munos, R., Lespiau, J.-B., Omidshafiei, S., Rowland, M., Ortega, P., Burch, N., Anthony, T., Balduzzi, D., De Vylder, B., et al. From Poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning*, pp. 8525–8535. PMLR, 2021.

Perolat, J., de Vylder, B., Hennes, D., Tarassov, E., Strub, F., de Boer, V., Muller, P., Connor, J. T., Burch, N., Anthony, T., et al. Mastering the game of stratego with model-free multiagent reinforcement learning. *arXiv preprint arXiv:2206.15378*, 2022.

Pinto, L., Davidson, J., Sukthankar, R., and Gupta, A. Robust adversarial reinforcement learning. In *International Conference on Machine Learning*, pp. 2817–2826. PMLR, 2017.

Schmid, M., Moravcik, M., Burch, N., Kadlec, R., Davidson, J., Waugh, K., Bard, N., Timbers, F., Lanctot, M., Holland, Z., et al. Player of games. *arXiv preprint arXiv:2112.03178*, 2021.

Serrino, J., Kleiman-Weiner, M., Parkes, D. C., and Tenenbaum, J. Finding friend and foe in multi-agent games. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Sokota, S., D'Orazio, R., Kolter, J. Z., Loizou, N., Lanctot, M., Mitliagkas, I., Brown, N., and Kroer, C. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. *arXiv preprint arXiv:2206.05825*, 2022.

Srinivasan, S., Lanctot, M., Zambaldi, V., Pérolat, J., Tuyls, K., Munos, R., and Bowling, M. Actor-critic policy optimization in partially observable multiagent environments. *Advances in neural information processing systems*, 31, 2018.

Steinberger, E. Single deep counterfactual regret minimization. *arXiv preprint arXiv:1901.07621*, 2019.

Steinberger, E., Lerer, A., and Brown, N. DREAM: Deep regret minimization with advantage baselines and model-free learning. *arXiv preprint arXiv:2006.10410*, 2020.

Sun, Y., Zheng, R., Liang, Y., and Huang, F. Who is the strongest enemy? towards optimal and efficient evasion attacks in deep RL. In *International Conference on Learning Representations*, 2022.

Sun, Y., Zheng, R., Hassanzadeh, P., Liang, Y., Feizi, S., Ganesh, S., and Huang, F. Certifiably robust policy learning against adversarial multi-agent communication. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dCOL0inGl3e.

Tan, K. L., Esfandiari, Y., Lee, X. Y., Sarkar, S., et al. Robustifying reinforcement learning agents via action space adversarial training. In *2020 American control conference (ACC)*, pp. 3959–3964. IEEE, 2020.

Tessler, C., Efroni, Y., and Mannor, S. Action robust reinforcement learning and applications in continuous control. In *International Conference on Machine Learning*, pp. 6215–6224. PMLR, 2019.

Thomas, G., Luo, Y., and Ma, T. Safe reinforcement learning by imagining the near future. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=vIDBSGl3vzl.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Wu, F., Li, L., Huang, Z., Vorobeychik, Y., Zhao, D., and Li, B. CROP: certifying robust policies for reinforcement learning through functional smoothing. In *International Conference on Learning Representations*, 2022.

Wurman, P. R., Barrett, S., Kawamoto, K., MacGlashan, J., Subramanian, K., Walsh, T. J., Capobianco, R., Devlic, A., Eckert, F., Fuchs, F., et al. Outracing champion gran turismo drivers with deep reinforcement learning. *Nature*, 602(7896):223–228, 2022.

Zha, D., Xie, J., Ma, W., Zhang, S., Lian, X., Hu, X., and Liu, J. Douzero: Mastering doudizhu with self-play deep reinforcement learning. In *International Conference on Machine Learning*, pp. 12333–12344. PMLR, 2021.

Zhang, H., Chen, H., Xiao, C., Li, B., Liu, M., Boning, D., and Hsieh, C.-J. Robust deep reinforcement learning against adversarial perturbations on state observations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 21024–21037. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/f0eb6568ea114ba6e293f903c34d7488-Paper.pdf.

Zhang, H., Chen, H., Boning, D. S., and Hsieh, C.-J. Robust reinforcement learning on state observations with learned optimal adversary. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=sCZbhBvqQaU.

A. Preliminary Algorithms

Here is the full game-theoretic algorithms of Double Oracle and Policy Space Response Oracles as introduced in 3.

Algorithm 2 Double Oracle (McMahan et al., 2003)

```
Result: Nash Equilibrium
Input: Initial population $\Pi^0$
repeat {for $t = 0, 1, \ldots$}
  $\pi^r \leftarrow$ NE in game restricted to strategies in $\Pi^t$
  for $i \in \{1, 2\}$ do
    Find a best response $\beta_i \leftarrow \mathbb{BR}_i(\pi^r_{-i})$
    $\Pi_i^{t+1} \leftarrow \Pi_i^t \cup \{\beta_i\}$
  end for
until No novel best response exists for either player
Return: $\pi^r$
```

Algorithm 3 Policy Space Response Oracles (Lanctot et al., 2017)

```
Result: Nash Equilibrium
Input: Initial population $\Pi^0$
repeat {for $t = 0, 1, \ldots$}
  $\pi^r \leftarrow$ NE in game restricted to strategies in $\Pi^t$
  for $i \in \{1, 2\}$ do
    Find a best response $\beta_i \leftarrow \mathbb{BR}_i(\pi^r_{-i})$
    $\Pi_i^{t+1} \leftarrow \Pi_i^t \cup \{\beta_i\}$
  end for
until Approximate exploitability is less than or equal to zero
Return: $\pi^r$
```

B. Additional Related Work

B.1. Robust Markov decision process and safe RL.

There are several lines of work that study RL under safety/risk constraints (Heger, 1994; Gaskett, 2003; Garcıa & Fernández, 2015; Bechtle et al., 2020; Thomas et al., 2021) or under intrinsic uncertainty of environment dynamics (Lim et al., 2013; Mankowitz et al., 2020). In particular, there are several works discussing coupled or non-rectangular uncertainty sets, which allow less conservative and more efficient robust policy learning by incorporating realistic conditions that naturally arise in practice. Mannor et al. (Mannor et al., 2012) propose to model coupled uncertain parameters based on the intuition that the total number of states with deviated parameters will be small. Mannor et al. (Mannor et al., 2016) identify "k-rectangular" uncertainty sets defined by the cardinality of possible conditional projections of uncertainty sets, which can lead to more tractable solutions. Another recent work by Goyal et al. (Goyal & Grand-Clement, 2023) propose to model the environment uncertainty with factor matrix uncertainty sets, which can efficiently compute an optimal robust policy.

B.2. Game-Theoretic Reinforcement Learning

Superhuman performance in two-player games usually involves two components: the first focuses on finding a model-free blueprint strategy, which is the setting we focus on in this paper. The second component improves this blueprint online via model-based subgame solving and search (Burch et al., 2014; Moravcik et al., 2016; Brown et al., 2018; 2020; Brown & Sandholm, 2017b; Schmid et al., 2021). This combination of blueprint strategies with subgame solving has led to state-of the art performance in Go (Silver et al., 2017), Poker (Brown & Sandholm, 2017a; 2018; Moravčík et al., 2017), Diplomacy (Gray et al., 2020), and The Resistance: Avalon (Serrino et al., 2019). Methods that only use a blueprint have achieved state-of-the-art performance on Starcraft (Vinyals et al., 2019), Gran Turismo (Wurman et al., 2022), DouDizhu (Zha et al., 2021), Mahjohng (Li et al., 2020), and Stratego (McAleer et al., 2020; Perolat et al., 2022). In the rest of this section we focus on other model-free methods for finding blueprints.

Deep CFR (Brown et al., 2019b; Steinberger, 2019) is a general method that trains a neural network on a buffer of counterfactual values. However, Deep CFR uses external sampling, which may be impractical for games with a large branching factor, such as Stratego and Barrage Stratego. DREAM (Steinberger et al., 2020) and ARMAC (Gruslys et al., 2020) are model-free regret-based deep learning approaches. ReCFR (Liu et al., 2022) propose a bootstrap method for estimating cumulative regrets with neural networks. ESCHER (McAleer et al., 2023) remove the importance sampling term of Deep CFR and show that doing so allows scaling to large games.

Neural Fictitious Self-Play (NFSP) (Heinrich & Silver, 2016) approximates fictitious play by progressively training a best response against an average of all past opponent policies using reinforcement learning. The average policy converges to an approximate Nash equilibrium in two-player zero-sum games.

There is an emerging literature connecting reinforcement learning to game theory. QPG (Srinivasan et al., 2018) shows that state-conditioned Q-values are related to counterfactual values by a reach weighted term summed over all histories in an infostate and proposes an actor-critic algorithm that empirically converges to a NE when the learning rate is annealed. NeuRD (Hennes et al., 2020), and F-FoReL (Perolat et al., 2021) approximate replicator dynamics and follow the regularized leader, respectively, with policy gradients. Actor Critic Hedge (ACH) (Fu et al., 2022) is similar to NeuRD but uses an information set based value function. All of these policy-gradient methods do not have theory proving that they converge with high probability in extensive form games when sampling trajectories from the policy. In practice, they often perform worse than NFSP and DREAM on small games but remain promising approaches for scaling to large games (Perolat et al., 2022).

C. Methodology

Difference between GRAD and ATLA While both GRAD and ATLA (Zhang et al., 2021) require training an adversary alongside the agent using RL, there is a key difference in their training approaches. In GRAD, both the agent and the adversary have two policy sets. During each training epoch, the agent aims to find an approximate best response to the fixed adversary, and vice versa for the adversary. This iterative process promotes the emergence of stable and robust policies. After each epoch, the trained policies are added to the respective policy sets. GRAD has the capability to continuously explore and learn new policies that are not present in the current policy set, thereby enabling ongoing improvement for both the agent and the adversary, which allows for a more thorough exploration of the policy space. In contrast, ATLA employs a limited number of iterations to train each agent in each round, which is not sufficient to allow the agent and adversary to find each other's best response within the policy space.

It is also worth noting that the original ATLA utilizes standard attack methods to train the adversary. However, several experimental observations indicate that agents trained with non-temporally-coupled adversaries tend to exhibit a conservative and overfitted behavior towards specific types of adversaries.

D. Experiment Details and Additional Results

D.1. Setup and Baselines

Our experiments are conducted on five various and challenging MuJoCo environments: Hopper, Walker2d, Halfcheetah, Ant, and Humanoid, all using the v2 version of MuJoCo. We use the Proximal Policy Optimization (PPO) algorithm as the policy optimizer and a Long Short-Term Memory (LSTM) network as the policy network for all of the robust training methods we evaluate. To maintain methodological consistency and minimize potential discrepancies arising from different PPO implementations across methods, we ensure highly similar benchmark results. For attack constraint ϵ, we use the commonly adopted values ϵ for each environment. For the temporally-coupled constraint $\bar{\epsilon}$, we set the optimal maximum $\bar{\epsilon}$ as $\epsilon/5$ (with minor adjustments in some environments). Other choices of $\bar{\epsilon}$ will be further discussed in the ablation studies.

We compare our approach GRAD with other robust RL baselines in this paper. Robust training frameworks can be categorized into two types. The first type requires training with a specified adversary during training, such as the alternating training framework (ATLA (Zhang et al., 2021)) and GRAD. The second type does not require training with an adversary, such as WocaR-PPO (Liang et al., 2022) and AR-PPO (PPO variant of AR-DDPG (Tessler et al., 2019)). The baselines we chose demonstrate state-of-the-art or great robustness in prior works. The first type of approaches require training agents with adversaries targeting specific attack domains and the second type of baselines can be evaluated directly for their robustness without the need for additional adversary training.

D.2. Implementation details

We provide detailed implementation information for our proposed method (GRAD) and baselines.

Reproducibility We train each agent configuration with 10 seeds and report the one with the median robust performance, rather than the best one. More implementation details are in Appendix D.2.

Training Steps For GRAD, we specify the number of training steps required for different environments. In the Hopper, Walker2d, and Halfcheetah environments, we train for 10 million steps. In the Ant and Humanoid environments, we extend the training duration to 20 million steps. For the ATLA baselines, we train for 2 million steps and 10 million steps in environments of varying difficulty.

Network Structure Our algorithm (GRAD) adopts the same PPO network structure as the baselines to maintain consistency. The network comprises a single-layer LSTM with 64 hidden neurons. Additionally, an input embedding layer is employed to project the state dimension to 64, and an output layer is used to project 64 to the output dimension. Both the agents and the adversaries use the same policy and value networks to facilitate training and evaluation. Furthermore, the network architecture for the best response and meta Nash remains consistent with the aforementioned configuration.

Schedule of ϵ and $\bar{\epsilon}$ During the training process, we gradually increase the values of ϵ and $\bar{\epsilon}$ from 0 to their respective target maximum values. This incremental adjustment occurs over the first half of the training steps. We reference the attack budget ϵ used in other baselines for the corresponding environments. This ensures consistency and allows for a fair comparison with existing methods. The target value of $\bar{\epsilon}$ is determined based on the adversary's training results, which is set as $\epsilon/5$. In some smaller dimensional environments, $\bar{\epsilon}$ can be set to $\epsilon/10$. We have observed that the final performance of the trained robust models does not differ by more than 5% when using these values for $\bar{\epsilon}$.

Training Time The training time for GRAD varies based on the specific environment and its associated difficulty. On a single V100 GPU, training GRAD typically requires over 20 hours for the Hopper, Walker2d, and Halfcheetah environments. For the more complex Ant and Humanoid environments, the training duration extends to approximately 40 hours. The training time required for defense against state adversaries or action adversaries is relatively similar.

Observation and Reward Normalization To ensure consistency with PPO implementation and maintain comparability across different codebases, we apply observation and reward normalization. Normalization helps to standardize the input observations and rewards, enhancing the stability and convergence of the training process. We have verified the performance of vanilla PPO on different implementations, and the results align closely with our implementation of GRAD based on Ray rllib.

Hyperparameter Selection Hyperparameters such as learning rate, entropy bonus coefficient, and other PPO-specific parameters are crucial for achieving optimal performance. Referring to the results obtained from vanilla PPO and the ATLA baselines as references, a small-scale grid search is conducted to fine-tune the hyperparameters specific to GRAD. Because of the significant training time and cost associated with GRAD, we initially perform a simplified parameter selection using the Inverted Pendulum as a test environment.

D.3. Full Results

Here are the comprehensive results for the robust models under different scenarios: no attack, the best standard attacks, and the best temporally-coupled attacks on the state space, action space, or both spaces.

D.4. Adversary Alogrithms

State Adversaries Aimed to introduce the attack methods utilized during training and testing in our experiments. When it comes to state adversaries, PA-AD as Alogrithm 4 stands out as the strongest attack compared to other state attacks. Therefore, we report the best state attack rewards under PA-AD attacks.

Action Adversaries In terms of action adversaries, an RL-based action adversary as Alogrithm 5 can inflict more severe damage on agents' rewards compared to OU noise and parameter noise in (Tessler et al., 2019).

Adaptable Adversaries For adaptable adversaries capable of perturbing both state and action spaces, considering the attack budget and cost, we prefer not to allow the adversary to perturb both spaces simultaneously at one timestep. Hence, it is necessary for the adversary to decide which space to perturb for each timestep. In alg:sa-ad, we introduce an additional

Environment		Hopper	Walker2d	Halfcheetah	Ant	Humanoid
ϵ / state dim		0.075 / 11	0.05 / 17	0.15 / 17	0.15 / 111	0.1 / 376
Natural Reward (no attack)	PA-ATLA-PPO	3425 ± 22	4153 ± 31	6175 ± 29	5340 ± 11	5792 ± 18
	PA-ATLA-PPO*	3476 ± 31	4032 ± 29	6291 ± 37	5431 ± 28	5659 ± 19
	WocaR-PPO	3588 ± 12	4102 ± 34	6032 ± 14	5576 ± 21	5781 ± 30
	GRAD (Ours)	3345 ± 15	4089 ± 24	6149 ± 27	5376 ± 21	5772 ± 23
Best Standard State Attack	PA-ATLA-PPO	2532 ± 31	2241 ± 17	3849 ± 28	2874 ± 25	1435 ± 22
	WocaR-PPO	2570 ± 23	2715 ± 16	4225 ± 20	3145 ± 19	2236 ± 26
	GRAD (Ours)	**2899 ± 27**	**3108 ± 25**	**4447 ± 28**	**3397 ± 18**	**2787 ± 25**
Best Temporally-coupled State Attack	PA-ATLA-PPO*	2286 ± 29	2095 ± 34	3572 ± 31	2594 ± 19	1462 ± 27
	WocaR-PPO	2298 ± 13	2562 ± 18	4136 ± 28	3046 ± 10	1974 ± 33
	GRAD (Ours)	**2867 ± 36**	**2945 ± 27**	**4526 ± 29**	**3519 ± 24**	**2864 ± 28**

Table 2. Average episode rewards ± standard error over 100 episodes for three state robust baselines and our **GRAD**. **Bold** numbers indicate the best results under different types of attacks on state spaces. The gray rows are the most robust agents.

Environment		Hopper	Walker2d	Halfcheetah	Ant	Humanoid
ϵ / action dim		0.2 / 3	0.2 / 6	0.2 / 6	0.15 / 8	0.15 / 17
Natural Reward (no attack)	AC-ATLA-PPO	3576 ± 43	4228 ± 26	5915 ± 28	5557 ± 33	6014 ± 49
	AC-ATLA-PPO*	3492 ± 28	4052 ± 29	5853 ± 32	5549 ± 21	5980 ± 30
	AR-PPO	3188 ± 28	3767 ± 15	5248 ± 35	5074 ± 34	5379 ± 42
	GRAD (Ours)	3482 ± 20	4159 ± 26	6047 ± 29	5512 ± 38	5894 ± 35
Best Standard Action Attack	AC-ATLA-PPO	2872 ± 18	3108 ± 22	3994 ± 41	2752 ± 33	3120 ± 39
	AR-PPO	1235 ± 26	1305 ± 31	1523 ± 24	1120 ± 10	1117 ± 37
	WocaR-PPO	2943 ± 18	3269 ± 28	3840 ± 29	3345 ± 21	3419 ± 37
	GRAD (Ours)	**3039 ± 20**	**3298 ± 35**	**4016 ± 28**	**3569 ± 26**	**4106 ± 32**
Best Temporally-Coupled Action Attack	AC-ATLA-PPO*	2625 ± 15	2780 ± 26	3815 ± 36	3382 ± 27	3092 ± 17
	AR-PPO	974 ± 15	1120 ± 28	1439 ± 21	679 ± 18	585 ± 29
	WocaR-PPO	2673 ± 28	2860 ± 32	4018 ± 37	3260 ± 27	3132 ± 35
	GRAD (Ours)	**3125 ± 26**	**3179 ± 20**	**4320 ± 27**	**3619 ± 34**	**4156 ± 29**

Table 3. Average episode rewards ± standard error for over 100 episodes action robust baselines and our **GRAD** under no attack and best action attacks.

Environment		Hopper	Walker2d	Halfcheetah	Ant	Humanoid
Natural Reward (no attack)	SA-ATLA-PPO	3498 ± 34	4195 ± 36	6078 ± 38	5442 ± 28	5913 ± 34
	SA-ATLA-PPO*	3372 ± 29	4219 ± 31	5979 ± 37	5510 ± 29	5949 ± 22
	GRAD (Ours)	3420 ± 18	4218 ± 27	6123 ± 18	5517 ± 21	5799 ± 27
Best Standard SA Attack	SA-ATLA-PPO	2294 ± 27	2503 ± 34	4028 ± 28	2750 ± 32	1981 ± 25
	WocaR-PPO	2469 ± 18	2594 ± 32	4012 ± 26	3024 ± 23	2263 ± 28
	GRAD (Ours)	**2846 ± 24**	**2897 ± 31**	**4168 ± 37**	**3286 ± 20**	**3042 ± 24**
Best Temporally-coupled SA Attack	SA-ATLA-PPO*	1994 ± 20	2492 ± 28	3694 ± 23	3145 ± 26	1972 ± 19
	WocaR-PPO	2297 ± 25	2497 ± 22	3935 ± 29	2887 ± 32	2180 ± 23
	GRAD (Ours)	**3051 ± 33**	**2932 ± 24**	**4096 ± 28**	**3336 ± 22**	**3295 ± 34**

Table 4. Average episode rewards ± standard error over 100 episodes for adaptable adversarial defense baselines and our **GRAD**.

dimension $\theta_t \in [-1, 1]$ in the adversary's action space to determine the attack domain. If $\theta_t 0$, the adversary perturbs the observation state s_t to $\tilde{s}_t$; otherwise, it attacks the agent's policy output action a_t to $\tilde{a}_t$. Building upon PA-AD (Sun et al., 2022), the adversary director only needs to learn $\hat{a}_t$, which is composed of the policy perturbation $\hat{d}_t$ concatenated with θ. Depending on the adversary's choice θ, different actors will craft state or action perturbations for a given policy perturbation direction $\hat{d}$. This means that the SA-AD attacker only requires an additional dimension for domain choice compared to the PA-AD attacker, without significantly increasing the complexity of adversary training, thereby minimizing the impact on adversary performance. We show the adaptable attack method in Algorithm 6.

Algorithm 4 Policy Adversarial Actor Director (PA-AD)

```
Input: Initialization of adversary director's policy $v$; victim policy $\pi$, the actor function $g$ for the state space $\mathcal{S}$, initial state $s_0$
for $t = 0, 1, 2, \dots$ do
    Director $v$ samples a policy perturbing direction and perturbed choice, $\widehat{a}_t \sim \nu(\cdot|s_t)$
    Actor perturbs $s_t$ to $\tilde{s}_t = g(\widehat{a}_t, s_t)$
    Victim takes action $a_t \sim \pi(\cdot|\tilde{s}_t)$, proceeds to $s_{t+1}$, receives $r_t$
    Director saves $(s_t, \widehat{a}_t, -r_t, s_{t+1})$ to the adversary buffer
    Director updates its policy $v$ using any RL algorithms
end for
```

D.5. Attack Budget ϵ

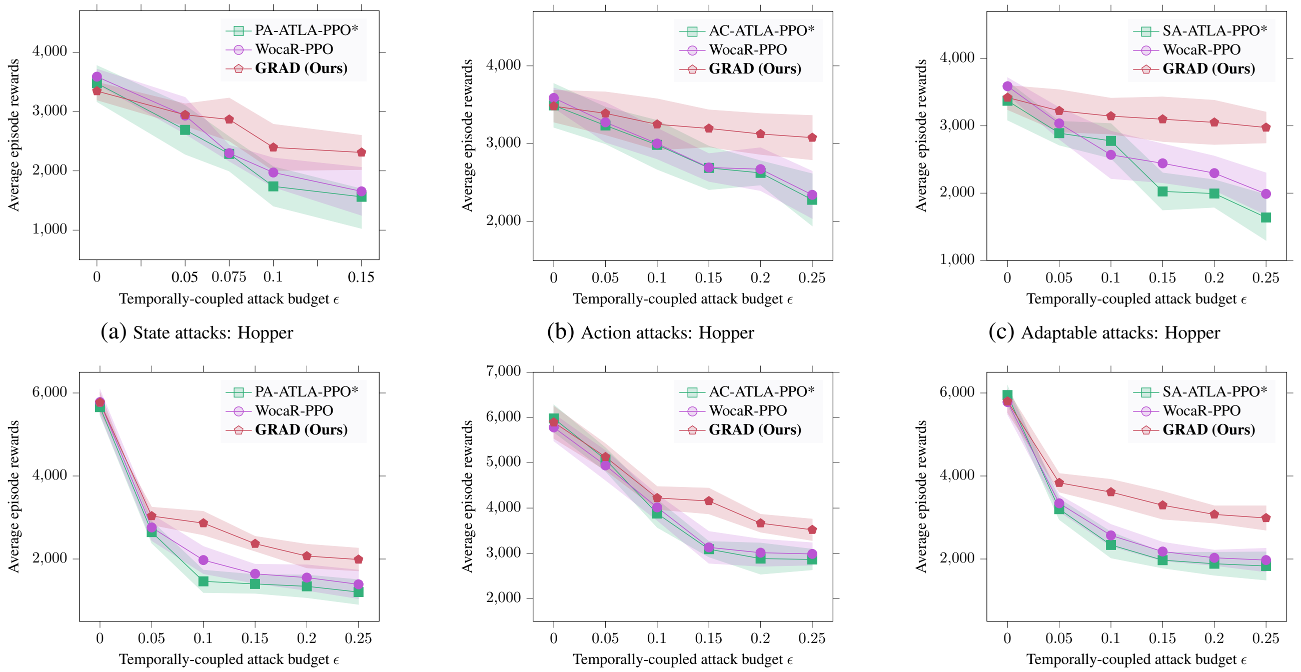
In Figure 4, we report the performance of baselines and GRAD under different attack budget ϵ. As the value of ϵ increases, the rewards of robust agents under different types of attacks decrease accordingly. However, our approach consistently demonstrates superior robustness as the attack budget changes.

Algorithm 5 Action Adversary (AC-AD)

```
Input: Initialization of action adversary policy $v$; victim policy $\pi$, initial state $s_0$
for $t = 0, 1, 2, \dots$ do
    adversary $v$ samples an action perturbations $\widehat{a}_t \sim \nu(\cdot|s_t)$,
    victim policy $\pi$ outputs action $a_t \sim \pi(\cdot|s_t)$
    the environment receives $\tilde{a}_t = a_t + \widehat{a}_t$, returns $s_{t+1}$ and $r_t$
    adversary saves $(s_t, \widehat{a}_t, -r_t, s_{t+1})$ to the adversary buffer
    adversary updates its policy $v$
end for
```

Algorithm 6 State or Action Adversary (SA-AD)

```
Input: Initialization of adversary director's policy $v$; victim policy $\pi$, the actor function $g_s$ for the state space $\mathcal{S}$ and $g_a$ for the action space $\mathcal{A}$, initial state $s_0$
for $t = 0, 1, 2, \dots$ do
    Director $v$ samples a policy perturbing direction and perturbed choice, $\widehat{a}_t \sim \nu(\cdot|s_t)$, where $\widehat{a}_t = (\widehat{d}_t, \theta_t), \theta_t \in [-1, 1]$
    if $\theta_t \geq 0$ then
        Actor perturbs $s_t$ to $\tilde{s}_t = g_s(\widehat{d}_t, s_t)$
        Victim takes action $a_t \sim \pi(\cdot|\tilde{s}_t)$, proceeds to $s_{t+1}$, receives $r_t$
    else
        victim policy outputs action $a_t \sim \pi(\cdot|s_t)$
        Actor perturbs $a_t$ to $\tilde{a}_t = g_a(\widehat{d}_t, a_t)$
        The environment receives $\tilde{a}_t$, returns $s_{t+1}$ and $r_t$
    end if
    Director saves $(s_t, \widehat{a}_t, -r_t, s_{t+1})$ to the adversary buffer
    Director updates its policy $v$ using any RL algorithms
end for
```



(a) State attacks: Hopper

(b) Action attacks: Hopper

(c) Adaptable attacks: Hopper

(d) State attacks: Humanoid

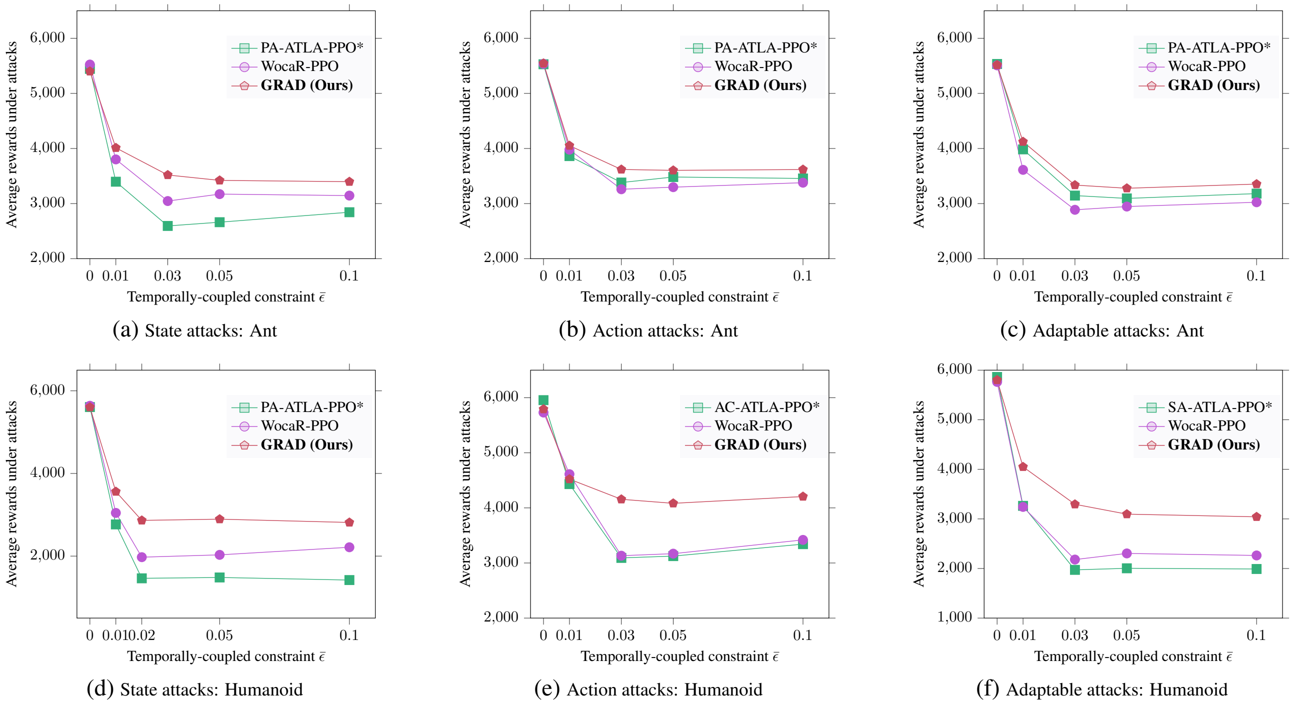
(e) Action attacks: Humanoid

(f) Adaptable attacks: Humanoid

Figure 4. Comparisons under state or action or adaptable temporally-coupled attacks w.r.t. diverse attack budgets ϵ's on Hopper and Humanoid.

D.6. Temporally-coupled ϵ

We also investigate the impact of temporally-coupled constraints $\bar{\epsilon}$ on attack performance, as we explained in our experiment section.



(a) State attacks: Ant

(b) Action attacks: Ant

(c) Adaptable attacks: Ant

(d) State attacks: Humanoid

(e) Action attacks: Humanoid

(f) Adaptable attacks: Humanoid

Figure 5. Comparisons under state or action or adaptable temporally-coupled attacks with diverse temporally-coupled constraints ϵ's on Ant and Humanoid.