TCRL: Temporal-Coupled Adversarial Training for Robust Constrained Reinforcement Learning in Worst-Case Scenarios

Extended Abstract

Wentao Xu
Northeastern University
Shenyang, China
20223441@stu.neu.edu.cn

Zhongming Yao*
Northeastern University
Shenyang, China
yaozming@stumail.neu.edu.cn

Weihao Li
Northeastern University
Shenyang, China
20223284@stu.neu.edu.cn

Zhenghang Song
Zhejiang University
Ningbo, China
22451060@zju.edu.cn

Yumeng Song
Aalborg University
Aalborg, Denmark
yumengs@cs.aau.dk

Tianyi Li
Aalborg University
Aalborg, Denmark
tianyi@cs.aau.dk

Yushuai Li
Aalborg University
Aalborg, Denmark
yusli@cs.aau.dk

ABSTRACT

Constrained Reinforcement Learning (CRL) aims to optimize decision-making policies under constraint conditions, making it highly applicable to safety-critical domains such as autonomous driving, robotics, and power grid management. However, existing robust CRL approaches predominantly focus on single-step perturbations and temporal-independent adversarial models, lacking explicit modeling of temporal-coupled perturbations robustness.To tackle these challenges, we propose TCRL, a novel **t**emporal-coupled adversarial training framework for robust **c**onstrained **r**einforcement **l**earning (TCRL) in worst-case scenarios. First, TCRL introduces a worst-case-perceived cost constraint function that estimates safety costs under temporal-coupled perturbations without the need to explicitly model adversarial attackers. Second, TCRL establishes a dual-constraint defense mechanism towards the reward to counter temporal-coupled adversaries while maintaining the unpredictability of the reward. The experimental results demonstrate that TCRL consistently outperforms existing methods in terms of robustness against temporal-coupled perturbation attacks across a variety of CRL tasks. A detailed version with full theoretical analysis, extended experiments, and additional implementation details is available at: https://github.com/biubiubiubiuhub/TCRL/tree/master.

KEYWORDS

Constrained reinforcement learning; Temporal-coupled; Adversarial training; Worst-case scenarios.

ACM Reference Format:
Wentao Xu, Zhongming Yao, Weihao Li, Zhenghang Song, Yumeng Song, Tianyi Li, and Yushuai Li. 2026. TCRL: Temporal-Coupled Adversarial Training for Robust Constrained Reinforcement Learning in Worst-Case Scenarios: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/GPHO5000

*Corresponding author

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/GPHO5000

1 INTRODUCTION

Constrained reinforcement learning (CRL) is an important branch of traditional reinforcement learning (RL). In recent years, CRL has attracted widespread attention due to its unique ability to optimize decision-making policies under constraint conditions, maximizing task rewards while ensuring that system constraints are not violated. However, when CRL is deployed in real-world physical environments, agents often encounter external adversarial perturbations, which expose potential inherent limitations in the robustness of existing methods [6].

Existing work [8, 10] has demonstrated that traditional CRL exhibits pronounced vulnerability to adversarial attacks. While robust RL has introduced some approaches to address environmental uncertainties [1, 2, 4, 7, 12], these approaches are not directly applicable to CRL. In CRL, satisfying safety constraints is typically prioritized over optimizing rewards. In contrast, agents in robust RL tend to aggressively pursue reward maximization, often violating critical safety constraints during exploration, especially under adversarial conditions. Existing studies [9, 13] in CRL largely address robustness by modeling adversarial input perturbations under constraint conditions. However, they are limited to addressing temporal-independent observational perturbations and fail to consider robustness under worst-case temporal-coupled attack scenarios. In real-world environments, attackers can generate strong temporal perturbations by learning the safety cost constraint function and reward function [5]. The attack intensity progressively increases over time, causing the agent to deviate from its target and potentially exhibit dangerous behaviors due to the influence of temporal-coupled perturbations, which can lead to significant reward degradation and potentially dangerous outcomes.

We propose TCRL, a robust CRL method enhancing agent robustness under temporally coupled state perturbations. To enable the cost constraint function to identify opponents with the worst-case temporal-coupled adversaries, it includes a safety cost constraint function for per-state worst-case cost estimation. To ensure CRL policy optimality under attack, it includes a dual-constraint reward-based defense to disrupt adversaries' temporal coupling while preserving reward unpredictability. We compare TCRL with three baselines. The experimental results show that TCRL effectively defends against worst-case temporal-coupled perturbations.

2 METHODOLOGY

First, we introduce a network to estimate the worst-case safety cost and construct a dedicated safety constraint function, which enables accurate worst-case cost estimation without explicitly modeling an attacker. We formulate the single-step worst-case action cost value $\underline{Q}_c^\pi(s,a) := \mathbb{E}_{\pi,P}\left[\sum_{t=0}^{\infty}\gamma^t c(s_t,a_t^*)\right]$, where a_t^* is the worst-case action taken by state s_t under a temporal-coupled perturbation attack and π is the policy of agents. The worst-case cost value $\underline{V}_c^\pi$ can be define in the same way. To estimate the worst-case costs, we introduce a new worst-case cost Bellman operator $\underline{\Gamma}^\pi Q_c(s,a) = \mathbb{E}_{s'\sim P(s'|s,a)}\left[c(s,a) + \gamma \max_{a'\in\Omega(s',\pi)} \underline{Q}_c^\pi(s',a')\right]$, where $\forall s \in \mathcal{S}$, $\Omega(s,\pi)$ is the set of all actions that the attacker may output by perturbing the inputs to $\tilde{s} \in \mathcal{B}_\epsilon(s')$ by perturbing the strategy π in state s'. The defined action set $\Omega_{\text{adv}}(s,\pi)$ involves identifying the potential actions that the policy π might produce when the state s undergoes temporal-coupled perturbations confined within $\mathcal{B}_\epsilon(s)$. To compute the worst-case cost value, we train a worst-case cost network $\underline{Q}_\varphi^\pi$. Given a mini-batch $\{s_t, a_t, r_t, s_{t+1}\}_{t=1}^N$, $\underline{Q}_\varphi^\pi$ is optimized using the following loss function: $\mathcal{L}_{\text{cost}}\left(\underline{Q}_\varphi^\pi\right) := \frac{1}{N}\sum_{t=1}^N \left(\underline{c}_t - \underline{Q}_\varphi^\pi(s_t,a_t)\right)^2$, where $\underline{c}_t = c_t + \gamma \min_{a'\in\Omega(s_{t+1},\pi)} \underline{Q}_\varphi^\pi(s_{t+1},a')$. The corresponding worst-cost value is computed as $\underline{V}_c^\pi(s) = \max_{a\in\Omega(s,\pi)} \underline{Q}_\varphi^\pi(s,a)$.

Second, we establish a dual-constraint defense mechanism on reward signals to effectively counteract temporal-coupled perturbation attacks. To limit reward temporal correlation to impede attackers from modeling the reward function, we propose a reward-based temporal decoupling optimization method with an autocorrelation constraint: $C_{\text{corr}} = \frac{1}{w^2}\sum_{k=1}^w \sum_{l=1}^k |\phi(\tilde{r}_t, \tilde{r}_{t-l})| \le \epsilon_{corr}$, where $\tilde{r}_t$ is the obtained reward value under attack and w is the time window. ϕ is the autocorrelation function. The autocorrelation function $\phi : \mathbb{R}^2 \to [-1,1]$ is in standardized covariance form: $\phi(\tilde{r}_t, \tilde{r}_{t-l}) = \frac{\mathbb{E}[(\tilde{r}_t-\mu_t)(\tilde{r}_{t-l}-\mu_{t-l})]}{\sigma_t\sigma_{t-l}}$, where $\mu_t = \mathbb{E}[\tilde{r}_t]$ and $\sigma_t^2 = Var(\tilde{r}_t)$ denote the expectation and and variance of the reward received by the agent under temporally coupled perturbation attacks, respectively.

To preserve reward function unpredictability and prevent attackers from forecasting subsequent rewards, in the reward signal discretization process, we divide the reward space $\mathcal{R} \in [\tilde{r}_{\min}, \tilde{r}_{\max}]$ over consecutive time steps into N equal-width bins, and the empirical distribution within a sliding window of size w is computed following the method in autocorrelation constraints: $p_i^{(w)}(t) = \frac{1}{w}\sum_{k=t-w+1}^t \mathbb{I}(\tilde{r}_k \in \mathcal{D}_i)$, where $\mathcal{D}_i = [\tilde{r}_{min} + (i-1)\Delta r, \tilde{r}_{min} + i\Delta r)$ is the i-th reward interval, and $\Delta i = (\tilde{r}_{max} - \tilde{r}_{min})/N$. The corresponding time-varying entropy function, based on this distribution, is expressed as follows: $H_t = -\sum_{i=1}^N p_i^{(w)}(t)\log p_i^{(w)}(t)$.

To prevent abrupt changes in the reward distribution caused by temporally coupled perturbation attacks, entropy variation rate constraints are introduced as follows:$C_{\text{ent}} = \|H_t - H_{t-1}\|_\infty \le \epsilon_{\text{ent}}$,where ϵ_{ent} is the maximum allowed rate of entropy change. This constraint ensures that entropy changes between consecutive time steps stay within a bounded range, promoting a smooth evolution of the reward distribution.

We optimize the strategy in the state affected by the attacker, denoted as $\tilde{\pi}$. The agent's interaction with the environment under attack yields a trajectory $\tilde{\tau} = \{\tilde{s}_0, \tilde{a}_0, \tilde{s}_1, \tilde{a}_1, \ldots\}$, where each action $\tilde{a}_t \in \Omega(\tilde{s}_t, \tilde{\pi})$. Accordingly, we formulate the CRL objective under the temporal- coupled perturbation attacker h as follows:

$$\pi_{\text{adv}} = \max_\pi \mathbb{E}\left[\sum_{t=0}^T \gamma^t r_t\right] \quad \text{s.t. } \underline{V}_c^{\tilde{\pi}} \le \eta,\ C_{\text{corr}} \le \epsilon_{\text{corr}},\ C_{\text{ent}} \le \epsilon_{\text{ent}}, \forall h. \tag{1}$$

3 EXPERIMENT

To assess the robustness of TCRL against temporal-coupled perturbations, we design a well-trained worst-case temporal-coupled attacker (Worst-TC). We selected robotic motion control as the test domain. The simulated environments are adopted from a previous benchmark [3]. To validate the effectiveness of TCRL, we integrate it with the PID-PPO-Lagrange (PPOL) framework [11] to form TCRL-PPOL and We present three baseline robust training methods, which are alternately trained alongside adversarial attackers to enhance robustness.

Table 1: The performance of training methods under Worst-TC attacks. Each experimental result is averaged over 50 episodes and 10 random seeds, and is reported as the mean ± standard deviation.

Method	*Reward*	*Cost*
PPOL-vanilla	630.73±234.36	75.44±6.73
PPOL-random	600.31±170.30	72.80±8.04
ADV-PPOL(MC)	521.57±374.80	25.32±27.69
TCRL-PPOL	**709.40±253.81**	**3.84±6.45**

Table 1 reports that TCRL-PPOL outperforms all three baselines in defending against state perturbations across the two tasks (Ball-Circle, Ball-Run). We observe that the safety performance of baselines degrades significantly when subjected to the Worst-TC attacker, which employs temporal-coupled perturbations. TCRL consistently demonstrates superior safety performance compared to all baselines, achieving the lowest safety cost even under the temporal-coupled perturbations introduced by the Worst-TC attacker. This strategy becomes advantageous under Worst-TC attack scenarios, allowing the policy to perform more robustly against adversarial perturbations, while maintaining higher rewards than baselines.

4 CONCLUSION

We propose TCRL, a robust CRL method enhancing agent robustness under temporally coupled state perturbations via a novel worst-case training framework. It includes a safety cost constraint function for per-state worst-case cost estimation and a dual-constraint reward-based defense to disrupt adversaries' temporal coupling while preserving reward unpredictability. Experiments show TCRL effectively defends against worst-case temporal-coupled perturbations and maintains safety under temporal-independent attacks. In future research, it is of interest to extend this work to multi-agent settings and real-world systems.

REFERENCES

[1] Lukas Brunke, Melissa Greeff, Adam W Hall, Zhaocong Yuan, Siqi Zhou, Jacopo Panerati, and Angela P Schoellig. 2022. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems* 5, 1 (2022), 411–444.

[2] Javier Garcıa and Fernando Fernández. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research* 16, 1 (2015), 1437–1480.

[3] Sven Gronauer. 2022. Bullet-safety-gym: A framework for constrained reinforcement learning. (2022).

[4] Xiangkun He and Chen Lv. 2023. Robotic control in adversarial and sparse reward environments: A robust goal-conditioned reinforcement learning approach. *IEEE Transactions on Artificial Intelligence* 5, 1 (2023), 244–253.

[5] Peter J Jin, Da Yang, and Bin Ran. 2013. Reducing the error accumulation in car-following models calibrated with vehicle trajectory data. *IEEE Transactions on Intelligent Transportation Systems* 15, 1 (2013), 148–157.

[6] Zeyang Li, Chuxiong Hu, Shengbo Eben Li, Jia Cheng, and Yunan Wang. 2023. Robust safe reinforcement learning under adversarial disturbances. In *2023 62nd IEEE Conference on Decision and Control (CDC)*. IEEE, 334–341.

[7] Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Steven Wu, Bo Li, and Ding Zhao. 2022. Constrained variational policy optimization for safe reinforcement learning. In *International Conference on Machine Learning*. PMLR, 13644–13668.

[8] Zuxin Liu, Zijian Guo, Zhepeng Cen, Huan Zhang, Jie Tan, Bo Li, and Ding Zhao. 2022. On the robustness of safe reinforcement learning under observational perturbations. *arXiv preprint arXiv:2205.14691* (2022).

[9] Daniel J Mankowitz, Dan A Calian, Rae Jeong, Cosmin Paduraru, Nicolas Heess, Sumanth Dathathri, Martin Riedmiller, and Timothy Mann. 2020. Robust constrained reinforcement learning for continuous control with model misspecification. *arXiv preprint arXiv:2010.10644* (2020).

[10] Jinling Meng, Fei Zhu, Yangyang Ge, and Peiyao Zhao. 2023. Integrating safety constraints into adversarial training for robust deep reinforcement learning. *Information Sciences* 619 (2023), 310–323.

[11] Adam Stooke, Joshua Achiam, and Pieter Abbeel. 2020. Responsive safety in reinforcement learning by pid lagrangian methods. (2020), 9133–9143.

[12] Long Yang, Jiaming Ji, Juntao Dai, Yu Zhang, Pengfei Li, and Gang Pan. 2022. Cup: A conservative update policy algorithm for safe reinforcement learning. *arXiv preprint arXiv:2202.07565* (2022).

[13] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh. 2020. Robust deep reinforcement learning against adversarial perturbations on state observations. *Advances in Neural Information Processing Systems* 33 (2020), 21024–21037.