Adversarially-Robust TD Learning with Markovian Data: Finite-Time Rates and Fundamental Limits

Sreejeet Maity **Aritra Mitra**

Abstract

One of the most basic problems in reinforcement learning (RL) is policy evaluation: estimating the long-term return, i.e., value function, corresponding to a given fixed policy. The celebrated Temporal Difference (TD) learning algorithm addresses this problem, and recent work has investigated finite-time convergence guarantees for this algorithm and variants thereof. However, these guarantees hinge on the reward observations being always generated from a well-behaved (e.g., sub-Gaussian) true reward distribution. Motivated by harsh, real-world environments where such an idealistic assumption may no longer hold, we revisit the policy evaluation problem from the perspective of *adversarial robustness*. In particular, we consider a Huber-contaminated reward model where an adversary can arbitrarily corrupt each reward sample with a small probability ϵ. Under this observation model, we first show that the adversary can cause the vanilla TD algorithm to converge to any arbitrary value function. We then develop a novel algorithm called `Robust-TD` and prove that its finite-time guarantees match that of vanilla `TD` with linear function approximation up to a small $O(\epsilon)$ term that captures the effect of corruption. We complement this result with a minimax lower bound, revealing that such an additive corruption-induced term is unavoidable. To our knowledge, these results are the first of their kind in the context of adversarial robustness of stochastic approximation schemes driven by Markov noise. The key new technical tool that enables our results is an analysis of the Median-of-Means estimator with corrupted, time-correlated data that might be of independent interest to the literature on robust statistics.

Keywords: Temporal difference learning, robust reinforcement learning, adversarial robustness, finite-time analysis.

1 Introduction

In recent years, a significant body of research has focused on understanding the effects of adversarial corruption on deep learning [1, 2]. While this line of work has contributed significantly to the design of reliable and trustworthy machine-learning models, the developments have primarily catered to supervised learning [3]. Much less is understood when an adversary poisons data arriving in an online manner in the context of reinforcement learning (RL). Arguably, one of the most fundamental problems in RL is that of *policy evaluation*, where a learner unaware of the true underlying model of a Markov Decision Process (MDP) seeks to evaluate the long-term return (i.e., the value function) associated with a given fixed policy. To do so, at each time step, it plays an action based on the policy to be evaluated, observes as data a reward, and transitions to a new state. Importantly, the rewards are always generated based on the (unknown) reward functions of the MDP. Departing from this paradigm, we consider a scenario where a small fraction of the reward samples can be *arbitrarily* corrupted by a powerful adversary possessing complete knowledge of the MDP. One would ideally like to obtain guarantees on value function estimation that *degrade gracefully* with the corruption fraction. Whether this is possible is a hitherto unexplored question that we resolve in this paper.

To provide context, in the absence of adversarial corruption, the classical Temporal Difference (TD) learning algorithm [4] solves the policy evaluation problem. An asymptotic analysis of `TD`(0) - the simplest TD learning algorithm - with linear function approximation was provided by the seminal work in [5]. More recently, a growing body of work has provided finite-time guarantees for TD learning with linear function approximation [6–10], and more general nonlinear stochastic approximation (SA) schemes [11–14]. *Crucially, the guarantees in each of the above papers assume that the rewards are always drawn from true reward distributions linked to the underlying MDP*. Moreover, the rewards are either assumed to be deterministic or generated from light-tailed sub-Gaussian distributions.

Motivation. Unfortunately, such assumptions do not adequately capture harsh, real-world environments. For instance, in large-scale, complex systems such as the power grid [15] or multi-robot networks [16], data is collected via imperfect sensors prone to unexpected failures and adversarial attacks. Motivated by the need to safeguard against such attacks that are common in cyber-physical systems [17], we consider a reward contamination model where, at each time step, with probability $1-\varepsilon$, the reward is generated from a true reward distribution, and with probability ε, from an arbitrary error distribution controlled by an adversary. Here, ε captures the power of the adversary. Our data poisoning model is directly inspired by the Huber model from robust statistics [18, 19]. Similar reward contamination models have also been widely studied in the context of multi-armed bandits [20–25]. However, beyond bandits, when it comes to *SA schemes in RL*, no prior work has provided a finite-time analysis of the effects of such attacks. Given this premise, we ask:

Is it possible to perform accurate value function estimation under the Huber-contaminated reward model? If so, what are the fundamental limits on performance imposed by this attack model?

Challenges. The main difficulty in answering these questions arises from the need to deal with two different forms of uncertainty: the lack of knowledge of the MDP, and the uncertainty injected by the adversary. Furthermore, other than requiring the true reward distributions to have finite first and second moments, we make no assumptions of sub-Gaussianity. This makes it particularly challenging for the learner to distinguish between time-correlated, potentially heavy-tailed clean rewards (inliers) and adversarial outliers.

2 Our Algorithm and Results

In this paper, we systematically study adversarial robustness in the context of policy evaluation with linear function approximation. Our specific contributions are as follows.

• **Vulnerability of `TD`**. We start with a simple result proving that under the Huber-contaminated reward model, an adversary can cause the vanilla `TD`(0) algorithm to converge to any arbitrary point. This is illustrated in Figure 1, where we observe that even when the corruption fraction ε is merely 0.001, the mean-square error of vanilla `TD`(0) can be large. Our finding in this regard directly motivates the need for adversarially robust variants of `TD`(0).

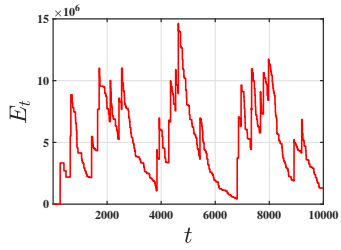

Figure 1: Plot of mean-square error E_t showing the effect of reward corruption on `TD`(0), with corruption probability $\varepsilon = 0.001$.

• **Robust-TD Algorithm**. On the algorithmic front, our main contribution is the development of an adversarially robust variant of `TD`(0) called `Robust-TD`. `Robust-TD` relies on two main new ideas that we briefly explain next.

Idea 1. Robust Mean Estimation. Our first idea is to use historical data of the reward observations to construct an instantaneous robust estimate $\hat{g}_t$ of the true TD update direction g_t at each time-step t. For this purpose, we adapt the median-of-means device from robust statistics to our needs, and exploit the affine structure of the TD update direction under linear function approximation. Unfortunately, this robust estimation step is not adequate on its own. To see why, let us note that since the robust TD update direction $\hat{g}_t$ is constructed using noisy data, all guarantees

associated with the estimation error $\|\hat{g}_t - g_t\|_2$ only hold with high probability, *but not deterministically.* Thus, there could very well be rare/extreme events on which such an estimation error is large. To safeguard against such rare events, we need an additional layer of safety; this leads to our next main algorithmic idea.

Idea 2. Dynamic Thresholding. Using a careful concentration analysis, we characterize a bound of the form $\|\hat{g}_t\|_2 \leq B_t$ that holds with high probability at each time-step t. Intuitively, the bound B_t informs us of a "typical region" that contains the estimated TD update direction $\hat{g}_t$. Then, the basic idea is the following: If $\|\hat{g}_t\|_2 > B_t$, we realize that an atypical event has taken place and, as such, no update to the TD parameter is made at time t; otherwise, we make an update. In this context, designing the dynamic threshold B_t is a delicate task: if the threshold is either too tight or too loose, then we end up with vacuous guarantees.

• **Main Convergence Result.** Our main convergence result for `Robust-TD` establishes a mean-square error bound of the form $\tilde{O}(\tau_{mix}/T) + O(\varepsilon)$, where τ_{mix} is the mixing time of the underlying data-generating Markov chain, T is the number of iterations, and ε is the corruption probability. When $\varepsilon = 0$, our result is consistent with prior finite-time guarantees for `TD`(0) with linear function approximation [7, 8]. Thus, `Robust-TD` is provably robust to adversarial reward contamination, and its guarantees match that of vanilla `TD`(0) up to a small $O(\varepsilon)$ term. Crucially, the additive $O(\varepsilon)$ term is *completely unaffected by the magnitude of the attack inputs* and depends only on instance-dependent parameters, such as the variance of the noise in the reward observation model. To our knowledge, this is the first result of its kind for SA schemes in RL driven by Markov noise and subject to adversarial outliers. Proving such a result is highly non-trivial, as we need to contend with the complex interplay between Markovian noise, adversarial perturbations, and function approximation.

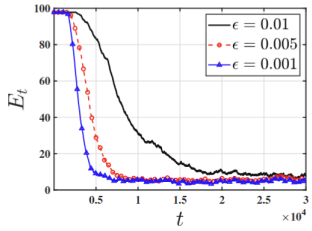

Figure 2: Plot of mean-square error E_t of proposed `Robust-TD` algorithm, with different corruption fractions ε.

• **Minimax Lower Bound.** To complement our upper-bound, we provide an algorithm-independent information-theoretic lower bound. This lower bound reveals that the additive $O(\varepsilon)$ term is *unavoidable*, and captures the fundamental price of adversarial contamination for the policy evaluation problem.

Overall, our algorithmic and theoretical contributions above provide a fairly complete characterization of the effects of reward contamination (under the Huber model) on policy evaluation in general, and TD learning in particular.

To corroborate our theory, we simulate the performance of `Robust-TD` on an MDP with 100 states, and report our observations in Figure 2. As can be seen from this figure, our proposed algorithm is robust to adversarial corruptions across a range of corruption probabilities.

• **Robust Mean Estimation with Markov Data**. A key ingredient in our algorithmic development is that of robust mean estimation. While the literature on robust statistics has made significant advances in this regard [26, 27], the results we know of all assume independent and identically (i.i.d.) distributed inliers. The same is true for some recent papers in RL [28, 29] that consider heavy-tailed i.i.d. rewards with no corruption. Since the reward samples in our setting are generated based on a Markov chain, we cannot directly appeal to such existing work. To overcome this difficulty, we provide the *first analysis of the Median-of-Means (`MoM`) estimator under Huber contamination and Markovian data.* In particular, our analysis carefully exploits the ergodicity of the underlying Markov chain along with a novel coupling argument. We also note that while the popular `MoM` scheme was known to be robust to heavy-tailed data, the fact that it is also robust to adversarial corruption appears to be new. As such, we believe that our main result in this context might be of independent interest to the broad area of robust statistics.

Related Work. We discuss the most relevant works on *adversarial robustness in RL* below. Data corruption in online finite-horizon episodic RL problems is studied by [30] and [31], where the notion of performance is measured by cumulative regret. Our setting is *fundamentally different* in that we consider an infinite horizon, discounted single-trajectory setting, where performance is captured by the mean-squared error w.r.t. the solution to the projected Bellman equation. Furthermore, our algorithm builds on stochastic approximation and differs significantly from the Upper-Confidence-Based (UCB)/Action-Elimination type algorithms employed in [30, 31]. Corruption-robust algorithms in the offline setting are considered in [32], where data tuples are collected offline in an i.i.d. manner, and the true rewards are assumed to be sub-Gaussian. In sharp contrast, data arrives sequentially in our setting, and, as such, *we need to contend with corruption under heavy-tailed Markovian data* - a much more challenging setting. Different from the SA problem we consider here, outlier-robust policy gradient (PG) algorithms have been explored in [33], where the issue of Markovian sampling does not arise. Finally, a very recent work [34] considers heavy-tailed rewards *with no adversarial corruption* in TD learning. The analysis in their paper requires a strong realizability assumption and relies on a projection step in the algorithm to control the iterates; we require neither, making it much more challenging to tackle both heavy-tailed data and adversarial perturbations. Moreover, our proposed algorithm differs considerably from that in [34]. **In summary, our work is the first to study the topic of adversarial reward corruption in the context of TD learning with function approximation and Markovian data.** We anticipate that the findings from this paper will facilitate the study of more involved corruption models in RL as future work.

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[3] Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

[4] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[5] John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. In *IEEE Transactions on Automatic Control*, 1997.

[6] Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[7] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.

[8] Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

[9] Gandharv Patil, LA Prashanth, Dheeraj Nagaraj, and Doina Precup. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. In *International Conference on Artificial Intelligence and Statistics*, pages 5438–5448. PMLR, 2023.

[10] Aritra Mitra. A simple finite-time analysis of td learning with linear function approximation. *IEEE Transactions on Automatic Control*, 70(2):1388–1394, 2024.

[11] Zaiwei Chen, Sheng Zhang, Thinh T Doan, Siva Theja Maguluri, and John-Paul Clarke. Performance of q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*, page 4, 2019.

[12] Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022.

[13] Adam Wierman Guannan Qu. Finite-time analysis of asynchronous stochastic approximation and q-learning. *Proceedings of Machine Learning Research*, 125:1–21, 2020.

[14] Zaiwei Chen, Siva T Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A lyapunov theory for finite-sample guarantees of markovian stochastic approximation. *Operations Research*, 72(4):1352–1367, 2024.

[15] Oliver Kosut, Liyan Jia, Robert J Thomas, and Lang Tong. Malicious data attacks on the smart grid. *IEEE Transactions on Smart Grid*, 2(4):645–658, 2011.

[16] Stephanie Gil, Swarun Kumar, Mark Mazumder, Dina Katabi, and Daniela Rus. Guaranteeing spoof-resilient multi-robot networks. *Autonomous Robots*, 41:1383–1400, 2017.

[17] Seyed Mehran Dibaji, Mohammad Pirani, David Bezalel Flamholz, Anuradha M Annaswamy, Karl Henrik Johansson, and Aranya Chakrabortty. A systems and control perspective of cps security. *Annual reviews in control*, 47:394–411, 2019.

[18] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer, 1992.

[19] Peter J Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

[20] Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Xiaojin Zhu. Adversarial attacks on stochastic bandits. *arXiv preprint arXiv:1810.12188*, 2018.

[21] Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 114–122, 2018.

[22] Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pages 4042–4050. PMLR, 2019.

[23] Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pages 1562–1578. PMLR, 2019.

[24] Sayash Kapoor, Kumar Kshitij Patel, and Purushottam Kar. Corruption-tolerant bandit learning. *Machine Learning*, 108(4):687–715, 2019.

[25] Shubhada Agrawal, Timothée Mathieu, Debabrota Basu, and Odalric-Ambrym Maillard. Crimed: Lower and upper bounds on regret for bandits with unbounded stochastic corruption. In *International Conference on Algorithmic Learning Theory*, pages 74–124. PMLR, 2024.

[26] Kevin A Lai, Anup B Rao, and Santosh Vempala. Agnostic estimation of mean and covariance. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 665–674. IEEE, 2016.

[27] Mengjie Chen, Chao Gao, and Zhao Ren. Robust covariance matrix estimation via matrix depth. *arXiv preprint arXiv:1506.00691*, 2015.

[28] Vincent Zhuang and Yanan Sui. No-regret reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 3385–3393. PMLR, 2021.

[29] Jin Zhu, Runzhe Wan, Zhengling Qi, Shikai Luo, and Chengchun Shi. Robust offline reinforcement learning with heavy-tailed rewards. In *International Conference on Artificial Intelligence and Statistics*, pages 541–549. PMLR, 2024.

[30] Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun. Corruption-robust exploration in episodic reinforcement learning. In *Conference on Learning Theory*, pages 3242–3245. PMLR, 2021.

[31] Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR, 2022.

[32] Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Corruption-robust offline reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5757–5773. PMLR, 2022.

[33] Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun. Robust policy gradient against strong data corruption. In *International Conference on Machine Learning*, pages 12391–12401. PMLR, 2021.

[34] Semih Cayci and Atilla Eryilmaz. Provably robust temporal difference learning for heavy-tailed rewards. *Advances in Neural Information Processing Systems*, 36, 2024.