Robust Q-Learning under Corrupted Rewards

Sreejeet Maity **Aritra Mitra**

Abstract

Recently, there has been a surge of interest in analyzing the non-asymptotic behavior of model-free reinforcement learning algorithms. However, the performance of such algorithms in non-ideal environments, such as in the presence of corrupted rewards, is poorly understood. Motivated by this gap, we investigate the robustness of the celebrated Q-learning algorithm to a strong-contamination attack model, where an adversary can arbitrarily perturb a small fraction of the observed rewards. We start by proving that such an attack can cause the vanilla Q-learning algorithm to incur arbitrarily large errors. We then develop a novel robust synchronous Q-learning algorithm that uses historical reward data to construct robust empirical Bellman operators at each time step. Finally, we prove a finite-time convergence rate for our algorithm that matches known state-of-the-art bounds (in the absence of attacks) up to a small inevitable $O(\varepsilon)$ error term that scales with the adversarial corruption fraction ε. Notably, our results continue to hold even when the true reward distributions have infinite support, provided they admit bounded second moments.

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

1 Introduction

We study reinforcement learning (RL) within a Markov decision process (MDP) setting where an agent sequentially interacts with an environment to maximize a long-term cumulative value function. To achieve this goal without knowledge of the dynamics of the MDP, the agent plays an action at each time, receives a reward in the form of feedback from the environment, and then transitions to a new state. This process then repeats itself. Using the sequence of observed rewards, the agent learns to take "better" actions over time. In this context, one of the most widely studied model-free RL algorithms is the celebrated Q-learning algorithm of Watkins and Dayan [1] that iteratively estimates the optimal state-action value function associated with the MDP. Viewing Q-learning through the lens of stochastic approximation (SA) theory, a rich body of work has investigated the asymptotic performance of this algorithm, i.e., its behavior as the number of iterations goes to infinity [2–5]. More recently, there has been a shift of interest towards providing a finer non-asymptotic/finite-time analysis of different RL algorithms [6–11]. That said, the literature discussed above exclusively pertains to scenarios where the reward feedback received from the environment is *always accurate*, i.e., it is generated by the true reward distribution of the underlying MDP. However, such an idealistic assumption is unlikely to hold when one seeks to deploy RL algorithms in real-world harsh environments. As such, since feedback in the form of rewards is crucial for the overall decision-making process, our main goal in this paper is to investigate the following questions:

Q1. *Are existing model-free RL algorithms robust to perturbations in the rewards?*
Q2. *Can we obtain reliable value-function estimates under corrupted rewards?*

Contributions. In response to the aforementioned questions, the main contributions of this work are as follows.

• **Problem Formulation.** To systematically investigate the question of robustness of RL algorithms, we consider a *strong-contamination reward attack model* where an adversary with complete knowledge of the MDP and the learner's observations is allowed to *arbitrarily* perturb ε-fraction of the rewards acquired over time. This attack model is directly inspired from the robust statistics literature [12, 13], where a small fraction of the data points in a static data set can be corrupted by an adversary.

• **Vulnerability of vanilla Q-Learning.**In [14], we prove that, for an attack model stricter than the one described there, vanilla Q-learning converges to the fixed point of a Bellman operator perturbed by reward contamination. By constructing an explicit example, we next establish that this incorrect fixed point can be arbitrarily far away from the optimal Q function. Furthermore, this continues to be the case even when the corruption fraction ε is small.

• **Robust Q-Learning Algorithm.** Motivated by the above findings, we develop a novel robust Q-learning algorithm in the synchronous sampling setting [15], i.e., when a generative model provides independent samples for all state-action pairs.

The key idea in our algorithm is to use historical data of rewards for each state-action pair to construct a robust empirical Bellman operator in each iteration. To do so, we leverage the recently developed robust trimmed mean-estimator in [13], along with a novel dynamic thresholding technique.

• **Near-optimal Rates.** Providing a rigorous finite-time convergence analysis of our algorithm is non-trivial since one needs to simultaneously contend with the benign randomness introduced by sampling and deliberate arbitrary adversarial injections. Nonetheless, in our main result, we establish a high-probability ℓ_∞-error-rate of $\tilde{O}(1/\sqrt{T}) + O(\sqrt{\varepsilon})$, where T is the number of iterations. In the absence of corruption (i.e., $\varepsilon = 0$), this rate matches those in recent works [9, 10]. Furthermore, we also provide an informal argument that the additive $O(\sqrt{\varepsilon})$ term appears to be inevitable.

Finally, we note that our proposed approach does not require the true reward distributions to be bounded or "light-tailed", i.e., we do not need to make any assumptions of sub-Gaussianity on the reward models. Instead, we only need the reward distributions to admit finite second moments. Thus, in principle, even the true reward samples (i.e., the inliers) can come from heavy-tailed distributions. This makes it non-trivial to tell apart clean reward data from corrupted rewards. Nonetheless, the approach developed in this paper overcomes this challenge.

Related Work. The synchronous Q-learning setting we consider here has been extensively studied in several works [15–17]. More recently, its finite-time performance has been characterized in [7, 9, 11]. Asynchronous versions of Q-learning have also been analyzed in [18, 10, 11]. Reward corruption models similar to what we consider here have been studied in the multi-armed bandits (MAB) literature [19–25], where an attacker with a finite budget can corrupt the rewards seen by the learner in a small fraction of the rounds. However, our setup, algorithm, and proof technique fundamentally differ from the MAB framework. That said, our main result has the same flavor as the findings in the above bandit papers: one can recover the performance bounds in the absence of corruptions up to an additive term that captures the attacker's budget; in our case, this is the additive $O(\sqrt{\varepsilon})$ term. Finally, some recent works have studied perturbed versions of Markovian stochastic approximation algorithms [26, 27], where the perturbations are due to communication constraints (e.g., delays and compression). Unlike the *structured* perturbations in these papers, the perturbations injected by the adversary in our work can be *arbitrary*.

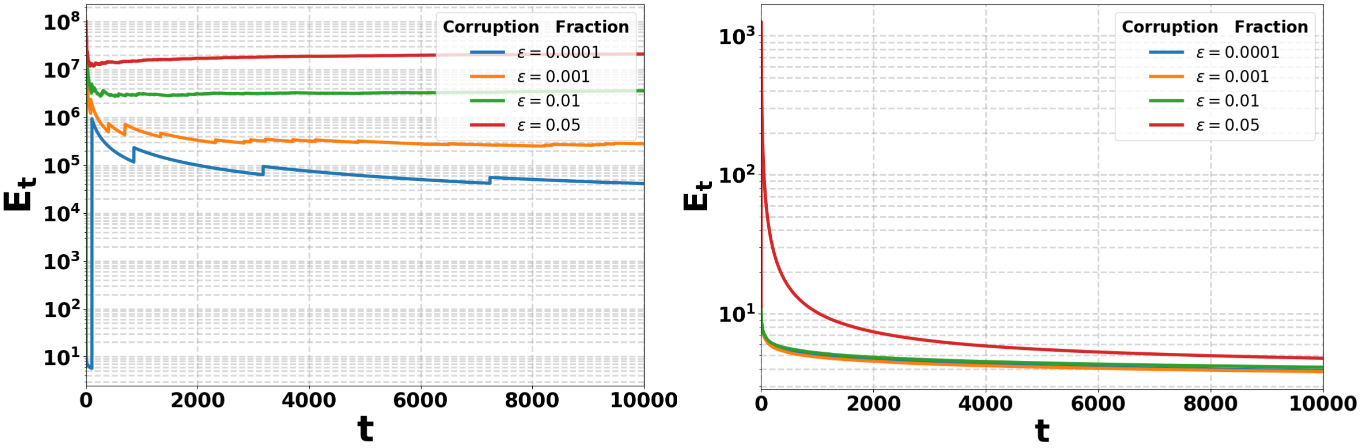

Figure 1: (**Left**) Plots of the ℓ_∞ error $E_t = \|Q_t - Q^*\|_\infty$ for `Vanilla-Q` with corruption fractions $\varepsilon \in \{0.0001, 0.001, 0.01, 0.05\}$ and reward variance $\sigma^2 = 10$, under a biasing attack of magnitude 10^5. (**Right**) Plots of the ℓ_∞ error for `Robust Sync-Q` with the same set of corruption fractions under the same attack. All curves are averaged over 100 independent runs.

Empirical validation. Our experiments mirror the theoretical results established in [14]. Under the corruption model described in , `Vanilla-Q` exhibits the predicted instability: even for modest ε, an adversarial bias significantly worsen the ℓ_∞ error $E_t = \|Q_t - Q^*\|_\infty$, culminating in divergence at longer horizons. In contrast, `Robust Sync-Q` achieves a substantial error reduction, even under the same adversarial interference. Across all tested ε and under a strong bias of 10^5, the robust variant maintains bounded error and shows the expected improvement with additional samples, whereas `Vanilla-Q` does not. Our results provide direct empirical support that vanilla Q-learning is vulnerable to adversaries, whereas the robust approach continues to learn reliably.

References

[1] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

[2] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[3] John N Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16:185–202, 1994.

[4] Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in neural information processing systems*, 6, 1993.

[5] Csaba Szepesvári. The asymptotic convergence-rate of Q-learning. *Advances in neural information processing systems*, 10, 1997.

[6] 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.

[7] Devavrat Shah and Qiaomin Xie. Q-learning with nearest neighbors. *Advances in Neural Information Processing Systems*, 31, 2018.

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

[9] Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

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

[11] Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236, 2024.

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

[13] Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *The Annals of Statistics*, 49(1):393–410, 2021.

[14] Sreejeet Maity and Aritra Mitra. Robust q-learning under corrupted rewards. In *2024 IEEE 63rd Conference on Decision and Control (CDC)*, pages 1181–1186, 2024.

[15] Michael Kearns and Satinder Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. *Advances in neural information processing systems*, 11, 1998.

[16] Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for Q-learning. *Journal of machine learning Research*, 5(1), 2003.

[17] Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. *Advances in Neural Information Processing Systems*, 31, 2018.

[18] Carolyn L Beck and Rayadurgam Srikant. Error bounds for constant step-size Q-learning. *Systems & control letters*, 61(12):1203–1208, 2012.

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

[20] 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.

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

[22] 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.

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

[24] Ilija Bogunovic, Andreas Krause, and Jonathan Scarlett. Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 1071–1081. PMLR, 2020.

[25] Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. Stochastic linear bandits robust to adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR, 2021.

[26] Aritra Mitra, George J Pappas, and Hamed Hassani. Temporal difference learning with compressed updates: Error-feedback meets reinforcement learning. *arXiv preprint arXiv:2301.00944*, 2023.

[27] Arman Adibi, Nicolò Dal Fabbro, Luca Schenato, Sanjeev Kulkarni, H Vincent Poor, George J Pappas, Hamed Hassani, and Aritra Mitra. Stochastic approximation with delayed updates: Finite-time rates under markovian sampling. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2024.