Proceedings of Machine Learning Research vol 134:1–4, 2021 34rd Annual Conference on Learning Theory

Corruption-robust exploration in episodic reinforcement learning

Thodoris Lykouris LYKOURIS@MIT.EDU
Massachusetts Institute of Technology

Max Simchowitz MSIMCHOW@BERKELEY.EDU
UC Berkeley

Aleksandrs Slivkins SLIVKINS@MICROSOFT.COM
Microsoft Research New York City

Wen Sun WS455@CORNELL.EDU
Cornell University

Editors: Mikhail Belkin and Samory Kpotufe

In reinforcement learning (RL), an agent encounters a particular state and decides which action to select; as a result, the agent transitions to a new state and collects some reward. Standard RL approaches assume that rewards and transition dynamics are drawn identically and independently from fixed (yet unknown) distributions that depend on the current state and the selected action. However, these techniques tend to be vulnerable to even a small amount of outliers from such i.i.d. patterns. Such outliers are prevalent in most RL applications *e.g.,* click fraud in online advertising, patients not following prescriptions in clinical trials, attacks against RL agents in computer gaming.

We focus on *episodic RL*, a basic paradigm in which time is partitioned into *episodes* of fixed length H, and the agent's state is reinitialized in each episode. The algorithm only observes the outcome of the chosen action, i.e., the next state and the reward received.We consider one model that captures outliers, that of *adversarial corruptions*. This model posits that most of the episodes display i.i.d. patterns but some of them are *corrupted*: have rewards and transitions that are selected by an adaptive adversary. The number of episodes that are corrupted, denoted by C, is not known to the agent. The adversary can choose an arbitrary sequence of episodes to be corrupted, *e.g.,* all corruptions may happen in the initial episodes, which causes irrevocable damage to standard algorithms that rely on these rounds for exploration. The goal is to design algorithms whose performance gracefully degrades as C becomes larger while retaining the i.i.d. bounds when $C = 0$. This model is well-understood in multi-armed bandits (MAB), where there is only one state. However, the main challenge in RL lies in effectively learning the transition dynamics; this is absent in MAB.

We initiate the study of episodic RL under adversarial corruptions in both the rewards and the transition probabilities of the underlying system extending recent results for multi-armed bandits. We provide a framework which modifies the aggressive exploration enjoyed by existing reinforcement learning approaches based on "optimism in the face of uncertainty", by complementing them with principles from "action elimination" and, in doing so, achieves the above desiderata. To showcase the generality of our approach, we derive results for both tabular settings as well as linear MDP settings. Notably, our work provides the first sublinear regret guarantee which accommodates any deviation from purely i.i.d. transitions in the bandit-feedback model for episodic RL. [1]

1. Extended abstract. Full version appears as [https://arxiv.org/abs/1911.08689, v3].

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, and Csaba Szepesvari. Online learning in markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems*. 2013.

Idan Amir, Idan Attias, Tomer Koren, Roi Livni, and Yishay Mansour. Prediction with corrupted expert advice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Peter Auer and Chao-Kai Chiang. An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *29th Conf. on Learning Theory (COLT)*, 2016.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fisher. Finite-time regret bounds for the multi-armed bandit problems. *Machine Learning*, 47:2–3, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002b.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Ilya Bogunovic, Andreas Krause, and Jonathan Scarlett. Corruption-tolerant gaussian process bandit optimization. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: stochastic and adversarial bandits. In *25th Conf. on Learning Theory (COLT)*, 2012.

Xi Chen, Akshay Krishnamurthy, and Yining Wang. Robust dynamic assortment optimization in the presence of outlier customers. *arXiv preprint arXiv:1910.04183*, 2019.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 2818–2826, 2015.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 5713–5723, 2017.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Simon S Du, Yuping Luo, Ruosong Wang, and Hanrui Zhang. Provably efficient q-learning with function approximation via distribution shift error checking oracle. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research (JMLR)*, 7:1079–1105, 2006.

Eyal Even-Dar, Sham M Kakade, and Yishay Mansour. Online markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Proceedings of the 32nd Annual Conference on Learning Theory (COLT)*, 2019.

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. *arXiv preprint arXiv:2011.11566*, 2020.

Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume (ICML)*, 2017.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning adversarial mdps with bandit feedback and unknown transition. In *Proceedings of the 33rd Annual Conference on Learning Theory (COLT)*, 2020a.

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of the 33rd Annual Conference on Learning Theory (COLT)*, 2020b.

Tiancheng Jin and Haipeng Luo. Simultaneously learning stochastic and adversarial episodic mdps with known transition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, 2003.

Akshay Krishnamurthy, Thodoris Lykouris, Chara Podimata, and Robert Schapire. Contextual search in the presence of irrational agents. In *Proceedings of the 53rd ACM Annual Symposium on Theory of Computing*, 2021.

Yingkai Li, Edmund Y Lou, and Liren Shan. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proc. of the 50th ACM Annual Symposium on Theory of Computing*, 2018.

Gergely Neu and Julia Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. *arXiv preprint arXiv:2007.01612*, 2020.

Gergely Neu and Ciara Pike-Burke. A unifying view of optimism in episodic reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Gergely Neu, Andras Antos, András György, and Csaba Szepesvári. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1804–1812, 2010.

Gergely Neu, Andras Gyorgy, and Csaba Szepesvári. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Artificial Intelligence and Statistics*, 2012.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial markov decision processes. In *International Conference on Machine Learning*, pages 5478–5486, 2019.

Daniel Russo. Worst-case regret bounds for exploration via randomized value functions. *Advances in Neural Information Processing Systems*, 2019.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *30th Conf. on Learning Theory (COLT)*, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *31th Intl. Conf. on Machine Learning (ICML)*, 2014.

Max Simchowitz and Kevin Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. In *International Conference on Learning Representations*, 2020.

Chen-Yu Wei and Haipeng Luo. More adaptive algorithms for adversarial bandits. In *Conference On Learning Theory, COLT 2018, Stockholm, Sweden, 6-9 July 2018.*, pages 1263–1291, 2018.

Lin F Yang and Mengdi Wang. Reinforcement leaning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, 2020.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, 2019.

Alexander Zimin and Gergely Neu. Online learning in episodic markovian decision processes by relative entropy policy search. In *Neural Information Processing Systems (NeurIPS)*, 2013.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits, 2019.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research (JMLR)*, 2021.

Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *36th Intl. Conf. on Machine Learning (ICML)*, 2019.