Adversarially Robust Multi-Armed Bandit Algorithm with Variance-Dependent Regret Bounds

Shinji Ito I-SHINJI@NEC.COM
NEC Corporation and RIKEN AIP

Taira Tsuchiya TSUCHIYA@SYS.I.KYOTO-U.AC.JP
Kyoto University and RIKEN AIP

Junya Honda HONDA@I.KYOTO-U.AC.JP
Kyoto University and RIKEN AIP

Editors: Po-Ling Loh and Maxim Raginsky

This paper[1] considers the multi-armed bandit (MAB) problem and provides a new best-of-both-worlds (BOBW) algorithm that works nearly optimally in both stochastic and adversarial settings. In stochastic settings, some existing BOBW algorithms achieve tight gap-dependent regret bounds of $O(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i})$ for suboptimality gap Δ_i of arm i and time horizon T. On the other hand, it is shown in Audibert et al. (2007) that the regret bound can be tightened to $O(\sum_{i:\Delta_i>0}(\frac{\sigma_i^2}{\Delta_i}+1)\log T)$ using the loss variance σ_i^2 of each arm i in the stochastic environments. In this paper, we propose an algorithm based on the follow-the-regularized-leader method, which employs adaptive learning rates that depend on the empirical prediction error of the loss. This is the first BOBW algorithm with gap-variance-dependent bounds, showing that the variance information can be used even in the possibly adversarial environment. Further, the leading constant factor in our gap-variance dependent bound is only (almost) twice the value for the lower bound. In addition, the proposed algorithm enjoys multiple data-dependent regret bounds in adversarial settings and works well in stochastic settings with adversarial corruptions. Table 1 summarizes the achievable bounds in comparison with UCB-V (Audibert et al., 2007), Tsallis-INF (Zimmert and Seldin, 2021) and LB-INF (Ito, 2021).

Table 1: Achievable regret bounds. $C > 0$ is the corruption level, L^* is the cumulative loss for the optimal arm, and Q_∞ is the variation of the loss.

Environment	Bound	UCB-V	Tsallis-INF	LB-INF	**Proposed**
Stochastic	Δ-dependent	✓	✓	✓	✓
	(Δ, σ^2)-dependent	✓			✓
Adversarial	Worst case: $\tilde{O}(\sqrt{KT})$		✓	✓	✓
	1st order: $\tilde{O}(\sqrt{KL^*})$			✓	✓
	2nd order: $\tilde{O}(\sqrt{KQ_\infty})$			✓	✓
Stochastic with adversarial corruption	(Δ, C)-dependent		✓	✓	✓
	(Δ, σ^2, C)-dependent				✓

1. Extended abstract. Full version appears as [arXiv:2206.06810, v1].

Acknowledgments

SI was supported by JST, ACT-I Grant Number JPMJPR18U5, Japan. TT was supported by JST, ACT-X Grant Number JPMJAX210E, Japan and JSPS, KAKENHI Grant Number JP21J21272, Japan. JH was supported by JSPS, KAKENHI Grant Number JP21K11747, Japan.

References

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Tuning bandit algorithms in stochastic environments. In International Conference on Algorithmic Learning Theory, pages 150–165. Springer, 2007.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In Conference on Learning Theory, pages 2552–2583. PMLR, 2021.

Julian Zimmert and Yevgeny Seldin. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits. Journal of Machine Learning Research, 22(28):1–49, 2021.