A Near-optimal, Scalable and Parallelizable Framework for Stochastic Bandits Robust to Adversarial Corruptions and Beyond

Zicheng Hu, Cheng Chen

East China Normal University

Multi-armed Bandits with Adversarial Corruptions

For $t = 1, \cdots, T$ **do**

- Environment generates an i.i.d. random reward vector $\{r_{t,k}\}_{k\in[K]}$ with means $\{\mu_k\}_{k\in[K]}$
- Adversary attacks the reward vector to produce the corrupted reward vector $\{\tilde{r}_{t,k}\}_{k\in[K]}$
- Agent selects an arm I_t and only observe the corrupted reward $\tilde{r}_{t,I_t}$

Goal: Minimize the **(pseudo-)regret**

$$R(T) = \sum_{t=1}^{T} \mu_{k^*} - \sum_{t=1}^{T} \mu_{I_t} = \sum_{t=1}^{T} \Delta_{I_t}$$

where $k^* \in \max_k \mu_k$ is one of the optimal arms and $\Delta_k = \mu_{k^*} - \mu_k$ is the suboptimality gap

Extended settings

- **Batched bandits:** The time horizon T is divided into L batches, agent can only observe the corrupted rewards from a batch after it has conclude
- **d-set semi bandits:** Each round, the agent selects a combinatorial action of d distinct arms and receives component-wise feedback for the chosen arms
- **Cooperative multi-agent bandits:** V agents collaboratively play a bandit game, and share messages to accelerate learning
- **Strongly observable graph bandits:** Pulling an arm may observe rewards of other arms, where the reward-feedback structure is represented by a directed graph.

Motivation

The FTRL framework can obtain optimal regret, but faces the following limitations:

- **Hard to parallelize:** the FTRL framework is unsuitable for batched bandits and cooperative multi-agent bandits
- **Computationally costly:** FTRL requires to solve an optimization problem in each round which usually does not have closed-form solutions
- **Unique assumption:** FTRL typically assumes a unique optimal action, except for the MAB setting

We propose a **near-optimal, efficient, parallelizable** framework which do not require the assumption of unique optimal action

Our Techniques

- Our framework proceeds in epochs, for each epoch m, we chooses a data-independent epoch length $N_m = K\lambda_m 2^{2(m-1)}$, which cannot be affected by the adversary

- In each epoch, we denote k_m as the arm with the maximum empirical reward in the previous epoch. Then the number of pulls for each arm is set to

$$\widetilde{n}_k^m = \begin{cases} \lambda_m(\Delta_k^{m-1})^{-2}, & k \neq k_m, \\ N_m - \sum_{k \neq k_m} \lambda_m(\Delta_k^{m-1})^{-2}, & k = k_m. \end{cases}$$

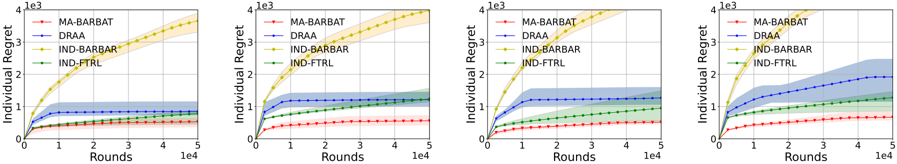
- We also adapts confidence levels across epochs to lower regret and avoid requiring the time horizon T in advance

Theoretical Results

Our regret bounds are tight up to a logarithmic factor $\log(T)$ for all settings we study except batched bandits:

Setting	Our regrets	Lower bound
MAB	$O\big(C + \sum_{\Delta_k>0} \frac{\log^2(T)}{\Delta_k}\big)$	$O\big(C + \sum_{\Delta_k>0} \frac{\log(T)}{\Delta_k}\big)$
Batched bandits	$O\big(CT^{\frac{1}{L+3}} + T^{\frac{4}{L+3}}(\sum_{\Delta_k>0} \frac{L\log(T)}{\Delta_k} + \frac{K\log(T)}{L\Delta})\big)$	$O\big(T^{\frac{1}{L}}(K + C^{1-\frac{1}{L}})\big)$
d-set semi bandits	$O\big(dC + \sum_{\Delta_k>0} \frac{\log^2(T)}{\Delta_k}\big)$	$O\big(dC + \sum_{\Delta_k>0} \frac{\log(T)}{\Delta_k}\big)$
Cooperative multi-agent bandits	$O\big(\frac{C}{V} + \sum_{\Delta_k>0} \frac{\log^2(T)}{V\Delta_k}\big)$	$O\big(\frac{C}{V} + \sum_{\Delta_k>0} \frac{\log(T)}{V\Delta_k}\big)$
Strongly observable graph bandits	$O\big(C + \sum_{k\in\mathcal{I}^*} \frac{\log^2(T)}{\Delta_k}\big)$	$O\big(C + \sum_{k\in\mathcal{I}^*} \frac{\log(T)}{\Delta_k}\big)$

Experiments



(a) K = 12, C = 2000 (b) K = 12, C = 5000 (c) K = 16, C = 2000 (d) K = 16, C = 5000

Comparison between MA-BARBAT, DRAA, IND-BARBAR and IND-FTRL in cooperative multi-agent bandits

Thank you for listening!