Distributed Learning with Bandit and Delayed Feedbacks

Jialin Yi and Milan Vojnovic

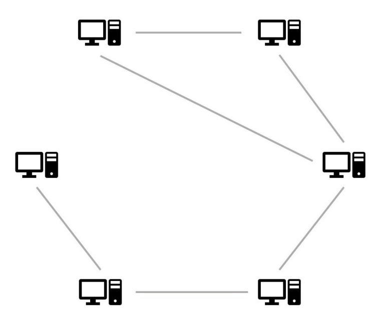
Department of Statistics, London School of Economics and Political Science

Motivation

Recommender system with geographically distributed servers

Accelerating the learning process with multiple machines

Communication model



- **communication graph** $G = (\mathcal{V}, \mathcal{E})$
- **delay**: $d \geq 0$ time steps per edge

Multi-agent multi-armed bandit

A *non-stochastic* bandit with K actions, at time step t, each agent $v \in \mathcal{V}$

- Selects an *action* $I_t(v) \in [K]$;
- Receives the *loss* $\ell_t(I_t(v))$;
- Exchanges *messages* with its neighbors $\mathcal{N}(v)$.

Learning objectives

✓ Minimize *individual regret*:

$$R_T^v = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t(I_t(v))\right] - \min_{i \in \mathcal{A}} \sum_{t=1}^{T} \ell_t(i).$$

✓ Minimize *average regret*:

$$R_T = \frac{1}{N} \sum_{v \in \mathcal{V}} R_T^v.$$

Decentralized FTRL (DFTRL)

Input: a sequence of *regularizers* $\{\psi_t\}$ and *loss estimator* $\hat{\ell}_t$.

For each time step t, each agent $v \in \mathcal{V}$

- Samples an action $I_t(v)$ from distribution p_t^v
- Computes a loss estimator $\hat{\ell}_t^{v,obs}$;
- Follows the regularized leading distribution

$$p_{t+1}^v = \arg\min_x \left\{ \left\langle \sum_{s=1}^{t} \hat{\ell}_t^{v,obs}, x \right\rangle + \psi_t(x) \right\}$$

Tsallis entropy: $\psi_t(x) = \sum_{i=1}^{K} -2\sqrt{x_i}/\eta_t$

negative entropy: $\psi_t(x) = \sum_{i=1}^{K} x_i \log(x_i)/\zeta_t$

Loss estimator

Let $q_t^v(i) = 1 - \prod_{u \in \mathcal{N}(v)} (1 - p_t^v(i))$

$$\hat{\ell}_t^{v,obs}(i) = \frac{\ell_{t-d}(i)}{q_{t-d}^v(i)} \mathbb{I}\{\exists u \in \mathcal{N}(v) : I_t(u) = i\}$$

when $t > d$ and 0 otherwise.

Regret upper bound

With a linear combination of *negative entropy* and *Tsallis entropy* regularizers, the average regret of DFTRL is

$$R_T = O\left(\left(\frac{\alpha(G)}{N}\right)^{1/4} \sqrt{KT} + \sqrt{d \log(K) T}\right)$$

where $\alpha(G)$ is the independence number of G.

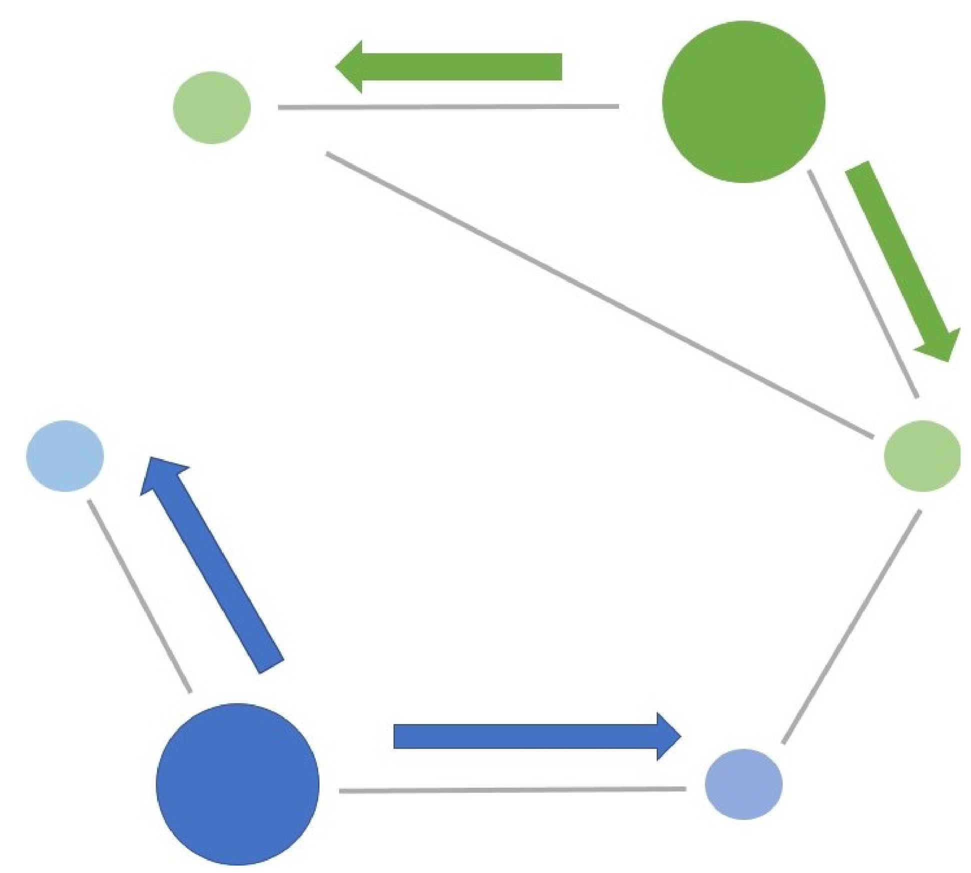
★ $O(\sqrt{d \log(K) T})$ is tight.

Center-based FTRL (CFTRL)

✓ Center agents & non-center agents.

✓ Center agent runs the DFTRL.

✓ Non-center agent copies the action distribution from its center.



Regret upper bound

With a *Tsallis entropy* regularizer, the individual regret of agent v of CFTRL is

$$R_T^v = O\left(\frac{\sqrt{KT}}{\sqrt{|\mathcal{N}(v)|}} + d \log(K) \sqrt{\max_u |\mathcal{N}(u)| T}\right).$$

Consequently, the average regret is

$$R_T = O\left(\frac{1}{N} \sum_{v \in V} \frac{1}{\sqrt{|\mathcal{N}(v)|}} \sqrt{KT} + d \log(K) \sqrt{T}\right).$$

★ $O(\sqrt{KT/|\mathcal{N}(v)|})$ in individual regret is tight.

The regret lower bound

For any learning algorithm, the worst-case individual regret of agent v is bounded as

$$R_T^v = \Omega\left(\max\left\{\frac{1}{\sqrt{|\mathcal{N}(v)|}}\sqrt{KT}, \sqrt{d \log(K) T}\right\}\right).$$

Numerical experiments

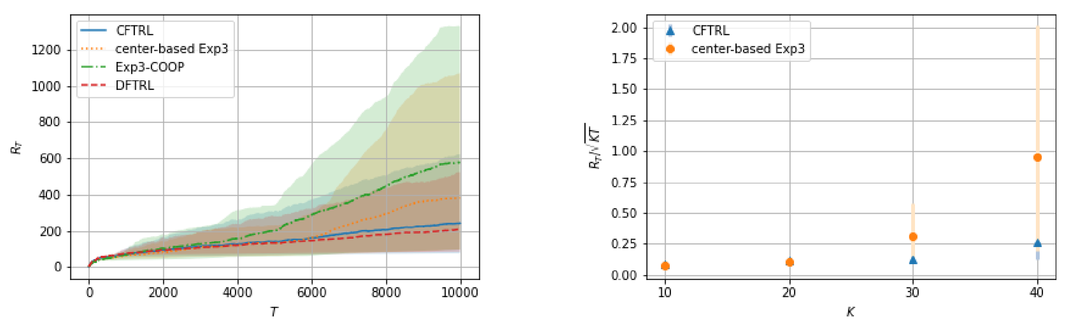


Figure 1. Regret comparisons on a 2-regular graph with and $N = 3$ agents and the delay $d = 1$.

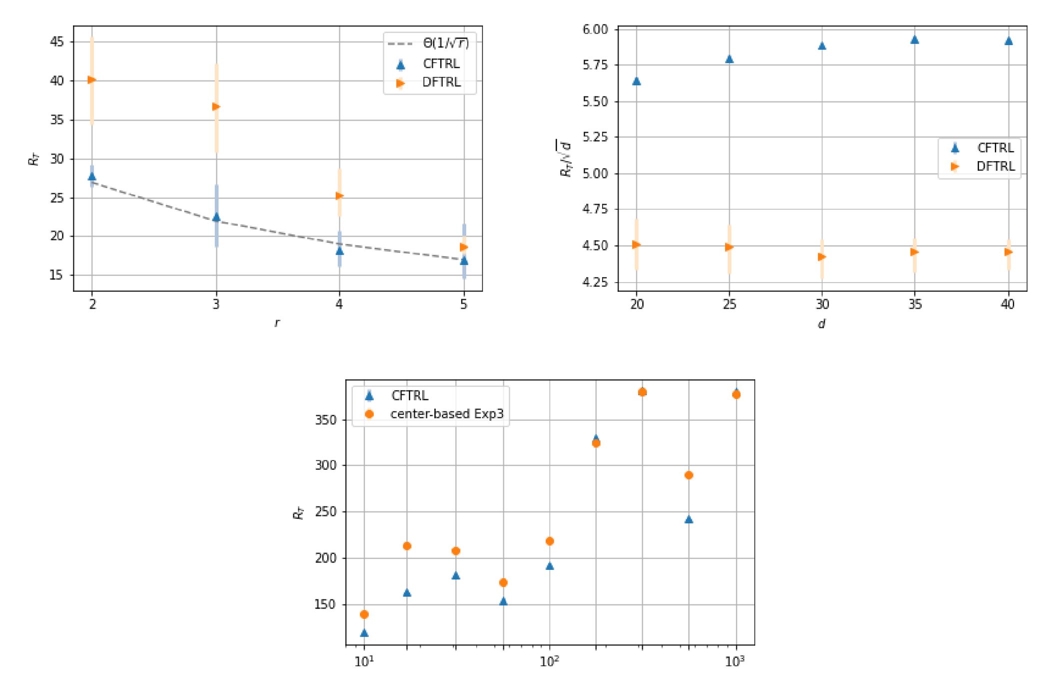


Figure 2. Regret comparison for the (normalized) average regrets on different communication networks: (upper left) r-regular graphs,, (upper right) a star graph, and (lower) Erdős–Rényi graphs.

To probe further ...

About this project

About the presenter