Multi-Agent Multi-Armed Bandits with Limited Communication

Mridul Agarwal[1,*], Vaneet Aggarwal[2], and Kamyar Azizzadenesheli[3,*]

[1]Amazon, [2]Purdue University, [3]Nvidia

[*]Work done when authors were at Purdue University

Background and Motivation

- Consider an IoT device swarm with small-scale devices deployed in different geographical locations. They can perform better if all the devices share their data. However, this data sharing is costly because of the frequency of transactions.
- Further the limited scale of the devices does not allow them relay information via multiple hops.
- Consider N workers, connected over a network with maximum degree K_G and diameter D, interacting with N i.i.d. K armed bandit environments.
- We ask, is there a way to reduce communication requirements and still achieve similar regret bounds.

Existing Algorithms and Learnings

- For single agent, or N =1, UCB algorithm(s) [1] achieves a regret bound of $\tilde{O}\left(\sqrt{KT}\right)$ and finds a good arm w.h.p.
- For $N > 1$, gossiping style algorithms [2] divide K arms among the N agents.
  - The agents identify their best arm and then communicate the arm index to others after epochs doubling in duration.
  - Other agents include this recommendation in their arm set and restart their bandit algorithm.

[1] Bubeck, et al. "Pure exploration in finitely-armed and continuous-armed bandits." *TCS* (2012).
[2] Chawla, et al. "The gossiping insert-eliminate algorithm for multi-agent bandits " *AISTATS* 2020.

Key Difficulties and Ideas

- The agents may wait for too long to identify the best arm with among the arms they are playing.
- The agents have guarantees about which arm are *good* or *bad* after every epoch.
- Once the agent with the best arm broadcast the best arm index, it may take multiple iterations for the all the agents to listen to it because of no-relay constraint.
- To ensure that the knowledge about the good arm propagates through the entire graph of diameter D, divide doubling length epochs into D sub-epochs of equal duration.
- One of the received arm after every sub-epoch is at most $\tilde{O}\left(\sqrt{D/T_j}\right)$ bad, where T_j is the duration of epoch j. Also, the regret of each sub-epoch is bounded by $\tilde{O}\left(\sqrt{DT_j}\right)$. Summation regret over all (sub-)epochs can still give $\tilde{O}\left(\sqrt{T}\right)$ guarantee.

LCC-UCB-GRAPH Algorithm

- N agents create sets by dividing K arms into $[\frac{N}{K}]$ sized sets and recommendations received from neighbors.
- Each agent interacts with the bandit environment with the arms they have and recommend the best arm to neighbors.
- Communicate after every $2^j/D$ time-steps and increment j after every 2^j time-steps.

Algorithm 3 LCC-UCB-GRAPH($\mathcal{S}_n, G, T_0, T$)

$t = 0, j = 0$
$\mathcal{R}_{n,1,0} = \emptyset$
for $t < T$ **do**
  $d = 1$
  for $d \leq D$ **do**
    Set augmented set $\mathcal{A}_{n,d,j} = \mathcal{S}_n \cup \mathcal{R}_{n,d,j}$
    $i^* = \text{UCB}(\mathcal{A}_{n,d,j}, \min(T - t, K'(K'+1)2^j))$
    $t = t + K'(K'+1)2^j$
    Send i^* to neighbors
    Receive most played arms of neighbors as $\mathcal{R}_{n,d,j}$
    $d = d + 1$
  end for
  $j = j + 1$
end for

Analysis - I

- N agents are connected with a network graph of diameter D and maximum degree K_G.
- Each agent receives K/N arms initially and at most K_G recommended arms from each each neighbor.
- At the end of each epoch, each agent is aware of, an arm which is at least $\Delta_j = D\sqrt{K'/T_{j-1}}$ close to the optimal arm.
- Regret analysis follows:
  - Regret from not playing the Δ_j-optimal arm in the entire epoch
  - Regret resulting from the imperfect ($\Delta_j \geq 0$) knowledge of the optimal arm
  - Summing over all the epochs.

Analysis - II

- Theorem [3]: The regret of any agent following the LCC-UCB-GRAPH algorithm is upper bounded by

$$\tilde{O}\left(D\sqrt{DK'T}\right), K' = (K/N + K_G)$$

- Theorem [3]: The number of bits exchanged are upper bounded by

$$\tilde{O}(K_G D \log K \log T)$$

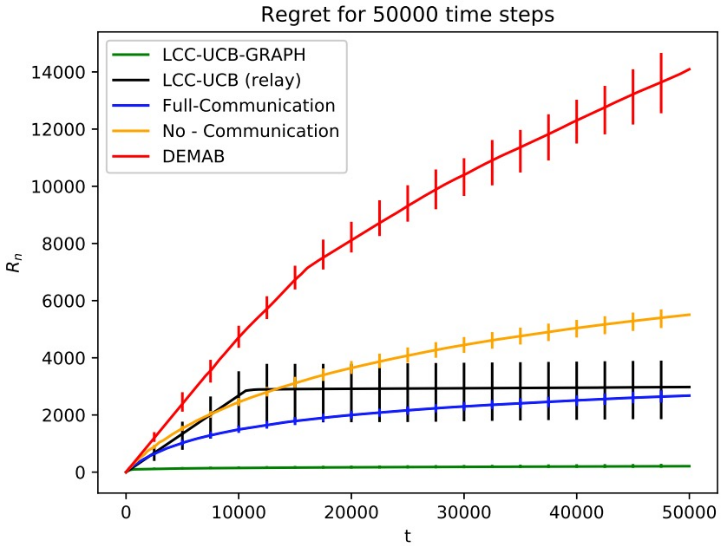
- Corollary: For a fully connected graph with $D = 1, K_G = N$, the regret follows:

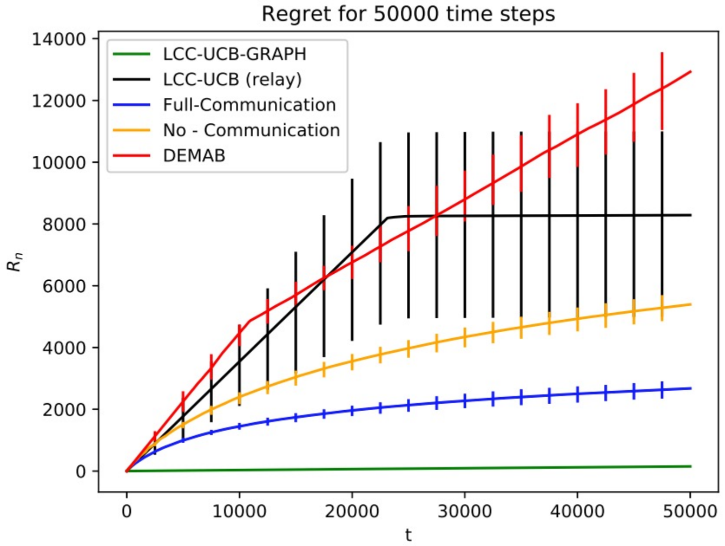
$$\tilde{O}\left(\sqrt{(N + K/N)T}\right)$$

[3] **Agarwal**, et al. "Multi-Agent Multi-Armed Bandits with Limited Communication" JMLR (21-138).

Empirical Analysis - I

- We evaluated the proposed LCC-UCB algorithm on sparse graphs. We considered (N,K) = (100, 250) and (150, 250).



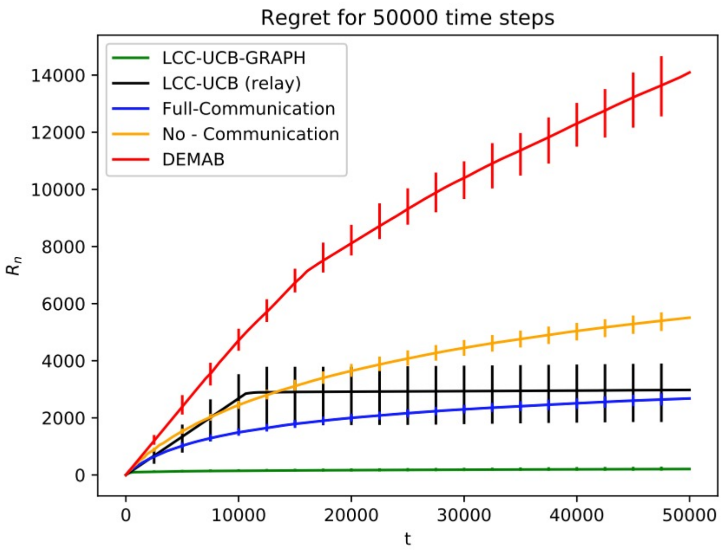


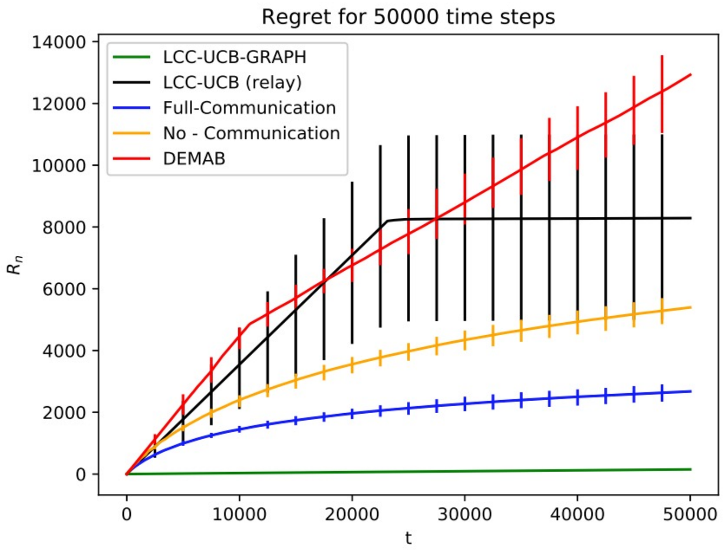
- We first note that LCC-UCB-GRAPH performs better than full communication strategy where agents communicate every time step. This is because the sparsity of graph does not allow efficient communication.

[4] Wang, et al., “Distributed bandit learning: Near-optimal regret with efficient communication.” , ICLR 2019

Empirical Analysis - II

- We evaluated the proposed LCC-UCB algorithm on sparse graphs. We considered (N,K) = (100, 250) (left Figure) and (150, 250) (right Figure).





- We then note that a relay based algorithm does not perform good as the number of agents increase as the number of arms K' available with an agent becomes $K/N + N$ instead of $K/N + K_G$

Summary:

- We consider a problem of multi-agent multi-armed bandits
- The agent are connected over a network with diameter D and maximum degree K_G
- Agents have limited computation resources and can only communicate limited bits
- Following LCC-UCB-GRAPH protocol, agents can
  - Achieve regret of $\tilde{O}\left(D\sqrt{DK'T}\right), K' = (K/N + K_G)$
  - By only communicating $\tilde{O}\left(\sqrt{(N + K/N\ )T}\right)$ bits