Collaborative Best Arm Identification in Multi-armed Bandits

Amit Anand Jha
IIT Madras

Nazal Mohamed
IIT Madras

Krishna Jagannathan
IIT Madras

***Abstract*—The paper [1] addresses collaborative best arm identification (BAI) in a fixed-budget multi-armed bandit (MAB) setting, where agents share information about selected arms and rewards with their one-hop neighbors. Focusing on star topology, the paper proposes two algorithms: UCB-Star, an exploratory variant of Upper Confidence Bound (UCB), and Follow Your Leader - Star (FYL-Star). These methods are extended to general networks using dominating set partitions (DSP), leading to UCB-DSP and FYL-DSP. Theoretical and empirical results show exponential decay in error probability with increasing budget and agent count.**

I. Introduction

The paper [1] addresses collaborative BAI under a fixed budget, introducing agent collaboration to the standard BAI framework. The setup involves a network of agents connected by a graph where nodes represent agents and edges their communication paths. When an agent plays an arm, both the arm index and obtained reward are shared with its one-hop neighbors, allowing them to use this information in planning their own actions.

This paper studies the impact of such network structure on identifying the best arm constrained within a fixed budget and proposes new algorithms that leverage collaboration between agents. Previous work includes [2] on collaborative regret minimization using the same mode of collaboration and [3] focus on stationary MAB in a fully connected graph assuming different reward distribution on each arm for different agents.

The paper focuses on star topology as a representative structure in social networks and proposes two algorithms: UCB-Star, an exploratory variant of UCB-E, and FYL-Star. For more general networks, they extend these algorithms to UCB-DSP and FYL-DSP using a dominating set partition (DSP). An ensemble learner (EL) aggregates samples from all partitions at the end of the budget to identify the best arm. Exponential decay in error probability with budget and number of agents is shown from theoretical analysis and empirical results.

II. Problem Setup

The model in this work is a stateless, stationary and stochastic MAB setup with K arms. Arm i has an unknown reward distribution P_i with mean μ_i. These distributions are bounded within $[0, 1]$ and rewards from different arms are independent. The BAI problem is examined under a fixed budget setting, where $n > K$ rounds are available to identify the arm with the highest mean, i.e., $i^* = \underset{i\in\{1,2,\ldots,K\}}{\operatorname{argmax}} \{\mu_i\}$, whose mean is μ^*. The sub-optimality gap for any sub-optimal arm i is defined as $\Delta_i = \mu^* - \mu_i$.

This paper considers m agents connected through a graph $G(V, E)$, where V is the set of nodes representing agents and E represents their connections. In each round $t \in \{1, 2, \ldots, n\}$, each agent $\nu \in \{1, 2, \ldots, m\}$ plays an arm $i \in \{1, 2, \ldots, K\}$, obtaining a reward $z^{\nu}_{i,t}$ from the arm's distribution P_i. Each agent's action A^{ν}_t and reward $z^{\nu}_{i,t}$ are immediately observed by itself and its one-hop neighbors, denoted $\mathcal{N}(\nu)$, which includes the agent itself. At any time t, an agent ν thus has access to all actions and rewards of agents in $\mathcal{N}(\nu)$ up to $(t-1)$. A policy π^{ν}_t determines the arm selection for agent ν based on available information. The network policy π^G includes policies from all agents.

Notation wise, $r^{\nu}_{i,t}$ denotes the total reward for arm i from all samples accessible to agent ν till time t. $s^{\nu}_{i,t}$ denotes the number of times agent ν plays the arm i till time t. The total number of samples from arm i available with agent ν at time t is denoted as $c^{\nu}_{i,t}$, and empirical mean for a particular agent ν from arm i till time t is denoted by $\mu^{\nu}_{i,t}$, i.e.,

$$r^{\nu}_{i,t} = \sum_{t'=1}^{t} \sum_{u\in\mathcal{N}(\nu)} z^{u}_{i,t'}, \quad c^{\nu}_{i,t} = \sum_{u\in\mathcal{N}(\nu)} s^{u}_{i,t}, \quad \mu^{\nu}_{i,t} = \frac{r^{\nu}_{i,t}}{c^{\nu}_{i,t}}.$$

The task is to estimate the best arm at the end of the budget n. Since each agent may have its perception of the best arm we need an 'aggregator' of some sort to declare the best arm estimate. An ensemble learner (EL) performs this job, adopting an aggregation policy denoted by π^{EL}. The aim is to devise policies for the network and EL such that the probability of error in the estimation of the best arm by the end of the budget n is minimised. If $\hat{i}_{\text{EL},n}$ is the estimated best arm given by the EL at the end of the budget n and δ be the probability that EL chooses a sub-optimal arm as the best arm. Then, the aim is to

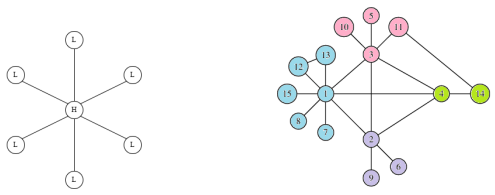
$$\min_{\pi^{\text{EL}},\pi^G} \delta = \min_{\pi^{\text{EL}},\pi^G} \mathbb{P}\Big(\hat{i}_{\text{EL},n} \neq i^*\Big).$$

The hardness of BAI in MAB problem is determined by two parameters H_1 and H_2 as:

$$H_1 = \sum_{i\neq i^*} \frac{1}{\Delta_i^2}, \quad H_2 = \max_i \frac{i}{\Delta^2_{<i>}},$$

where $\Delta_{<i>}$ denotes i^{th} element in ascending order of sub-optimality gaps, i.e., $\Delta_{<1>} \leq \Delta_{<2>} \leq \ldots \Delta_{<K>}$.

Report prepared by Srikar Babu Gadipudi (EE21B138) and Shreya .S. Ramanujam (EE21B126) as a part of the CS6046: Multi-armed Bandits course.



(a) Star Network (b) General Network

Fig. 1: Two networks discussed in the paper: a) a 6-node star network, where hub nodes are denoted by H and leaf nodes are denoted by L and b) generic network, where 1, 2, 3, 4 act as hub nodes for partitioned star networks.

III. Policies for Star Networks

In this section, we focus on collaborative BAI in a m-node star network. An m-node star network is one in which a hub agent (say ν=1) is connected to (m-1) leaf agents, as shown in Figure 1a. The star topology serves as a building block for more general network topologies discussed in the next section. The paper proposes two policies for star networks, discussed below.

1) UCB-Star: This is a highly exploratory policy in which all nodes (hub and leaf) explore independently using samples from their own history and from their one-hop neighbours, as shown in Algorithm 1. At the end of budget n, the hub node estimates the best arm, since it receives samples of all agents.

In round t, an agent ν plays the arm i for which the metric $B^{\nu}_{i,t}$ maximum, where

$$B^{\nu}_{i,t} := \hat{x}^{\nu}_{i,(t-1)} + \sqrt{\frac{a}{c^{\nu}_{i,(t-1)}}} \tag{1}$$

Here, $a > 0$ is the *exploration parameter* set by the user. The paper claims it can be shown that the upper bound for probability of error in BAI exponentially decays in both budget n and number of agents m.

Algorithm 1 UCB-Star

Exploration parameter: $a > 0$ Hub Node: $\nu = 1$

for $1 \le t \le K$ **do**
  $A^{\nu}_t = t \quad \forall \nu \in \{1, 2, 3, \ldots, m\}$
end for
for $K < t \le n$ **do**
  $A^{\nu}_t = \underset{i \in \{1,2,\ldots,K\}}{\text{argmax}} \; B^{\nu}_{i,t}$
end for
Best Arm for Star Network G is $A^G_n = \underset{i \in \{1,2,\ldots,K\}}{\text{argmax}} \; \hat{\mu}^1_{i,n}$

2) FYL-Star: Here, all leaf nodes follow the actions of the hub node after a particular number of rounds, as shown in Algorithm 2. The hub node plays using $B^{\nu}_{i,t}$ [1] in all rounds. This policy relies heavily on the hub and does not allow leaf nodes to explore independently. An upper bound for the policy is stated below (refer to the paper for proof).

Theorem 1. *The probability of error in estimating the best arm for an m-node star network $G = (V, E)$ using FYL-Star policy is upper bounded as*

$$\mathbb{P}(\mathcal{E}) \le 2K\,(m(n-K)+1)\,e^{-\frac{2a}{25}}, \tag{2}$$

for some positive constant a such that, $a \le \frac{25m(n-K)}{36H_1}$.

Algorithm 2 FYL-Star

Exploration parameter: $a > 0$ Hub Node: $\nu = 1$

for $1 \le t \le K$ **do**
  $A^{\nu}_t = t \quad \forall \nu \in \{1, 2, 3, \ldots, m\}$
end for
for $K + 1 \le t \le n$ **do**
  $A^1_t = \underset{i \in \{1,2,\ldots,K\}}{\text{argmax}} \; B^1_{i,t}$
  if $t = K + 1$ **then**
    $A^{\nu}_t = \text{random}(1, 2, \ldots, K), \quad \forall \nu \in \{2, 3, \ldots, m\}$
  else
    $A^{\nu}_t = A^1_{t-1}, \quad \forall \nu \in \{2, 3, \ldots, m\}$
  end if
end for
Best Arm for Network G is $A^G_n = \underset{i \in \{1,2,\ldots,K\}}{\text{argmax}} \; \hat{\mu}^1_{i,n}$

IV. Policies for General Networks

Though star networks are prevalent, most real world networks are more complex. This paper uses DSP to partition any generic network into a set of star networks and employs the above-mentioned algorithms for each partition. We first recall the following definitions:

Dominating Set: A dominating set D of a network $G = (V, E)$ is a subset of V such that every node in $V \backslash D$ is connected to at least one of the nodes in D.

Dominating Set Partition: Let D be a dominating set of G. A dominating set partition based on D is obtained by partitioning V into $|D|$ components such that each component contains a node in D and a subset of its one-hop neighbour.

We present the proposed algorithms for general networks:

1) UCB-DSP: This policy extends UCB-Star to general networks by partitioning the network into star structures based on a dominating set, then running UCB-Star on each partition, as shown in Algorithm 3. During the first K rounds, each agent explores by sampling each arm once; thereafter, they follow Equation 3 until the budget is exhausted. The EL aggregates rewards and play counts from each hub node to estimate the best arm. The upper bound on estimation error remains as in Theorem 2.

$$P^{\nu_p}_{i,t} := \hat{\chi}^{\nu_p}_{i,(t-1)} + \sqrt{\frac{a}{\beta^{\nu_p}_{i,(t-1)}}} \tag{3}$$

where $\hat{\chi}^{\nu^p}_{i,t}$ is the sample mean of all the samples of arm i till time t accessible to agent ν_p in partition S_p and $\beta^{\nu^p}_{i,t}$ is the number of times arm i is played by agents in $\mathcal{N}'(\nu_p)$.

Algorithm 3 UCB-DSP

Exploration parameter: $a > 0$
Run a subroutine on $G(V, E)$ to obtain a Dominating Set D and partitions $\{S_1, S_2, \ldots, S_{|D|}\}$
for $1 \leq t \leq K$ **do**
  $A_t^{\nu} = t \quad \forall \nu \in \{1, 2, \ldots, m\}$
end for
for $K+1 \leq t \leq n$ **do**
  for each $S_p \in \{S_1, S_2, \ldots, S_{|D|}\}$ **do**
    $A_t^{\nu_p} = \underset{i \in \{1,2,\ldots,K\}}{\text{argmax}} \; P_{i,t}^{\nu_p}$
  end for
end for
Best Arm estimate by EL is

$$\hat{i}_{\text{EL},n} = \underset{i \in \{1,2,\ldots,K\}}{\text{argmax}} \frac{\sum_{p=1}^{|D|} \hat{\chi}_{i,n}^{h_p} \cdot \beta_{i,n}^{h_p}}{\sum_{p=1}^{|D|} \beta_{i,n}^{h_p}}$$

2) FYL-DSP: Similar to the above extension, FYL-DSP is an extension FYL-Star. The network is partitioned using DSP and each star partition plays using FYL-Star policy, as shown in Algorithm 4. In this policy, the EL ensembles the samples and the number of times each arm is played from the hub nodes of each star partition to estimate the best arm. The upper bound for the probability of error in estimation of the best arm is given by Theorem 2 (refer to the paper for proof).

Theorem 2. *The probability of error in estimating the best arm for an m-node star network $G = (V, E)$, by the EL using the FYL-DSP policy, is upper bounded as:*

$$\mathbb{P}(\mathcal{E}) \leq 2K \left(m(n-K)+1\right) e^{-2b\lambda^2}, \tag{4}$$

for exploration parameter b and λ such that $b\lambda^2 \leq \frac{m(n-1)}{4H_1}$.

V. Empirical Analysis

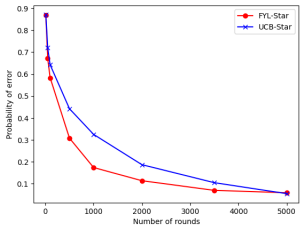
Following the paper, we simulate the UCB-Star and FYL-Star algorithms (Algorithms 1 and 2) using a network with 15 agents and 10 Bernoulli-distributed arms, where $\{\mu_1 = 0.20, \mu_2 = 0.21, \ldots, \mu_{10} = 0.29\}$. We vary the number of rounds and compute success probabilities over 1000 simulations, setting the exploration parameter $a = 1$. As shown in Figure 2a, FYL-Star performs better than UCB-Star, with both exhibiting exponential behavior in line with theoretical bounds.

Figure 2b compares UCB-DSP and FYL-DSP in a general network with the same configuration. We simulate on partitions based on different dominating sets to observe the effect of partitioning on policy performance. The paper shows UCB-DSP performance varies with partitioning, while FYL-DSP remains stable. They also discuss a UCB-DSP variant where samples of neighboring nodes from different partitions are included. Although theoretical bounds are not proven due to high complexity, empirical results show an exponential decay in this variant as well.

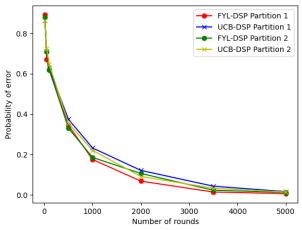
Algorithm 4 FYL-DSP

Exploration parameter: $a > 0$
Run a subroutine on $G(V, E)$ to obtain a Dominating Set D and partitions $\{S_1, S_2, \ldots, S_{|D|}\}$
for $1 \leq t \leq K$ **do**
  $A_t^{\nu} = t \quad \forall \nu \in \{1, 2, \ldots, m\}$
end for
for $K+1 \leq t \leq n$ **do**
  for each $S_p \in \{S_1, S_2, \ldots, S_{|D|}\}$ **do**
    $A_t^{h_p} = \underset{i \in \{1,2,\ldots,K\}}{\text{argmax}} \; P_{i,t}^{h_p}$
    if $t = K+1$ **then**
      $A_t^{\nu_p} = \text{random}(1, 2, \ldots, K) \quad \forall \nu_p \in S_p \setminus \{h_p\}$
    else
      $A_t^{\nu_p} = A_{t-1}^{h_p} \quad \forall \nu_p \in S_p \setminus \{h_p\}$
    end if
  end for
end for
Best Arm estimate by EL is

$$\hat{i}_{\text{EL},n} = \underset{i \in \{1,2,\ldots,K\}}{\text{argmax}} \frac{\sum_{p=1}^{|D|} \hat{\chi}_{i,n}^{h_p} \cdot \beta_{i,n}^{h_p}}{\sum_{p=1}^{|D|} \beta_{i,n}^{h_p}}$$



(a) Performance on Star Network



(b) Performance on General Network

Fig. 2: 15 agent star network - a) UCB-Star and FYL-Star b) UCB-DSP and FYL-DSP with 2 partitions, detailed in paper.

VI. Conclusion

The paper presented collaborative algorithms for BAI in MAB problems. The proposed algorithms—UCB-Star, FYL-Star, and their generalized DSP variants leverage network structures to minimize error probability within a fixed budget. Theoretical upper bounds on error probability are shown to be exponentially decaying with number of rounds and number of agents. Experimental results confirm theoretical bounds, with FYL-DSP proving stable across various network partitions.

References

[1] A. A. Jha, N. Mohamed, and K. Jagannathan, "Collaborative best arm identification in multi-armed bandits," in *2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 2022, pp. 335–343.

[2] R. K. Kolla, K. Jagannathan, and A. Gopalan, "Collaborative learning of stochastic bandits over a social network," *IEEE/ACM Transactions on Networking*, vol. 26, no. 4, pp. 1782–1795, 2018.

[3] S. Shahrampour, A. Rakhlin, and A. Jadbabaie, "Multi-armed bandits in multi-agent networks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2786–2790.