Bayesian Algorithms for Decentralized Stochastic Bandits

Anusha Lalitha[1] and Andrea Goldsmith[1,2]

[1]Stanford University, [2]Princeton University

Abstract

We consider a network of N agents playing the same instance of a K-armed Multi-Armed Bandit (MAB) problem. The goal is to minimize cumulative regret averaged over the entire network. We propose a decentralized Bayesian multi-armed bandit framework that extends single-agent Bayesian bandit algorithms to the decentralized setting. Using this, we propose a decentralized Thompson Sampling algorithm and a decentralized Bayes-UCB algorithm. We analyze the decentralized Thompson Sampling algorithm under Bernoulli rewards and establish a problem-dependent upper bound on the cumulative regret. We show that regret incurred scales logarithmically over the time horizon with constants that match those of an optimal centralized agent.

Background and Motivation

Consider a multi-agent MAB problem with N agents connected through an undirected graph $\mathcal{G}$.

- Agent i sequentially chooses arms $\{A_t^{(i)}\}_{t\geq 1}$ from a finite set of arms $\{1,\dots,K\}$.
- When arm k is played, agent receives a reward $Y_t^{(i)} \sim p_{\theta_k^*}$ with mean $\mu_k := \mathbb{E}[Y_t^{(i)} \mid A_t^{(i)} = k]$.
- True underlying reward parameters $\theta^* = [\theta_1^*, \dots, \theta_K^*]$ and expected values of rewards are unknown to the agents.

We assume without loss of generality that $\mu_1 \geq \mu_2 \geq \dots \mu_K$. Agents aim to minimize the per-agent regret over the network

$$R(T) := \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\theta^*}\left[\sum_{t=1}^{T}\left(Y_{t,A_1}^{(i)} - Y_{t,A_t^{(i)}}^{(i)}\right)\right],$$

while exchanging at most $\mathsf{poly}(K)$ messages with their neighbors per iteration.

Two extremes: If the agents do not interact at all, then $R(T) \leq O(\frac{K}{\Delta}\log T)$. If we assume perfect collaboration, then $R(T) \leq O(\frac{K}{N\Delta}\log NT) = O(\frac{K}{N\Delta}\log T + \frac{\log N}{N\Delta})$. We aim to design a strategy which incurs a per agent regret close to the one incurred by the optimal centralized algorithm.

Decentralized Bayesian Multi-agent MAB Algorithms

We assume that agents take a Bayesian approach. Specifically, each agent i starts with a prior distribution on θ_k^* for each k and at every time t maintains a posterior distribution $q_{k,t}^{(i)}(\cdot)$ for each k. Locally, agents are selects arms via Bayesian single-agent algorithm: `Bayes_MAB`.

Algorithm 1: Decentralized Bayesian Multi-agent MAB Algorithm

Input: initial prior q_0, learning rate $\eta > 0$, communication matrix W

Initialize: $q_{k,1}^{(i)} \leftarrow q_0$ for all $i \in [N]$ and $k \in [K]$

for $t = 1, 2, \dots$ **do**

  for $i = 1, \dots, N$ **do**

    Select arm $A_t^{(i)} \leftarrow$ `Bayes_MAB`$\left(q_{1,t}^{(i)}, \dots, q_{K,t}^{(i)}\right)$

    Play $A_t^{(i)}$ and Observe $Y_t^{(i)} \sim p_{\theta^*_{A_t^{(i)}}}$

    Update posterior distribution:

$$\tilde{q}_{A_t^{(i)},t+1}^{(i)}(\theta) \leftarrow \frac{q_{A_t^{(i)},t}^{(i)}(\theta)\, p_\theta^\eta\left(Y_t^{(i)}\right)}{\int_{\phi\in\Theta} q_{A_t^{(i)},t}^{(i)}(\phi)\, p_\phi^\eta\left(Y_t^{(i)}\right) d\mu(\phi)}, \ \forall \theta \in \Theta$$

    Send messages $\{\tilde{q}_{k,t+1}^{(i)}\}_{k\in[K]}$ to all $j \in \mathcal{N}(i)$

  for $i = 1, \dots, N$ **do**

   for $k = 1, \dots, K$ **do**

    Merge posteriors for all $\theta \in \Theta$:

$$q_{k,t+1}^{(i)}(\theta) \leftarrow \frac{\exp\left(\sum_{j=1}^N W_{ij}\log \tilde{q}_{k,t+1}^{(j)}(\theta)\right)}{\int_{\phi\in\Theta}\exp\left(\sum_{j=1}^N W_{ij}\log\tilde{q}_{k,t+1}^{(j)}(\phi)\right)d\mu(\phi)}$$

We obtain a decentralized Thompson Sampling and a decentralized Bayes-UCB algorithm by substituting `Bayes_MAB` in the decentralized Bayesian MAB algorithm (Alg 1) with Thompson Sampling (Alg 2) and Bayes-UCB (Alg 3) respectively.

Algorithm 2: Thompson Sampling

Input: $q_{k,t}^{(i)}$, $\forall k \in \{1,\dots,K\}$

for $k = 1, \dots, K$ **do**

  Sample $\theta_{k,t}^{(i)} \sim q_{k,t}^{(i)}$

Select arm $A_t^{(i)} \leftarrow \arg\max_k \mathbb{E}_{\theta_{k,t}^{(i)}}[Y_t^{(i)}]$

return Arm $A_t^{(i)}$

Algorithm 3: Bayes-UCB

Input: $q_{k,t}^{(i)}$, $\forall k \in \{1,\dots,K\}$, time horizon T, parameters of the quantile c

Definition: Denote $\{\rho_{k,t}^{(i)}\}_{k\in[K]}$ as posterior over means $[\mu_1, \dots, \mu_K]$

Denote `Quantile`(κ, ρ) as quantile function associated with distribution ρ such that $\mathsf{P}_\rho(X \leq \texttt{Quantile}(\kappa,\rho)) = \kappa$

for $k = 1, \dots, K$ **do**

  Compute:

  $C_k^{(i)}(t) \leftarrow \texttt{Quantile}\left(1 - \frac{1}{t(\log T)^c}, \rho_{k,t}^{(i)}\right)$

Select arm $A_t^{(i)} \leftarrow \arg\max_k C_k^{(i)}(t)$

return Arm $A_t^{(i)}$

Regret Analysis for Decentralized Thompson Sampling

Theorem 1. Consider the decentralized multi-armed bandit problem with N agents, K arms and Bernoulli rewards. Let W be a doubly stochastic communication matrix. For any $\epsilon > 0$, choosing $\eta = N$, and prior as Beta$(1,1)$, i.e., uniform distribution, the per-agent cumulative regret incurred by decentralized Thompson Sampling (dec-TS) after T rounds of play can be upper bounded as

$$R(T) \leq \sum_{k=2}^{K}\Delta_k(1+\epsilon)^2\frac{\log NT}{Nd(\mu_k,\mu_1)} + \frac{3\left(1+\frac{8}{\epsilon}\right)\log N}{1-\lambda_2(W)}\sum_{k=2}^{K}\Delta_k + O\left(\frac{1}{\epsilon^{\tilde{N}}}\right),$$

where $d(a,b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$ denotes the KL-divergence between two Bernoulli distributions, $\lambda_2(W)$ denotes the second largest eigenvalue of matrix W in absolute value and $\tilde{N} = \frac{N\log N}{1-\lambda_2(W)}$. Asymptotically, the per-agent regret incurred by dec-TS scales logarithmically with the time horizon T which satisfies

$$\lim_{T\to\infty}\frac{R(T)}{\log T} \leq \sum_{k=2}^{K}\frac{\Delta_k}{Nd(\mu_k,\mu_1)}.$$

Empirical Results

We compare proposed algorithms with prior work: coop-UCB algorithm [1], and DDUCB algorithm [2]

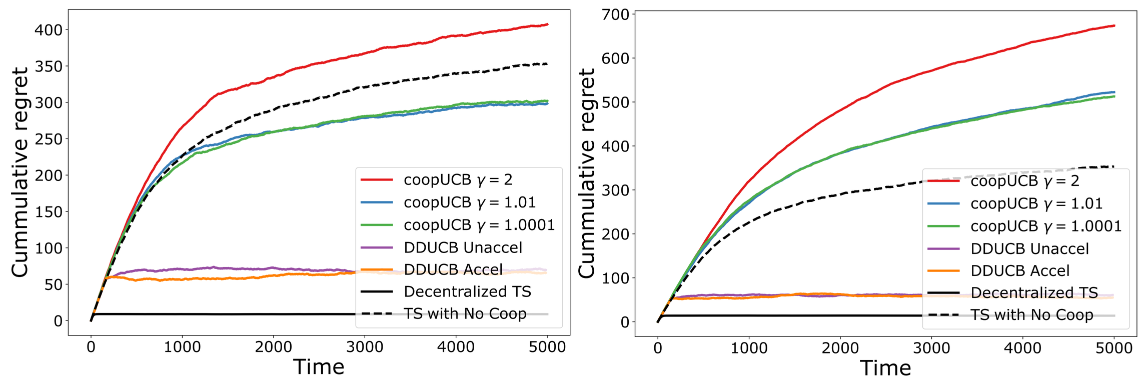

Figure 1: Per-agent regret for a network of 100 agents with cycle topology (left) and grid topology (right). Agents have 17 Gaussian arms with means $\{0.5, 0.1, \dots, 0.1\}$ and variance $\sigma^2 = 1$.

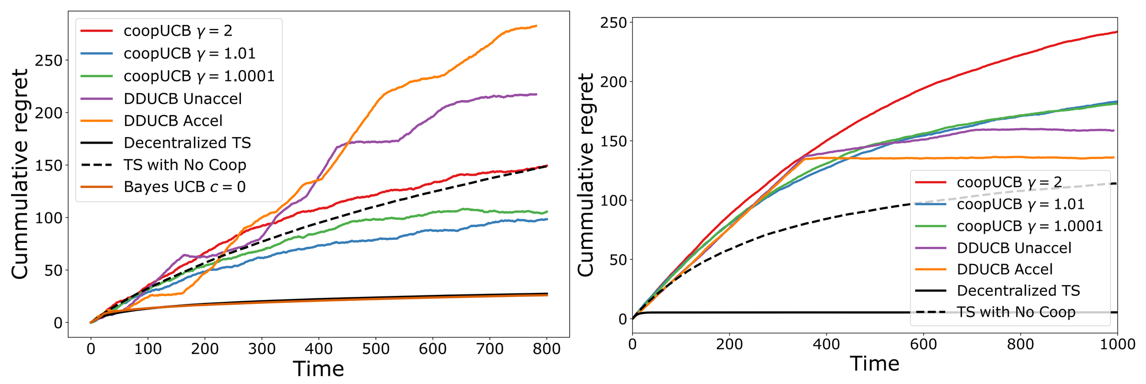

Figure 2: Per-agent regret for a network with cycle topology with 20 agents with 20 arms (left) and 200 agents with 10 arms (right).

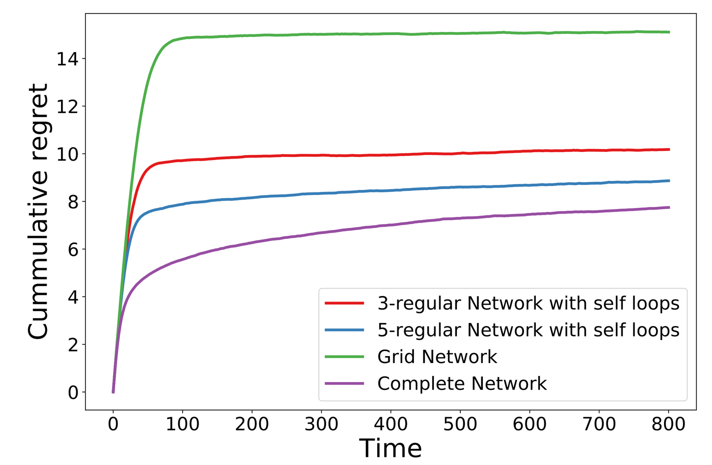

Figure 3: Per-agent cumulative regret over time for 64 agents with 17 Bernoulli arms with mean $\{0.5, 0.1, \dots, 0.1\}$ for varying network topology.

References

[1] P. Landgren, V. Srivastava, and N. E. Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *IEEE CDC*, 2016.

[2] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic multi-armed bandits. *CoRR*, abs/1810.04468, 2018.