Wireless Parameter Tuning with Clustering Multi-Agents in Linear Stochastic Bandit

Hamza Cherkaoui
Noark's Ark Paris
Huawei Technologies France
Boulogne-Billancourt, 92100, France
hamza.cherkaoui@huawei.com

Merwan Barlier
Noark's Ark Paris
Huawei Technologies France
Boulogne-Billancourt, 92100, France
merwan.barlier@huawei.com

Igor Colin
Noark's Ark Paris
Huawei Technologies France
Boulogne-Billancourt, 92100, France
igor.colin@huawei.com

Abstract

The constant demand of massive data exchange brings our actual networks on theirs limits. With the aim to optimally configure a wireless network, the collaboration of antennas with similar behavior can drastically improves their performance. With the formalism of multi-agents linear stochastic bandit, we propose a novel clustering approach based on an overlapping ellipsoids test coupled with a bipartite graph labeling strategy to regroup similar antennas and improve their performance. Our approach proposes a good recovery of the underlying antennas similarity graph which improve the overall performance of the network. We demonstrate the potential of this technique with experiments on synthetic and real data. We also discuss the impact of the agent policy on the clustering quality.

1 Introduction

Context The increasing demand of massive data exchanges brings our actual networks on theirs limits. Considering a network of wireless antennas, a promising direction of improvement would be to allow them to automatically set theirs parameters. An efficient approach would be to allow collaboration between antennas. The network will then benefit from a wider exploration of parameters settings. However, the collaboration needs to take place when the group seek to find the –same– optimal set of parameters. Thus, a challenging difficulty is to group the antennas adequately. In this paper we study the clustering of collaborative agents in the exploration-exploitation dilemma. We formalize our contribution by studying the problem of clustering multi-agents linear stochastic bandit. Linear stochastic bandit is decision making problem where, at each iteration, each agent needs to choose an action within a set and receive a stochastic reward. The objective is to maximize the expected value of the cumulative reward.

Related work The linear stochastic bandit has a well developed literature. The setting was first introduced by Abe and Long [1999], Auer [2002]. The corresponding algorithms could be divided in two parts: the cumulative reward maximization ones [Li et al., 2010, Abbasi-Yadkori et al., 2011] and the pure exploration setting for which the most rewarded action need to be estimated regardless of the expenses during this exploration phase [Hoffman et al., 2014, Xu et al., 2018]. The multi-agents setting was widely study the stochastic linear bandit [Ban and He, 2013, Gentile et al., 2014].

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Recently, the clustering of agents regained in interest [Gentile et al., 2014, Nguyen and Lauw, 2014, Ban and He, 2021, Moradipari et al., 2022]. However, for each of theses contributions, the criterion for clustering the agents (i.e. the antennas) does not take into account the exact confidence ellipsoid toward their own unknown bandit parameter (i.e. their own optimal antenna settings).

2 Setting

Modeling We consider a set of K vectors $(x_k)_{k=1}^K$ in $\mathbb{R}^d$ modeling the set of possible parametrization of the antenna. The N antennas are divided in –two clusters– labelled with $l \in \{0, 1\}$, we denoted by l_i the label for each antenna. Each cluster has an unknown bandit parameter θ_l (i.e. the shared optimal parameter setting). At each iteration t, each antenna i chooses an arms $k_{i,t}$, following a policy π, and observe the reward $r_{i,t} = \theta_l^\top x_{k_{i,t}} + \eta_{i,t}$ (i.e. the effect of the communication quality). We suppose that, for each antenna i, $(\eta_t)_{t=1}^T$ are conditionally R-subgaussian. The objective is to maximize, in T rounds, for each antenna, the expected cumulative reward or equivalently to minimize the regret $r_i = \mathbb{E}\left[\sum_{t=0}^T \max_k \theta_{l_i}^\top x_k - r_{i,t}\right]$. Classical bandit policy π uses the matrix $V_i = \sum_{t=0}^T x_{k_{i,t}} x_{k_{i,t}}^\top$ to estimate the unknown bandit parameter $\hat{\theta}_i$. At each iteration t, a network controller C manages the information sharing between antennas. As a baseline we consider the *decentralized controller* $C_{\texttt{dec}}$ which does not share any information between antennas as described by Algorithm 1.

Clustering module We introduce a novel *clustering controller* $C_{\texttt{clu}}$. At each round t, the controller estimates the underlying antenna similarity graph $\mathcal{G}$ with N vertices and the corresponding adjacency matrix denoted G, label each vertices (antenna) and shared the observations between antennas within the same cluster. Each edge of $\mathcal{G}$, $\{i, j\}$, are binary weighted taking 1 if the two confidence ellipsoids of the unknown bandit parameter estimation of the two antennas are overlapping, 0 if not. At each iteration, the controller estimates the edge value between each pair of antennas. To this aim, we take advantage of the ellipsoid overlapping test introduce in Gilitschenski and Hanebeck [2012]. To avoid clustering error, we keep a track of the local observations of each antenna –without sharing– to base the overlapping test on it. The test detects the overlapping if the minimum of a convex function $f_{i,j}$ is negative. This function involves the pair of local unknown bandit parameter estimations $\hat{\theta}_i^{\texttt{loc}}$ and $\hat{\theta}_j^{\texttt{loc}}$, the local observation matrices $V_i^{\texttt{loc}}$ and $V_j^{\texttt{loc}}$ along with their local confidence level $\epsilon_i^{\texttt{loc}}$ and $\epsilon_j^{\texttt{loc}}$. The confidence level of the ellipsoids could be tightly derived from Abbasi-Yadkori et al. [2011]. Note that, in contrast, the policy π is based on the shared observations too increase the performance of the antenna. Following the update of the matrix G, we label the vertices to cluster the antennas. In this contribution, we make the hypothesis of *–only two underlying clusters among the antennas–*. With this hypothesis, we simply apply a coloring algorithm Asratian et al. [1998] on the assumed bipartite $\mathcal{G}$ graph. Once all the antennas have a label, they share their observations within their cluster. The overall procedure of the *clustering controller* is described in Algorithm 2.

Algorithm 1: Decentralized controller algorithm

Input : T, a policy π
initialization: $t = 0$;
repeat
  % Choose a parameter setting to explore for each antenna
  for $\forall i \in \{1..N\}$ **do**
    Apply the policy π
until *while* $t < T$;

Algorithm 2: Clustering controller algorithm

Input : T, a policy π
initialization: $t = 0$;
repeat
  % $\mathcal{G}$ estimation
  for $\forall (i, j)$ *with* $i, j \in \{1..N\}$ **do**
    Edge-$(i, j) = 1$ if $\min_\lambda f_{i,j}(\lambda) < 0$ else 0
  % Coloring of $\mathcal{G}$
  Apply the coloring algorithm on $\mathcal{G}$
  % Choose a parameter setting to explore for each antenna
  for $\forall i \in \{1..N\}$ **do**
    Apply the policy π with the shared observations
until *while* $t < T$;

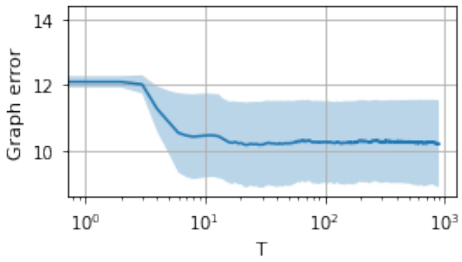

Figure 1: **Evolution of the average estimation error of the graph $\mathcal{G}$.** The error is defined as the ℓ_2-norm of the difference between the estimated and the true adjacency matrix G. The solid line corresponds to the average on the number of trials and the transparent line represent the standard-deviation across trials. The x-axis is logarithmic.

Algorithm 1 The algorithm of the decentralized controller which do not shared any observation between antennas. – **Algorithm 2** The algorithm of the clustering controller which shared the observations between antennas of the same cluster. At each iteration, the controller re-estimates the underlying graph $\mathcal{G}$ and labels each antenna.

Remarks Several notes could be done. *First*, by considering only the observations of the antenna –without sharing–, if the arms $(x_k)_{k=1}^d$ spawn $\mathbb{R}^d$ and the noise is centered, the estimation of the unknown bandit parameter is unbiased with the ordinary least square error. This means that, after a certain iteration, all the ellipsoids should departure one from each others. *Second*, the choice of policy π is determinant for the estimation quality of the adjacency matrix G. The more the policy explores all the dimension of the problem, the more the confidence ellipsoids are reduced which make them departure from each others if the corresponding antennas do not belong to the same cluster. *Third*, Our approach has no constraint on the policy π, this means that our technique can be coupled with policies such as *LinUCB* Li et al. [2010], *LinGapE* Xu et al. [2018] or *$\mathcal{E}$-Optimal Design* Boyd et al. [2004].

3 Experiments

We demonstrate the benefit of our clustering approach with synthetic and real data experiments. All the results are reproducible with the package `Bandpy`[1] along with the repository[2] to reproduce the figures of this paper.

Synthetic data We first validate the proposed approach on synthetic data to demonstrate the gain achieved by clustering the antennas together. We consider the scenario of $N = 20$ antennas equally divided in two clusters with their corresponding unknown bandit parameter $\theta_1, \theta_2 \in \mathbb{R}^d$, with $d = 15$, drawn randomly from a centered and normalized multivariate Gaussian distribution. Similarly, we generate $K = 10$ arms. The noise distribution is also a centered and normalized Gaussian. We couple our clustering method with the *LinUCB* Li et al. [2010] algorithm to observe the effect of our approach on the regret minimization. We consider $T = 100$ total iterations to properly underline the different phases of the algorithm and fix the α-*LinUCB* parameter to 0.1. The agents start to shared their observations at the iteration $m = 100$. We run the experiment 14 times to average the evolution of the mean –across agents– regret

Figure 1 depicts the evolution of the average –across trials– estimation error of the graph $\mathcal{G}$. The error is defined as the ℓ_2-norm of the difference between the estimated and the true adjacency matrix G. The x-axis is logarithmic and correspond to the number iteration t starting on the observations sharing at $t = m$. We observe a decreasing of the estimation error which converge in average to a constant error. Considering the remark 1, we explain that constant error by a number of total iterations value T set too low, which however underlines how conservative but slow our graph estimation is.

Figure 2 displays the regret minimization for the baseline –without the clustering– and our approach. We notice that both regrets decreases. However, when the antennas from the *clustering controller* starts sharing their observations at $t = m$ it produces a faster minimization of the regret. Remark that the regret values does not start from the same point since the arms and the unknown bandit parameters are drawn randomly which makes the worst regret different for each cluster.

[1] Available at `https://github.com/hcherkaoui/bandpy/`

[2] Available at `https://github.com/hcherkaoui/neurips_2022_workshop/`

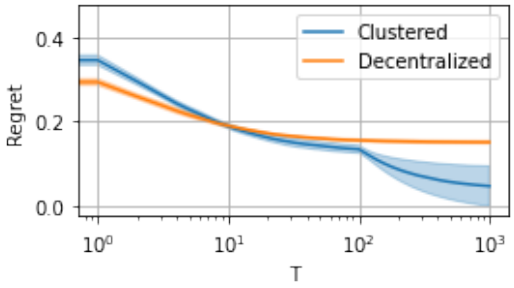

Figure 2: **Evolution of the average regret comparison between the clustered and decentralized controller**. The solid line corresponds to the average on the number of trials and the transparent line represent the standard-deviation across trials. The x-axis is logarithmic.

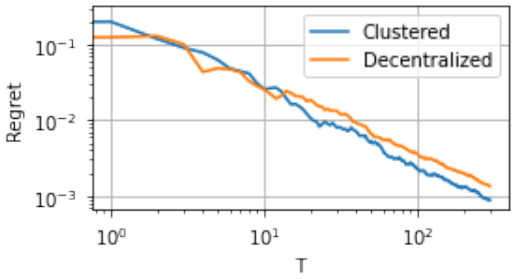

Figure 3: **Evolution of the average regret comparison between the clustered and decentralized controller**. The solid line corresponds to the average on the number of trials and the fine transparent line represent the small standard-deviation across trials. All the axes are logarithmic.

Real data To demonstrate the performance of our approach on real data. For the sake of replicability, we consider the –public– dataset MovieLens[3]. We retain the $N = 15$ users –agents– with the highest number of ratings and ensure that each movie was rated by all the retained users, by doing so we obtain $K = 19$ movies –arms–. At each iteration, the agents choose an arm which corresponding reward will be fetched from the dataset with a centered Gaussian noise of variance $\sigma = 4$ is added. We couple our clustering method with the *LinUCB* Li et al. [2010] algorithm to observe the effect of our approach on the regret minimization. We consider $T = 300$ total iterations and fix the α-*LinUCB* parameter to 0.1. The agents starts to shared their observations at the iteration $m = 30$. We run the experiment 7 times to average the evolution of the mean –across agents– regret.

Figure 3 exhibits the regret minimization for the baseline –without the clustering– and our approach. We notice a slightly better regret minimization for the cluster controller. However, as opposed to Figure 2, we do not distinguish a steep change in the regret minimization once the agents shared their observations. Thus, the effect of the sharing in that case is not clear and a possible explanation is that the hypothesis of two main clusters in the agents is too crude.

4 Conclusions

Conclusion In this work, we have introduced a clustering approach based on the estimation of the underlying graph that describe the similarity between antennas. We took advantage of an overlapping ellipsoids test and propose to use a simple bipartite graph labelling to recover the clusters. Despite the limitation of two underlying clusters our approach offers a flexible usage with different bandit policies. We demonstrate the potential of our approach with experiments on synthetic and real data.

Discussion Several aspects remains challenging. One of the difficulty with our approach is to label the antennas in our underlying graph. With only two clusters a coloring algorithm suffices to label accurately. However, an obvious extension would be to consider more than two clusters which makes the labelling not trivial. Moreover, as underlined in this paper, the quality of estimation of the underlying graph heavily depends on how well all the dimensions of the problem are explored. This creates a dilemma since most of the state-of-the-art policies focus on the most uncertain directions of the problem. A potential improvement would be to constraint the policy to reduce the uncertainty in the direction of every eigen-vectors of the each ellipsoids to improve their separation.

[3] Available at `https://grouplens.org/datasets/movielens/`

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11, 1999.

A. S. Asratian, T. M. Denley, and R. Häggkvist. *Bipartite graphs and their applications*, volume 131. Cambridge university press, 1998.

P. Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Y. Ban and J. He. A gang of bandits. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Y. Ban and J. He. Local clustering in contextual multi-armed bandits. In *Proceedings of the Web Conference 2021*, pages 2335–2346, 2021.

S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

C. Gentile, S. Li, and G. Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, volume 32, pages 757–765, 2014.

I. Gilitschenski and U. D. Hanebeck. A robust computational test for overlap of two arbitrary-dimensional ellipsoids in fault-detection of kalman filters. In *International Conference on Information Fusion*, pages 396–401, 2012.

M. D. Hoffman, B. Shahriari, and N. de Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *AISTATS*, 2014.

L. Li, W. Chu, J. Langford, and R. E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *International conference on World wide web*, pages 661–670, 2010.

A. Moradipari, M. Ghavamzadeh, and M. Alizadeh. Collaborative multi-agent stochastic linear bandits. *arXiv preprint*, 2022.

T. T. Nguyen and H. W. Lauw. Dynamic clustering of contextual multi-armed bandits. In *International Conference on Conference on Information and Knowledge Management*, pages 1959–1962, 2014.

L. Xu, J. Honda, and M. Sugiyama. A fully adaptive algorithm for pure exploration in linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 843–851, 2018.