Multi-Agent Combinatorial-Multi-Armed-Bandit framework for the Submodular Welfare Problem under Bandit Feedback

Extended Abstract

Subham Pokhriyal
Indian Institute of Technology Ropar
Rupnagar, India
subham.22csz0002@iitrpr.ac.in

Shweta Jain
Indian Institute of Technology Ropar
Rupnagar, India
shwetajain@iitrpr.ac.in

Vaneet Aggarwal
Purdue University
West Lafayette, USA
vaneet@purdue.edu

ABSTRACT

We study the *Submodular Welfare Problem* (SWP), where items are partitioned among agents with monotone submodular utilities to maximize total welfare under *bandit feedback*, i.e., only aggregate outcomes are observable. Classical SWP assumes full value-oracle access, achieving $1/2$ and $(1 - 1/e)$ approximations via greedy and continuous-greedy algorithms, respectively. We extend this to a *multi-agent combinatorial bandit* framework (MA-CMAB), where actions are partitions under full-bandit feedback with non-communicating agents. Unlike prior single-agent or separable multi-agent CMAB models, our setting couples agents through shared allocation constraints. We propose an explore–then–commit strategy with randomized assignments, achieving $\tilde{O}(T^{2/3})$ regret against a $(1-1/e)$ benchmark—the first such guarantee for partition-based submodular welfare under bandit feedback.

KEYWORDS

Multi-Agent Combinatorial Multi-Arm Bandits; Resilience guarantees; Submodular Welfare; Submodular optimization

ACM Reference Format:
Subham Pokhriyal, Shweta Jain, and Vaneet Aggarwal. 2026. Multi-Agent Combinatorial-Multi-Armed-Bandit framework for the Submodular Welfare Problem under Bandit Feedback: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/NBHM8571

1 INTRODUCTION

Combinatorial multi-armed bandit (CMAB) problems involve selecting subsets of base arms at each round and receiving stochastic rewards. The challenge lies in the structured yet exponentially large action space, which complicates exploration and optimization. Two feedback models are typically studied: under *semi-bandit feedback*, individual arm rewards are observed, whereas in the *full-bandit* setting, only the total reward from the chosen subset is revealed. The latter, though harder, better reflects real-world constraints such as privacy, limited instrumentation, or aggregated observations. Applications include recommender systems, data summarization, fair division, and influence maximization [8, 9, 15, 16].

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/NBHM8571

This work addresses a *multi-agent* variant of CMAB grounded in the *Submodular Welfare Problem* (SWP), which seeks to allocate items among agents with monotone submodular utilities to maximize total welfare. Classical SWP admits a $1/2$-approximation via the greedy algorithm [7, 12] and an optimal $(1-1/e)$ approximation through the continuous-greedy method with pipage rounding [20]. Online extensions of submodular optimization by adapting offline approximations to stochastic bandit feedback [8, 9, 15] achieve sublinear regret but are confined to *single-agent* or *communicating multi-agent* models with separable objectives.

This work studies a *non-communicating multi-agent* setting where each agent i has a distinct submodular utility $w_i(\cdot)$, and the learner selects a *partition* of items—assigning disjoint subsets to agents based only on aggregate reward feedback. This captures practical scenarios such as cohort-based recommendation, group influence maximization, and equitable allocation of indivisible goods, where agents compete for shared resources and item exclusivity couples their decisions. We introduce the **Multi-Agent Combinatorial Multi-Armed Bandit (MA-CMAB)** framework, which unifies partition-based submodular welfare maximization with online learning under full-bandit feedback. Our approach connects offline approximation guarantees to online regret minimization via a discrete randomized policy inspired by continuous greedy optimization, yielding resilience guarantees and a $\tilde{O}(T^{2/3})$ regret bound—the first such result for partition-based submodular welfare in a decentralized multi-agent bandit setting.

2 RELATED WORK

Early approximation guarantees on SWP include the $1/2$-approximation via greedy allocation [7, 12], while the optimal $(1 - 1/e)$-approximation was achieved through the continuous greedy algorithm combined with pipage rounding [2, 5, 20]. Hardness results show that improving beyond $(1 - 1/e)$ is NP-hard [10], and matching this ratio with value queries alone requires exponentially many oracle calls [13]. Online extensions of submodular maximization under bandit feedback have been developed for single-agent combinatorial settings [8, 14, 15], establishing offline-to-online reductions and robustness-based regret guarantees. In multi-agent learning, decentralized and collaborative multi-armed bandit models have been studied under various communication structures [1, 6, 11, 18], as well as collision-based and heterogeneous bandit formulations [3, 4, 17].Federated combinatorial multi-agent bandits have enabled coordination through shared information [9]. However, these works either assume separable objectives or communication among agents. In contrast, the MA-CMAB framework for SWP under full-bandit feedback considers non-communicating agents jointly selecting a

partition under matroid constraints, connecting the resilience of continuous greedy [5, 20] to stochastic regret minimization.

3 PROBLEM STATEMENT

We consider M agents $P = \{1, \dots, M\}$ and a finite collection of N indivisible items, denoted by Q. The ground set of assignments is $X = P \times Q$, where (i, j) denotes assigning item j to agent i. A feasible allocation $S \subseteq X$ satisfies the *partition matroid* constraint

$$\mathcal{D} = \{S \subseteq X : \ |S \cap (P \times \{j\})| \le 1, \ |S \cap (\{i\} \times Q)| \le b_i, \ \forall i, j\},$$

with optional quotas $b_i \in \mathbb{Z}_{\ge 0}$ ($b_i = N$ removes capacity limits). Each agent i has a monotone submodular valuation $w_i : 2^Q \to [0, 1]$, and the total welfare is $f(S) = \sum_{i=1}^{M} w_i(s_i)$, where $s_i = \{j : (i, j) \in S\}$. Let $\mathrm{OPT} \in \arg\max_{A \in \mathcal{D}} f(A)$ be the optimal allocation. Since exact maximization is NP-hard, we assume access to an α-approximate oracle $\mathcal{A}_{\text{off}}$ such that $f(A_{\mathcal{A}_{\text{off}}}) \ge \alpha f(\mathrm{OPT})$ for $\alpha \in (0, 1]$, which serves as the benchmark. Across T rounds, a policy π selects allocations $A_t \in \mathcal{D}$ and receives stochastic *full-bandit* feedback $r_t = f_t(A_t)$ with $\mathbb{E}[f_t(A_t)] = f(A_t)$. Only aggregate rewards are observed—no per-agent signals. The cumulative α-regret, which generalizes CMAB pseudo-regret to partition-based submodular welfare relative is

$$\mathbb{E}[\mathcal{R}(T)] = \alpha T f(\mathrm{OPT}) - \mathbb{E}\left[\sum_{t=1}^{T} f_t(A_t)\right]. \tag{1}$$

4 RESILIENCE GUARANTEE FOR THE SUBMODULAR WELFARE PROBLEM

We first define resilience, which means that even if utility evaluations for each agent are noisy within ϵ, the algorithm still achieves near-optimal performance up to an additive error proportional to ϵ. Next, we show that Continuous Greedy algorithm satisfies the resilience guarantees (Theorem 4.2).

Definition 4.1 ((α, δ, η)-Resilient Approximation). An offline algorithm $\mathcal{A}$ is (α, δ, η)-*resilient* if, given approximate oracle access $\hat{f}$ satisfying $|f(S) - \hat{f}(S)| \le \varepsilon$ for all feasible S, it returns $S_{\mathcal{A}}$ such that $\mathbb{E}[f(S_{\mathcal{A}})] \ge \alpha f(\mathrm{OPT}) - \delta\varepsilon$, where η bounds oracle calls and δ quantifies noise sensitivity.

Theorem 4.2 (Resilience of Continuous Greedy). *Under oracle noise $|\hat{w}_i(s_i) - w_i(s_i)| \le \varepsilon \le 1/(MN)^2 \ \forall s_i$, Continuous Greedy[20] is (α, δ, η)-resilient with $\alpha = 1 - \frac{1}{e}$, $\delta = (4MN + N)(1 - \frac{1}{M})$, $\eta = (MN)^8$. Hence, $\mathbb{E}[f(S)] \ge (1 - \frac{1}{e}) f(\mathrm{OPT}) - (4MN + 2M)\varepsilon$.*

The proof of Theorem 4.2 proceeds by first deriving a noise-aware upper bound on the optimal welfare in terms of the multilinear extension and expected marginal gains, explicitly accounting for additive oracle errors. It then analyzes the iterative progress of the Continuous Greedy updates and shows that the standard exponential decay of the optimality gap is preserved up to a controlled additive loss. This establishes that the fractional solution maintains the optimal $(1 - 1/e)$ approximation structure under noisy evaluations. The solution is subsequently converted to an integral allocation via pipage rounding, incurring only a negligible discretization loss. The detailed statements and complete proofs of the supporting technical results are deferred to the full version[1].

[1] http://arxiv.org/abs/2602.16183

Offline-to-Online Regret. We next connect offline resilience to stochastic regret in the online setting, where the learner interacts with the environment for T rounds, selecting feasible allocations and observing only aggregate (*full-bandit*) rewards. The goal is to minimize cumulative α-regret.

Let η denote the number of explored allocations and C the aggregation constant ($C = M$ for multi-agent or 1 otherwise). The (α, δ)-robust offline routine satisfies $\mathbb{E}[f(S)] \ge \alpha f(\mathrm{OPT}) - \delta\varepsilon$. During exploration, each allocation is played m times, yielding confidence radius $\mathrm{rad} = \sqrt{\frac{\log T}{2m}}$. Balancing exploration cost and estimation error gives $m^\star = \Theta((\delta C)^{2/3} \eta^{-2/3} T^{2/3} \log(T)^{1/3})$.

Theorem 4.3 (Final Regret Bound). *With $m = \lceil m^\star \rceil$, the expected α-regret satisfies*

$$\mathbb{E}[\mathcal{R}(T)] = O\left(\delta^{2/3} \eta^{1/3} C^{2/3} T^{2/3} \log(T)^{1/3}\right). \tag{2}$$

[Interpretation] The regret scales as $\tilde{O}(T^{2/3})$, matching the optimal stochastic rate for combinatorial bandits. Here, δ quantifies offline robustness to oracle noise, C the aggregation complexity, and η the number of explored allocations. This yields the first sublinear regret guarantee for partition-based submodular welfare under full-bandit feedback in a decentralized multi-agent setting.

5 ALGORITHMIC FRAMEWORK AND ANALYSIS

5.1 MA-CMAB for Social Welfare

The proposed MA-CMAB algorithm addresses the challenge of *multi-agent submodular welfare maximization* in a stochastic setting via a principled offline-to-online reduction. It operates in two phases: an **exploration phase**, collecting empirical estimates over $m = \left\lceil \frac{\delta^{2/3} T^{2/3} M^{2/3} (\log T)^{1/3}}{2\eta^{2/3}} \right\rceil$ rounds, and an **exploitation phase**, where a (α, δ, η)-resilient offline algorithm $\mathcal{A}_{\text{off}}$ outputs a high-welfare partition $S = \{s_1, \dots, s_M\}$. Empirical rewards $\bar{w}_i(s_i)$ and aggregate estimates $\bar{f}(S)$ guide the oracle, enabling robust allocation under uncertainty. The design extends the classical ETC framework [19] to a multi-agent, combinatorial, and submodular regime. Regret analysis is based on a high-probability clean event $\mathcal{E}$ ensuring concentration of empirical estimates, yielding near-optimal Utilitarian Social Welfare (USW) during exploitation. Compared to prior single-agent or separable approaches (e.g., C-ETC [15]), MA-CMAB tackles the harder problem of partitioned, non-communicating allocation with submodular utilities and full-bandit feedback, establishing a general framework for online social welfare maximization in structured stochastic environments.

6 CONCLUSION

We introduced **MA-CMAB**, a unified framework for fair division of indivisible goods among agents with submodular valuations under bandit feedback. Our offline analysis established resilience of welfare maximization with noisy oracles, and the online reduction achieved $\tilde{O}(T^{2/3})$ regret. Future work includes reducing oracle complexity, extending to semi-bandit and contextual settings, and tightening resilience–regret tradeoffs under stronger noise and richer combinatorial constraints.

ACKNOWLEDGMENTS

This work is supported by Anusandhan National Research Foun dation (ANRF)/ Science and Engineering Research Board (SERB)- Purdue University Overseas Visiting Doctoral Fellowship (Award No. SB/S9/Z-03/2017-XVII (2024)) and ANRF under grant MTR/2022/00 0818.

REFERENCES

[1] Mridul Agarwal, Vaneet Aggarwal, and Kamyar Azizzadenesheli. 2022. Multi-agent multi-armed bandits with limited communication. *Journal of Machine Learning Research* 23, 212 (2022), 1–24.

[2] Alexander A Ageev and Maxim I Sviridenko. 2004. Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *Journal of Combinatorial Optimization* 8, 3 (2004), 307–328.

[3] Animashree Anandkumar, Nithin Michael, Ao Kevin Tang, and Ananthram Swami. 2011. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications* 29, 4 (2011), 731–745.

[4] Ilai Bistritz and Amir Leshem. 2018. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems* 31 (2018).

[5] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. 2007. Maximizing a submodular set function subject to a matroid constraint. In *International Conference on Integer Programming and Combinatorial Optimization*. Springer, 182–196.

[6] Ronshee Chawla, Daniel Vial, Sanjay Shakkottai, and R Srikant. 2023. Collaborative multi-agent heterogeneous multi-armed bandits. In *International Conference on Machine Learning*. PMLR, 4189–4217.

[7] ML Fisher, GL Nemhauser, and LA Wolsey. 1978. An analysis of approximations for maximizing submodular set functions—II. *Mathematical Programming Studies* 8 (1978), 73–87.

[8] Fares Fourati, Vaneet Aggarwal, Christopher Quinn, and Mohamed-Slim Alouini. 2023. Randomized greedy learning for non-monotone stochastic submodular maximization under full-bandit feedback. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 7455–7471.

[9] Fares Fourati, Mohamed-Slim Alouini, and Vaneet Aggarwal. 2024. Federated combinatorial multi-agent multi-armed bandits. In *Proceedings of the 41st International Conference on Machine Learning*. 13760–13782.

[10] Subhash Khot, Richard J Lipton, Evangelos Markakis, and Aranyak Mehta. 2005. Inapproximability results for combinatorial auctions with submodular utility functions. In *International Workshop on Internet and Network Economics*. Springer, 92–101.

[11] Ravi Kumar Kolla, Krishna Jagannathan, and Aditya Gopalan. 2018. Collaborative learning of stochastic bandits over a social network. *IEEE/ACM Transactions on Networking* 26, 4 (2018), 1782–1795.

[12] Benny Lehmann, Daniel Lehmann, and Noam Nisan. 2001. Combinatorial auctions with decreasing marginal utilities. In *Proceedings of the 3rd ACM conference on Electronic Commerce*. 18–28.

[13] Vahab Mirrokni, Michael Schapira, and Jan Vondrák. 2008. Tight information-theoretic lower bounds for welfare maximization in combinatorial auctions. In *Proceedings of the 9th ACM conference on Electronic commerce*. 70–77.

[14] Rad Niazadeh, Negin Golrezaei, Joshua R Wang, Fransisca Susan, and Ashwinkumar Badanidiyuru. 2021. Online learning via offline greedy algorithms: Applications in market design and optimization. In *Proceedings of the 22nd ACM Conference on Economics and Computation*. 737–738.

[15] Guanyu Nie, Yididiya Y Nadew, Yanhui Zhu, Vaneet Aggarwal, and Christopher John Quinn. 2023. A framework for adapting offline algorithms to solve combinatorial multi-armed bandit problems with bandit feedback. In *International Conference on Machine Learning*. PMLR, 26166–26198.

[16] Subham Pokhriyal, Shweta Jain, Ganesh Ghalme, and Vaneet Aggarwal. 2025. Anytime Fairness Guarantees in Stochastic Combinatorial MABs: A Novel Learning Framework. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. 1660–1669.

[17] Jonathan Rosenski, Ohad Shamir, and Liran Szlak. 2016. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*. PMLR, 155–163.

[18] Abishek Sankararaman, Ayalvadi Ganesh, and Sanjay Shakkottai. 2019. Social learning in multi agent multi armed bandits. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 3, 3 (2019), 1–35.

[19] Aleksandrs Slivkins. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends® in Machine Learning* 12, 1-2 (2019), 1–286.

[20] Jan Vondrák. 2008. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*. 67–74.