Federated Combinatorial Multi-Agent Multi-Armed Bandits

Fares Fourati [1] **Mohamed-Slim Alouini** [1] **Vaneet Aggarwal** [2]

Abstract

This paper introduces a federated learning framework tailored for online combinatorial optimization with bandit feedback. In this setting, agents select subsets of arms, observe noisy rewards for these subsets without accessing individual arm information, and can cooperate and share information at specific intervals. Our framework transforms any offline resilient single-agent $(\alpha - \epsilon)$-approximation algorithm—having a complexity of $\tilde{\mathcal{O}}\left(\frac{\psi}{\epsilon^{\beta}}\right)$, where the logarithm is omitted, for some function ψ and constant β—into an online multi-agent algorithm with m communicating agents and an α-regret of no more than $\tilde{\mathcal{O}}\left(m^{-\frac{1}{3+\beta}}\psi^{\frac{1}{3+\beta}}T^{\frac{2+\beta}{3+\beta}}\right)$. Our approach not only eliminates the ϵ approximation error but also ensures sublinear growth with respect to the time horizon T and demonstrates a linear speedup with an increasing number of communicating agents. Additionally, the algorithm is notably communication-efficient, requiring only a sublinear number of communication rounds, quantified as $\tilde{\mathcal{O}}\left(\psi T^{\frac{\beta}{\beta+1}}\right)$. Furthermore, the framework has been successfully applied to online stochastic submodular maximization using various offline algorithms, yielding the first results for both single-agent and multi-agent settings and recovering specialized single-agent theoretical guarantees. We empirically validate our approach to a stochastic data summarization problem, illustrating the effectiveness of the proposed framework, even in single-agent scenarios.

[1]Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, KSA. [2]School of Industrial Engineering, Purdue University, West Lafayette, IN 47907, USA. Correspondence to: Fares Fourati <fares.fourati@kaust.edu.sa>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

The *Multi-Armed Bandits* (MAB) (Slivkins et al., 2019; Lattimore & Szepesvári, 2020) model online decision-making, where at every time step, an agent plays an arm and observes its associated reward. In combinatorial MAB, the agent can play a set of arms, instead of one arm, at each time step and receive a reward for that selection. When the agent only learns about the reward linked to the played set, it is known as *full-bandit feedback* or *bandit feedback*. If the agent gains additional insights into how each arm contributes to the overall reward, it is called *semi-bandit feedback*. Dealing with bandit feedback is more challenging since agents have less knowledge than in the semi-bandit feedback setting. In this work, we consider bandit feedback (Fourati et al., 2023a; Nie et al., 2023; Fourati et al., 2024), which has several applications, such as recommender systems, revenue maximization (Fourati et al., 2023a), influence maximization (Nie et al., 2023; Fourati et al., 2024), and data summarization, as shown in this work.

Federated learning (FL), an emerging machine learning paradigm, involves collaborative learning among multiple agents. In this process, selected agents share their local updates with a central server, which then aggregates these updates and sends the consolidated output back to each participating agent (Konečný et al., 2016; McMahan et al., 2017; Li et al., 2020; Hosseinalipour et al., 2020; Elgabli et al., 2022; Fourati et al., 2023b). While the problem has been studied in the context of continuous optimization, we address combinatorial optimization (Korte et al., 2011) in a federated online stochastic setting with bandit feedback. For example, the multi-agent setting under consideration can be applied to recommender systems, where each agent aims to recommend a set of products and then shares its findings with a server. We provide a general FL framework to adapt combinatorial offline single-agent approximation algorithms to online multi-agent settings, presenting the first results for the regret bounds in this setup.

We consider a setting with $m' \geq 1$ agents connected through a network. Among them, only a randomly selected subset of $m \leq m'$ agents can cooperate and share information in communication rounds, possibly through a server. This setup accommodates scenarios of partial participation, which are more practical in some real-world settings due to inherent

availability and communication constraints. When $m = m'$, the scenario reverts to a full-participation setting. All agents aim to solve the same online stochastic combinatorial problem within a time horizon of T. They conduct local exploration and then, if selected, share local estimations. Each agent i can play any subset of arms in parallel with others and receives noisy rewards for that set. In a combinatorial setting, the number of possible actions becomes exponentially large with the number of base arms, making sharing estimations for all possible actions prohibitive. Therefore, we consider a more practical approach wherein, in each communication round, selected agents share only a single action (subset) estimation.

Our work is the first to provide a general multi-agent framework for adapting combinatorial offline single-agent approximation algorithms to a FL setting to solve stochastic combinatorial multi-agent MAB (C-MA-MAB) problems with only bandit feedback. The proposed approach provably achieves a regret of at most $\tilde{\mathcal{O}}(m^{-\frac{1}{3+\beta}}\psi^{\frac{1}{3+\beta}}T^{\frac{2+\beta}{3+\beta}})$, which is sub-linear in the time horizon T and decreases with an increasing number of agents m, for some constant β and some function ψ that govern the complexity of the considered offline approximation algorithm. The framework does not require the agents to communicate every step; instead, it only needs $\tilde{\mathcal{O}}(\psi T^{\frac{\beta}{\beta+1}})$ communication times. Our proposed framework can serve to solve online combinatorial problems for both single-agent ($m^{'} = m = 1$) and multi-agent ($m^{'} > 1$) scenarios. Notably, our framework enjoys a linear speedup with an increasing number of agents, translating to decreased regrets.

We note that for a single agent, under bandit feedback, various offline-to-online transformations have been proposed for submodular maximization (Nie et al., 2022; Fourati et al., 2023a; 2024). Additionally, Nie et al. (2023) studied a framework for general combinatorial optimization in which any resilient offline algorithm with an α-approximation guarantee can be adapted to an online algorithm with sublinear α-regret guarantees. Similarly, an offline resilient $(\alpha-\epsilon)$-approximation algorithm can provide sublinear $(\alpha-\epsilon)$-regret guarantees for an online algorithm for any $\epsilon \geq 0$. This approach leads to a linear α-regret online algorithm when $\epsilon > 0$. Recently, Fourati et al. (2024) addressed this issue of linear regret in the case of monotone submodular optimization with a cardinality constraint. They adapted a sub-optimal offline $(1-1/e-\epsilon)$-approximation algorithm—whose complexity grows with $\log(\frac{1}{\epsilon})$—to a sub-linear $(1-1/e)$-regret online algorithm with bandit feedback, successfully eliminating the ϵ approximation error while ensuring sub-linearity with respect to the horizon T.

In this work, we generalize and extend the results previously established by Nie et al. (2023) and Fourati et al. (2024). We demonstrate that any sub-optimal single-agent offline approximation algorithm, with an approximation factor of $(\alpha-\epsilon)$ where $\epsilon \geq 0$, and ensuring resilience, can be adapted to a multi-agent setting with bandit feedback. This adaptation provides sub-linear α-regret guarantees, in contrast to the sub-linear $(\alpha-\epsilon)$-regret guarantees provided by Nie et al. (2023). This approach eliminates the ϵ error while maintaining sub-linearity with respect to T across any combinatorial problem, any reward, and under any constraints.

We note that in contrast to previous works that dealt with specific assumptions about reward functions and constraints on actions—such as assuming stochastic submodular rewards (Fourati et al., 2023a) or monotone rewards with cardinality constraints (Nie et al., 2022; Fourati et al., 2024)—our work considers general reward functions without making any assumptions about the reward type or constraints. Furthermore, our work explores the use of any offline algorithm $\mathcal{A}(\epsilon)$ as a subroutine with complexity in the general form of $\mathcal{O}\left(\frac{\psi}{\epsilon^{\beta}}\right)$ or $\mathcal{O}\left(\frac{\psi}{\epsilon^{\beta}}\log\left(\frac{1}{\epsilon}\right)\right)$, where ψ is a function of the problem characteristics, $\beta \geq 0$ some constant, and ϵ is an approximation error factor, ensuring that our results apply to any algorithm with these characteristics.

In addition, unlike previous works that focused on combinatorial single-agent scenarios (Fourati et al., 2023a; 2024; Nie et al., 2022; 2023), we address a combinatorial multi-agent setting where collaboration is permitted, possibly with partial participation, and optimize the worst-case regret for any agent, thereby recovering and generalizing the single-agent setting. Additionally, although our proposed algorithm has parameters, such as $\epsilon^{\star}$ and $r^{\star}$, none are considered hyperparameters. We derive closed-form values for these parameters as functions of the problem parameters, such as the range T, the number of available agents m, and the subroutine-related value of the constant β, and the function ψ, to minimize the expected cumulative α-regret.

Contributions: We introduce a novel FL framework for online combinatorial optimization, adapting single-agent offline algorithms to tackle online multi-agent problems with bandit feedback. The paper demonstrates the adaptability of any single-agent sub-optimal offline $(\alpha-\epsilon)$-approximation algorithm, ensuring resilience, for a C-MA-MAB with α-regret guarantees of $\tilde{\mathcal{O}}(m^{-\frac{1}{3+\beta}}T^{\frac{2+\beta}{3+\beta}})$, lifting the ϵ error, providing sub-linearity in the horizon T, and ensuring linear speedup with an increasing number of agents m. The framework only requires sub-linear communication rounds of $\tilde{\mathcal{O}}\left(T^{\frac{\beta}{\beta+1}}\right)$, which becomes at most logarithmic with respect to T when $\beta = 0$. Furthermore, we leverage our theoretical results to address online submodular maximization, present the first specialized regret for the non-monotone cardinality constraint case, and recover bounds for different offline algorithms, demonstrating tighter regrets than previous specialized works. Finally, we showcase the applicability of

our framework to both single-agent and multi-agent stochastic data summarization problems against MAB baselines, highlighting its effectiveness in practical scenarios.

2. Problem Statement

We formally present the problem as follows. We denote Ω as the ground set of n base arms. We consider a set $\mathbb{A}'$ with $m' \geq 1$ agents, where only a randomly selected subset $\mathbb{A}$ of $m \geq 1$ agents communicates in the communication rounds. We examine decision-making problems within a fixed period T, where, at every time step t, agent i chooses a subset $S_{i,t} \subseteq \Omega$. Let $\mathbb{S} \subseteq 2^{\Omega}$ represent the set of all permitted subsets, depending on the problem constraints.

In each time step t, agent i plays an action $S_{i,t} \in \mathbb{S}$ and acquires a noisy reward $f_t(S_{i,t})$. We assume that the reward f_t is a realization of a stochastic function with a mean of f, bounded in $[0, 1]$ [1], and i.i.d. conditioned on a given action. Thus, over a horizon T, agent i achieves a cumulative reward of $\sum_{t=1}^{T} f_t(S_{i,t})$. We define the expected reward function for a given action S as $f(S) = \mathbb{E}[f_t(S)]$, hence $S^\star = \arg\max_{S \subseteq \mathbb{S}} f(S)$ denote the optimal set in expectation.

In offline settings, attention is on the algorithm's complexity and worst-case expected output approximation guarantees. Conversely, in the online setting, attention is on cumulative rewards, where agents seek to minimize their expected cumulative regrets over time. One standard metric to assess the algorithm's online performance is to contrast the agent to an ideal learner that knows and consistently plays the best choice in expectation $S^\star$. However, the significance of such a comparison becomes questionable if the optimization of f over $\mathbb{S}$ is NP-hard, if the horizon is not exponentially large in the problem parameters (Fourati et al., 2023a; 2024; Nie et al., 2023). Hence, if a polynomial time offline algorithm $\mathcal{A}(\epsilon)$, happen to be an $(\alpha - \varepsilon)$-approximation[2] algorithm, for a given error $\epsilon \geq 0$, and an approximation ratio of $(\alpha - \epsilon) \leq 1$, for specific combinatorial objectives, a common approach involves comparing the agent's cumulative reward to $\sum_{t=1}^{T}(\alpha - \epsilon) f(S^\star)$ and denoting the difference as the $(\alpha - \epsilon)$-regret (Nie et al., 2023). In this work, we compare the learner's cumulative reward to a (tighter) agent that achieves $\sum_{t=1}^{T} \alpha f(S^\star)$, and we denote the difference as the α-regret, which is defined for every agent i as follows:

$$\mathcal{R}_i(T) = \sum_{t=1}^{T} (\alpha f_t(S^\star) - f_t(S_{i,t})). \tag{1}$$

[1]Results can be directly extended to a general function $f(\cdot)$ with a minimum value f_{min} and a maximum value f_{max} by considering a normalized function $g(\cdot) = (f(\cdot) - f_{min})/(f_{max} - f_{min})$.

[2]An algorithm $\mathcal{A}(\epsilon)$ is an $(\alpha - \epsilon)$-approximation algorithm for maximizing a deterministic function $f : \mathbb{S} \to \mathbb{R}$ after N oracle calls to f satisfies $\mathbb{E}[f(\Theta)] \geq (\alpha - \epsilon) f(S^\star)$, where $S^\star$ the optimal subset under f and the expectation is over the randomness of $\mathcal{A}(\epsilon)$.

The cumulative α-regret $\mathcal{R}_i(T)$ is random, given that it is a sum of α-regrets, which are functions of stochastic rewards and depend on the chosen subsets. In this work, we aim to minimize the expected cumulative α-regret, where the expectation encompasses the oracle noise and the randomness of the series of actions.

3. Related Work

In Table 1, we provide a comprehensive overview comparing our α-regret guarantees with existing ones across various adaptations of offline approximations for different combinatorial problems. Each row corresponds to a case involving an offline algorithm or a general class of algorithms with a specified complexity form, offline approximation factor, and target online regret factor α. The last two columns present previously established α-regret bounds and our own for adapting the proposed offline algorithm or class of algorithms as a subroutine to the online setting. The third row presents the GENERAL transformation, which generalizes the previous two cases, with a complexity of the form $\mathcal{O}(\psi)$ for any function ψ. The fifth row presents the MORE GENERAL case, with a complexity of the form $\mathcal{O}(\psi \log^{\gamma}(\frac{1}{\epsilon}))$, where $\gamma \in \{0, 1\}$, further generalizing the previous cases. The last row, presenting the MOST GENERAL, encompasses all previous rows, with a complexity of the form $\mathcal{O}(\frac{\psi}{\epsilon^{\beta}} \log^{\gamma}(\frac{1}{\epsilon}))$, where $\beta \geq 0$, including the GENERAL when $\beta = \gamma = 0$, the MORE GENERAL when $\beta = 0$, and the example in the sixth row.

Combinatorial Single-Agent Examples: We note that online submodular optimization with bandit feedback has been considered in (Nie et al., 2022; Fourati et al., 2023a; 2024). These results are differentiated from our work in Table 1. For single-agent non-monotone submodular rewards (See Section 6) under bandit feedback, Fourati et al. (2023a) adapts the RANDOMIZEDUSM in (Buchbinder et al., 2015), achieving sub-linear $\frac{1}{2}$-regret. Additionally, for single-agent monotone submodular rewards under bandit feedback with a cardinality constraint k, Nie et al. (2022) adapts the GREEDY in (Nemhauser et al., 1978), and Fourati et al. (2024) adapts the STOCHASTIC-GREEDY in (Mirzasoleiman et al., 2015), with both achieving sub-linear $(1 - \frac{1}{e})$-regret. Our work not only recovers all the results above in the single-agent setting but also generalizes them to the multi-agent setting, showing a decreasing regret with a factor of $m^{-\frac{1}{3}}$.

Multi-Agent: Previous works have proposed solutions for solving offline distributed submodular maximization. For example, partitioning the arms among the agents and running a greedy algorithm on each agent was proposed in previous works (Barbosa et al., 2015; Mirzasoleiman et al., 2013). While this is practical in some settings, it is designed for deterministic objectives and leads to lower approximation guarantees (half the ratio for monotone sub-

Offline Algorithm	Offline Complexity	Offline Factor	Online Factor α	Prior α-regret	Our α-regret Bound
RANDOMIZEDUSM (Buchbinder et al., 2015)	$\mathcal{O}(n)$	$1/2$	$1/2$	$\tilde{\mathcal{O}}\left(nT^{\frac{2}{3}}\right)^{*}$	$\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}}nT^{\frac{2}{3}}\right)$
GREEDY (Nemhauser et al., 1978)	$\mathcal{O}(nk)$	$1-\frac{1}{e}$	$1-\frac{1}{e}$	$\tilde{\mathcal{O}}\left(kn^{\frac{1}{3}}T^{\frac{2}{3}}\right)^{**}$	$\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}}kn^{\frac{1}{3}}T^{\frac{2}{3}}\right)$
GENERAL Excluding the next rows	$\mathcal{O}(\psi)$	α	α	$\tilde{\mathcal{O}}\left(\delta^{\frac{2}{3}}\psi^{\frac{1}{3}}T^{\frac{2}{3}}\right)^{**}$	$\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}}\delta^{\frac{2}{3}}\psi^{\frac{1}{3}}T^{\frac{2}{3}}\right)$
STOCHASTIC-GREEDY (Mirzasoleiman et al., 2015)	$\mathcal{O}\left(n\log(\frac{1}{\epsilon})\right)$	$1-\frac{1}{e}-\epsilon$	$1-\frac{1}{e}$	$\tilde{\mathcal{O}}\left(k^{\frac{2}{3}}n^{\frac{1}{3}}T^{\frac{2}{3}}\right)^{\dagger}$	$\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}}k^{\frac{2}{3}}n^{\frac{1}{3}}T^{\frac{2}{3}}\right)$
MORE GENERAL Excluding the next rows	$\mathcal{O}(\psi\log^{\gamma}(\frac{1}{\epsilon}))$	$\alpha-\epsilon$	α	**None**	$\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}}\delta^{\frac{2}{3}}\psi^{\frac{1}{3}}T^{\frac{2}{3}}\right)$
RANDOMSAMPLING (Buchbinder et al., 2017)	$\mathcal{O}\left(\frac{n}{\epsilon^2}\log(\frac{1}{\epsilon})\right)$	$\frac{1}{e}-\epsilon$	$\frac{1}{e}$	**None**	$\tilde{\mathcal{O}}\left(m^{-\frac{1}{5}}k^{\frac{2}{5}}n^{\frac{1}{5}}T^{\frac{4}{5}}\right)$
MOST GENERAL Including the previous rows	$\mathcal{O}(\frac{\psi}{\epsilon^{\beta}}\log^{\gamma}(\frac{1}{\epsilon}))$	$\alpha-\epsilon$	α	**None**	$\tilde{\mathcal{O}}\left(m^{-\frac{1}{3+\beta}}\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}}T^{\frac{2+\beta}{3+\beta}}\right)$

Table 1: The table summarizes the results of combinatorial optimization under bandit feedback. We use $\tilde{\mathcal{O}}$ to simplify expressions. Key parameters include horizon T, number of communicating agents m, base arm count n, and cardinality constraint k. Each row presents a specific offline algorithm or a class transformation with a given complexity, offline approximation factor, and a target online factor α. For the general rows, we consider classes of offline algorithms, with an approximation error factor ϵ, with general complexity forms, with general constants, $\psi \geq 0$, $\beta \geq 0$, $\gamma \in \{0, 1\}$, and $\delta \geq 0$. * (Fourati et al., 2023a), ** (Nie et al., 2023), † (Fourati et al., 2024).

modular maximization). Moreover, regret analysis for (non-combinatorial) multi-agent MAB (MA-MAB) problems has been investigated (Chawla et al., 2020; Wang et al., 2020; Agarwal et al., 2022). In (Chawla et al., 2020), gossip style communication approach is used between agents to achieve $\tilde{\mathcal{O}}((\frac{n}{m}+2)^{\frac{1}{3}}T^{\frac{2}{3}})$ regret. Wang et al. (2020) present an algorithm for a distributed bandit setting where all the agents communicate with a central node, which is shown to achieve a regret of $\tilde{\mathcal{O}}(\sqrt{nT/m})$. Agarwal et al. (2022) propose another algorithm, which splits the arms among the different agents, such that each learner plays arms only within a subset of arms and the best-communicated arm indices from other agents in the previous round. This achieves a regret of $\tilde{\mathcal{O}}(\sqrt{(\frac{n}{m}+m)T})$ while reducing the communication significantly. We note that these works do not directly extend to combinatorial bandits since the confidence-bound based approaches here cannot work for combinatorial bandits since $\mathcal{O}(2^n)$ sets cannot be explored. Recent works have considered FL for contextual bandits (Li & Wang, 2022; He et al., 2022; Li et al., 2023); however, these works do not apply to our setting. Our work is the first to present an FL framework for general combinatorial multi-agent MAB with bandit feedback.

Combinatorial Single-Agent Frameworks: Previous frameworks have been proposed for combinatorial single-agent MAB problems (Niazadeh et al., 2021; Nie et al., 2023). Niazadeh et al. (2021) proposes a framework aiming to adjust an iterative greedy offline algorithm into an online version within an adversarial bandit setting. However, their approach requisites the offline algorithm to possess an iterative greedy structure. In contrast, our framework treats the offline algorithm as a black-box algorithm. Furthermore, unlike our work, Niazadeh et al. (2021) imposes a condition known as Blackwell reproducibility on the offline algorithm, in addition to the resilience property. A closely related work is the single-agent framework by Nie et al. (2023) called C-ETC, which adapts offline algorithms with robustness guarantees to stochastic combinatorial single-agent MAB, generalizing some previous works for submodular maximization (Nie et al., 2022; Fourati et al., 2023a). However, C-ETC fails to generalize the more recent work of Fourati et al. (2024). Our framework generalizes and outperforms the results of Nie et al. (2023) in many ways. First, it extends to the multi-agent scenario involving $m \geq 1$ communicating agents, including the single-agent scenario. Additionally, while the C-ETC framework cannot replicate the results of Fourati et al. (2024) for submodular maximization, ours not only recovers all these previous works, including the framework, but also achieves even tighter regret guarantees that decrease with an increasing number of selected agents.

4. Combinatorial MA-MAB Framework

We first define resilient approximation algorithms and then present our proposed offline-to-online framework.

4.1. Offline Resilient-Approximation

We introduce the concept of resilient approximation, a metric that allows us to evaluate how an offline approximation algorithm reacts to controlled variations in function evaluations. We demonstrate that this specific characteristic alone can ensure that the offline algorithm can be modified to tackle stochastic C-MA-MAB settings, with only bandit feedback and achieving a sub-linear regret. Moreover, this adaptation does not rely on the algorithm's structure but treats it as a black-box algorithm.

We define the ξ-controlled-estimation $\bar{f}$ of a reward function f to deal with controlled variations.

Definition 4.1 (ξ-controlled-estimation)**.** For a set function $f : \mathbb{S} \to \mathbb{R}$ defined over a finite domain $\mathbb{S} \subseteq 2^{\Omega}$, a set function $\bar{f} : \mathbb{S} \to \mathbb{R}$, for a $\xi \geq 0$, is a ξ-controlled estimation of f if $|f(S) - \bar{f}(S)| \leq \xi$ for all $S \in \mathbb{S}$.

An estimation $\bar{f}$ is a ξ-controlled-estimation for a set function f if it consistently stays within a small positive range ξ of the actual values across all possible sets in $\mathbb{S}$. Given such a ξ-controlled-estimation for a set function f, we define $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation as follows:

Definition 4.2 ($(\alpha, \beta, \gamma, \psi, \delta)$-resilient approximation)**.** For any $\epsilon \geq 0$, an algorithm $\mathcal{A}(\epsilon)$ is an $(\alpha, \beta, \gamma, \psi, \delta)$-resilient approximation for maximizing a function $f : \mathbb{S} \subseteq 2^{\Omega} \to \mathbb{R}$ if, after making a number of calls $N(\beta, \gamma, \psi, \epsilon)$ to a ξ-controlled estimator $\bar{f}$, its output Θ satisfies the following condition: $\mathbb{E}[f(\Theta)] \geq (\alpha - \epsilon) f(S^{\star}) - \delta\xi$, where $S^{\star}$ is the optimal set under f, and the expectation is over the randomness of $\mathcal{A}(\epsilon)$. Here, $N(\beta, \gamma, \psi, \epsilon)$ is defined as ψ when $\epsilon = 0$, and as $\psi \frac{1}{\epsilon^{\beta}} \log^{\gamma}(\frac{1}{\epsilon})$ otherwise.

Remark 4.3. Several offline approximation algorithms are designed to solve specific combinatorial problems under particular reward and constraint assumptions. In Appendix E, we study various ones and demonstrate their resilience, characterizing each by their corresponding parameters: $\alpha, \beta, \gamma, \psi$, and δ. Given the resilience of an offline algorithm, one can directly apply Theorem 5.3 to ascertain the corresponding worst-case online guarantees when extending that offline algorithm to online settings.

Remark 4.4. An equivalent definition to resilient approximation was proposed by Nie et al. (2023) as a robust approximation. However, the resilient approximation provides more information about the algorithm's complexity. Specifically, an $(\alpha - \epsilon, \delta)$-robust-approximation algorithm that requires a number of oracle calls equal to $\frac{\psi}{\epsilon^{\beta}} \log^{\gamma}(\frac{1}{\epsilon})$, then it is an $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation algorithm. Furthermore, an $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation algorithm is an $(\alpha - \epsilon, \delta)$-robust-approximation algorithm, with $\epsilon \geq 0$.

Generally, the offline algorithms are designed for deterministic (noiseless) functions. However, in real-world applications, access to a noiseless reward function is not always possible, often due to the inherently stochastic nature of the problem. For example, recommending the same set of products to different people, or even to the same person at different times, may not yield consistent outcomes and rewards. This noise could also stem from using a stochastic approximation of the oracle function for computational efficiency, as seen in the case of stochastic data summarization discussed in Section 7. Therefore, in such cases, resilience against noisy rewards is necessary.

We demonstrate in Theorem 5.3 that resilience is, in fact, sufficient to ensure that the offline algorithm can be adjusted to achieve online sub-linear regret guarantees. In what follows, we explain how we use a resilient approximation algorithm $\mathcal{A}(\epsilon)$ as a subroutine in our proposed framework.

Algorithm 1 C-MA-MAB

Input: Horizon T, actions $\mathbb{S}$, agents $\mathbb{A}'$, number m, $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation algorithm $\mathcal{A}(\epsilon)$

Initialize $r_i^{\star} \leftarrow \left\lceil m^{-1} \left(\delta \sqrt{\log(T)} \left(\frac{Tm}{\psi} \right)^{\frac{1}{\beta+1}} \right)^{\frac{2+2\beta}{3+\beta}} \right\rceil$

Initialize $\epsilon^{\star} \leftarrow (\frac{\psi r_i^{\star}}{T})^{\frac{1}{\beta+1}} \mathbf{1}_{\{\beta>0 \text{ OR } \gamma>0\}}$.

Multi-Agent Exploration Time
Server starts running $\mathcal{A}(\epsilon^{\star})$, $j \leftarrow 0$
while $\mathcal{A}(\epsilon^{\star})$ queries the value of some $A \subseteq \mathbb{S}$ **do**
  Server broadcasts action A to the agents, $j \leftarrow j + 1$
  Server randomly selects a set $\mathbb{A} \subseteq \mathbb{A}'$ of m agents
  for agent i in $\mathbb{A}'$ in parallel **do**
    For $r_i^{\star}$ times, play action A
    If $i \in \mathbb{A}$, agent tracks & upload local mean $\bar{f}_i$
  end for
  Server calculates the mean $\bar{f}$ and feeds it to $\mathcal{A}(\epsilon^{\star})$
end while

Multi-Agent Exploitation Time
Server broadcasts Θ, the final output of $\mathcal{A}(\epsilon^{\star})$
for agent i in $\mathbb{A}'$ in parallel **do**
  for *remaining time of the agent* i **do**
    Play action Θ
  end for
end for

4.2. Offline-to-Online Multi-Agent Framework

We present our proposed C-MA-MAB Framework; see Algorithm 1. This framework is applicable for both single-agent and multi-agent settings, wherein the design of our algorithm ensures that the single-agent setting is simply a special case of the multi-agent scenario. For a set $\mathbb{S}$ of possible actions, a set of $m' \geq 1$ agents $\mathbb{A}'$, a number $m \geq 1$ of communicating agents, and a time horizon T, our algorithm can adapt any off-the-shelf, offline single-agent combinatorial $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation algorithm $\mathcal{A}(\epsilon)$ to the C-MA-MAB setting. This adaptation comes with theoretical guarantees, as detailed in Theorem 5.3, applicable under any reward type or action constraint.

While the offline algorithm may be a function of some variable $\epsilon \geq 0$, which trades off its complexity and approximation guarantees, our proposed algorithm finds $\epsilon^{\star}$ which optimizes this trade-off based on the problem parameters and its complexity to minimize regret and uses $\mathcal{A}(\epsilon^{\star})$ for exploration instead. Furthermore, in the exploration time,

whenever the offline algorithm $\mathcal{A}(\epsilon^\star)$ requests the value oracle for action A, each agent of the $m' \geq 1$ agents plays the action A for $r_i^\star = r^\star/m$ times, where $r^\star$ is another parameter chosen based on the setting and the subroutine complexity to minimize regret, with

$$r^\star = m \left\lceil m^{-1} \left(\delta \sqrt{\log(T)} \left(\frac{Tm}{\psi} \right)^{\frac{1}{\beta+1}} \right)^{\frac{2+2\beta}{3+\beta}} \right\rceil . \quad (2)$$

Furthermore, we choose $\epsilon^\star$ as follows:

$$\epsilon^\star = (\frac{\psi r^\star}{Tm})^{\frac{1}{\beta+1}} \mathbf{1}_{\{\beta>0 \text{ OR } \gamma>0\}}. \quad (3)$$

Only the randomly selected m agents track and broadcast their local estimations to the server, i.e., each agent $i \in \mathbb{A}$ sends its local estimation $\bar{f}_i$, then the server aggregates these estimations in one global estimation $\bar{f}$ of rewards for A and then returns $\bar{f}$ to $\mathcal{A}(\epsilon^\star)$. Finally, in the exploitation phase, all the agents in $\mathbb{A}'$ play Θ, the output from algorithm $\mathcal{A}(\epsilon^\star)$, for the remaining time.

Remark 4.5. C-MA-MAB has low storage complexity. In every step, an agent needs to store, at most, n indices and a real value representing the empirical mean of one action. Only the action Θ is stored during exploitation time, and no additional computation is required. During exploration time, each agent needs to store only the proposed action A and update its associated empirical mean; everything is deleted once the server proposes another action. Furthermore, the proposed framework does not require the combinatorial offline algorithm $\mathcal{A}(\epsilon)$ to have any particular structure and employs $\mathcal{A}(\epsilon^\star)$ as a black-box algorithm. Consequently, it shares the same complexity as the subroutine $\mathcal{A}(\epsilon^\star)$. Moreover, $\mathcal{A}(\epsilon^\star)$ is executed on the server, alleviating the computational overhead for the agents.

Remark 4.6. In the multi-agent setting, we assume that the server can either be a separate entity or one of the agents playing the role of the server. In the single-agent setting, without loss of generality, the sole agent can be considered as the server. The server orchestrates communication by selecting clients and recommending which actions to explore, based on the offline subroutine, in a synchronized manner. Recent studies have explored federated linear and kernelized contextual bandits with asynchronous communication (Li & Wang, 2022; He et al., 2022; Li et al., 2023). Future research might investigate general combinatorial optimization with asynchronous communication.

Remark 4.7. The proposed C-MA-MAB uses the time horizon T to compute $r^\star$ and $\epsilon^\star$. When the exact time horizon is unknown, the results can be enhanced by employing the concept of an anytime algorithm through the use of the geometric doubling trick by establishing a geometric sequence of time intervals, denoted as T_i, where $T_i = T_0 2^i$ for $i \in \mathbb{N}$, where T_0 is a sufficiently large value to ensure proper initialization. From Theorem 4 in (Besson & Kaufmann, 2018), it follows that the regret bound preserves the $T^{\frac{2+\beta}{3+\beta}}$ dependence with only changes in constant factors.

5. Theoretical Analysis

We upper-bound the required communication rounds, we lower-bound the probability of the empirical mean being ξ-controlled-estimation of the expectation (*good event*), and upper-bound the expected cumulative α-regret.

5.1. Communication Analysis

By the design of the algorithm, the m randomly selected agents, after locally estimating the quality of a suggested action A, communicate only one value, representing the local estimation $\bar{f}_i$. The agents exploit the decided set Θ during the exploitation phase without further communication. During exploration, the selected agents must communicate their local estimation for every requested action A. Therefore, the number of communication rounds is upper-bounded by the number of requested actions, i.e., the required oracle calls $N(\beta, \gamma, \psi, \epsilon^\star)$, which we upper-bound in the following lemma, which we prove in Appendix B.

Lemma 5.1. *The number of communication times, i.e., the number of oracle queries $N(\beta, \gamma, \psi, \epsilon^\star)$ of the subroutine $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation algorithm $\mathcal{A}(\epsilon^\star)$ satisfies: $N(\beta, \gamma, \psi, \epsilon^\star) \leq \mathcal{O}(\psi T^{\frac{\beta}{\beta+1}} \log^\gamma(T))$.*

From Lemma 5.1 it follows that for cases where $\beta = 0$, which applies to several offline algorithms (Nemhauser et al., 1978; Khuller et al., 1999; Sviridenko, 2004; Buchbinder et al., 2015; Mirzasoleiman et al., 2015; Yaroslavtsev et al., 2020), the communication rounds are at most $\tilde{\mathcal{O}}(\psi)$, scaling at most logarithmically with T. For example, as shown in Corollary 6.1, with n arms and using RandomizedUSM (Buchbinder et al., 2015) as a subroutine (where ψ is n, β is zero, and γ is zero), our framework guarantees a communication complexity of $\mathcal{O}(n)$, not scaling with T.

By design of the C-MA-MAB algorithm, after every action queried by the subroutine, the agents have to explore and estimate the values of the proposed action. To do that, each agent has to play the proposed action for $r_i^\star$ times. Therefore, using the result from Lemma 5.1 on the number of required communication rounds and by the definition of $r_i^\star$, we can derive that the required exploration steps for every agent, i.e., $\mathcal{O}(r_i^\star \psi T^{\frac{\beta}{\beta+1}} \log^\gamma(T))$, which is decreasing with an increasing number of agents m.

5.2. Estimation Analysis

In C-MA-MAB, every agent plays each action queried by the subroutine $\mathcal{A}(\epsilon^\star)$ the same number of times. These repetitions provide an estimation of the action values. We

define a *good event* $\mathcal{E}$ when the empirical mean estimation $\bar{f}$ is a ξ-controlled-estimation of the reward expectation f, on the played actions during exploration time, with $\xi := \sqrt{\log(T)/r^\star}$. For every communication round j, each action A_j queried by the $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation $\mathcal{A}(\epsilon^\star)$, where $j \in \{1, \cdots, N(\beta, \gamma, \psi, \epsilon^\star)\}$, we define the event $\mathcal{E}_j$ as:

$$\mathcal{E}_j \triangleq \{\left|\bar{f}(A_j) - f(A_j)\right| \leq \xi\}. \tag{4}$$

Therefore, the *good event* $\mathcal{E}$, which considers the empirical mean estimation $\bar{f}$ is a ξ-controlled estimate of the reward expectation f for every communication round j, i.e., considers the realization of $\mathcal{E}_j$ for every $j \in \{1, \cdots, N(\beta, \gamma, \psi, \epsilon^\star)\}$, which is expressed as follows:

$$\mathcal{E} = \mathcal{E}_1 \cap \cdots \cap \mathcal{E}_{N(\beta,\gamma,\psi,\epsilon^\star)}. \tag{5}$$

Each action A queried by the offline algorithm have been explored for $r^\star$ number of times among the agents. These $r^\star$ rewards are i.i.d. with expectation $f(A)$ and confined within the $[0, 1]$ range. Consequently, we can bound the deviation of the empirical mean $\bar{f}(A_j)$ from the expected value $f(A_j)$ for every action undertaken. Thus, we upper bound the probability of the *good event* in the following lemma, which we prove in Appendix C.

Lemma 5.2. *The probability of the good event $\mathcal{E}$, (5), when using an $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation algorithm $\mathcal{A}(\epsilon^\star)$ as a subroutine satisfies:*

$$\mathbb{P}(\mathcal{E}) \geq 1 - 2N(\beta, \gamma, \psi, \epsilon^\star)T^{-2}.$$

Combining both results of Lemma 5.1 and Lemma 5.2 it follows that the *bad event* happens with a probability of at most $\tilde{\mathcal{O}}(\psi T^{\frac{\beta}{\beta+1}-2})$, decreasing as T increases.

5.3. Regret Analysis

We analyze the expected cumulative α-regret for the C-MA-MAB (Algorithm 1), with m communicating agents.

Theorem 5.3. *For the sequential combinatorial decision-making problem defined in Section 2, with $T \geq \max\{\psi m, \psi m^{\frac{1+\beta}{2}}/\delta^{\beta+1}\}$, the expected cumulative α-regret of the C-MA-MAB presented in Algorithm 1 using an $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation algorithm $\mathcal{A}(\epsilon^\star)$ as subroutine is at most $\tilde{\mathcal{O}}\left(m^{-\frac{1}{3+\beta}}\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}}T^{\frac{2+\beta}{3+\beta}}\right)$.*

The above theorem implies that an offline algorithm does not need to have an α-approximation guarantee to be adapted to achieve sublinear α-regret guarantees. In fact, an $(\alpha - \epsilon)$-approximation algorithm, denoted as $\mathcal{A}(\epsilon)$ with $\epsilon > 0$, if it is a $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation, can be extended to a sub-linear α-regret algorithm. Later, in Section 6, we apply the above theorem to special combinatorial cases.

Remark 5.4. Linear speedup is evident in our approach, as the collective regret across m agents is $\tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}}(Tm)^{\frac{2+\beta}{3+\beta}}\right)$. This mirrors the regret one would observe if all m agents collaborated, sharing a total of Tm time for the central agent. Consequently, the distributed setup incurs no loss, with each agent interacting with the environment T times, and the combined regret reflects that of a single agent allocated Tm time for interaction.

Remark 5.5. The δ function depends on the offline algorithm and refers to a general function of the combinatorial problem parameters, such as the number of main arms, n, or the cardinality constraint, k. This function serves as the scaling factor for the radius ξ in the lower bound on the expected reward, as given by $(\mathbb{E}[f(\Theta)] \geq (\alpha - \epsilon)f(S^\star) - \delta\xi)$, as defined in Definition 4.2 for $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation algorithms. In some offline algorithms, we have demonstrated that this scaling function depends on the cardinality constraint k. For example, Lemma E.4 shows that δ equals $4k$ for RANDOMSAMPLING. In other cases, the function may depend solely on the number of main arms n. For instance, Lemma E.1 establishes that δ is $\frac{5}{2}n$ for the RANDOMIZEDUSM. Our analysis accommodates any δ function defined in terms of the problem parameters.

Remark 5.6. When algorithm approximations do not depend on ϵ, it implies that their complexity is of the form $\tilde{\mathcal{O}}(\psi)$, hence $\beta = \gamma = 0$. It follows that using such an $(\alpha, 0, 0, \psi, \delta)$-resilient-approximation algorithm as a subroutine achieves a regret of at most $\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}}\delta^{\frac{2}{3}}\psi^{\frac{1}{3}}T^{\frac{2}{3}}\right)$.

Remark 5.7. A lower bound remains an open question for general combinatorial stochastic rewards under bandit feedback. A lower bound is missing even for the special cases of stochastic submodular rewards under bandit feedback. Some lower bounds have been proposed in restrictive special settings. For example, Niazadeh et al. (2021) showed a $\tilde{\Omega}(T^{\frac{2}{3}})$ lower bound for adversarial submodular rewards, where the reward could only be observed in user-specified exploration rounds. Moreover, Tajdini et al. (2023) demonstrated that for monotone stochastic submodular bandits with a cardinality constraint, for small time horizon T, a regret scaling like $T^{\frac{2}{3}}$ is inevitable when compared to the greedy algorithm in (Nemhauser et al., 1978). However, it does not provide a lower bound on $(1 - 1/e)$-regret.

In the following, we provide a sketch of the proof and leave a detailed one in Appendix D. We separate the proof into two cases. One case when the *good event* $\mathcal{E}$ happens, which we show in Lemma 5.2 happens with high probability and then we generalize the result under any event.

5.3.1. REGRET OF AN AGENT UNDER THE GOOD EVENT

We upper-bound the expected α-regret conditioned on the *good event* $\mathcal{E}$. However, for simplicity in notation, we employ $\mathbb{E}[\cdot]$ rather than $\mathbb{E}[\cdot|\mathcal{E}]$ in certain instances. We de-

compose and then bound the expected α-regret into two components: one addressing the regret stemming from exploration (P_1) and the other from exploitation (P_2).

$$\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] = \sum_{t=1}^{T} (\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(S_{i,t})]) \tag{6}$$

$$= \underbrace{\sum_{j=1}^{N(\beta,\gamma,\psi,\epsilon^\star)} r_i^\star \left(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(A_j)]\right)}_{\alpha\text{-regret from exploration } (P_1)}$$

$$+ \underbrace{\sum_{t=T_{N(\beta,\gamma,\psi,\epsilon^\star)}+1}^{T} (\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(\Theta)])}_{\alpha\text{-regret from exploitation } (P_2)}. \tag{7}$$

We begin with bounding regret from exploration and using that the rewards are within the interval $[0, 1]$,

$$P_1 \leq \sum_{j=1}^{N(\beta,\gamma,\psi,\epsilon^\star)} \frac{r^\star}{m}\alpha \leq N(\beta,\gamma,\psi,\epsilon^\star)\frac{r^\star}{m}. \tag{8}$$

When the *good event* $\mathcal{E}$ occurs, we know that $|\bar{f}(A) - f(A)| \leq \xi$ for all considered action A. Using an $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation $\mathcal{A}(\epsilon^\star)$, with output Θ, we have

$$\alpha f(S^\star) - \mathbb{E}[f(\Theta)] \leq \delta\xi + \epsilon^\star f(S^\star). \tag{9}$$

Therefore, with $f(\mathrm{S}^\star) < 1$, we have:

$$P_2 \leq \sum_{t=T_{N(\beta,\gamma,\psi,\epsilon^\star)}+1}^{T} (\delta\xi + \epsilon^\star) \leq T(\delta\xi + \epsilon^\star). \tag{10}$$

Therefore, using Eq. (8) and Eq. (10), the total expected cumulative regret in Eq. (7) can be bounded as:

$$\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] \leq N(\beta,\gamma,\psi,\epsilon^\star)\frac{r^\star}{m} + T(\delta\xi + \epsilon^\star). \tag{11}$$

Using the confidence radius $\xi = \sqrt{\log(T)/r^\star}$ and $N(\beta,\gamma,\psi,\epsilon^\star) = \psi\frac{1}{\epsilon^{\star\beta}}\log^\gamma(\frac{1}{\epsilon^\star})$, we have

$$\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] \leq \frac{\psi r^\star \log^\gamma(\frac{1}{\epsilon^\star})}{\epsilon^{\star\beta} m} + T(\sqrt{\frac{\delta^2\log(T)}{r^\star}} + \epsilon^\star).$$

We note that the above inequality is correct for all values of $r^\star \geq m$ and $\epsilon^\star \geq 0$ with the convention that $0^0 = 1$. In our algorithm, we choose the values of $r^\star$ and $\epsilon^\star$ as functions of the problem parameters and on the subroutine complexity parameters. We choose $r^\star$ as defined in Eq. (2) and $\epsilon^\star$ as defined in Eq. (3).

Recall that $\beta \geq 0$ and $\gamma \in \{0, 1\}$. Therefore, we consider all the possible cases, the first when $\beta = \gamma = 0$, the second when $\beta = 0$ and $\gamma = 1$, and the third when $\beta > 0$. For all the cases, when the *good event* $\mathcal{E}$ happens, the expected α-regret of our C-MA-MAB with an $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation as subroutine we have

$$\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] \leq \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right). \tag{12}$$

5.3.2. Regret of an Agent under Any Event

Given that the reward $f_t(\cdot)$ is upper bound by 1, the expected cumulative α-regret when the event $\bar{\mathcal{E}}$ happens over a range T is upper-bounded as follows: $\mathbb{E}[\mathcal{R}_i(T)|\bar{\mathcal{E}}] \leq T$. Combining the results when *good event* happens and does not happen, using the law of total expectation, and using Eq. (12), Lemma 5.1, Lemma 5.2, and $T \geq m\psi$:

$$\mathbb{E}[\mathcal{R}_i(T)] = \mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] \cdot \mathbb{P}(\mathcal{E}) + \mathbb{E}[\mathcal{R}_i(T)|\bar{\mathcal{E}}] \cdot \mathbb{P}(\bar{\mathcal{E}})$$
$$\leq \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right).$$

This establishes the result in Theorem 5.3.

6. Application to Submodular Maximization

We use our C-MA-MAB framework to address scenarios involving stochastic submodular[3] rewards, with bandit feedback. Submodular maximization (SM) is an NP-hard problem (Nemhauser et al., 1978; Feige et al., 2011), which recently has shown growing interest in studying combinatorial MAB (Chen et al., 2018; Niazadeh et al., 2021; Nie et al., 2022; Fourati et al., 2023a; 2024). In the following, we present our results for SM for monotone[4] and non-monotone rewards with and without a cardinality constraint and leave the knapsack constraint in Appendix E.4.

For unconstrained SM (USM), Buchbinder et al. (2015) proposed RandomizedUSM, achieving a $\frac{1}{2}$-approximation. In Lemma E.1 in Appendix E, we generalize Corollary 2 in (Fourati et al., 2023a) to show its resilience and present the following corollary, which recovers the guarantees of the online algorithm in (Fourati et al., 2023a) for a single agent and generalizes it for multi-agent setting.

Corollary 6.1. *C-MA-MAB, using the* RandomizedUSM *as a subroutine, needs at most $\mathcal{O}(n)$ communication times and its $\frac{1}{2}$-regret is at most $\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}} n T^{\frac{2}{3}}\right)$ for USM.*

For SM under a cardinality constraint k (SMC) with monotone rewards, the Greedy in (Nemhauser & Wolsey, 1978) achieves $1 - 1/e$. In contrast, the Stochastic-Greedy

[3] A function $f : 2^\Omega \to \mathbb{R}$, over a finite set Ω, is considered submodular if it discloses the characteristic of diminishing returns: for any $A \subseteq B \subset \Omega$ and $v \in \Omega \backslash B$, the inequality $f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$ is verified.

[4] A function $f : 2^\Omega \to \mathbb{R}$, over a finite set Ω, is considered monotone if for any $A \subseteq B \subseteq \Omega$ we have $f(A) \leq f(B)$.

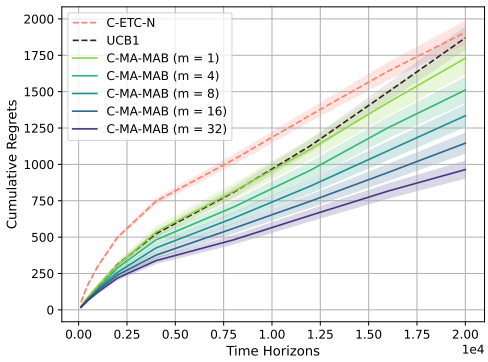


Figure 1: Cumulative regrets of summarizing images from CIFAR10 for different horizons T using our C-MA-MAB framework with different number of agents m, against C-ETC-N and UCB1.

in (Mirzasoleiman et al., 2015) achieves $1 - 1/e - \epsilon$, where ϵ is a parameter balancing accuracy and complexity. We provide the resilience of these two algorithms in Appendix E and present the following results.

Corollary 6.2. *C-MA-MAB, using the* GREEDY *as a subroutine, needs $\mathcal{O}(nk)$ communication times and its $(1 - 1/e)$-regret is at most $\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}} k n^{\frac{1}{3}} T^{\frac{2}{3}}\right)$ for monotone SMC.*

Corollary 6.3. *C-MA-MAB, using the* STOCHASTIC-GREEDY*, needs $\tilde{\mathcal{O}}(n)$ communication times and its $(1 - \frac{1}{e})$-regret is at most $\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}} k^{\frac{2}{3}} n^{\frac{1}{3}} T^{\frac{2}{3}}\right)$ for monotone SMC.*

The STOCHASTIC-GREEDY algorithm has sub-optimal approximation guarantees of $(1 - 1/e - \epsilon)$; thus, using the C-ETC framework from (Nie et al., 2023) can only guarantee sub-linear $(1 - 1/e - \epsilon)$-regret. Consequently, $(1 - 1/e)$-regret will be linear in T. However, our C-MA-MAB guarantees sublinear $(1 - 1/e)$-regret, recovering the single agent's results in (Fourati et al., 2024). Furthermore, the two corollaries above demonstrate that employing a sub-optimal approximation algorithm in terms of the approximation factor, rather than one that achieves optimal approximation guarantees, does not necessarily imply lower regret guarantees, as shown in (Fourati et al., 2024).

For non-monotone SMC, the RANDOMSAMPLING algorithm in (Buchbinder et al., 2017) achieves $1/e - \epsilon$, where ϵ is a parameter balancing accuracy and complexity. We provide the resilience of this algorithm in Appendix E and derive the first result for single-agent and multi-agent online stochastic non-monotone SMC with sublinear regrets.

Corollary 6.4. *C-MA-MAB, using the* RANDOM SAMPLING *in (Buchbinder et al., 2017) as a subroutine, needs $\tilde{\mathcal{O}}(nT^{\frac{2}{3}})$ communication times and its $(\frac{1}{e})$-regret is at most $\tilde{\mathcal{O}}\left(m^{-\frac{1}{5}} k^{\frac{2}{5}} n^{\frac{1}{5}} T^{\frac{4}{5}}\right)$ for non-monotone SMC.*

7. Experiments with Data Summarization

We employ our C-MA-MAB on data summarization, a primary challenge in machine learning (Mirzasoleiman et al., 2013), mainly when dealing with a large dataset. While this problem has been widely studied with access to a deterministic oracle (Lin & Bilmes, 2011; Mirzasoleiman et al., 2013; 2015; 2020; Sivasubramanian et al., 2024), this work is the first to address online data summarization under a stochastic objective function. We run experiments on FMNIST (Xiao et al., 2017) and CIFAR10 (Krizhevsky et al., 2009), present the latter in the main paper, and relegate more details and results to Appendix F.

In data summarization, an action A consists of a set of at most k images to summarize a large dataset $\mathcal{D}$. Adding more images achieves better summarization but follows a diminishing return property. Thus, it falls in the monotone SMC (Mirzasoleiman et al., 2015). Evaluating a given action A against a dataset $\mathcal{D}$ may become expensive with a large dataset. Thus, we consider a stochastic objective where the chosen subset is compared only to a random subset $\mathcal{R} \subseteq \mathcal{D}$ drawn uniformly at random from $\mathcal{D}$, using a similarity metric C, resulting in noisier but lower complexity evaluations. We do not solve the problem for a given realization $\mathcal{R}$, but we solve it in expectation: $\underset{\mathcal{A} \subseteq \mathcal{D}: |\mathcal{A}| \leq k}{\arg\max}\ \mathbb{E}_{\mathcal{R}}\left[\sum_{i \in \mathcal{R}} \max_{v \in \mathcal{A}} C(i, v)\right]$.

We test our method when using the STOCHASTIC-GREEDY algorithm as a subroutine (Mirzasoleiman et al., 2015) for one agent and multiple agents and compare it to the proposed algorithm in C-ETC framework for SMC (C-ETC-N) (Nie et al., 2023), and the upper confidence bound (UCB1) algorithm (Auer et al., 2002).

The C-MA-MAB demonstrates sub-linear regret guarantees as depicted in Fig. 1. Additionally, it is apparent that, for varying values of m, the C-MA-MAB consistently outperforms both C-ETC-N and UCB1, even with a single agent, exhibiting lower regrets over diverse time horizons. Notably, an increase in the number of agents correlates with a reduction in regret for these agents. These observations reinforce the same conclusions drawn from the theoretical analysis.

Conclusion

We introduce C-MA-MAB, a framework for single-agent and multi-agent online stochastic combinatorial problems, which adapts resilient offline $(\alpha - \epsilon)$-approximation algorithms to online algorithms under bandit feedback, achieving sublinear α-regret bounds with respect to the time horizon T, eliminating the ϵ error, and ensuring a linear speedup. We also present specialized bounds for SM with and without constraints and apply C-MA-MAB to online stochastic data summarization.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, M., Aggarwal, V., and Azizzadenesheli, K. Multi-agent multi-armed bandits with limited communication. *The Journal of Machine Learning Research*, 23(1):9529–9552, 2022.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002.

Balakrishnan, R., Li, T., Zhou, T., Himayat, N., Smith, V., and Bilmes, J. Diverse client selection for federated learning via submodular maximization. In *International Conference on Learning Representations*, 2022.

Barbosa, R., Ene, A., Nguyen, H., and Ward, J. The power of randomization: Distributed submodular maximization on massive datasets. In *International Conference on Machine Learning*, pp. 1236–1244. PMLR, 2015.

Besson, L. and Kaufmann, E. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Buchbinder, N., Feldman, M., Seffi, J., and Schwartz, R. A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing*, 44(5):1384–1402, 2015.

Buchbinder, N., Feldman, M., and Schwartz, R. Comparing apples and oranges: Query trade-off in submodular maximization. *Mathematics of Operations Research*, 42(2): 308–329, 2017.

Chawla, R., Sankararaman, A., Ganesh, A., and Shakkottai, S. The gossiping insert-eliminate algorithm for multi-agent bandits. In *International conference on artificial intelligence and statistics*, pp. 3471–3481. PMLR, 2020.

Chen, L., Xu, J., and Lu, Z. Contextual combinatorial multi-armed bandits with volatile arms and submodular reward. *Advances in Neural Information Processing Systems*, 31: 3247–3256, 2018.

Edmonds, J. Submodular functions, matroids, and certain polyhedra. In *Combinatorial Optimization—Eureka, You Shrink!*, pp. 11–26. Springer, 2003.

Elgabli, A., Issaid, C. B., Bedi, A. S., Rajawat, K., Bennis, M., and Aggarwal, V. Fednew: A communication-efficient and privacy-preserving newton-type method for federated learning. In *International Conference on Machine Learning*, pp. 5861–5877. PMLR, 2022.

Feige, U. A threshold of ln n for approximating set cover. *Journal of the ACM (JACM)*, 45(4):634–652, 1998.

Feige, U., Mirrokni, V. S., and Vondrák, J. Maximizing non-monotone submodular functions. *SIAM Journal on Computing*, 40(4):1133–1153, 2011.

Fourati, F., Aggarwal, V., Quinn, C., and Alouini, M.-S. Randomized greedy learning for non-monotone stochastic submodular maximization under full-bandit feedback. In *International Conference on Artificial Intelligence and Statistics*, pp. 7455–7471. PMLR, 2023a.

Fourati, F., Kharrat, S., Aggarwal, V., Alouini, M.-S., and Canini, M. FilFL: Client filtering for optimized client participation in federated learning. *arXiv preprint arXiv:2302.06599*, 2023b.

Fourati, F., Quinn, C. J., Alouini, M.-S., and Aggarwal, V. Combinatorial stochastic-greedy bandit. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 12052–12060, 2024.

Goemans, M. X. and Williamson, D. P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

He, J., Wang, T., Min, Y., and Gu, Q. A simple and provably efficient algorithm for asynchronous federated contextual linear bandits. *Advances in neural information processing systems*, 35:4762–4775, 2022.

Hoeffding, W. Probability inequalities for sums of bounded random variables. In *The collected works of Wassily Hoeffding*, pp. 409–426. Springer, 1994.

Hosseinalipour, S., Brinton, C. G., Aggarwal, V., Dai, H., and Chiang, M. From federated to fog learning: Distributed machine learning over heterogeneous wireless networks. *IEEE Communications Magazine*, 58(12):41–47, 2020.

Iwata, S., Fleischer, L., and Fujishige, S. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM (JACM)*, 48(4): 761–777, 2001.

Khuller, S., Moss, A., and Naor, J. S. The budgeted maximum coverage problem. *Information processing letters*, 70(1):39–45, 1999.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Korte, B. H., Vygen, J., Korte, B., and Vygen, J. *Combinatorial optimization*, volume 1. Springer, 2011.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. *Canadian Institute for Advanced Research, 2009. URL http://www.cs.toronto.edu/ kriz/cifar.html*, 2009.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Li, C. and Wang, H. Asynchronous upper confidence bound algorithms for federated linear bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 6529–6553. PMLR, 2022.

Li, C., Wang, H., Wang, M., and Wang, H. Learning kernelized contextual bandits in a distributed and asynchronous environment. In *International Conference on Learning Representation*, 2023.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Lin, H. and Bilmes, J. A class of submodular functions for document summarization. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pp. 510–520, 2011.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mirzasoleiman, B., Karbasi, A., Sarkar, R., and Krause, A. Distributed submodular maximization: Identifying representative elements in massive data. *Advances in Neural Information Processing Systems*, 26, 2013.

Mirzasoleiman, B., Badanidiyuru, A., Karbasi, A., Vondrák, J., and Krause, A. Lazier than lazy greedy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Nemhauser, G. L. and Wolsey, L. A. Best algorithms for approximating the maximum of a submodular set function. *Mathematics of operations research*, 3(3):177–188, 1978.

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978.

Niazadeh, R., Golrezaei, N., Wang, J. R., Susan, F., and Badanidiyuru, A. Online learning via offline greedy algorithms: Applications in market design and optimization. In *Proceedings of the 22nd ACM Conference on Economics and Computation*, pp. 737–738, 2021.

Nie, G., Agarwal, M., Umrawal, A. K., Aggarwal, V., and Quinn, C. J. An explore-then-commit algorithm for submodular maximization under full-bandit feedback. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022.

Nie, G., Nadew, Y. Y., Zhu, Y., Aggarwal, V., and Quinn, C. J. A framework for adapting offline algorithms to solve combinatorial multi-armed bandit problems with bandit feedback. *International Conference on Machine Learning*, 2023.

Sivasubramanian, D., Nagalapatti, L., Iyer, R., and Ramakrishnan, G. Gradient coreset for federated learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2648–2657, 2024.

Slivkins, A. et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Sviridenko, M. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters*, 32(1):41–43, 2004.

Tajdini, A., Jain, L., and Jamieson, K. Minimax optimal submodular optimization with bandit feedback. *arXiv preprint arXiv:2310.18465*, 2023.

Takemori, S., Sato, M., Sonoda, T., Singh, J., and Ohkuma, T. Submodular bandit problem under multiple constraints. In *Conference on Uncertainty in Artificial Intelligence*, pp. 191–200. PMLR, 2020.

Wang, P.-A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 4120–4129. PMLR, 2020.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yaroslavtsev, G., Zhou, S., and Avdiukhin, D. "bring your own greedy"+ max: Near-optimal 1/2-approximations for submodular knapsack. In *International Conference on Artificial Intelligence and Statistics*, pp. 3263–3274. PMLR, 2020.

A. Notation

In the following, for a given agent i when we discuss any feasible action denoted as A, we use $f_t(A)$ to represent the realization of the stochastic reward at time t when taking that action. We denote the expectation of the reward of playing that action as $f(A)$. We also introduce $\bar{f}_t(A)$, which is the empirical mean of rewards received from playing action A up to and including time t.

We omit the subscript t when we write $\bar{f}(A)$, assuming it is clear that action A has been played $r^\star$ times. We use A_j, with j ranging from 1 to the total number of the approximation algorithm queries $N(\beta, \gamma, \psi, \epsilon^\star)$, to refer to the j-th action the algorithm queries. Additionally, we define T_j, where j also varies from 1 to the total number of the offline algorithm queries $N(\beta, \gamma, \psi, \epsilon^\star)$, as the time step when A_j has been played $r^\star$ times.

B. Communication Rounds Analysis

In this subsection we prove Lemma 5.1 which upperbounds the number of oracle queries $N(\beta, \gamma, \psi, \epsilon^\star)$ of the offline algorithm $\mathcal{A}(\epsilon^\star)$ as follows:

$$N(\beta, \gamma, \psi, \epsilon^\star) \leq \mathcal{O}(\psi T^{\frac{\beta}{\beta+1}} \log^\gamma(T))$$

We provide examples of the resulting number of oracle queries for different offline algorithms or class transformations in Table 2, as shown in Lemma 5.1.

Proof. Using an $(\alpha, \beta, \gamma, \psi, \delta)$-robust approximation as subroutine, we have after $N(\beta, \gamma, \psi, \epsilon) = \psi \frac{1}{\epsilon^\beta} \log^\gamma(\frac{1}{\epsilon})$ oracle calls, for $\epsilon \geq 0$, $\beta \geq 0$, and $\gamma \in \{0, 1\}$.

Recall that $\epsilon^\star$ is as follows:

$$\epsilon^\star = (\frac{\psi r^\star}{Tm})^{\frac{1}{\beta+1}} \mathbf{1}_{\{\beta>0 \text{ OR } \gamma>0\}}. \tag{13}$$

Recall that $\beta \geq 0$ and $\gamma \in \{0, 1\}$. Therefore, we consider all the possible cases (three), the first when $\beta = \gamma = 0$, the second when $\beta = 0$ and $\gamma > 0$, and the third when $\beta > 0$.

Case 1: $\beta = \gamma = 0$**.** Using Eq.(3), $\epsilon^\star$ becomes the following

$$\epsilon^\star = 0. \tag{14}$$

$$\begin{aligned} N(\beta, \gamma, \psi, \epsilon^\star) &= 2\psi \\ &= \mathcal{O}(\psi T^{\frac{\beta}{\beta+1}} \log^\gamma(T)) \end{aligned} \tag{15}$$

Case 2: $\beta = 0$ **and** $\gamma \neq 0$**.** We have $\gamma \neq 0$ and $\gamma \in \{0, 1\}$, thus $\gamma = 1$.

Using Eq.(3), $\epsilon^\star$ becomes the following

$$\epsilon^\star = \frac{\psi r^\star}{Tm}. \tag{16}$$

Offline Algorithm	Offline Complexity	C-MA-MAB Resulted Communication Complexity
RANDOMIZEDUSM (Buchbinder et al., 2015)	$\mathcal{O}(n)$	$\mathcal{O}(n)$
GREEDY (Nemhauser et al., 1978)	$\mathcal{O}(nk)$	$\mathcal{O}(nk)$
GENERAL Excluding the next rows	$\mathcal{O}(\psi)$	$\mathcal{O}(\psi)$
STOCHASTIC-GREEDY (Mirzasoleiman et al., 2015)	$\mathcal{O}\left(n\log(\frac{1}{\epsilon})\right)$	$\mathcal{O}(n\log(T))$
MORE GENERAL Excluding the next rows	$\mathcal{O}(\psi\log^{\gamma}(\frac{1}{\epsilon}))$	$\mathcal{O}(\psi\log^{\gamma}(T))$
RANDOMSAMPLING (Buchbinder et al., 2017)	$\mathcal{O}\left(\frac{n}{\epsilon^2}\log(\frac{1}{\epsilon})\right)$	$\mathcal{O}(\psi T^{\frac{2}{3}}\log(T))$
MOST GENERAL Including the previous rows	$\mathcal{O}(\frac{\psi}{\epsilon^{\beta}}\log^{\gamma}(\frac{1}{\epsilon}))$	$\mathcal{O}(\psi T^{\frac{\beta}{\beta+1}}\log^{\gamma}(T))$

Table 2: The table shows the resulted communication complexity with different offline algorithms. We use $\tilde{\mathcal{O}}$ to simplify expressions. Key parameters include horizon T, number of communicating agents m, base arm count n, and cardinality constraint k. Each row presents a specific offline algorithm or a class transformation with a given offline complexity. For the general rows, we consider classes of offline algorithms, with an approximation error factor ϵ, with general complexity forms, with general constants, $\psi \geq 0$, $\beta \geq 0$, and $\gamma \in \{0, 1\}$.

$$
\begin{aligned}
N(\beta,\gamma,\psi,\epsilon^{\star}) &= 2\psi(\frac{1}{\epsilon^{\star}})^{\beta}\log^{\gamma}(\frac{1}{\epsilon^{\star}}) \\
&= 2\psi\log(\frac{1}{\epsilon^{\star}}) \\
&= 2\psi\log(\frac{Tm}{\psi r^{\star}}) \\
&= 2\psi\log(\frac{T}{\psi r_i^{\star}}) \\
&\leq 2\psi\log(T) \\
&= \mathcal{O}(\psi\log(T)) \\
&= \mathcal{O}(\psi T^{\frac{\beta}{\beta+1}}\log^{\gamma}(T))
\end{aligned} \tag{17}
$$

Case 3: $\beta > 0$**.** Using Eq.(3), $\epsilon^{\star}$ becomes the following

$$
\epsilon^{\star} = (\frac{\psi r^{\star}}{Tm})^{\frac{1}{\beta+1}}. \tag{18}
$$

Therefore,

$$
\begin{aligned}
N(\beta,\gamma,\psi,\epsilon^{\star}) &= 2\psi(\frac{1}{\epsilon^{\star}})^{\beta}\log^{\gamma}(\frac{1}{\epsilon^{\star}}) \\
&= 2\psi(\frac{1}{\beta+1})^{\gamma}(\frac{Tm}{\psi r^{\star}})^{\frac{\beta}{\beta+1}}\log^{\gamma}(\frac{Tm}{\psi r^{\star}}) \\
&= 2\psi(\frac{1}{\beta+1})^{\gamma}(\frac{T}{\psi r_i^{\star}})^{\frac{\beta}{\beta+1}}\log^{\gamma}(\frac{T}{\psi r_i^{\star}}) \\
&\leq 2\psi(\frac{1}{\beta+1})^{\gamma}(T)^{\frac{\beta}{\beta+1}}\log^{\gamma}(T) \\
&= \mathcal{O}(\psi T^{\frac{\beta}{\beta+1}}\log^{\gamma}(T))
\end{aligned} \tag{19}
$$

□

C. Estimation Analysis

Hoeffding's inequality (Hoeffding, 1994) is a powerful technique for bounding probabilities of bounded random variables. We state the inequality, then we use it to show that $\mathcal{E}$ happens with high probability.

Lemma C.1 (Hoeffding's inequality)**.** *Let $X_1, \cdots, X_n$ be independent random variables bounded in the interval $[0, 1]$, and let $\bar{X}$ denote their empirical mean. Then we have for any $\xi > 0$,*

$$\mathbb{P}\left(\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq \xi\right) \leq 2\exp\left(-2n\xi^2\right). \tag{20}$$

We use the above Lemma to prove Lemma 5.2 which bounds the probability of the *good event* $\mathcal{E}$ as follows:

$$\mathbb{P}(\mathcal{E}) \geq 1 - \frac{2N(\beta, \gamma, \psi, \epsilon^\star)}{T^2}.$$

Proof. Applying the Hoeffding bound in Lemma C.1 to the empirical mean $\bar{f}(A_i)$ of A_i resulted from $r^\star$ independent rewards and with $\xi = \sqrt{\log(T)/r^\star}$ provides

$$\begin{aligned}
\mathbb{P}(\bar{\mathcal{E}}_i) &= \mathbb{P}\left[\left|\bar{f}(A_i) - f(A_i)\right| \geq \xi\right] \\
&\leq 2\exp\left(-2r^\star\xi^2\right) \\
&= 2\exp\left(-2r^\star(\log(T)/r^\star)\right) \\
&= 2\exp\left(-2\log(T)\right) \\
&= \frac{2}{T^2}.
\end{aligned} \tag{21}$$

Then, we can bound the probability of the *good event* as follows:

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &= \mathbb{P}(\mathcal{E}_1 \cap \cdots \cap \mathcal{E}_{N(\beta,\gamma,\psi,\epsilon^\star)}) \\
&= 1 - \mathbb{P}(\bar{\mathcal{E}}_1 \cup \cdots \cup \bar{\mathcal{E}}_{N(\beta,\gamma,\psi,\epsilon^\star)}) && \text{(De Morgan's Law)} \\
&\geq 1 - \sum_{i=1}^{N(\beta,\gamma,\psi,\epsilon^\star)} \mathbb{P}(\bar{\mathcal{E}}_i) && \text{(union bounds)} \\
&\geq 1 - \frac{2N(\beta, \gamma, \psi, \epsilon^\star)}{T^2}. && \text{(using (21))}
\end{aligned}$$

□

D. regret Analysis

We establish Theorem 5.3 outlined in Section 4.1 of the main paper. This theorem asserts that for the sequential decision-making scenario delineated in Section 2, the expected cumulative α-regret of the C-MA-MAB, employing an $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation algorithm $\mathcal{A}(\epsilon)$ as a component, is bounded by $\tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}}m^{-\frac{1+\beta}{3+\beta}}T^{\frac{2+\beta}{3+\beta}}\right)$.

We separate the proof into two cases. One case when the *good event* $\mathcal{E}$ happens, which we show in Lemma 5.2 happens with high probability and then we generalize the result under any event.

D.1. regret of an Agent under the Good Event

We upper-bound the expected α-regret given the occurrence of the *good event* $\mathcal{E}$. All expectations are conditioned on $\mathcal{E}$ throughout this section. However, for the sake of simplicity in notation, we employ $\mathbb{E}[\cdot]$ rather than $\mathbb{E}[\cdot|\mathcal{E}]$ in certain instances.

We break down the anticipated α-regret given $\mathcal{E}$ into two parts: one dealing with regret arising from exploration and the other from exploitation. It is crucial to remember that $f_t(A_t)$ denotes the stochastic reward obtained by selecting action A_t,

a variable dependent on the historical means of actions in previous rounds.

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] &= \sum_{t=1}^{T}(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f_t(S_{i,t})]) \\
&= \sum_{t=1}^{T}(\alpha f(\mathrm{S}^\star) - \mathbb{E}[\mathbb{E}[f_t(S_{i,t})|S_{i,t}]]) && (22)\\
&= \sum_{t=1}^{T}(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(S_{i,t})]) && (f(\cdot) \text{ defined as expected reward})\\
&= \sum_{j=1}^{N(\beta,\gamma,\psi,\epsilon^\star)} r_i^\star\left(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(A_j)]\right) + \sum_{t=T_{N(\beta,\gamma,\psi,\epsilon^\star)}+1}^{T}(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(A_t)])\\
&= \underbrace{\sum_{j=1}^{N(\beta,\gamma,\psi,\epsilon^\star)} r_i^\star\left(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(A_j)]\right)}_{\text{Agent } i \text{ Exploration regret}} + \underbrace{\sum_{t=T_{N(\beta,\gamma,\psi,\epsilon^\star)}+1}^{T}(\alpha f(S^\star) - \mathbb{E}[f(\Theta)])}_{\text{Agent } i \text{ Exploitation regret}}. && (23)
\end{aligned}$$

We establish distinct bounds for the regret stemming from exploration and exploitation individually.

Bounding the agent i exploration regret:

$$\begin{aligned}
\sum_{j=1}^{N(\beta,\gamma,\psi,\epsilon^\star)} r_i\left(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(A_i)]\right) &\le \sum_{j=1}^{N(\beta,\gamma,\psi,\epsilon^\star)} \frac{r^\star}{m}(\alpha - 0) && (\text{rewards in } [0,1])\\
&\le N(\beta,\gamma,\psi,\epsilon^\star)\frac{r^\star}{m}. && (24)
\end{aligned}$$

Bounding exploitation regret: When the *good event* $\mathcal{E}$ occurs, we know that $|\bar{f}(A) - f(A)| \le \xi$ for all considered action A. For C-MA-MAB framework, with exploitation set Θ, with an $(\alpha,\beta,\gamma,\psi,\delta)$-resilient-approximation $\mathcal{A}(\epsilon)$ as a subroutine, we have after $N(\beta,\gamma,\psi,\epsilon) = \psi\frac{1}{\epsilon^\beta}\log^\gamma(\frac{1}{\epsilon})$ oracle calls, for $\epsilon \ge 0$, $\beta \ge 0$, and $\gamma \in \{0,1\}$,

$$\mathbb{E}[f(\Theta)] \ge (\alpha - \epsilon)f(\mathrm{S}^\star) - \delta\xi. \tag{25}$$

Therefore, using $\epsilon^\star$, we have

$$\alpha f(S^\star) - \mathbb{E}[f(\Theta)] \le \delta\xi + \epsilon^\star f(S^\star). \tag{26}$$

Therefore, we can bound the exploitation regret as follows:

$$\begin{aligned}
\sum_{t=T_{N(\beta,\gamma,\psi,\epsilon^\star)}+1}^{T}(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(S)]) &\le \sum_{t=T_{N(\beta,\gamma,\psi,\epsilon^\star)}+1}^{T}(\delta\xi + \epsilon^\star f(\mathrm{S}^\star)) && (\text{using } (26))\\
&\le \sum_{t=T_{N(\beta,\gamma,\psi,\epsilon^\star)}+1}^{T}(\delta\xi + \epsilon^\star) && (\text{rewards are bounded in } [0,1])\\
&\le T(\delta\xi + \epsilon^\star). && (27)
\end{aligned}$$

Bounding total regret:

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] &= \sum_{i=1}^{N(\beta,\gamma,\psi,\epsilon^\star)} \frac{r^\star}{m}\left(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(A_i)]\right) + \sum_{t=T_{N(\beta,\gamma,\psi,\epsilon^\star)}+1}^{T}(\alpha f(\mathrm{S}^\star) - \mathbb{E}[f(S)]) && (\text{copying } (23))\\
&\le N(\beta,\gamma,\psi,\epsilon^\star)\frac{r^\star}{m} + T(\delta\xi + \epsilon^\star) && (\text{using } (24), (27))
\end{aligned}$$

Using the defined confidence radius $\xi = \sqrt{\log(T)/r^\star}$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] &\leq N(\beta,\gamma,\psi,\epsilon^\star)\frac{r^\star}{m} + \epsilon^\star T + T\delta\sqrt{\log(T)/r^\star} \\ &\leq \psi\frac{1}{\epsilon^{\star\beta}}\log^\gamma(\frac{1}{\epsilon^\star})\frac{r^\star}{m} + \epsilon^\star T + T\delta\sqrt{\log(T)/r^\star} \end{aligned} \tag{28}$$

We note that the inequality in Eq.(28) is correct for all values of $r^\star \geq m$ and $\epsilon^\star \geq 0$ with the convention that $0^0 = 1$.

In our algorithm, we choose the following values of $r^\star$ and $\epsilon^\star$ as functions of the problem parameters such as the range T and the number of available agents m as well as on the subroutine algorithm approximation resilience parameter guarantees, that are β, γ, ψ, and δ. Specifically, using an $(\alpha,\beta,\gamma,\psi,\delta)$-robust approximation as subroutine, we choose $r^\star$ as follows:

$$r^\star = m\left\lceil m^{-1}\left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)^{\frac{1}{\beta+1}}\right)^{\frac{2+2\beta}{3+\beta}}\right\rceil, \tag{29}$$

For clarity in the following steps we further define z as follows:

$$z = \left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)^{\frac{1}{\beta+1}}\right)^{\frac{2+2\beta}{3+\beta}}, \tag{30}$$

Moreover, we choose $\epsilon^\star$ as follows:

$$\epsilon^\star = (\frac{\psi r^\star}{Tm})^{\frac{1}{\beta+1}}\mathbf{1}_{\{\beta>0 \text{ OR } \gamma>0\}}. \tag{31}$$

Therefore, from Eq. (28),

$$\begin{aligned}\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] &\leq \psi(\frac{1}{\epsilon^\star})^\beta\log^\gamma(\frac{1}{\epsilon^\star})\frac{r^\star}{m} + \epsilon^\star T + T\delta\sqrt{\log(T)/r^\star} && (32)\\ &\leq \psi(\frac{1}{\epsilon^\star})^\beta\log^\gamma(\frac{1}{\epsilon^\star})\left\lceil\frac{z}{m}\right\rceil + \epsilon^\star T + T\delta\sqrt{\log(T)/(m\left\lceil\frac{z}{m}\right\rceil)} && (33)\\ &\leq \psi(\frac{1}{\epsilon^\star})^\beta\log^\gamma(\frac{1}{\epsilon^\star})\left\lceil\frac{z}{m}\right\rceil + \epsilon^\star T + T\delta\sqrt{\log(T)/(m\frac{z}{m})} && \text{(Since } \lceil z/m\rceil \geq z/m) \\ &\leq 2\psi(\frac{1}{\epsilon^\star})^\beta\log^\gamma(\frac{1}{\epsilon^\star})\frac{z}{m} + \epsilon^\star T + T\delta\sqrt{\log(T)/z} && \text{(Since } z \geq m,\ \lceil z/m\rceil \leq 2z)\end{aligned}$$

where the last inequality follows when $z \geq m$, which holds for $T \geq \frac{\psi m^{\frac{1+\beta}{2}}}{\delta^{\beta+1}}$.

Therefore,

$$\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] \leq 2\psi(\frac{1}{\epsilon^\star})^\beta\log^\gamma(\frac{1}{\epsilon^\star})\frac{z}{m} + \epsilon^\star T + T\delta\sqrt{\log(T)/z} \tag{34}$$

Recall that $\beta \geq 0$ and $\gamma \in \{0,1\}$. Therefore, we consider all the possible cases (three), the first when $\beta = \gamma = 0$, the second when $\beta = 0$ and $\gamma > 0$, and the third when $\beta > 0$.

Case 1: $\beta = \gamma = 0$**.**

In this case using Eq. (30), z becomes the following

$$z = \left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)\right)^{\frac{2}{3}}. \tag{35}$$

Moreover, using Eq.(3), $\epsilon^\star$ becomes the following

$$\epsilon^\star = 0. \tag{36}$$

From Eq.(34) we have

$$\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] \leq \frac{2\psi}{m}\left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)\right)^{\frac{2}{3}} + T\delta\sqrt{\log(T)}\left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)\right)^{-\frac{1}{3}} \tag{37}$$

$$\leq 2\frac{\psi}{\psi^{\frac{2}{3}}}\frac{m^{\frac{2}{3}}}{m}\delta^{\frac{2}{3}}\log(T)^{\frac{1}{3}}T^{\frac{2}{3}} + T^{\frac{2}{3}}\delta^{\frac{2}{3}}\log(T)^{\frac{1}{3}}m^{-\frac{1}{3}}\psi^{\frac{1}{3}} \tag{38}$$

$$= \mathcal{O}\left(\delta^{\frac{2}{3}}\psi^{\frac{1}{3}}m^{-\frac{1}{3}}T^{\frac{2}{3}}\log(T)^{\frac{1}{3}}\right). \tag{39}$$

$$= \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3}}\psi^{\frac{1}{3}}m^{-\frac{1}{3}}T^{\frac{2}{3}}\right) \tag{40}$$

$$= \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}}m^{-\frac{1}{3+\beta}}T^{\frac{2+\beta}{3+\beta}}\right). \quad \text{(Given that } \beta = 0\text{)}$$

Case 2: $\beta = 0$ **and** $\gamma \neq 0$**.**

We have $\gamma \neq 0$ and $\gamma \in \{0, 1\}$, thus $\gamma = 1$. In this case using Eq. (30), z becomes the following

$$z = \left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)\right)^{\frac{2}{3}}. \tag{41}$$

Moreover, using Eq.(3), $\epsilon^\star$ becomes the following

$$\epsilon^\star = \frac{\psi r^\star}{Tm}. \tag{42}$$

From Eq. (34) we have

$$\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] \leq \frac{2\psi}{m}\log(\frac{Tm}{\psi r^\star})z + \frac{\psi r^\star}{Tm}T + T\delta\sqrt{\log(T)}\left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)\right)^{-\frac{1}{3}} \tag{43}$$

$$\leq \frac{2\psi}{m}\log(\frac{Tm}{\psi r^\star})z + \frac{2\psi}{m}z + T\delta\sqrt{\log(T)}\left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)\right)^{-\frac{1}{3}} \tag{44}$$

$$\leq \frac{2\psi}{m}\log(T)z + \frac{2\psi}{m}z + T\delta\sqrt{\log(T)}\left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)\right)^{-\frac{1}{3}} \tag{45}$$

$$\leq \frac{4\psi}{m}\log(T)\left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)\right)^{\frac{2}{3}} + T\delta\sqrt{\log(T)}\left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)\right)^{-\frac{1}{3}} \tag{46}$$

$$= \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3}}\psi^{\frac{1}{3}}m^{-\frac{1}{3}}T^{\frac{2}{3}}\right) \tag{47}$$

$$= \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}}m^{-\frac{1}{3+\beta}}T^{\frac{2+\beta}{3+\beta}}\right). \quad \text{(Given that } \beta = 0\text{)}$$

Case 3: $\beta > 0$**.** In this case using Eq. (30), z becomes the following

$$z = \left(\delta\sqrt{\log(T)}\left(\frac{Tm}{\psi}\right)^{\frac{1}{\beta+1}}\right)^{\frac{2+2\beta}{3+\beta}}. \tag{48}$$

Moreover, using Eq.(3), $\epsilon^\star$ becomes the following

$$\epsilon^\star = (\frac{\psi r^\star}{Tm})^{\frac{1}{\beta+1}}. \tag{49}$$

From Eq.(34) we have

$$\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] \leq \frac{2\psi}{m}(\frac{1}{\epsilon^\star})^\beta\log^\gamma(\frac{1}{\epsilon^\star})z + \epsilon^\star T + T\delta\sqrt{\log(T)}\,(z)^{-\frac{1}{2}} \tag{50}$$

Therefore, with $\beta > 0$ it exists a constant C sufficiently large such as:

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] &\leq \frac{2C\psi}{m}(\frac{1}{\epsilon^\star})^\beta z + \epsilon^\star T + T\delta\sqrt{\log(T)}\,(z)^{-\frac{1}{2}} && (51)\\
&\leq \frac{2C\psi}{m}(\frac{Tm}{\psi r^\star})^{\frac{\beta}{\beta+1}} z + (\frac{\psi r^\star}{Tm})^{\frac{1}{\beta+1}} T + T\delta\sqrt{\log(T)}\,(z)^{-\frac{1}{2}} && (52)\\
&\leq \frac{2C\psi}{m}(\frac{Tm}{\psi z})^{\frac{\beta}{\beta+1}} z + (\frac{2\psi z}{Tm})^{\frac{1}{\beta+1}} T + T\delta\sqrt{\log(T)}\,(z)^{-\frac{1}{2}} && (53)\\
&\leq D\frac{\psi}{m}(\frac{Tm}{\psi z})^{\frac{\beta}{\beta+1}} z + T\delta\sqrt{\log(T)}\,(z)^{-\frac{1}{2}} && (54)\\
&= \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right).
\end{aligned}$$

In conclusion, for all the cases, if the *bad event* $\mathcal{E}$ happens, the expected α-regret of our C-MA-MAB with an $(\alpha, \beta, \gamma, \psi, \delta)$-resilient-approximation as a subroutine is upper bounded by as follows:

$$\mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}] \leq \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right). \tag{55}$$

D.2. Regret of an Agent under Any Event

By Lemma 5.2, the probability of the *bad event* is upper-bounded as follows

$$\mathbb{P}(\bar{\mathcal{E}}) \leq \frac{2N(\beta, \gamma, \psi, \epsilon^\star)}{T^2}. \tag{56}$$

Since the reward function $f_t(\cdot)$ is upper bounded by 1, the expected α-regret incurred under $\bar{\mathcal{E}}$ for a range T is limited to a maximum of T,

$$\mathbb{E}[\mathcal{R}_i(T)|\bar{\mathcal{E}}] \leq T. \tag{57}$$

Combining the results when *bad event* happens and does not happen, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{R}_i(T)] &= \mathbb{E}[\mathcal{R}_i(T)|\mathcal{E}]\cdot\mathbb{P}(\mathcal{E}) + \mathbb{E}[\mathcal{R}_i(T)|\bar{\mathcal{E}}]\cdot\mathbb{P}(\bar{\mathcal{E}}) && \text{(Law of total expectation)}\\
&\leq \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right) 1 + T\cdot\frac{2N(\beta, \gamma, \psi, \epsilon^\star)}{T^2} && \text{(using (55), (56), and (57))}\\
&\leq \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right) 1 + \frac{2N(\beta, \gamma, \psi, \epsilon^\star)}{T} && (58)\\
&\leq \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right) 1 + \frac{\tilde{\mathcal{O}}(\psi^{\frac{1}{3+\beta}}\psi^{\frac{2+\beta}{3+\beta}} T^{\frac{\beta}{\beta+1}})}{T} && \text{(using Lemma 5.1)}\\
&= \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right) + \tilde{\mathcal{O}}(\psi^{\frac{1}{3+\beta}}\psi^{\frac{2+\beta}{3+\beta}} T^{-\frac{1}{\beta+1}}) && (59)\\
&= \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right) + \tilde{\mathcal{O}}(\psi^{\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}} m^{-\frac{2+\beta}{3+\beta}} T^{-\frac{1}{\beta+1}}) && \text{(using } T \geq m\psi)\\
&= \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right) + \tilde{\mathcal{O}}(\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}-\frac{1+\beta}{3+\beta}} T^{\frac{2+\beta}{3+\beta}-\frac{1}{\beta+1}}) && (60)\\
&= \tilde{\mathcal{O}}\left(\delta^{\frac{2}{3+\beta}}\psi^{\frac{1}{3+\beta}} m^{-\frac{1}{3+\beta}} T^{\frac{2+\beta}{3+\beta}}\right).
\end{aligned}$$

This concludes the proof.

E. Application to Submodular Maximization

We apply our general framework to online stochastic submodular maximization. In the following, we provide examples of submodular maximization problems. We provide results for unconstrained, under cardinality, and under knapsack constraints, by showing the resilience of different offline algorithms in these settings.

E.1. Submodular Maximization Examples

Submodularity arises in critical contexts within combinatorial optimization, such as graph cuts (Goemans & Williamson, 1995; Iwata et al., 2001), rank functions of matroids (Edmonds, 2003), and set covering problems (Feige, 1998). Furthermore, recent works have demonstrated that various real-world problems exhibit submodularity, including data summarization and coreset selection for model training (Mirzasoleiman et al., 2015; 2020), client participation optimization in FL (Balakrishnan et al., 2022; Fourati et al., 2023b), recommendation systems (Takemori et al., 2020), crowdsourcing, crowdsensing, and influence maximization (Fourati et al., 2024).

E.2. Unconstrained Submodular Maximization

Lemma E.1 (Generalization of Corollary 2 of (Fourati et al., 2023a))**.** RANDOMIZEDUSM *(Buchbinder et al., 2015) is a* $(\frac{1}{2}, 0, 0, 4n, \frac{5}{2}n)$*-resilient-approximation algorithm for unconstrained non-monotone SM.*

Proof. The offline RandomizedUSM $\frac{1}{2}$-approximation algorithm proposed in (Buchbinder et al., 2015) requires $4n$ oracle calls, thus $\alpha = 1/2$, $\psi = 4n$, $\gamma = 0$, and $\beta = 0$. Furthermore, as shown in Corollary 2 of (Fourati et al., 2023a), using a ξ-controlled-estimation $\bar{f}$ of the function f,

$$\mathbb{E}[f(X_n)] \geq \frac{1}{2}\mathbb{E}[f(OPT)] - \frac{5}{2}n\,\xi. \tag{61}$$

Therefore, the RandomizedUSM is an $(\frac{1}{2}, 0, 0, 4n, \frac{5}{2}n)$-resilient-approximation algorithm for unconstrained non-monotone submodular maximization problem. □

E.3. Submodular Maximization with Cardinality Constraint

Lemma E.2 (Generalization of Corollary 4.3 of (Nie et al., 2022))**.** GREEDY *in (Nemhauser et al., 1978) is a* $(1 - \frac{1}{e}, 0, 0, nk, 2k)$*-resilient-approximation algorithm for monotone SM under a cardinality constraint* k*.*

Proof. The offline GREEDY $(1 - 1/e)$-approximation algorithm proposed in (Nemhauser et al., 1978) requires nk oracle calls, thus $\alpha = (1 - 1/e)$, $\psi = nk$, $\gamma = 0$, and $\beta = 0$. Furthermore, as shown in Corollary 4.3 of (Nie et al., 2022), using a ξ-controlled-estimation $\bar{f}$ of the function f,

$$\mathbb{E}[f(X_n)] \geq (1 - 1/e)\mathbb{E}[f(OPT)] - 2k\,\xi. \tag{62}$$

Therefore, the GREEDY is an $(1 - \frac{1}{e}, 0, 0, nk, 2k)$-resilient-approximation algorithm algorithm for monotone submodular maximization under a k-cardinality constraint. □

Lemma E.3 (Generalization of Corollary 1 of (Fourati et al., 2024))**.** STOCHASTIC-GREEDY *in (Mirzasoleiman et al., 2015) is a* $(1 - \frac{1}{e}, 0, 1, n, 2k)$*-resilient-approximation algorithm for monotone SM under a cardinality constraint* k*.*

Proof. The offline STOCHASTIC-GREEDY $(1 - 1/e - \epsilon)$-approximation algorithm proposed in (Mirzasoleiman et al., 2015) requires $n\log(\frac{1}{\epsilon})$ oracle calls, thus $\alpha = (1 - 1/e)$, $\psi = n$, $\gamma = 1$, and $\beta = 0$. Furthermore, as shown in Corollary 1 of (Fourati et al., 2024), using a ξ-controlled-estimation $\bar{f}$ of the function f,

$$\mathbb{E}[f(X_n)] \geq (1 - 1/e)\mathbb{E}[f(OPT)] - 2k\,\xi. \tag{63}$$

Therefore, the STOCHASTIC-GREEDY is an $(1 - \frac{1}{e}, 0, 1, n, 2k)$-resilient-approximation algorithm algorithm for monotone submodular maximization under a k-cardinality constraint.

□

Lemma E.4. RANDOMSAMPLING *in (Buchbinder et al., 2017) is a* $(\frac{1}{e}, 2, 1, n, 4k)$*-resilient-approximation algorithm for non-monotone SM under a cardinality constraint* k*.*

Proof. The offline RANDOMSAMPLING $(1/e - \epsilon)$-approximation algorithm proposed in (Buchbinder et al., 2017) requires $n\frac{1}{\epsilon^2}\log(\frac{1}{\epsilon})$ oracle calls, thus $\alpha = (1/e)$, $\psi = n$, $\gamma = 1$, and $\beta = 2$.

By design of the algorithm, we have

$$\begin{aligned} f(S_i) - f(S_{i-1}) &\geq \bar{f}(S_i) - \bar{f}(S_{i-1}) - 2\xi && \text{(using } \xi\text{-controlled estimation)} \\ &= \max\left\{\bar{f}(S_{i-1} \cup \{u_i\}) - f(S_{i-1}), 0\right\} - 2\xi && \text{(by algorithm)} \end{aligned}$$

Similar to the steps in (Buchbinder et al., 2017), let A_i represent an event encompassing all random decisions made by the algorithm up to iteration i (excluding it). In the initial segment of the proof, we select a specific iteration $1 \leq i \leq k$ and an associated event A_i. All probabilities and expectations within this proof portion are implicitly conditioned on A_i. It is important to note that, when conditioned on A_i, the set S_{i-1} becomes deterministic. We use $v_1, v_2, \ldots, v_k$ to denote the k elements of $\mathcal{N}$ (including S_{i-1}) with the highest marginal contribution to S_{i-1}, arranged in non-increasing order of marginal contribution. Additionally, let X_j be an indicator for the event $u_i = v_j$.

$$\mathbb{E}\left[f(S_i) - f(S_{i-1})\right] = \mathbb{E}\left[\max\left\{\bar{f}(S_{i-1} \cup \{u_i\}) - \bar{f}(S_{i-1}), 0\right\}\right] - 2\xi \tag{64}$$

$$\geq \sum_{j=1}^{k}\left[\mathbb{E}\left[X_j\right] \cdot \max\left\{\bar{f}(S_{i-1} \cup \{v_j\}) - \bar{f}(S_{i-1}), 0\right\}\right] - 2\xi \tag{65}$$

$$\geq \frac{\sum_{j=1}^{k}\mathbb{E}\left[X_j\right] \cdot \sum_{j=1}^{k}\max\left\{\bar{f}(S_{i-1} \cup \{v_j\}) - \bar{f}(S_{i-1}), 0\right\}}{k} - 2\xi. \tag{66}$$

Where the last inequality holds by Chebyshev's sum inequality since $\max\left\{\bar{f}(S_{i-1} \cup \{v_j\}) - \bar{f}(S_{i-1}), 0\right\}$ is non-increasing in j by definition and $\mathbb{E}\left[X_j\right]$ is non-increasing in j by Lemma 4.2 in (Buchbinder et al., 2017). Therefore,

$$\begin{aligned} \sum_{j=1}^{k}\max\left\{\bar{f}(S_{i-1} \cup \{v_j\}) - \bar{f}(S_{i-1}), 0\right\} &\geq \sum_{u \in OPT}\bar{f}(S_{i-1} \cup \{u\}) - \bar{f}(S_{i-1}) && \text{(by definition of } v_j) \\ &\geq \sum_{u \in OPT}\left(f(S_{i-1} \cup \{u\}) - f(S_{i-1}) - 2\xi\right) && \text{(by } \xi\text{-controlled estimation)} \\ &\geq \sum_{u \in OPT}\left(f(S_{i-1} \cup \{u\}) - f(S_{i-1})\right) - 2k\xi && (|OPT| \leq k) \\ &\geq f(OPT \cup S_{i-1}) - f(S_{i-1}) - 2k\xi. && \text{(by submodularity of } f) \end{aligned}$$

By Lemma 4.1 in (Buchbinder et al., 2017), we have $\mathbb{E}\left[X_j\right] \geq 1 - \varepsilon$. Therefore,

$$\mathbb{E}\left[f(S_i) - f(S_{i-1})\right] \geq (1 - \varepsilon) \cdot \frac{f(OPT \cup S_{i-1}) - f(S_{i-1})}{k} - 4\xi. \tag{67}$$

First, since the above equation holds for every given event A_i, it also holds in expectation unconditionally. More formally, we get for every $1 \leq i \leq k$,

$$\mathbb{E}\left[f(S_i) - f(S_{i-1})\right] \geq (1 - \varepsilon) \cdot \frac{\mathbb{E}\left[f(OPT \cup S_{i-1})\right] - \mathbb{E}\left[f(S_{i-1})\right]}{k} - 4\xi.$$

Let us lower bound $\mathbb{E}\left[f(OPT \cup S_{i-1})\right]$. RANDOMSAMPLING adds each element to its solution with probability at most $(\lceil pn \rceil / n)/s = 1/k$. Hence, each element belongs to S_{i-1} with probability at most $1 - (1 - 1/k)^{i-1}$. Let $h(S) = h(S \cup OPT)$. Since h is a nonnegative submodular function, we get by Lemma 2.2,

$$\mathbb{E}\left[f(OPT \cup S_i)\right] = \mathbb{E}\left[h(S_i)\right] \geq (1 - 1/k)^i \cdot h(\varnothing) = (1 - 1/k)^i \cdot f(OPT).$$

Combining the two above inequalities yields,

$$\begin{aligned}\mathbb{E}\left[f\left(S_i\right)-f\left(S_{i-1}\right)\right] &\geq (1-\varepsilon)\cdot\frac{(1-1/k)^{i-1}\cdot f(OPT)-\mathbb{E}\left[f\left(S_{i-1}\right)\right]}{k}-4\xi\\ &\geq \frac{\left[(1-1/k)^{i-1}-\varepsilon\right]\cdot f(OPT)-\mathbb{E}\left[f\left(S_{i-1}\right)\right]}{k}-4\xi.\end{aligned}$$

We prove by induction the following:

$$\mathbb{E}\left[f\left(S_i\right)\right]\geq\frac{i}{k}\cdot\left[(1-1/k)^{i-1}-\varepsilon\right]\cdot f(OPT)-4i\xi.$$

For $i=0$, the corollary holds since $f\left(S_0\right)\geq 0=(0/k)\cdot\left[(1-1/k)^{-1}-\varepsilon\right]\cdot f(OPT)$ - 0. Assume the corollary holds for $i-1\geq 0$, let us prove it for i :

$$\begin{aligned}\mathbb{E}\left[f\left(S_i\right)\right] &\geq \mathbb{E}\left[f\left(S_{i-1}\right)\right]+\frac{\left[(1-1/k)^{i-1}-\varepsilon\right]\cdot f(OPT)-\mathbb{E}\left[f\left(S_{i-1}\right)\right]}{k}-4\xi\\ &=(1-1/k)\cdot\mathbb{E}\left[f\left(S_{i-1}\right)\right]+\frac{\left[(1-1/k)^{i-1}-\varepsilon\right]\cdot f(OPT)}{k}-4\xi\\ &\geq(1-1/k)\cdot\left(\frac{i-1}{k}\cdot\left[(1-1/k)^{i-2}-\varepsilon\right]\cdot f(OPT)-4(i-1)\xi\right)+\frac{\left[(1-1/k)^{i-1}-\varepsilon\right]\cdot f(OPT)}{k}-4\xi\\ &\geq(1-1/k)\cdot\left(\frac{i-1}{k}\cdot\left[(1-1/k)^{i-2}-\varepsilon\right]\cdot f(OPT)\right)+\frac{\left[(1-1/k)^{i-1}-\varepsilon\right]\cdot f(OPT)}{k}-4(i-1)\xi-4\xi\\ &\geq\frac{i}{k}\cdot\left[(1-1/k)^{i-1}-\varepsilon\right]\cdot f(OPT)-4i\xi.\end{aligned}$$

Plugging $i=k$ into the above corollary yields

$$\mathbb{E}\left[f\left(S_k\right)\right]\geq\left[(1-1/k)^{k-1}-\varepsilon\right]\cdot f(OPT)-4k\xi\geq\left(e^{-1}-\varepsilon\right)\cdot f(OPT)-4k\xi.$$

Therefore, the RANDOMSAMPLING is an $(\frac{1}{e},2,1,n,4k)$-resilient-approximation algorithm algorithm for non-monotone submodular maximization under a k-cardinality constraint.

□

E.4. Submodular Maximization with Knapsack Constraint

We assume that the cost function $c:\Omega\rightarrow R_{>0}$ is both known and linear regarding knapsack constraints. In this framework, the cost linked with a subset is the total of the costs of its elements, denoted as $c(A)=\sum_{v\in A}c(v)$. The *marginal density* is symbolized as $\rho(e|A)=\frac{f(A\cup e)-f(A)}{c(e)}$ for any subset $A\subseteq\Omega$ and element $e\in\Omega\setminus A$. We don't consider scenarios with large budgets $B>\sum_{v\in\Omega}c(v)$ and presume that all items have non-zero costs, meeting $0<c(v)\leq B$. A cardinality constraint represents a particular case with unit costs. Additionally, $q=\frac{B}{c_{\min}}$.

Lemma E.5 (Generalization of Theorem in (Nie et al., 2023))**.** PARTIALENUMERATION *in (Sviridenko, 2004; Khuller et al., 1999) is a* $(1-\frac{1}{e},0,0,n^5,4+2\tilde{K}+2q)$*-resilient-approximation algorithm for monotone SM under a knapsack constraint.*

Proof. As shown in (Nie et al., 2023), the PARTIALENUMERATION (Sviridenko, 2004; Khuller et al., 1999) is a $(1-\frac{1}{e},4+2\tilde{K}+2q)$-robust approximation algorithm for monotone SM under a knapsack constraint. Furthermore, it requires $\mathcal{O}(n^5)$ oracle calls, thus it is a $(1-\frac{1}{e},0,0,n^5,4+2\tilde{K}+2q)$-resilient-approximation algorithm. □

Corollary E.6. *C-MA-MAB using the* PARTIALENUMERATION *in (Sviridenko, 2004; Khuller et al., 1999) as a subroutine requires at most* $\tilde{\mathcal{O}}(n^5)$ *communication times and yields a* $(1-\frac{1}{e})$*-regret of at most* $\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}}\tilde{K}^{\frac{2}{3}}q^{\frac{2}{3}}n^{\frac{5}{3}}T^{\frac{2}{3}}\right)$ *for monotone SM under a knapsack constraint.*

Lemma E.7 (Generalization of Theorem in (Nie et al., 2023))**.** GREEDY+MAX *in (Yaroslavtsev et al., 2020) is a* $(\frac{1}{2}, 0, 0, n\tilde{K}, \frac{1}{2} + \tilde{K} + 2q)$*-resilient-approximation algorithm for monotone SM problem under a knapsack constraint.*

Proof. As shown in (Nie et al., 2023), GREEDY+MAX (Yaroslavtsev et al., 2020) is a $(\frac{1}{2}, \frac{1}{2} + \tilde{K} + 2q)$-robust approximation algorithm for monotone SM under a knapsack constraint. Furthermore, it requires at most $\mathcal{O}(n\tilde{K})$ oracle calls, thus it is a $(\frac{1}{2}, 0, 0, n\tilde{K}, \frac{1}{2} + \tilde{K} + 2q)$-resilient-approximation algorithm. □

Corollary E.8. *C-MA-MAB using the* GREEDY+MAX *in (Yaroslavtsev et al., 2020) as a subroutine requires at most* $\tilde{\mathcal{O}}(n\tilde{K})$ *communication times and yields a* $(1/2)$*-regret of at most* $\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}} q^{\frac{2}{3}} \tilde{K} n^{\frac{1}{3}} T^{\frac{2}{3}}\right)$ *for monotone SM under a knapsack constraint.*

Lemma E.9 (Generalization of Theorem in (Nie et al., 2023))**.** GREEDY+ *in (Khuller et al., 1999) is a* $(\frac{1}{2}(1-\frac{1}{e}), 0, 0, n^2, 2+\tilde{K}+q)$*-resilient-approximation algorithm for monotone SM problem under a knapsack constraint.*

Proof. As shown in (Nie et al., 2023), the GREEDY+ (Khuller et al., 1999) is a $(\frac{1}{2}(1-\frac{1}{e}), 2+\tilde{K}+q)$-robust approximation algorithm for monotone SM under a knapsack constraint. Furthermore, it requires at most n^2 oracle calls, thus it is a $(\frac{1}{2}(1-\frac{1}{e}), 0, 0, n^2, 2+\tilde{K}+q)$-resilient-approximation algorithm. □

Corollary E.10. *C-MA-MAB using the* GREEDY+ *in (Khuller et al., 1999) as a subroutine requires at most* $\tilde{\mathcal{O}}(n^2)$ *communication times and yields a* $(\frac{1}{2}(1-\frac{1}{e}))$*-regret of at most* $\tilde{\mathcal{O}}\left(m^{-\frac{1}{3}} q^{\frac{2}{3}} \tilde{K}^{\frac{2}{3}} n^{\frac{2}{3}} T^{\frac{2}{3}}\right)$ *for monotone SM under a knapsack constraint.*

F. Experiments on Data Summarization

We conduct experiments on stochastic data summarization.

F.1. Motivation for Stochastic Data Summarization

Data summarization is a primary challenge in machine learning, mainly when dealing with a large dataset. While this problem has been extensively studied in the deterministic case, i.e., with access to a deterministic oracle (Mirzasoleiman et al., 2015; 2020; Sivasubramanian et al., 2024), this work is the first to address online data summarization under a noisy stochastic objective function.

F.1.1. DETERMINISTIC DATA SUMMARIZATION

In data summarization, agents must play an action A of at most k images to summarize a large dataset $\mathcal{D}$. Based on a similarity metric C between two images, the deterministic objective is:

$$\underset{\mathcal{A}\subseteq\mathcal{D}:|\mathcal{A}|\leq k}{\arg\max} \sum_{i\in\mathcal{D}} \max_{v\in\mathcal{A}} C(i,v). \tag{68}$$

Adding more images always increases this objective, and it follows a diminishing return property. Thus, it falls in the monotone SM under a cardinality constraint k (Mirzasoleiman et al., 2015).

F.1.2. STOCHASTIC DATA SUMMARIZATION

Notice that even evaluating the above deterministic objective for a given action A may become very expensive with a large dataset $\mathcal{D}$; more precisely, the evaluation for one action has a complexity of $\mathcal{O}(|\mathcal{A}||\mathcal{D}|)$. We propose a stochastic version of the above optimization problem and solve it through our framework. We consider a stochastic objective where the chosen subset is compared only to a random subset $\mathcal{R} \subseteq \mathcal{D}$ drawn uniformly at random from the large dataset $\mathcal{D}$, resulting in a noisier but lower complexity evaluations of $\mathcal{O}(|\mathcal{A}||\mathcal{R}|)$. We do not solve the problem for a given realization $\mathcal{R}$, but we solve it in expectation:

$$\underset{\mathcal{A}\subseteq\mathcal{D}:|\mathcal{A}|\leq k}{\arg\max}\ \mathbb{E}_{\mathcal{R}}\left[\sum_{i\in\mathcal{R}} \max_{v\in\mathcal{A}} C(i,v)\right]. \tag{69}$$

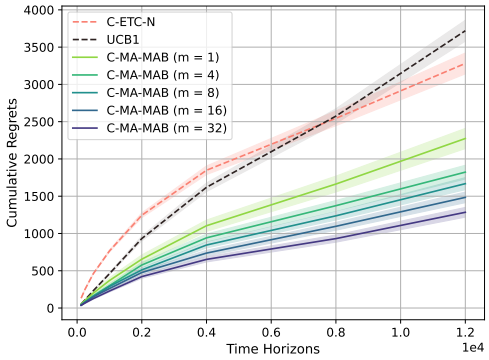
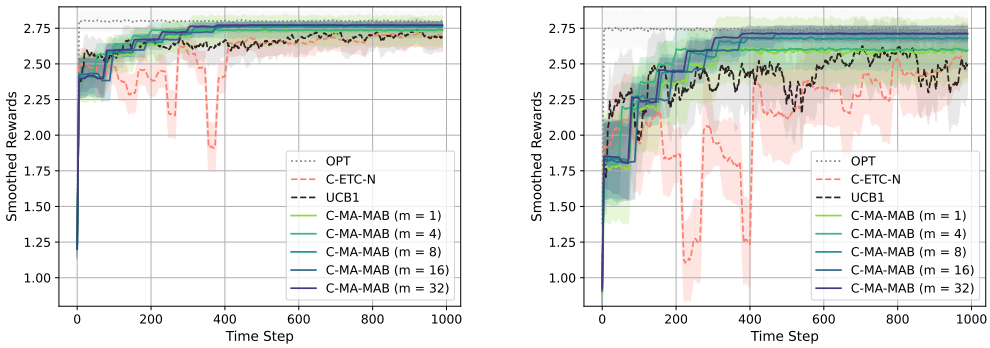


Figure 2: Cumulative regrets of summarizing images from FMNIST for different time horizons T using our framework with different number of agents against C-ETC framework and UCB.



(a) CIFAR10 (b) FMNIST

Figure 3: Compare the instantaneous rewards of summarizing images from CIFAR10 and FMNIST for different time horizons T using our framework with one, two, four, sixteen, and thirtytwo agents against C-ETC-N, UCB1, and OPT.

F.2. Experimental Details

We test our method when using the STOCHASTIC-GREEDY algorithm as a subroutine (Mirzasoleiman et al., 2015) for one agent and multiple agents and compare it to the proposed algorithm in C-ETC framework for SMC (C-ETC-N) (Nie et al., 2023), and the upper confidence bound (UCB1) algorithm (Auer et al., 2002). We consider settings with one, four, eight, sixteen, and thirty-two agents. We resize the images to have sixteen pixels. We set a cardinality constraint of $k = 5$. Our goal is to summarize information from fifteen images, and instead of comparing it to all the images, we only consider a random batch $\mathcal{R}$ of 3 images. We run the experiments 100 times. We test for several time horizons in $\{125, 250, 500, 1000, 2000, 4000, 8000, 12000, 16000, 20000\}$.

F.3. Additional Results

As depicted in Fig. 1, the C-MA-MAB demonstrates sub-linear regret guarantees across different agent scenarios within a time horizon T. Additionally, it is apparent that, for varying values of m, including the single-agent scenario, the C-MA-MAB consistently outperforms both C-ETC-N and UCB1, exhibiting lower regrets over diverse time horizons. Notably, an increase in the number of agents correlates with a reduction in regret for these agents. These observations reinforce the same conclusions drawn from the theoretical analysis.