Minimax Optimization: The Case of Convex-Submodular

Arman Adibi
University of Pennsylvania

Aryan Mokhtari
University of Texas at Austin

Hamed Hassani
University of Pennsylvania

Abstract

Minimax optimization has been central in addressing various applications in machine learning, game theory, and control theory. Prior literature has thus far mainly focused on studying such problems in the continuous domain, e.g., convex-concave minimax optimization is now understood to a significant extent. Nevertheless, minimax problems extend far beyond the continuous domain to mixed continuous-discrete domains or even fully discrete domains. In this paper, we study mixed continuous-discrete minimax problems where the minimization is over a continuous variable belonging to Euclidean space and the maximization is over subsets of a given ground set. We introduce the class of convex-submodular minimax problems, where the objective is convex with respect to the continuous variable and submodular with respect to the discrete variable. Even though such problems appear frequently in machine learning applications, little is known about how to address them from algorithmic and theoretical perspectives. For such problems, we first show that obtaining saddle points are hard up to any approximation, and thus introduce new notions of (near-) optimality. We then provide several algorithmic procedures for solving convex and monotone-submodular minimax problems and characterize their convergence rates, computational complexity, and quality of the final solution according to our notions of optimally. Our proposed algorithms are iterative and combine tools from both discrete and continuous optimization. Finally, we provide numerical experiments to showcase the effectiveness of our purposed methods.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1 INTRODUCTION

The problem of solving a minimax optimization problem, also known as the saddle point problem, appears in many domains such as robust optimization (Ben-Tal et al., 2009), game theory (Osborne and Rubinstein, 1994), and robust control (Zhou and Doyle, 1998; Hast et al., 2013). It has also recently attracted a lot of attention in the machine learning community due to the rise of generative adversarial networks (GANs) (Goodfellow et al., 2014) and robust learning (Bertsimas et al., 2011; Lanckriet et al., 2002; Li et al., 2019). There has been an extensive literature on the design of convergent methods for solving minimax problems for the case that both minimization and maximization variables belong to continuous domains (Tseng, 1995; Nesterov, 2007; Li and Lin, 2015; Ouyang and Xu, 2019; Thekumparampil et al., 2019; Zhao, 2019; Hamedani and Aybat, 2018; Alkousa et al., 2019; Daskalakis et al., 2017; Ibrahim et al., 2020; Nouiehed et al., 2019; Mokhtari et al., 2020a; Lin et al., 2020a; Murty and Kabadi, 1985). In particular, for the case that the loss function is (strongly) convex with respect to the minimization variable and (strongly) concave with respect to maximization variable several efficient algorithms have been studied (Nesterov, 2007; Li and Lin, 2015; Mokhtari et al., 2020b), including the extra-gradient method (Korpelevich, 1976; Nemirovski, 2004) that is known to be optimal for this setting. However, all these methods suffer from two major limitations: (i) they are provably convergent only in convex-concave settings; (ii) they are designed for the settings that both minimization and maximization variables belong to continuous domains.

There has been some effort to address the first limitation by finding a first-order stationary point or locally stable point for the problems that are not convex-concave (Lin et al., 2020b; Diakonikolas et al., 2021; Yang et al., 2020; Sanjabi et al., 2018). However, these approaches fail to guarantee any global optimality as it is known that finding a saddle point in a general nonconvex-nonconcave setting is NP-hard (Jin et al., 2020). Nonetheless, it might be possible to achieve global approximation guarantees for *structured* saddle minimax problems. Addressing the second limita-

tions and developing methods for discrete-continuous domains or fully discrete domains requires exploiting tools from discrete optimization. Several recent works have considered applications involving specific discrete-continuous minimax problems and proposed structure-informed algorithms (Zhou and Bilmes, 2018). However, to our knowledge, there is no work that provides a principled algorithmic or theoretical framework to study minimax problems with mixed discrete-continuous components and it is not even clear if such problems allow for tractable solutions with global guarantees.

In this paper, we tackle these two issues and present iterative methods with theoretical guarantees to solve structured non convex-concave minimax problems, where the minimization variable is from a continuous domain and the maximization variable belongs to a discrete domain. Concretely, for a non-negative function $f : \mathbb{R}^d \times 2^V \to \mathbb{R}_+$, consider the minimax problem

$$\text{OPT} \triangleq \min_{\mathbf{x}\in\mathcal{X}} \max_{S\in\mathcal{I}} f(\mathbf{x}, S), \tag{1}$$

where $\mathbf{x}$ belongs to a convex set $\mathcal{X} \subset \mathbb{R}^d$ and S is a subset of the ground set V with n elements that is constrained to be inside a matroid $\mathcal{I}$. Given a fixed S, the function $f(\cdot, S)$ is convex with respect to the continuous (minimization) variable. Further, given a fixed $\mathbf{x}$, the function $f(\mathbf{x}, \cdot)$ is submodular with respect to the discrete (maximization) variable. We refer to this problem as *convex-submodular minimax problem*.

The convex-submodular minimax problem in (1) encompasses various applications. In Section 4, we describe specific optimization problems, such as convex-facility-location, as well as applications such as designing adversarial attacks on recommender systems. There are various other applications that can be cast into Problem (1), in particular, when convex models have to be learned while data points are selected or changed according to notions of summarization, diversity, and deletion. Examples include learning under data deletion (Ginart et al., 2019; Neel et al., 2020; Wu et al., 2020), robust text classification (Lei et al., 2018), minimax curriculum learning (Zhou and Bilmes, 2018; Zhou et al., 2021, 2020; Soviany et al., 2021), minimax supervised learning (Farnia and Tse, 2016), and minimax active learning (Ebrahimi et al., 2020).

1.1 Our Contributions

In this paper, we provide a principled study of the problem defined in (1), from both theoretical and algorithmic perspectives, when f is convex in the minimization variable and submodular as well as *monotone* in the maximization variable[1]. We introduce efficient iterative algorithms for solving this problem and develop a theoretical framework for analyzing such algorithms with guarantees on the quality of the resulting solutions according to the notions of optimality that we define.

Notions of (near-)optimality and hardness results. For minimax problems, the strongest notion of optimality is defined through saddle points or their approximate versions. We first provide a negative result that shows finding a saddle point or any approximate version of it (which we term as an (α, ϵ)-saddle point) is NP-hard for general convex-submodular problems (Theorem 1). We thus introduce a slightly weaker notion of optimality that we call (α, ϵ)-approximate minimax solutions for Problem (1). Roughly speaking, the quality of the minimax objective at such solutions is at most $\frac{1}{\alpha}(\text{OPT} + \epsilon)$, and hence they are near-optimal when $\alpha < 1$. We show in Theorem 2 that obtaining such solutions for $\alpha > 1 - 1/e$ is NP-hard. This is a non-trivial result that does not readily follow from known hardness results in submodular maximization. Consequently, we focus on efficiently finding solutions in the regime of $\alpha \leq (1 - 1/e)$. We present several algorithms that achieve this goal and theoretically analyze their complexity and quality of their solution.

Algorithms with guarantees on convergence rate, complexity, and solution quality. Our proposed algorithms are as follows (see also Table 1): (i) *Greedy-based methods.* We first present Gradient-Greedy (GG), a method alternating between gradient descent for minimization and greedy for maximization. We further introduce Extra-Gradient-Greedy (EGG) that uses an extra-gradient step instead of gradient step for the minimization variable. We prove that both algorithms achieve a $((1 - 1/e), \epsilon)$-approximate minimax solution after $\mathcal{O}(1/\epsilon^2)$ iterations when $\mathcal{I}$ is a cardinality constraint. Importantly, EGG does not require the bounded gradient norm condition as opposed to GG. Our results for the case that $\mathcal{I}$ is a matroid constraint (see Table 1) are provided in the supplementary material. (ii) *Replacement greedy-based methods.* The greedy-based methods require $\mathcal{O}(nk)$ function computations at each iteration. To improve this complexity, we present alternating methods that use replacement greedy for the maximization part to reduce the cost of each iteration to $\mathcal{O}(n)$. The Gradient Replacement-Greedy (GRG) algorithm achieves a $(1/2, \epsilon)$-approximate minimax solution after $\mathcal{O}(1/\epsilon^2)$ iterations and Extra-Gradient Replacement-Greedy (EGRG) achieves a $(1/2, \epsilon)$-approximate minimax solution after $\mathcal{O}(1/\epsilon^2)$, when $\mathcal{I}$ is a cardinality constraint. (iii) *Continuous extension-based methods.* Note that all mentioned methods achieve a convergence rate of

[1]For completeness, a function $g : 2^V \to \mathbb{R}$ is called submodular if for any two subsets $S, T \subseteq V$ we have: $g(S \cap T) + g(S \cup T) \leq g(S) + g(T)$. Moreover, g is called monotone if for any $S \subseteq T$ we have $g(S) \leq g(T)$.

Table 1: Algorithms performance guarantee. Here c_f is the cost of single computation of f, $c_{P_{\mathbf{x}}}$ and $c_{P_{\mathbf{y}}}$ are cost of projection in $\mathcal{X}$ and $\mathcal{Y}$, $c_{\nabla_{\mathbf{x}} f}$ is the cost of computing gradient of f with respect to $\mathbf{x}$, and $c_{\nabla_{\mathbf{x}} F}$ and $c_{\nabla_{\mathbf{y}} F}$ are the cost of computing gradient of multilinear extension F with respect to $\mathbf{x}$ and $\mathbf{y}$, respectively. k is the cardinality constraint ($|S| \leq k$) and n is size of the ground set $|V| = n$.

Alg.	Number of iterations	Approx. ratio	Cost per iteration	Cardinality const.	Matroid const.	Unbounded gradient
GG	$\mathcal{O}(1/\epsilon^2)$	$1-1/e$	$nk.c_f + c_{\nabla_{\mathbf{x}} f + c_{P_{\mathbf{x}}}}$	✓	×	×
GG	$\mathcal{O}(1/\epsilon^2)$	$1/2$	$nk.c_f + c_{\nabla_{\mathbf{x}} f + c_{P_{\mathbf{x}}}}$	×	✓	×
GRG	$\mathcal{O}(1/\epsilon^2)$	$1/2$	$(n+k)c_f + c_{\nabla_{\mathbf{x}} f + c_{P_{\mathbf{x}}}}$	✓	×	×
EGG	$\mathcal{O}(1/\epsilon^2)$	$1-1/e$	$2nk.c_f + 2c_{\nabla_{\mathbf{x}} f + 2c_{P_{\mathbf{x}}}}$	✓	×	✓
EGG	$\mathcal{O}(1/\epsilon^2)$	$1/2$	$2nk.c_f + 2c_{\nabla_{\mathbf{x}} f + 2c_{P_{\mathbf{x}}}}$	✓	✓	✓
EGRG	$\mathcal{O}(1/\epsilon^2)$	$1/2$	$2(n+k)c_f + 2c_{\nabla_{\mathbf{x}} f + 2c_{P_{\mathbf{x}}}}$	✓	×	×
EGCE	$\mathcal{O}(1/\epsilon)$	$1/2$	$2c_{\nabla_{\mathbf{x}} F} + 2c_{\nabla_{\mathbf{y}} F + 2c_{P_{\mathbf{x}}} + 2c_{P_{\mathbf{y}}}}$	✓	✓	✓

$\mathcal{O}(1/\epsilon^2)$. To improve this convergence rate, we further introduce the extra-gradient on continuous extension (EGCE) method that runs extra-gradient update on the continuous extension of the submodular function. We show that EGCE is able to achieve an $(1/2, \epsilon)$-approximate minimax solution after at most $\mathcal{O}(1/\epsilon)$ iterations, when $\mathcal{I}$ is a general matroid constraint.

1.2 Related Work

Several recent works have considered specific applications that require solving Problem (1) when f is nonconvex-submodular (Zhou and Bilmes, 2018; Lei et al., 2018; Mirzasoleiman et al., 2020). Zhou and Bilmes (2018) consider the problem of minimax curriculum learning which is a special case of minimax strongly convex-submodular optimization, and propose an algorithm similar to gradient-greedy (GG). They provide an upper bound on the distance between their obtained solution and the optimal solution when f is strongly convex in $\mathbf{x}$ and monotone-submodular in S with non-zero curvature. Moreover, Lei et al. (2018) study designing an adversarial attack in text classification and show that for some specific neural network structures, the task of designing an adversarial attack can be formulated as submodular maximization, leading to a minimax nonconvex-submodular problem. An algorithm similar to gradient-greedy is then proposed by Lei et al. (2018) for designing attacks and it has led to successful experimental results. In contrast, this paper is the first to introduce a principled study of Problem (1) for general functions f with newly developed notions of optimality, algorithmic frameworks, and theoretical guarantees.

Another line of work is the literature on differentiable submodular maximization (Tschiatschek et al., 2018; Wilder et al., 2019; Sakaue, 2021) in which the goal is to find a smooth maximization oracle for submodular maximization to compute the gradient of the objective function. Another related work is "Submodular+ Concave"(Mitra et al., 2021) in which authors studied the problems that can be written as a summation of submodular and concave function. Both of these works consider fundamentally different problems from our setting.

Another relevant line of work is the literature on robust submodular optimization (Krause et al., 2008; Bogunovic et al., 2017b; Mirzasoleiman et al., 2017; Kazemi et al., 2018; Bogunovic et al., 2018; Iyer, 2021; Orlin et al., 2018; Chen et al., 2017; Anari et al., 2019; Wilder, 2018; Bogunovic et al., 2017a; Mitrović et al., 2017). This setting corresponds to solving a *max-min* optimization problem which involves only *discrete variables*, and hence, it is different from our setting with fundamentally different methods. For such problems, finding discrete solutions with any approximation factor is NP-hard; and consequently, the literature has mostly focused on obtaining solutions that satisfy a bi-criteria approximation guarantee. Another related work is distributionally robust submodular maximization in (Staib et al., 2019) which is a special case of *max-min* version of Problem (1). In this setting, the inner minimization has a special structure that allows for a closed form solution, and hence, the problem can be solved by using appropriate techniques from continuous submodular optimization. We will derive the implication of our results on the max-min version of Problem (1) in the supplementary material.

2 CONVEX-SUBMODULAR MINIMAX OPTIMIZATION

For the minimax problem in (1), a natural goal is to find a so-called saddle point. Next, we formally define the notion of saddle point for Problem (1).

Definition 1. *A pair $(\mathbf{x}^*, S^*)$ is a saddle point of the function f if the following condition holds:*

$$\forall \mathbf{x} \in \mathcal{X}, S \in \mathcal{I}: \; f(\mathbf{x}^*, S) \leq f(\mathbf{x}^*, S^*) \leq f(\mathbf{x}, S^*) \quad (2)$$

Based on this definition, $(\mathbf{x}^*, S^*)$ is a saddle point of Problem (1), if there is no incentive to modify the minimization variable $\mathbf{x}^*$ when the maximization variable is fixed and equal to S^*, and, conversely, there is no incentive to change the maximization variable from S^* when the minimization variable is $\mathbf{x}^*$. In other words, a saddle point can be interpreted as an equilibrium.

There is a rich literature on efficient approaches for finding an ϵ-saddle point for convex-concave minimax optimization, where ϵ is an arbitrary positive constant (Thekumparampil et al., 2019). To define an ϵ-saddle point, we first need to define the duality gap, which is given by $D(\mathbf{x}, S) := \bar{\phi}(\mathbf{x}) - \underline{\phi}(S)$, where

$$\bar{\phi}(\mathbf{x}) := \max_{S \in \mathcal{I}} f(\mathbf{x}, S), \qquad \underline{\phi}(S) := \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, S).$$

Considering these definitions, we call a pair of solution ϵ-saddle point if their duality gap is at most ϵ.

Definition 2. *A pair $(\bar{\mathbf{x}}, \bar{S})$ is called an $\epsilon-$saddle point of f if it satisfies*

$$D(\bar{\mathbf{x}}, \bar{S}) = \bar{\phi}(\bar{\mathbf{x}}) - \underline{\phi}(\bar{S}) \leq \epsilon \tag{3}$$

One can verify that if we set $\epsilon = 0$, then Definitions 1 and 2 coincide, i.e., $(\mathbf{x}^*, S^*)$ satisfies (3) for $\epsilon = 0$ if and only if $(\mathbf{x}^*, S^*)$ satisfies (2). Hence, to derive a finite time analysis we often aim for finding an ϵ-saddle point. For instance, for smooth and convex-concave problems extra-gradient obtains an ϵ-saddle point after $\mathcal{O}(1/\epsilon)$ iterations (which is the optimal complexity).

However, for our convex-submodular setting, one cannot expect to find an ϵ-saddle point efficiently, as the special case of finding an ϵ-accurate solution for submodular maximization is in general NP-hard (Nemhauser and Wolsey, 1978; Wolsey, 1982; Krause and Golovin, 2014). Although solving the problem of maximizing a monotone submodular function subject to a matroid constraint is hard, one can find α-approximate solution of that in polynomial time, i.e., finding a solution that its function value is at least αOPT, where $\alpha \in (0, 1)$. Inspired by this observation, we introduce the notion of (α, ϵ)-saddle point for our convex-submodular setting.

Definition 3. *A pair $(\hat{\mathbf{x}}, \hat{S})$ is called an $(\alpha, \epsilon)-$saddle point of f if it satisfies*

$$\alpha\bar{\phi}(\hat{\mathbf{x}}) - \underline{\phi}(\hat{S}) \leq \epsilon. \tag{4}$$

Our first result is a negative result that shows even finding an (α, ϵ)-saddle point is not tractable.

Theorem 1. *Finding (α, ϵ)-saddle point for Problem (1) is NP-hard for any $\alpha > 0$.*

While this result shows intractability of finding (approximate) saddle-points for Problem (1), one avenue to provide solutions with guaranteed quality is to see whether we can find solutions that achieve a fraction of OPT. We thus proceed to introduce the notion of *approximate minimax solution*.

Definition 4. *We call a point $\hat{\mathbf{x}}$ an (α, ϵ)-approximate minimax solution of Problem (1) if it satisfies*

$$\alpha\bar{\phi}(\hat{\mathbf{x}}) \leq \text{OPT} + \epsilon, \tag{5}$$

where $\text{OPT} = \min_{\mathbf{x} \in \mathcal{X}} \bar{\phi}(\mathbf{x}) = \min_{\mathbf{x} \in \mathcal{X}} \max_{S \in \mathcal{I}} f(\mathbf{x}, S)$.

Next, we describe the notion of an (α, ϵ)-approximate minimax solution for Problem (1). The minimax problem in (1) can be interpreted as a sequential game, where we first select an action $\mathbf{x}$ and then an adversary chooses a set S to maximize our loss $f(\mathbf{x}, S)$. In this case, our goal is to find $\mathbf{x}$ that minimizes the loss obtained by the worst possible action by the adversary, i.e., we aim to minimize the function $\bar{\phi}(\hat{\mathbf{x}}) := \max_{S \in \mathcal{I}} f(\hat{\mathbf{x}}, S)$ over the choice of $\hat{\mathbf{x}}$. Indeed, finding the exact minimizer is also hard and we should seek approximate solutions. Hence, our goal is to find solutions $\hat{\mathbf{x}}$ whose worst-case loss $\bar{\phi}(\hat{\mathbf{x}})$ is only a factor larger than the best possible loss $\text{OPT} = \min_{\mathbf{x} \in \mathcal{X}} \bar{\phi}(\mathbf{x})$. That said, by finding an (α, ϵ)-approximate minimax solution for Problem (1) we obtain a solution whose loss is at most $(\text{OPT} + \epsilon)/\alpha$, where $0 < \alpha \leq 1$ and $\epsilon > 0$.

The task of finding an $\hat{\mathbf{x}}$ that is (α, ϵ)-approximate minimax solution is easier than finding a pair $(\tilde{\mathbf{x}}, \tilde{S})$ that is an (α, ϵ)-saddle point, since if the pair $(\tilde{\mathbf{x}}, \tilde{S})$ satisfies (4), then $\tilde{\mathbf{x}}$ satisfies (5):

$$\alpha\bar{\phi}(\tilde{\mathbf{x}}) - \underline{\phi}(\tilde{S}) \leq \epsilon \;\Rightarrow\; \alpha\bar{\phi}(\tilde{\mathbf{x}}) - \max_{S \in \mathcal{I}} \underline{\phi}(S) \leq \epsilon$$
$$\Rightarrow\; \alpha\bar{\phi}(\tilde{\mathbf{x}}) - \min_{\mathbf{x} \in \mathcal{X}} \bar{\phi}(\mathbf{x}) \leq \epsilon$$

Hence, the condition in (4) is more strict compared to (5). In fact, in the next section, we show that unlike the task of finding an (α, ϵ)-saddle point of Problem (1) that is NP-hard for any $\alpha \in (0, 1]$, one can find an (α, ϵ)-approximate minimax solution of Problem (1) in poly-time for $\alpha \in (0, 1 - 1/e]$. Alas, the problem is still NP-hard for $\alpha \in (1 - 1/e, 1]$ as we show in Theorem 2.

Theorem 2. *Let $\alpha = 1 - 1/e + \gamma$ for a positive constant $\gamma > 0$. If there exists a polynomial time algorithm and a polynomial time oracle that can achieve an (α, ϵ)-approximate solution for any choice of the function $f(\mathbf{x}, S)$ in problem (1), then $P = NP$.*

We emphasize that Theorem 2 does not follow directly from that fact that submodular maximization beyond $(1 - 1/e)$-approximation is hard, and hence it is non-trivial. Indeed, one naive way to argue for the proof of this theorem (which is incorrect) is to consider functions $f(\mathbf{x}, S)$ whose output does not depend on the variable

$\mathbf{x}$, i.e. $f(\mathbf{x}, S) = f(S)$, and use the hardness results for submodular optimization. But for such functions *any* point $\mathbf{x}$ is an optimal solution (with $\alpha = 1$). Hence, the proof of the theorem (provided in the appendix) requires a novel idea beyond trivial consequences of known results for submodularity.

So far we have shown two results: (i) Finding an approximate (α, ϵ)-saddle point is hard for $\alpha > 0$. (ii) We introduced the notion of (α, ϵ)-approximate solution and showed that for $\alpha > 1–1/e$ finding an (α, ϵ)-approximate solution is hard. The only missing piece is showing whether or not it is possible to *efficiently* find an (α, ϵ)-approximate minimax solution when $\alpha \leq 1–1/e$. In the rest of the paper, we provide an affirmative answer to this question and present methods achieving this goal.

3 ALGORITHMS

In this section, we present a set of algorithms that are able to find an (α, ϵ)-approximate minimax solution of Problem (1). To present these algorithms, we first present two subroutines that we use in the implementation of our algorithms[2]: (i) greedy update and (ii) replacement greedy update.

Greedy subroutine. In the greedy update, for a fixed minimization variable $\mathbf{x}$, we select a subset S with k elements in a greedy fashion, i.e., we sequentially pick k elements that maximize the marginal gain. Specifically, if we define $\Delta_e f(\mathbf{x}, S) = f(\mathbf{x}, S \cup \{e\}) - f(\mathbf{x}, S)$ as the marginal gain of element e, in the greedy update, for a given variable $\mathbf{x}$ we perform the update

$$S_{i+1} = S_i \cup \{\arg\max \Delta_e f(\mathbf{x}, S)\}, \tag{6}$$

for $i = 0, \ldots, k-1$, where S_0 is the empty set. The output of this process is S_k with k elements. We use the notation $\textsc{Greedy}(f, k, \mathbf{x})$ for the greedy subroutine, which takes function f, cardinality constraint parameter k, and variable $\mathbf{x}$ as inputs, and returns a set S by performing (6) for k steps.

Replacement greedy subroutine. In the replacement greedy update (Mitrovic et al., 2018; Schrijver, 2003; Stan et al., 2017a), for a given variable $\mathbf{x}$ and set S, the output is an updated set S^+ whose function value at $\mathbf{x}$ is larger than the one for S, i.e., $f(\mathbf{x}, S) \leq f(\mathbf{x}, S^+)$. The procedure for finding the new set S^+ is relatively simple. If the size of the input set S is less than k, we add one more element to the set S that maximizes the marginal gain and the resulted set would be S^+. In other words, if $|S| < k$,

$$S^+ = S \cup \{\arg\max \Delta_e f(\mathbf{x}, S)\}. \tag{7}$$

[2]For better exposition, we consider the case that $\mathcal{I}$ is k-carnality constraint and refer to Appendix for matroids.

Algorithm 1
Option I: Gradient Greedy (GG)
Option II: Extra-Gradient Greedy (EGG)

Initialize the set S_1 to $\emptyset$ and variable $\mathbf{x}_1$ to zero.
for $t = 1$ to T **do**
  Option I: $\mathbf{x}_{t+1} = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, S_t))$
    $S_{t+1} = \textsc{Greedy}(f, k, \mathbf{x}_{t+1})$
  Option II: $\hat{\mathbf{x}}_t = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, S_t))$
    $\hat{S}_t = \textsc{Greedy}(f, k, \hat{\mathbf{x}}_t)$
    $\mathbf{x}_{t+1} = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla f(\hat{\mathbf{x}}_t, \hat{S}_t))$
    $S_{t+1} = \textsc{Greedy}(f, k, \mathbf{x}_{t+1})$
end for
Option I: $\mathbf{x}_{sol} = (\sum_{t=1}^T \gamma_t)^{-1} \sum_{t=1}^T \gamma_t \mathbf{x}_t$
Option II: $\mathbf{x}_{sol} = (\sum_{t=1}^T \gamma_t)^{-1} \sum_{t=1}^T \gamma_t \hat{\mathbf{x}}_t$

If the size of the input set S is k, we first remove one element of the set S that leads to minimum decrease in the function value (denoted by e^*) and then replace it with another element of the ground set that maximizes the marginal gain. Hence, if $|S| = k$, we have

$$S^+ = (S \setminus \{e^*\}) \cup \{\arg\max \Delta_e f(\mathbf{x}, S \setminus \{e^*\})\}, \tag{8}$$

where $e^* = \arg\max_{e \in S} \{f(\mathbf{x}, S \setminus e)\}$.

We use the notation $\textsc{RepGreedy}(f, k, \mathbf{x}, S)$ for the replacement greedy subroutine. Note that replacement greedy is computationally cheaper than greedy, as it requires only one pass over the ground set, while greedy requires k passes.

3.1 Greedy-based Algorithms

Next, we present greedy-based methods to find (α, ϵ)-approximate minimax solutions for Problem (1).

Gradient Greedy. The first algorithm that we present is Gradient Greedy (GG), which uses a projected gradient descent step to update the minimization iterate $\mathbf{x}_t$ at each iteration, i.e., $\mathbf{x}_{t+1} = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, S_t))$, and then uses a greedy procedure to update the maximization variable S_t. This update is performed in an alternating fashion, where we first use $\mathbf{x}_t$ and S_t to find $\mathbf{x}_{t+1}$ and then we use the updated variable $\mathbf{x}_{t+1}$ to compute S_{t+1}. Note that the final output of this process is a weighted average of all variables $\mathbf{x}_t$ that are observed from time $t = 1$ to $t = T$, defined as $\mathbf{x}_{sol} = (\sum_{t=1}^T \gamma_t)^{-1} \sum_{t=1}^T \gamma_t \mathbf{x}_t$. The steps of GG are summarized in Algorithm 1 option I.

Next, we show that GG is able to find a $(1 - 1/e, \epsilon)$-approximate minimax solution after $\mathcal{O}(1/\epsilon^2)$ iterations. To prove this claim we require the following assumptions on the objective function f.

Assumption 1. *The function f is L-smooth with respect to $\mathbf{x}$, i.e., for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d, S \in \mathcal{I}$, we have $\|\nabla_{\mathbf{x}} f(\mathbf{x}, S) - \nabla_{\mathbf{x}} f(\mathbf{x}', S)\| \leq L \|\mathbf{x} - \mathbf{x}'\|$.*

Assumption 2. *The gradient of function f with respect to $\mathbf{x}$ is uniformly bounded by a constant M, i.e., for any $\mathbf{x} \in \mathbb{R}^d, S \in 2^V$, we have $\|\nabla_{\mathbf{x}} f(\mathbf{x}, S)\| \leq M$.*

Theorem 3. *Consider Gradient Greedy (GG) in Algorithm 1 option I. If f is convex-submodular and Assumptions 1-2 hold, then the output of this algorithm after $\mathcal{O}(1/\epsilon^2)$ iterations with step size $\mathcal{O}(\epsilon)$, is a $((1-1/e), \epsilon)$-approximate minimax solution of Problem (1) .*

The smoothness assumption (Assumption 1) is required to guarantee convergence of gradient-based methods at the rate of $1/\epsilon^2$. The bounded gradient assumption (Assumption 2), however, comes from the fact that even in convex-concave problems gradient descent-ascent algorithms only converge when the gradient norm is uniformly bounded. This issue has been addressed in the convex-concave setting by the update of extra-gradient method which converges to a saddle point only under smoothness assumption. However, this improvement is not for free and it requires two gradient computations per update, instead of one. Next, we leverage this technique to present an alternating method that obtains the approximation factor and iteration complexity of GG without requiring Assumption 2.

Extra-gradient greedy. We now present the Extra-Gradient Greedy (EGG) algorithm, which consists of two gradient updates as suggested by extra-gradient and two greedy steps to find the auxiliary set $\hat{S}_t$ and the updated set S_{t+1}. In the extra-gradient method, we take a preliminary step to find a middle/auxiliary point and then compute the next iterate using the gradient information of the middle point. If we consider $\mathbf{x}_t$ and S_t as the current iterates, we first run a gradient step to find the auxiliary minimization variable according to the update $\hat{\mathbf{x}}_t = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, S_t))$ then we compute the auxiliary set $\hat{S}_t$ by performing a greedy step based on the auxiliary iterate $\hat{\mathbf{x}}_t$, i.e., $\hat{S}_t = \textsc{Greedy}(f, k, \hat{\mathbf{x}}_t)$. Once $\hat{\mathbf{x}}_t$ and $\hat{S}_t$ are computed, we update the minimization variable $\mathbf{x}_{t+1}$ by descending towards a gradient evaluated at $(\hat{\mathbf{x}}_t, \hat{S}_t)$, i.e., $\mathbf{x}_{t+1} = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla f(\hat{\mathbf{x}}_t, \hat{S}_t))$. Lastly, we compute the new set S_{t+1} by running a greedy update based on the new iterate $\mathbf{x}_{t+1}$, i.e., $S_{t+1} = \textsc{Greedy}(f, k, \mathbf{x}_{t+1})$. Steps of EGG are outlined in Algorithm 1 (option II).

Next we establish our theoretical result for Extra-gradient Greedy and show that only under smoothness assumption it finds an $(1-1/e, \epsilon)$-approximate minimax solution after $\mathcal{O}(1/\epsilon^2)$ iterations.

Theorem 4. *Consider Extra-Gradient Greedy (EGG) outlined in Algorithm 1 option II. If f is convex-submodular and Assumption 1 holds, then the output of this algorithm after $\mathcal{O}(1/\epsilon^2)$ iterations with step size $\mathcal{O}(\epsilon)$, is a $((1-1/e), \epsilon)$-approximate minimax solution of Problem (1).*

Algorithm 2
Option I: Gradient Replacement-greedy (GRG)
Option II: Extra-gradient Replacement-greedy (EGRG)

Initialize the set S_1 to $\emptyset$ and variable $\mathbf{x}_1$ to zero.
for $t = 1$ to T **do**
  Option I: $\mathbf{x}_{t+1} = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, S_t))$
    $S_{t+1} = \textsc{RepGreedy}(f, k, \mathbf{x}_{t+1}, S_t)$
  Option II: $\hat{\mathbf{x}}_t = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla f(\mathbf{x}_t, S_t))$
    $\hat{S}_t = \textsc{RepGreedy}(f, k, \mathbf{x}_t, S_t)$
    $\mathbf{x}_{t+1} = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla f(\hat{\mathbf{x}}_t, \hat{S}_t))$
    $S_{t+1} = \textsc{RepGreedy}(f, k, \hat{\mathbf{x}}_t, \hat{S}_t)$
end for
Option I: $\mathbf{x}_{sol} = (\sum_{t=1}^T \gamma_t)^{-1} \sum_{t=1}^T \gamma_t \mathbf{x}_t$
Option II: $\mathbf{x}_{sol} = (\sum_{t=1}^T \gamma_t)^{-1} \sum_{t=1}^T \gamma_t \hat{\mathbf{x}}_t$

Remark 1. *Note that as both GG and EGG are greedy based methods, they can also be used for the case of general matroid constraint. However, the approximation guarantee would be $1/2$ instead of $1-1/e$. The details are provided in the supplementary material.*

3.2 Replacement Greedy-based Methods

As we showed earlier, for the cardinality constraint problem GG and EGG achieve the optimal approximation guarantee of $1-1/e$ for the minimax problem in (1). However, they both require running greedy updates at each iteration which makes their per iteration complexity $\mathcal{O}(nk)$. To resolve this issue, we propose the use of replacement-greedy in lieu of greedy update. This modification reduces the complexity of each iteration to $\mathcal{O}(n+k)$ at the cost of lowering the approximation factor.

Gradient replacement-greedy. We first present the Gradient Replacement-Greedy (GRG) algorithm which alternates between a gradient update and a replacement greedy update. As shown in Algorithm 2 option I, the only difference between GRG and GG algorithms is the substitution of greedy update with replacement greedy. Next, we establish the theoretical guarantee of GRG.

Theorem 5. *Consider the Gradient Replacement-Greedy (GRG) algorithm in Algorithm 2 option I. If f is convex-submodular and Assumptions 1-2 hold, then the output of this algorithm after $\mathcal{O}(1/\epsilon^2)$ iterations with step size $\mathcal{O}(\epsilon)$, is a $(1/2, \epsilon)$-approximate minimax solution of Problem (1).*

Extra-gradient replacement-greedy. The GRG algorithm requires the bounded gradient assumption similar to GG. To address this issue, a natural idea is to exploiting the extra-gradient approach for updating $\mathbf{x}$ and introducing the Extra-gradient Replacement-greedy (EGRG) algorithm, outlined in Algorithm 2 option II. However, unlike the case of Greedy-based

methods, here we can not drop the bounded gradient assumption by exploiting the idea of extra-gradient update. Next, we elaborate on this issue.

Note that to prove that EGRG finds a $(1/2, \epsilon)$-approximate minimax solution we need to find an upper bound on $f(\hat{\mathbf{x}}_t, S) - 2f(\hat{\mathbf{x}}_t, \hat{S}_t)$ for every S. To establish such a bound, we need to relate $f(\hat{\mathbf{x}}_t, \hat{S}_t)$ to $f(\mathbf{x}_t, \hat{S}_t)$ which requires the function f to be Lipschitz with respect to $\mathbf{x}$, which is equivalent to the bounded gradient condition in Assumption 2; see proof of Theorem 2 in the appendix for more details. Note that such argument is not required for the EGG method, as in greedy based method we always have the following inequality $f(\hat{\mathbf{x}}_t, S) - (1 - 1/e)^{-1} f(\hat{\mathbf{x}}_t, \hat{S}_t) \leq 0$ for every S. As a result, the required conditions for the convergence of GRG and EGRG are similar and we only state EGRG results for completeness.

Theorem 6. *Consider Extra-Gradient Replacement Greedy(EGRG) in Algorithm 2 option II. If f is convex-submodular and Assumptions 1-2 hold, then the output of EGRG after $\mathcal{O}(1/\epsilon^2)$ iterations with step size $\mathcal{O}(\epsilon)$, is a $(1/2, \epsilon)$-approximate minimax solution of (1).*

3.3 Extra-gradient on Continuous Extension

So far all proposed algorithms achieve (α, ϵ)-approximate minimax solutions in $\mathcal{O}(1/\epsilon^2)$ iterations. In this section, we investigate the possibility of achieving a faster rate of $\mathcal{O}(1/\epsilon)$. Note that, in the discussed algorithms, the update for the discrete variable is not smooth and the iterates jump from one set to another in consecutive iterations, which results in slowing down the convergence. To overcome this limitation, we introduce the continuous multi-linear extension of Problem (1); for introduction to multi-linear extension of submodular maximization problems see (Vondrák, 2007; Calinescu et al., 2011; Badanidiyuru and Vondrák, 2014; Feldman et al., 2011; Hassani et al., 2017; Sadeghi and Fazel, 2020). As we will show, the continuous extension of Problem (1) is equivalent to its original version, and by extending the extra-gradient methodology to this setting we achieve a convergence rate of $\mathcal{O}(1/\epsilon)$ for the case that $\mathcal{I}$ is a matroid.

Definition 5. *The* continuous extension *of a function $f : \mathbb{R}^d \times 2^V \to \mathbb{R}_+$ is the function $F : \mathbb{R}^d \times [0,1]^n \to \mathbb{R}_+$ defined as $F(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{S \sim \mathbf{y}}[f(\mathbf{x}, S)]$, where $S \sim \mathbf{y}$ is a random set wherein each element i is included with probability y_i independently.*

We show that for convex-submodular problems we have (see Proposition 1 in Appendix A.8):

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{S \in \mathcal{I}} f(\mathbf{x}, S) = \min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{K}} F(\mathbf{x}, \mathbf{y}), \tag{9}$$

where $\mathcal{I}$ is assumed to be a matroid constraint and $\mathcal{K}$ is the corresponding base polytope($\mathcal{K} = \mathbf{conv}\{1_S : S \in \mathcal{I}\}$). We present Extra-Gradient on Continuous Extension (EGCE) in Algorithm 3 which applies the updates of extra-gradient on the continuous extension function $F(\mathbf{x}, \mathbf{y})$.

Algorithm 3 Extra-gradient on Continuous Extension

Initialize the variables $\mathbf{y}_1$ and $\mathbf{x}_1$ to zero.
for $t = 1$ to T **do**
$\hat{\mathbf{x}}_t = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla_x F(\mathbf{x}_t, \mathbf{y}_t))$
$\hat{\mathbf{y}}_t = \pi_{\mathcal{K}}(\mathbf{y}_t + \gamma_t \nabla_y F(\mathbf{x}_t, \mathbf{y}_t))$
$\mathbf{x}_{t+1} = \pi_{\mathcal{X}}(\mathbf{x}_t - \gamma_t \nabla_x F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t))$
$\mathbf{y}_{t+1} = \pi_{\mathcal{K}}(\mathbf{y}_t + \gamma_t \nabla_y F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t))$
end for
Return solution $\mathbf{x}_{sol} = (\sum_{t=1}^T \gamma_t)^{-1} \sum_{t=1}^T \gamma_t \hat{\mathbf{x}}_t$

Theorem 7. *Consider the Extra-Gradient On Continuous Extension (EGCE) algorithm outlined in Algorithm 3. If f is convex-submodular and Assumptions 1-2 hold, then the output of this algorithm after $\mathcal{O}(1/\epsilon)$ iterations with constant step size is a $(1/2, \epsilon)$-approximate minimax solution of Problem (1).*

4 EXPERIMENTS

In this section, we study two specific instances of Problem (1): (i) convex-facility location functions along with a synthetic experimental setup, and (ii) designing adversarial attacks for item recommendation which is a real world application of our framework.

Convex-facility location functions. Consider the function $f : \mathbb{R}^d \times 2^V \to \mathbb{R}_+$ defined as $f(\mathbf{x}, S) = \sum_{i=1}^n \max_{j \in S} f_{i,j}(\mathbf{x}) + g(\mathbf{x})$, where $g : \mathbb{R}^d \to \mathbb{R}$ and $f_{i,j} : \mathbb{R}^d \to \mathbb{R}$ are convex. Indeed, $f(\mathbf{x}, S)$ is convex with respect to $\mathbf{x}$. Also, for a fixed $\mathbf{x}$, we recover the objective of the facility location problem, which is submodular and monotone. To introduce our setup, suppose $\mathbf{x} \in \mathbb{R}_+^d$ can be written as the concatenation of $n = d/m$ vectors $\mathbf{x}_i \in \mathbb{R}_+^m$ of size m, i.e., $\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_n]$. In our experiments, we assume that the function $f_{i,j}(\mathbf{x})$ is defined as $f_{i,j}(\mathbf{x}) = \mathbf{x}_i^T Q_{i,j} \mathbf{x}_j$, where $Q_{i,j} \in \mathcal{S}_{++}^m$ is a positive definite matrix and all of its elements are also positive, i.e., $Q_{i,j} > 0$. Moreover, we consider the case that the regularization function g is defined as $g(\mathbf{x}) := \lambda(\sum_{i=1}^n \|\mathbf{x}_i\|^2)^{-1}$, and the constraint set for the minimization variable $\mathbf{x}$ is defined as $\mathcal{C} := \{\mathbf{x} = [\mathbf{x}_1; \ldots; \mathbf{x}_n] | \|\mathbf{x}_i\| \leq 1, \text{ for } i = 1, \ldots, n\}$. Considering these definitions the convex-submodular minimax optimization problem that we aim to solve can be written as

$$\min_{\mathbf{x}_i \in \mathcal{C}} \max_{|S| \leq k} \sum_{i=1}^n \max_{j \in S} \mathbf{x}_i^T Q_{i,j} \mathbf{x}_j + \lambda \Big(\sum_{i=1}^n \|\mathbf{x}_i\|^2 \Big)^{-1} \tag{10}$$

where the constraint on the maximization variable S is a cardinality constraint of size k. For our numerical

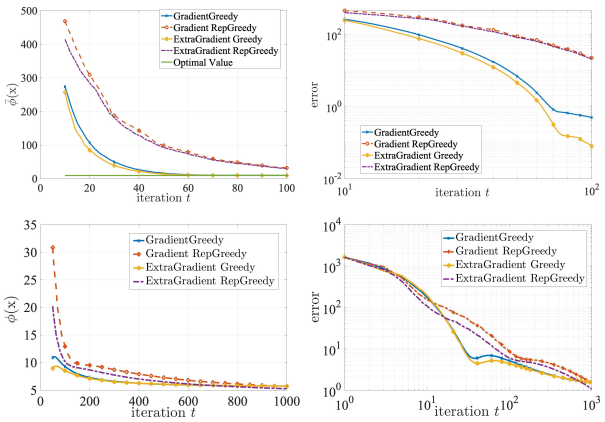

Figure 1: Comparison of our proposed methods for convex-facility location functions(case I and Case II)

experiments, we tested two cases, in the first case we set the problem parameters as $m = 10$, $n = 30$, $k = 5$, and $\lambda = 1$ and in the second case we set the problem parameters as $m = 10$, $n = 50$, $k = 10$, and $\lambda = 1$.

Case I ($m = 10$, $n = 30$, $k = 5$, $\lambda = 1$). In this case, we choose m, n to be small so that we can solve the inner max in problem (10) and compute $\bar{\phi}(\mathbf{x}) = \max_{|S|\leq k} f(\mathbf{x}, S)$ exactly using search over all the subsets of size k. We report $\bar{\phi}(\mathbf{x}_t)$ as well as optimal value of problem (10). Results in Figure 1, (first plot) show that the algorithms converge to the optimal minimax value. We also demonstrate the relative error of these algorithms $\text{error}_t := \phi(\mathbf{x}_t) - \text{OPT}$ in second plot. As we observe in Figure 1 (second plot), greedy based methods converge faster than replacement greedy based algorithms in terms of iteration complexity.

Case II ($m = 10$, $n = 50$, $k = 10$, $\lambda = 1$). We now investigate the behavior of our proposed methods for solving (10) in the second case when k, n are relatively larger. Note that exact computation of $\bar{\phi}(\mathbf{x}) = \max_{|S|\leq k} f(\mathbf{x}, S)$ is not computationally tractable for this case, since it requires solving a submodular maximization problem. Hence, in third plot in figure 1, we report the value of the function $\phi(\mathbf{x}) := f(\mathbf{x}, \textsc{Greedy}(f, k, \mathbf{x}))$ which is an approximation for $\bar{\phi}(\mathbf{x})$. In other words, instead of computing $\bar{\phi}(\mathbf{x})$ which is the maximum of $f(\mathbf{x}, S)$ over the choice of S, we report $\phi(\mathbf{x})$ which is the value of $f(\mathbf{x}, S)$ when S is obtained via the greedy method. The convergence paths of $\phi(\mathbf{x}_t)$ for our proposed methods are reported in the third plot of Figure 1. We further show the relative error of these algorithms defined as $\text{error}_t := \phi(\mathbf{x}_t) - \phi(\mathbf{x}_T)$ in the fourth plot to better compare their convergence rates.

Adversarial Attack for Item Recommendation.

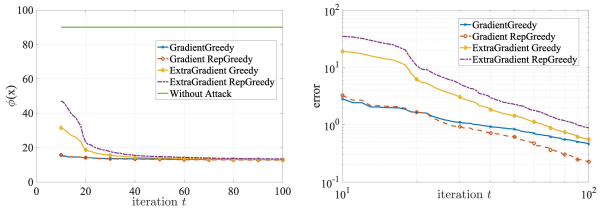

Figure 2: Comparison of our proposed methods for for Problem (11) (the green line is the performance of the recommender system when there is no adversary)

In this section, we study the application of designing an adversarial attack for a movie recommendation task. Consider a (completed) rating matrix X whose entries $X_{i,j}$ correspond to the estimated rating that user i has given to movie j. Given a rating matrix X, the recommender system chooses k movies via maximizing the utility function $\max_{|S|\leq k} h(X, S) := \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \max_{j\in S} X_{u,j}$ where $\mathcal{U}$ is the set of all users. The attacker's goal is to slightly perturb the rating matrix X to a matrix X' such that the utility $\max_{|S|\leq l} h(X', S)$ is minimized. Therefore, the attacker aims at solving the minimax problem

$$\min_{\substack{\|X'-X\|_F\leq\epsilon \\ 0\leq X'_{i,j}\leq 5}} \max_{|S|\leq k} h(X', S), \tag{11}$$

where $\|.\|_F$ is the Frobenius norm. Note that $h(X, S)$ is convex-submodular (convexity in $\mathbf{x}$ is clear, and the function $h(\mathbf{x}, S)$ is a facility location function in S). Hence, this problem is an instance of Problem (1). To evaluate the performance of our methods, we consider movie recommendation on the Movielens dataset (Harper and Konstan, 2015). We pick 2000 most rated movies with 200 users with highest number of rates for these movies (similar to (Stan et al., 2017b; Adibi et al., 2020)) and we set $k = 10$. The adversary has a power to manipulate up to 0.5% of movies ratings on average (i.e. $\epsilon = 0.5 \times 0.01 \times 200 \times 2000$). We plot $\phi(X_{\text{alg}})$ in each iteration as a measure of effectiveness of our algorithms and compare it to the case that there is no attack. Figure 2 shows the comparison of our algorithms. As we can see in Figure 2, the facility location based recommendation systems are extremely vulnerable to adversarial attacks and the performance drops from 90 (when there is no adversary) to around 12 when we have attacks.

5 CONCLUSION

In this paper, we introduced and studied the convex-submodular minimax problem in (1). We defined multiple notions of (near-) optimality and provided hardness results regarding these notions in various regimes. In

particular, one of the notions was (α, ϵ)-approximate minimax solution. We showed that for $\alpha > 1 - 1/e$ finding an (α, ϵ)-approximate minimax solution is hard. For $\alpha \leq 1 - 1/e$, we proposed five algorithms and characterized their theoretical guarantees in different settings. The main take-away message from our algorithmic procedures is that, if the function f has bounded gradient, then one can use the GG Algorithm, or the GRG algorithm which has a better complexity albeit it has a worse approximation factor. If the gradient of f is not uniformly bounded, then one has to resort to the proposed EGG algorithm.

An interesting future direction is to find more computationally efficient algorithms for harder constraints such as matroid constraint with $1 - 1/e$ factor. We believe gradient continuous-greedy will achieve the $1 - 1/e$ factor on matroid constraints; However, it needs to run the continuous greedy algorithm each iteration which is computationally costly.

Acknowledgement

The work of A. Adibi and H. Hassani is funded by NSF award CPS-1837253, NSF CAREER award CIF-1943064, and Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award FA9550-20-1-0111. This research of A. Mokhtari is supported in part by NSF Grants 2127697, 2019844, and 2112471, ARO Grant W911NF2110226, the Machine Learning Lab (MLL) at UT Austin, and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program.

References

A. Adibi, A. Mokhtari, and H. Hassani. Submodular meta-learning. *arXiv preprint arXiv:2007.05852*, 2020.

M. Alkousa, D. Dvinskikh, F. Stonyakin, A. Gasnikov, and D. Kovalev. Accelerated methods for composite non-bilinear saddle point problem. *arXiv preprint arXiv:1906.03620*, 2019.

N. Anari, N. Haghtalab, S. Naor, S. Pokutta, M. Singh, and A. Torrico. Structured robust submodular maximization: Offline and online algorithms. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3128–3137. PMLR, 2019.

A. Badanidiyuru and J. Vondrák. Fast algorithms for maximizing submodular functions. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1497–1514. SIAM, 2014.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton university press, 2009.

D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011.

I. Bogunovic, S. Mitrović, J. Scarlett, and V. Cevher. A distributed algorithm for partitioned robust submodular maximization. In *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5. IEEE, 2017a.

I. Bogunovic, S. Mitrović, J. Scarlett, and V. Cevher. Robust submodular maximization: A non-uniform partitioning approach. In *International Conference on Machine Learning*, pages 508–516. PMLR, 2017b.

I. Bogunovic, J. Zhao, and V. Cevher. Robust maximization of non-submodular objectives. In *International Conference on Artificial Intelligence and Statistics*, pages 890–899. PMLR, 2018.

G. Calinescu, C. Chekuri, M. Pal, and J. Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. *SIAM Journal on Computing*, 40(6):1740–1766, 2011.

R. Chen, B. Lucier, Y. Singer, and V. Syrgkanis. Robust optimization for non-convex objectives. *arXiv preprint arXiv:1707.01047*, 2017.

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

J. Diakonikolas, C. Daskalakis, and M. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2746–2754. PMLR, 2021.

S. Ebrahimi, W. Gan, D. Chen, G. Biamby, K. Salahi, M. Laielli, S. Zhu, and T. Darrell. Minimax active learning. *arXiv preprint arXiv:2012.10467*, 2020.

F. Farnia and D. Tse. A minimax approach to supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4240–4248, 2016.

M. Feldman, J. Naor, and R. Schwartz. A unified continuous greedy algorithm for submodular maximization. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 570–579. IEEE, 2011.

A. Ginart, M. Y. Guan, G. Valiant, and J. Zou. Making ai forget you: Data deletion in machine learning. *arXiv preprint arXiv:1907.05012*, 2019.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

E. Y. Hamedani and N. S. Aybat. A primal-dual algorithm for general convex-concave saddle point problems. *arXiv preprint arXiv:1803.01401*, 2, 2018.

F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

H. Hassani, M. Soltanolkotabi, and A. Karbasi. Gradient methods for submodular maximization. *arXiv preprint arXiv:1708.03949*, 2017.

M. Hast, K. J. Åström, B. Bernhardsson, and S. Boyd. Pid design by convex-concave optimization. In *2013 European Control Conference (ECC)*, pages 4460–4465. IEEE, 2013.

A. Ibrahim, W. Azizian, G. Gidel, and I. Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593. PMLR, 2020.

R. Iyer. A unified framework of constrained robust submodular optimization with applications, 2021.

C. Jin, P. Netrapalli, and M. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*, pages 4880–4889. PMLR, 2020.

E. Kazemi, M. Zadimoghaddam, and A. Karbasi. Scalable deletion-robust submodular maximization: Data summarization with privacy and fairness constraints. In *International conference on machine learning*, pages 2544–2553. PMLR, 2018.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

A. Krause and D. Golovin. Submodular function maximization. *Tractability*, 3:71–104, 2014.

A. Krause, H. B. McMahan, C. Guestrin, and A. Gupta. Robust submodular observation selection. *Journal of Machine Learning Research*, 9(12), 2008.

G. R. Lanckriet, L. E. Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3 (Dec):555–582, 2002.

Q. Lei, L. Wu, P.-Y. Chen, A. G. Dimakis, I. S. Dhillon, and M. Witbrock. Discrete adversarial attacks and submodular optimization with applications to text classification. *arXiv preprint arXiv:1812.00151*, 2018.

H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. *Advances in neural information processing systems*, 28:379–387, 2015.

S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, and S. Russell. Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4213–4220, 2019.

T. Lin, C. Jin, and M. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020a.

T. Lin, C. Jin, and M. I. Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020b.

B. Mirzasoleiman, A. Karbasi, and A. Krause. Deletion-robust submodular maximization: Data summarization with "the right to be forgotten". In *International Conference on Machine Learning*, pages 2449–2458. PMLR, 2017.

B. Mirzasoleiman, K. Cao, and J. Leskovec. Coresets for robust training of neural networks against noisy labels. *arXiv preprint arXiv:2011.07451*, 2020.

S. Mitra, M. Feldman, and A. Karbasi. Submodular+ concave. *Advances in Neural Information Processing Systems*, 34, 2021.

M. Mitrovic, E. Kazemi, M. Zadimoghaddam, and A. Karbasi. Data summarization at scale: A two-stage submodular approach. *arXiv preprint arXiv:1806.02815*, 2018.

S. Mitrović, I. Bogunovic, A. Norouzi-Fard, J. Tarnawski, and V. Cevher. Streaming robust submodular maximization: A partitioned thresholding approach. *arXiv preprint arXiv:1711.02598*, 2017.

A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020a.

A. Mokhtari, A. E. Ozdaglar, and S. Pattathil. Convergence rate of o(1/k) for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.

K. G. Murty and S. N. Kabadi. Some np-complete problems in quadratic and nonlinear programming. Technical report, 1985.

S. Neel, A. Roth, and S. Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. *arXiv preprint arXiv:2007.02923*, 2020.

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

A. Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz

continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

M. Nouiehed, M. Sanjabi, T. Huang, J. D. Lee, and M. Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *arXiv preprint arXiv:1902.08297*, 2019.

J. B. Orlin, A. S. Schulz, and R. Udwani. Robust monotone submodular function maximization. *Mathematical Programming*, 172(1):505–537, 2018.

M. J. Osborne and A. Rubinstein. *A course in game theory*. MIT press, 1994.

Y. Ouyang and Y. Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, pages 1–35, 2019.

O. Sadeghi and M. Fazel. Online continuous dr-submodular maximization with long-term budget constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 4410–4419. PMLR, 2020.

S. Sakaue. Differentiable greedy algorithm for monotone submodular maximization: Guarantees, gradient estimators, and applications. In *International Conference on Artificial Intelligence and Statistics*, pages 28–36. PMLR, 2021.

M. Sanjabi, M. Razaviyayn, and J. D. Lee. Solving non-convex non-concave min-max games under polyak-lojasiewicz condition. *arXiv preprint arXiv:1812.02878*, 2018.

A. Schrijver. *Combinatorial optimization: polyhedra and efficiency*, volume 24. Springer Science & Business Media, 2003.

P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe. Curriculum learning: A survey. *arXiv preprint arXiv:2101.10382*, 2021.

M. Staib, B. Wilder, and S. Jegelka. Distributionally robust submodular maximization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 506–516. PMLR, 2019.

S. Stan, M. Zadimoghaddam, A. Krause, and A. Karbasi. Probabilistic submodular maximization in sub-linear time. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3241–3250. PMLR, 06–11 Aug 2017a. URL http://proceedings.mlr.press/v70/stan17a.html.

S. Stan, M. Zadimoghaddam, A. Krause, and A. Karbasi. Probabilistic submodular maximization in sub-linear time. In *International Conference on Machine Learning*, pages 3241–3250. PMLR, 2017b.

K. K. Thekumparampil, P. Jain, P. Netrapalli, and S. Oh. Efficient algorithms for smooth minimax optimization. *arXiv preprint arXiv:1907.01543*, 2019.

S. Tschiatschek, A. Sahin, and A. Krause. Differentiable submodular maximization. *arXiv preprint arXiv:1803.01785*, 2018.

P. Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2): 237–252, 1995.

J. Vondrák. Submodularity in combinatorial optimization. 2007.

B. Wilder. Equilibrium computation and robust optimization in zero sum games with submodular structure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

B. Wilder, B. Dilkina, and M. Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1658–1665, 2019.

L. A. Wolsey. An analysis of the greedy algorithm for the submodular set covering problem. *Combinatorica*, 2(4):385–393, 1982.

Y. Wu, E. Dobriban, and S. Davidson. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*, pages 10355–10366. PMLR, 2020.

J. Yang, N. Kiyavash, and N. He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020.

R. Zhao. Optimal stochastic algorithms for convex-concave saddle-point problems. *arXiv preprint arXiv:1903.01687*, 2019.

K. Zhou and J. C. Doyle. *Essentials of robust control*, volume 104. Prentice hall Upper Saddle River, NJ, 1998.

T. Zhou and J. A. Bilmes. Minimax curriculum learning: Machine teaching with desirable difficulties and scheduled diversity. In *ICLR (Poster)*, 2018.

T. Zhou, S. Wang, and J. A. Bilmes. Curriculum learning by dynamic instance hardness. *Advances in Neural Information Processing Systems*, 33, 2020.

T. Zhou, S. Wang, and J. Bilmes. Curriculum learning by optimizing learning dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 433–441. PMLR, 2021.

A SUPPLEMENTARY MATERIALS

A.1 Proof of Theorem 1

Consider the function $f : \mathbb{R}^d \times 2^V \to \mathbb{R}_+$, where $f(\mathbf{x}, .)$ is submodular for every $\mathbf{x}$ and $f(., S)$ is convex for every S. Then, the maxmin convex-submodular problem is an optimization problem where the maximization is over continuous variable and minimization is over a discrete variable as

$$\text{OPT}_{maxmin} \triangleq \max_{S\in\mathcal{I}} \min_{x\in\mathcal{X}} f(\mathbf{x}, S), \tag{12}$$

Let us define the notion of approximate solution for maxmin problem as follows:

Definition 6. *We call a point $\hat{S}$ an (α, ϵ)-approximate maxmin solution of Problem (12) if it satisfies*

$$\alpha\underline{\phi}(\hat{S}) \geq \text{OPT}_{\text{maxmin}} - \alpha\epsilon, \tag{13}$$

We know any $(\alpha, \epsilon)-$saddle point, denoted by $(\bar{\mathbf{x}}, \bar{S})$, has the following properties:

1. $\underline{\phi}(\bar{S}) > \alpha.\text{OPT}_{maxmin} - \epsilon$
2. $\bar{\phi}(\bar{\mathbf{x}}) < \frac{1}{\alpha}.\text{OPT}_{minmax} + \frac{\epsilon}{\alpha}$

This is due to the fact that we have:

1. $\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, \bar{S}) \leq \min_{\mathbf{x}\in\mathcal{X}} \max_{S\in\mathcal{I}} f(\mathbf{x}, S) = \text{OPT}_{minmax}$
2. $\text{OPT}_{maxmin} = \max_{S\in\mathcal{I}} \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}, S) \leq \max_{S\in\mathcal{I}} f(\bar{\mathbf{x}}, S)$

these two conditions imply that by finding an $(\alpha, \epsilon)-$saddle point we find an $\alpha-$approximate solution for the minimax problem (1) and a $\frac{1}{\alpha}-$approximate solution for the max-min problem (12). In order to prove finding $(\alpha, \epsilon)-$ saddle point is NP-hard, we will prove that finding approximate solution for maxmin convex-submodular is NP-hard. We do this establishing a connection between this problem and the problem of robust submodular maximization through following result stated and proved in (Krause et al., 2008).

Consider monotone-submodular functions $f_1, f_2, \dots f_n$ and the following robust submodular maximization problem:

$$\text{OPT}_1 = \max_{|S|\leq k} \min_{i\in[n]} f_i(S). \tag{14}$$

Solving this problem up to approximation factor is NP-hard, i.e. finding a solution S such that $\max_i f_i(S) \geq \alpha\text{OPT}_1$ is an NP-hard task for any $\alpha > 0$.

Now, consider the following problem:

$$\text{OPT}_2 = \max_{|S|\leq k} \min_{\substack{x\in\mathbb{R}^n, x\geq 0 \\ x^T\mathbb{1}=1}} \sum_{i=1}^{n} x_i.f_i(S) \tag{15}$$

where $\mathbb{1}$ is vector of all ones and $x = [x_1, x_2, x_3 \dots x_n]^T$. For this problem, it is easy to verify that $\text{OPT}_1 = \text{OPT}_2$ since for every set $S \in V$ we have $\min_{i\in[n]} f_i(S) = \min_{\substack{x\in\mathbb{R}^n, x\geq 0 \\ x^T\mathbb{1}=1}} \sum_{i=1}^{n} x_i.f_i(S)$. Therefore, finding a $\alpha-$approximate solution for problem in (15) is NP-hard. Problem (15) is max-min convex-submodular optimization which means max-min convex-submodular optimization is NP-hard in general. We show that by finding $(\alpha, \epsilon)-$saddle point we can provide $\frac{1}{\alpha}-$approximate solution for max-min problem; therefore, since we proved finding $\frac{1}{\alpha}-$approximate solution for max-min problem is NP-hard, finding $(\alpha, \epsilon)-$saddle point is NP-hard too.

A.2 Proof of Theorem 2

Before stating this proof, let us explain what we mean by "NP-hard" for the considered setting. We note that an algorithm for Problem (1) is supposed to search for an approximate solution only in $\mathcal{X}$ (i.e., in terms of the variable $\mathbf{x}$), and for this, it will require some information about the values $f(\mathbf{x}, S)$. However, for every fixed $\mathbf{x}$, there may be restrictions on obtaining some specific values of $f(\mathbf{x}, S)$. For example, finding the exact value of $\bar{\phi}(\mathbf{x})$ can in general be NP-hard (as maximizing a monotone-submodular function beyond $(1-1/e)$ approximation is hard). In order to appropriately address these restrictions, we will view our setting as a procedure between the algorithm and an oracle that we now describe below.

Upon receiving an input point $\mathbf{x}_{in} \in \mathcal{X}$, the oracle chooses based on this input a set S_{out} such that $|S_{out}| \leq k$, and returns all the following information: the set S_{out}, the value $f(\mathbf{x}_{in}, S_{out})$, and the gradient of $f(\mathbf{x}_{in}, S_{out})$ with respect to $\mathbf{x}$ at the point $(\mathbf{x}_{in}, S_{out})$. The only restriction on the oracle is that it is a polynomial-time oracle, i.e. the oracle's procedure to find the set S_{out} requires poly-time complexity in terms of the size of the ground-set $|V|$. More precisely, there exists an integer $q > 0$ such that for any ground set V, the oracle uses at most $\mathcal{O}(|V|^q)$ operations to find the output set S_{out} corresponding to an input $\mathbf{x}_{in}$. Note that we do not put any restriction on what the oracle does apart from having poly-time complexity; e.g. it could output a greedy solution, or it could output a random set S_{out}, or it could do any other procedure. We call such an oracle a polynomial-time oracle. Given this choice of the oracle, the algorithm proceeds in m rounds, and in each round $r \in [m]$, it chooses an input point $\mathbf{x}_r \in \mathcal{X}$ to query from the oracle. Importantly, we consider algorithms that require a polynomial number of rounds in terms of the size of the ground set V. More precisely, for any ground set V and α, ϵ, the number of rounds of the algorithm is at most $c(\alpha, \epsilon)|V|^q$ where $q > 0$ is an absolute constant and $c(\alpha, \epsilon)$ is another constant that only depends on α and ϵ. We call such an algorithm a **polynomial-time algorithm**. Next, we show that no polynomial-time algorithm is capable of finding an (α, ϵ)-approximate of (1) for $\alpha > 1 - 1/e$.

In the following, we assume for simplicity that $\epsilon = 0$. The proof can be trivially extended to any value of ϵ, as we will explain at the end of the proof. Recall from the statement of the theorem that $\alpha = 1 - 1/e + \gamma$ for a fixed constant $\gamma > 0$.

We know for a fact that monotone-submodular maximization beyond the $(1-1/e)$-approximation in NP-hard (Krause and Golovin, 2014). I.e. unless P = NP, for any integer $q > 0$ there exists a monotone submodular function $g_1 : 2^V \to \mathbb{R}_+$ and an integer $k \leq |V|$ such that finding a set S with cardinally k where $g_1(S) \geq (1 - 1/e + \gamma/3) \max_{|S| \leq k} g_1(S)$ requires computing more than $|V|^q$ function values (i.e. complexity is larger than $|V|^q$). Consider such a function g_1 and the choice of k, and define $\text{OPT}_{g_1} = \max_{|S| \leq k} g_1(S)$. We also define another function $g_2 : 2^V \to \mathbb{R}_+$ as follows: $g_2(S) = \min\{g_1(S), (1 - 1/e + \gamma/3)\text{OPT}_{g_1}\}$. It is important to note that finding a set $|S| \leq k$ such that $g_1(S) \neq g_2(S)$ requires complexity larger than $|V|^q$ (also note that the choice of q is arbitrary here, i.e. for every q there exists a g_1, etc.).

Consider the integer $n = \lceil 4/\gamma + 1 \rceil$ and let $\mathcal{X}$ be the n-dimensional simplex, i.e.

$$\mathcal{X} = \{\mathbf{x} = (x_1, \cdots, x_n) \text{ s.t. } \sum_{i=1}^{n} x_i = 1 \,\&\, x_i \geq 0 \ \forall i = 1, \cdots, n\}.$$

For $j \in \{1, \cdots, n\}$ we define $f_j(\mathbf{x}, S) : \mathcal{X} \times 2^V \to \mathbb{R}_+$ as

$$f_j(\mathbf{x}, S) = \sum_{i \in [n], i \neq j} x_i g_1(S) + x_j g_2(S)$$

We note a few facts about each of the functions $f_j(\mathbf{x}, S)$:

(i) Any $(\alpha, 0)$ approximate solution for the function f_j has the property that $x_j > 1/n + \gamma/4$. This is simply because for any $(\alpha, 0)$-approximate solution $\mathbf{x} = (x_1, \cdots, x_n)$ we have $\alpha\bar{\phi}(\mathbf{x}) \leq \min_{\mathbf{x} \in \mathcal{X}} \max_{|S| \leq k} f_j(\mathbf{x}, S)$, and hence

$$\alpha \left(x_j(1 - 1/e + \gamma/3) + (1 - x_j)\right) \text{OPT}_{g_1} \leq (1 - 1/e + \gamma/3)\text{OPT}_{g_1}$$

From the above inequality (and by noting that $\gamma \in (0, 1]$) we can always deduce that $x_j > \gamma/2$, and thus $x_j > 1/n + \gamma/4$.

(ii) Given a polynomial-time oracle, we can not distinguish between the functions $f_1, \ldots, f_n$ using a query from the oracle. This is because the oracle can not find a set S with carnality at most k for which $g_1(S) \neq g_2(S)$ (as finding that set by the oracle is intractable), and thus, for the set S_{out} that the oracle finds, the outcome of the oracle will be the function value $f(\mathbf{x}_{\text{in}}, S_{\text{out}}) = g_1(S_{\text{out}})$ and $\nabla_{\mathbf{x}} f(\mathbf{x}_{\text{in}}, S_{\text{out}}) = g_1(S_{\text{out}})\mathbf{1}_n$ where $\mathbf{1}_n$ is the all-ones vector of dimension n. These outputs bear absolutely no information about the index j.

Given the above facts, we are now ready to finalize the proof. Consider the scenario where the index j is chosen uniformly at random inside the set $[n]$, and the algorithm aims at finding an approximate solution of the function f_j. Note that the choice of j is hidden to the algorithm. Now, given fact (ii) above, if both the algorithm and oracle are polynomial-time, then in all the rounds and queries, there will be absolutely no information revealed about the index j. As a result, the mutual information between the outcome of the queries and the index j will be zero.

On the other hand, from fact (i) above, if an algorithm can find an $(\alpha, 0)$-approximate solution, we claim that the solution is informative about the index j. More precisely, given the solution that the algorithm has found, we can infer the hidden index j using the following procedure: the algorithm's solution $\mathbf{x} = (x_1, \cdots, x_n)$ can be viewed as a probability distribution over the set $\{1, \cdots, n\}$. As a result, if we use this probability distribution to draw an integer $\hat{j}$ from the set $\{1, \cdots, n\}$, then we have $\Pr\{\hat{j} = j\} = x_j >= 1/n + \gamma/4$. Thus, we can decode the index j with a probability that is strictly larger than a random guess. This means that the mutual information of the solution found by the algorithm and the index j is strictly lower-bounded by a positive constant (which only depends on γ). This contradicts the result of the previous paragraph.

Note that in the above we have assumed that $\epsilon = 0$. For general ϵ, we note that we can always choose the function g_1 such that OPT_{g_1} is sufficiently large. As a result, we can write $\epsilon = \epsilon' \times \text{OPT}_{g_1}$ where ϵ' can be made arbitrarily small. Hence, proving hardness for obtaining an (α, ϵ)-approximate becomes equivalent to proving harness for obtaining an $(\alpha/(1+\epsilon'), 0)$ approximate solution. The conclusion is now immediate since out proof above works for any $\alpha = 1 - 1/e + \gamma$ and ϵ' can be made arbitrarily small by making OPT_{g_1} sufficiently large.

A.3 Proof of Theorem 3: Gradient Greedy(GG) Convergence

Let us pick $\gamma_t = \alpha$.Then, we can write the following based on update $x_{t+1} = \pi_{\mathcal{X}}(x_t - \gamma_t \nabla f(x_t, S_t))$ in algorithm 1 and assumption 2.

$$\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \leq \|\mathbf{x}_t - \alpha\nabla f(\mathbf{x}_t, S_t) - \mathbf{x}\|^2 = \|\mathbf{x}_t - \mathbf{x}\|^2 + \|\alpha\nabla f(\mathbf{x}_t, S_t)\|^2 - 2\langle \alpha\nabla f(\mathbf{x}_t, S_t), \mathbf{x}_t - \mathbf{x}\rangle \tag{16}$$

$$\leq \|\mathbf{x}_t - \mathbf{x}\|^2 + \alpha^2 M^2 - 2\langle \alpha\nabla f(\mathbf{x}_t, S_t), \mathbf{x}_t - \mathbf{x}\rangle \tag{17}$$

which results in

$$2\alpha(f(\mathbf{x}_t, S_t) - f(\mathbf{x}, S_t)) \leq 2\langle \alpha\nabla f(\mathbf{x}_t, S_t), \mathbf{x}_t - \mathbf{x}\rangle \leq \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 + \alpha^2 M^2 \tag{18}$$

and finally

$$f(\mathbf{x}_t, S_t) - f(\tilde{\mathbf{x}}, S_t) \leq \frac{1}{2\alpha}(-\|\mathbf{x}_t - \tilde{\mathbf{x}}\|^2 + \|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}\|^2) + \frac{1}{2}\alpha M^2 \tag{19}$$

summing up over t we have:

$$\sum_{t=1}^{T} f(\mathbf{x}_t, S_t) - f(\tilde{\mathbf{x}}, S_t) \leq \frac{1}{2\alpha}(\|\mathbf{x}_0 - \tilde{\mathbf{x}}\|^2) + \frac{1}{2}\alpha T M^2 \tag{20}$$

our set of continuous variable is bounded which means $\|\mathbf{x}\|^2 \leq H$; this results:

$$\sum_{t=1}^{T} f(\mathbf{x}_t, S_t) - f(\tilde{\mathbf{x}}, S_t) \leq \frac{H}{2\alpha} + \frac{1}{2}\alpha T M^2 \tag{21}$$

Also, from greedy update we have for every S(check (Krause and Golovin, 2014)):

$$f(\mathbf{x}_{t-1}, S) - \frac{f(\mathbf{x}_{t-1}, S_t)}{1 - \frac{1}{e}} \leq 0 \tag{22}$$

Now, using the Lipschitz condition (consequence of Assumption 2):

$$|f(\mathbf{x}_t, S_t) - f(\mathbf{x}_{t-1}, S_t)| \leq M\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \leq M\alpha\|\nabla f(\mathbf{x}_{t-1}, S_t)\| \leq M^2\alpha \tag{23}$$

Putting (22) and (23) together:

$$(1 - \frac{1}{e})f(\mathbf{x}_t, S) - f(\mathbf{x}_t, S_t) \leq 2M^2\alpha \tag{24}$$

and summing over t we have:

$$\sum_{t=1}^{T}(1 - \frac{1}{e})f(\mathbf{x}_t, S) - f(\mathbf{x}_t, S_t) \leq 2M^2\alpha T \tag{25}$$

From (21) and (25) we can then obtain the following:

$$\sum_{t=1}^{T}(1 - \frac{1}{e})f(\mathbf{x}_t, S) - f(\tilde{\mathbf{x}}, S_t) \leq 2M^2\alpha T + \frac{H}{2\alpha} + \frac{1}{2}\alpha T M^2 \tag{26}$$

and finally:

$$\frac{\sum_{t=1}^{T}\alpha((1 - \frac{1}{e})f(\mathbf{x}_t, S) - f(\tilde{\mathbf{x}}, S_t))}{\sum_{t=1}^{T}\alpha} \leq \frac{3M^2\alpha T + \frac{H}{2\alpha}}{T} \tag{27}$$

From convexity we have:

$$f(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t, S) \leq \frac{\sum_{t=1}^{T} f(\mathbf{x}_{t-1}, S)}{T} \tag{28}$$

which results in:

$$(1 - \frac{1}{e})f(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t, S) - \frac{\sum_{t=1}^{T}\alpha(f(\tilde{\mathbf{x}}, S_t))}{\sum_{t=1}^{T}\alpha} \leq \frac{3M^2\alpha T + \frac{H}{2\alpha}}{T} \tag{29}$$

Defining $\mathbf{x}^* = \arg\min\max f(\mathbf{x}, S)$, we know that $\min_{\mathbf{x}}\max_S f(\mathbf{x}, S) = \max f(\mathbf{x}^*, S) \geq f(\mathbf{x}^*, S_t)$. Now in (29) we let $\tilde{\mathbf{x}} = \mathbf{x}^*$ and write:

$$(1 - \frac{1}{e})\max_S f(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t, S) - \min_x \max_S f(\mathbf{x}, S) \leq \frac{3M^2\alpha T + \frac{H}{2\alpha}}{T} \tag{30}$$

Letting $\alpha = \frac{1}{\sqrt{T}}$ we obtain:

$$(1 - \frac{1}{e})\max_S f(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t, S) - \min_{\mathbf{x}} \max_S f(\mathbf{x}, S) \leq \frac{3M^2 + \frac{H}{2}}{\sqrt{T}} \tag{31}$$

Finally, if we define $K = 3M^2 + \frac{H}{2}$ and let $T = \frac{K^2}{\epsilon^2}$; then $\mathbf{x}_{sol} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t$ is a $(1 - 1/e, \epsilon)$- approximate minimax solution.

A.4 Proof of Theorem 5: Gradient Replacement-greedy(GRG) Convergence

Let g be a monotone-submodular function, and consider sets $B, S \subseteq V$ with size k. Define $e^* = \arg\max_{e \in S} g(S \setminus e) - g(S)$, and $v^* = \arg\max_{v \in V} g(S \cup v \setminus e^*) - g(S \cup v \setminus e^*)$. We have:

$$g(S \cup v^* \setminus e^*) - g(S) \geq \frac{1}{k} \sum_{v \in B} g(S \cup v \setminus e^*) - g(S)$$
$$= \frac{1}{k} \sum_{v \in B} (g(S \cup v \setminus e^*) - g(S \cup v) + g(S \cup v) - g(S)) \tag{32}$$

where the first inequality comes from the definition of v^*. We know that for a monotone-submodular function g we have $g(B \cup S) - g(S) \leq \sum_{v \in B} (g(S \cup v) - g(S))$ for any choice of B, S (Stan et al., 2017b); which results in:

$$\frac{1}{k} \sum_{v \in B} (g(S \cup v) - g(S)) \geq \frac{1}{k} (g(B \cup S) - g(S)) \geq \frac{1}{k} (g(B) - g(S)) \tag{33}$$

Here, the first inequality is due to submodularity and the second inequalities is due to monotonicity. Also, we have:

$$\frac{1}{k} \sum_{v \in B} (g(S \cup v) - g(S \cup v \setminus e^*)) \leq \frac{1}{k} \sum_{v \in B} (g(S) - g(S \setminus e^*)) = g(S) - g(S \setminus e^*)$$
$$\leq \frac{1}{k} \sum_{e \in S} (g(S) - g(S \setminus e)) \leq \frac{1}{k} g(S) \tag{34}$$

where the first and second inequality comes from submodularity. Combining (32),(33), and (34) we have that for every set B of size k:

$$g(S \cup v^* \setminus e^*) - g(S) \geq \frac{1}{k} (g(B) - 2g(S)) \tag{35}$$

If we apply (35) for the replacement greedy update in Gradient Replacement-greedy(GRG) algorithm, we obtain:

$$f(\mathbf{x}_t, S_t) - f(\mathbf{x}_t, S_{t-1}) \geq \frac{1}{k} \left(f(\mathbf{x}_t, S) - 2f(\mathbf{x}_t, S_{t-1}) \right) \tag{36}$$

and hence

$$f(\mathbf{x}_t, S) - 2f(\mathbf{x}_t, S_t) \leq (1 - \frac{2}{k})(f(\mathbf{x}_t, S) - 2f(\mathbf{x}_t, S_{t-1})) \tag{37}$$

Note that as f is M-Lipschitz we have for every S (consequence of Assumption 2):

$$|f(\mathbf{x}_t, S) - f(\mathbf{x}_{t-1}, S)| \leq M \|\mathbf{x}_t - \mathbf{x}_{t-1}\| \leq M\alpha \|\nabla f(\mathbf{x}_{t-1}, S_t)\| \leq M^2 \alpha \tag{38}$$

Combining (37) and (38) we obtain that

$$f(\mathbf{x}_t, S) - 2f(\mathbf{x}_t, S_t) \leq (1 - \frac{2}{k})(f(\mathbf{x}_{t-1}, S) - 2f(\mathbf{x}_{t-1}, S_{t-1}) + 3M^2\alpha) \tag{39}$$

Using a recursive argument we can show that

$$f(\mathbf{x}_t, S) - 2f(\mathbf{x}_t, S_t) \leq (1 - \frac{2}{k})^t (f(\mathbf{x}_0, S) - 2f(\mathbf{x}_0, S_0)) + \sum_{m=1}^{t} (1 - \frac{2}{k})^m 3M^2\alpha \tag{40}$$

Now since $f(\mathbf{x}_0, S_0)$ is non-negative, we can eliminate $-f(\mathbf{x}_0, S_0)$ from the right hand side. Using this observation and by simplifying the geometric sum we obtain that

$$f(\mathbf{x}_t, S) - 2f(\mathbf{x}_t, S_t) \leq (1 - \frac{2}{k})^t f(\mathbf{x}_0, S) + 3M^2\alpha \frac{k}{2} \tag{41}$$

Now, note that $(1-\frac{2}{k})^t$ is bounded above by $e^{-\frac{2t}{k}}$ and therefore we have

$$f(\mathbf{x}_t, S) - 2f(\mathbf{x}_t, S_t) \leq Ae^{-\frac{2t}{k}} + 3M^2\alpha\frac{k}{2}, \tag{42}$$

where A is an upper bound for function value at point zero, $f(0, S) \leq A$. Now, from the analysis of gradient descent similar to (21), we have:

$$\sum_{t=1}^{T} f(\mathbf{x}_t, S_t) - f(\tilde{\mathbf{x}}, S_t) \leq \frac{H}{2\alpha} + \frac{1}{2}\alpha T M^2 \tag{43}$$

Combining this inequality with (42) we have:

$$\sum_{t=1}^{T} \frac{1}{2} f(\mathbf{x}_t, S) - f(\tilde{\mathbf{x}}, S_t) \leq \frac{H}{2\alpha} + \frac{\sum_{t=1}^{T} Ae^{-\frac{2t}{k}}}{2} + \frac{5TM^2\alpha k}{4} \leq \frac{K}{2\alpha} + \frac{Ae^{-\frac{2}{k}}}{2(1-e^{-\frac{2}{k}})} + \frac{5TM^2\alpha k}{4} \tag{44}$$

Thus, choosing the parameters $\alpha = \frac{1}{\sqrt{T}}$ will lead to

$$\frac{1}{T}\sum_{t=1}^{T} \frac{1}{2} f(\mathbf{x}_t, S) - f(\tilde{\mathbf{x}}, S_t) \leq \frac{H}{2\sqrt{T}} + \frac{Ae^{-\frac{2}{k}}}{2T(1-e^{-\frac{2}{k}})} + \frac{5M^2 k}{4\sqrt{T}} \leq \frac{K}{\sqrt{T}} \tag{45}$$

where K is some constant.

In summary, we have obtained the following relation that will be used to drive the guarantee for the minimax problem:

$$\frac{1}{T}\sum_{t=1}^{T} \frac{1}{2} f(\mathbf{x}_t, S) - f(\tilde{\mathbf{x}}, S_t) \leq \frac{K}{\sqrt{T}} \tag{46}$$

We know because of convexity we have:

$$f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{t-1}, S\right) \leq \frac{\sum_{t=1}^{T} f(\mathbf{x}_{t-1}, S)}{T} \tag{47}$$

Now combining (46) and (47) we have:

$$\frac{1}{2} f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t, S\right) - \frac{1}{T}\sum_{t=1}^{T} f(\tilde{\mathbf{x}}, S_t) \leq \frac{K}{\sqrt{T}} \tag{48}$$

Also for $\mathbf{x}^* = \arg\min_x \max_S f(\mathbf{x}, S)$ we have $\min_{\mathbf{x}} \max_S f(\mathbf{x}, S) = \max_S f(\mathbf{x}^*, S) \geq f(\mathbf{x}^*, S_t)$. By using $\tilde{\mathbf{x}} = \mathbf{x}^*$ we can write:

$$\frac{1}{2} f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t, S\right) - \min_{\mathbf{x}} \max_{S} f(\mathbf{x}, S) \leq \frac{1}{2} f\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t, S\right) - \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}^*, S_t) \leq \frac{K}{\sqrt{T}} \tag{49}$$

Let $T = \frac{K^2}{\epsilon^2}$; then $\mathbf{x}_{sol} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$ is a $(1/2, \epsilon)$- approximate minimax solution.

A.5 Proof of Theorem 4: Extra-gradient Greedy(EGG) Convergence

Consider the Extra-gradient Greedy method, we can write the following equations to find the bound on convergence of $\mathbf{x}$:

$$\begin{aligned}
&\|\hat{\mathbf{x}}_t - \mathbf{x}\|^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x} - \gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)\|^2 \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\mathbf{x}_t - \mathbf{x}) + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_t) \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + 2(\mathbf{x}_t - \hat{\mathbf{x}}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_t) \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 - 2\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2
\end{aligned} \tag{50}$$

Hence, we have

$$2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) \leq \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \tag{51}$$

Similarly we can show that

$$\begin{aligned}
&\|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x} - \gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)\|^2 \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\mathbf{x}_t - \mathbf{x}) + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}) + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2(\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_{t+1} - \mathbf{x}_t) \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}) + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - 2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2
\end{aligned} \tag{52}$$

Hence, we have

$$2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}) \leq \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \tag{53}$$

Now note that we can write $2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x})$ as

$$\begin{aligned}
& 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) && (54)\\
=\ & 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}) + 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}) && (55)\\
=\ & 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}) + 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}) && (56)\\
& + 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}) - 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}) && (57)\\
=\ & 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}) + 2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\mathbf{x}_{t+1} - \mathbf{x}) && (58)\\
& + 2\gamma_t \left(\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)\right)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}) && (59)\\
\leq\ & \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 && (60)\\
& + \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 && (61)\\
& + 2\gamma_t \left(\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1})\right) && (62)\\
=\ & -\|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 && (63)\\
& + 2\gamma_t \left(\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)\right)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}), && (64)
\end{aligned}$$

where the inequality follows from the results in (51) and (53).

Next we derive an upper bound for the inner product $\left(\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)\right)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1})$ using the smoothness of the function f, i.e.,

$$\begin{aligned}
&\left(\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)\right)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}) \\
&\leq \|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)\| \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\| \\
&\leq \left(\|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t) - \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, S_t)\| + \|\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, S_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)\|\right) \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\| \\
&\leq \left(L_{\mathbf{x},S} \|\hat{S}_t - S_t\| + L_{\mathbf{x},\mathbf{x}} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|\right) \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|
\end{aligned}$$

Now to complete our upper bound we need to bound $\|\hat{S}_t - S_t\|$ which can be done as

$$\|\hat{S}_t - S_t\| \leq \phi \|\hat{\mathbf{x}}_t - \mathbf{x}_t\| + \sigma \tag{65}$$

The above relation holds because for every two feasible set we have $\|A - B\| \leq 2k$;therefore, if we let $\sigma = 2k$ and $\phi = 1$ the above condition is always true. Considering this result we obtain that

$$\Big(\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, S_t)\Big)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}) \leq (L_{\mathbf{x},S}\phi + L_{\mathbf{x},\mathbf{x}}) \|\hat{\mathbf{x}}_t - \mathbf{x}_t\| \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\| + L_{\mathbf{x},S}\sigma \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|$$

Applying this upper bound into (54) implies that

$$\begin{aligned}
&2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) \\
&\leq -\|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \\
&\quad + 2\gamma_t (L_{\mathbf{x},S}\phi + L_{\mathbf{x},\mathbf{x}}) \|\hat{\mathbf{x}}_t - \mathbf{x}_t\| \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\| + 2\gamma_t L_{\mathbf{x},S}\sigma \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\| \\
&\leq \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 \\
&\quad + \big[-\|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + 2\gamma_t (L_{\mathbf{x},S}\phi + L_{\mathbf{x},\mathbf{x}}) \|\hat{\mathbf{x}}_t - \mathbf{x}_t\| \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|\big] \\
&\quad + 2\gamma_t L_{\mathbf{x},S}\sigma \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\| \\
&\leq \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \\
&\quad + \gamma_t (L_{\mathbf{x},S}\phi + L_{\mathbf{x},\mathbf{x}}) \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \gamma_t (L_{\mathbf{x},S}\phi + L_{\mathbf{x},\mathbf{x}}) \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 + 4\gamma_t^2 L_{\mathbf{x},S}^2 \sigma^2 \\
&\quad + \frac{1}{4}\|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 + 4\gamma_t^2 L_{\mathbf{x},S}^2 \sigma^2
\end{aligned}$$

where the third inequality holds because of the fact that $2ab \leq a^2 + b^2$, and the last inequality holds since we assume $\gamma_t(L_{\mathbf{x},S}\phi + L_{\mathbf{x},\mathbf{x}}) \leq 3/4$.

Using this result we have that

$$2\gamma_t \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) \leq \|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2 + 4\gamma_t^2 L_{\mathbf{x},S}^2 \sigma^2$$

Now by convexity of f with respect to $\mathbf{x}$ we have

$$\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_t, \hat{S}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) \geq f(\hat{\mathbf{x}}_t, \hat{S}_t) - f(\mathbf{x}, \hat{S}_t)$$

and therefore

$$f(\hat{\mathbf{x}}_t, \hat{S}_t) - f(\mathbf{x}, \hat{S}_t) \leq \frac{1}{2\gamma_t} \left(\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2\right) + 2\gamma_t L_{\mathbf{x},S}^2 \sigma^2$$

Moreover we know that

$$f(\hat{\mathbf{x}}_t, S) - \frac{1}{1 - 1/e} f(\hat{\mathbf{x}}_t, \hat{S}_t) \leq 0$$

Hence,

$$\begin{aligned}
&f(\hat{\mathbf{x}}_t, \hat{S}_t) - f(\mathbf{x}, \hat{S}_t) + (1 - 1/e) f(\hat{\mathbf{x}}_t, S) - f(\hat{\mathbf{x}}_t, \hat{S}_t) \\
&\qquad \leq \frac{1}{2\gamma_t} \left(\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2\right) + 2\gamma_t L_{\mathbf{x},S}^2 \sigma^2
\end{aligned} \tag{66}$$

Let $\gamma_t = \frac{1}{\sqrt{T}}$. Then, since $\|\mathbf{x}\|^2 \leq H$, we have

$$\frac{1}{T} \sum_{t=1}^{T} -f(\mathbf{x}, \hat{S}_t) + (1 - 1/e) f(\hat{\mathbf{x}}_t, S) \leq \frac{H}{2\sqrt{T}} + \frac{2L_{\mathbf{x},S}^2 \sigma^2}{\sqrt{T}} \tag{67}$$

Let $K = 2L_{\mathbf{x},S}^2 \sigma^2 + K/2$ then we have

$$\frac{1}{T} \sum_{t=1}^{T} -f(\mathbf{x}, \hat{S}_t) + (1 - 1/e) f(\hat{\mathbf{x}}_t, S) \leq \frac{K}{\sqrt{T}} \tag{68}$$

We know because of convexity

$$f(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{x}}_{t-1}, S) \leq \frac{\sum_{t=1}^{T} f(\hat{\mathbf{x}}_{t-1}, S)}{T} \tag{69}$$

Now combining (68) and (69) we have

$$(1-1/e)f(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{x}}_t, S) - \frac{1}{T}\sum_{t=1}^{T} f(\tilde{\mathbf{x}}, \hat{S}_t) \leq \frac{K}{\sqrt{T}} \tag{70}$$

Also for $\mathbf{x}^* = \arg\min_{\mathbf{x}} \max_S f(\mathbf{x}, S)$, we have $\min_{\mathbf{x}} \max_S f(\mathbf{x}, S) = \max_S f(\mathbf{x}^*, S) \geq f(\mathbf{x}^*, S_t)$. We let $\tilde{\mathbf{x}} = \mathbf{x}^*$ and write

$$(1-1/e)f(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{x}}_t, S) - \min_{\mathbf{x}}\max_{S} f(\mathbf{x}, S) \tag{71}$$

$$\leq (1-1/e)f(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{x}}_t, S) - \frac{1}{T}\sum_{t=1}^{T} f(\mathbf{x}^*, \hat{S}_t) \leq \frac{K}{\sqrt{T}} \tag{72}$$

Let $T = \frac{K^2}{\epsilon^2}$; then, $\mathbf{x}_{sol} = \frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{x}}_t$ is an $((1-1/e), \epsilon)$ approximate minimax solution.

A.6 Extra-gradient Replacement-greedy(EGRG) Convergence

For the analysis with respect to $\mathbf{x}$, we can show that

$$f(\hat{\mathbf{x}}_t, \hat{S}_t) - f(\mathbf{x}, \hat{S}_t) \leq \frac{1}{2\gamma_t}\left(\|\mathbf{x}_t - \mathbf{x}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}\|^2\right) + 2\gamma_t L_{\mathbf{x},S}^2\sigma^2$$

therefore:

$$\frac{\sum_{t=1}^{T}\gamma_t\left[f(\hat{\mathbf{x}}_t, \hat{S}_t) - f(\mathbf{x}, \hat{S}_t)\right]}{\sum_{t=1}^{T}\gamma_t} \leq \frac{\|\mathbf{x} - \mathbf{x}_1\|^2}{2\sum_{t=1}^{T}\gamma_t} + \frac{\sum_{t=1}^{T} 2\gamma_t^2 L_{x,S}^2\sigma^2}{\sum_{t=1}^{T}\gamma_t} \leq \frac{K_1}{\sqrt{T}} \tag{73}$$

It remains to derive an upper bound for $f(\hat{\mathbf{x}}_t, S) - 2f(\hat{\mathbf{x}}_t, \hat{S}_t)$. According to the update of replacement-greedy method, we can write the following inequalities:

$$f(\mathbf{x}_t, \hat{S}_t) - f(\mathbf{x}_t, S_t) \geq \frac{1}{k}\left(f(\mathbf{x}_t, S) - 2f(\mathbf{x}_t, S_t)\right) \tag{74}$$

and

$$f(\hat{\mathbf{x}}_t, S_{t+1}) - f(\hat{\mathbf{x}}_t, \hat{S}_t) \geq \frac{1}{k}\left(f(\hat{\mathbf{x}}_t, S) - 2f(\hat{\mathbf{x}}_t, \hat{S}_t)\right) \tag{75}$$

Using the second expression, we can write

$$\frac{1}{k}\left(f(\hat{\mathbf{x}}_{t-1}, S) - 2f(\hat{\mathbf{x}}_{t-1}, \hat{S}_{t-1})\right) \leq f(\hat{\mathbf{x}}_{t-1}, S_t) - f(\hat{\mathbf{x}}_{t-1}, \hat{S}_{t-1}) \tag{76}$$

Let $\bar{\phi}(\mathbf{x}_t) = \max_{|S|\leq k} f(\mathbf{x}_t, S)$; if we assume for every $\mathbf{x}, S$, $\|\nabla_{\mathbf{x}} f(\mathbf{x}, S)\| \leq G$ then we have:

$$|\bar{\phi}(\mathbf{x}) - \bar{\phi}(\mathbf{y})| \leq G||\mathbf{x} - \mathbf{y}|| \tag{77}$$

hence

$$\bar{\phi}(\hat{\mathbf{x}}_{t-1}) - 2f(\hat{\mathbf{x}}_{t-1}, S_t) \leq (1 - \frac{2}{k})(\bar{\phi}(\hat{\mathbf{x}}_{t-1}) - 2f(\hat{\mathbf{x}}_{t-1}, \hat{S}_{t-1})) \tag{78}$$

Note that

$$f(\hat{\mathbf{x}}_{t-1}, S_t) \leq f(\mathbf{x}_t, S_t) + L_x||\mathbf{x}_t - \hat{\mathbf{x}}_{t-1}||$$
$$\leq f(\mathbf{x}_t, \hat{S}_t) + \gamma_t G^2 \tag{79}$$
$$\leq f(\hat{\mathbf{x}}_t, \hat{S}_t) + 2\gamma_t G^2, \tag{80}$$

therefore,

$$\bar{\phi}(\hat{\mathbf{x}}_{t-1}) - 2f(\hat{\mathbf{x}}_t, \hat{S}_t) \leq (1 - \frac{2}{k})(\bar{\phi}(\hat{\mathbf{x}}_{t-1}) - 2f(\hat{\mathbf{x}}_{t-1}, \hat{S}_{t-1})) + 4\gamma_t G^2 \tag{81}$$

Also, note that

$$|\bar{\phi}(\hat{\mathbf{x}}_t) - \bar{\phi}(\hat{\mathbf{x}}_{t-1})| \leq G^2 \gamma_t. \tag{82}$$

Putting (81), (80) and (82) together, we obtain:

$$\bar{\phi}(\hat{\mathbf{x}}_t) - 2f(\hat{\mathbf{x}}_t, \hat{S}_t) \leq (1 - \frac{2}{k})(\bar{\phi}(\hat{\mathbf{x}}_{t-1}) - 2f(\hat{\mathbf{x}}_{t-1}, \hat{S}_{t-1})) + 5\gamma_t G^2 \tag{83}$$

let $\gamma_t = \frac{1}{\sqrt{T}}$ and $\bar{\phi}(\hat{\mathbf{x}}_0) - 2f(\hat{\mathbf{x}}_0, \hat{S}_0) = A_0$, then

$$\sum_{t=1}^{T} \gamma_t(\bar{\phi}(\hat{\mathbf{x}}_t) - 2f(\hat{\mathbf{x}}_t, \hat{S}_t)) \leq \sum_{t=1}^{T} \frac{1}{\sqrt{T}}(\sum_{t=0}^{t-1} \frac{5G^2}{\sqrt{T}}(1 - \frac{2}{k})^t + (1 - \frac{2}{k})^t A_0)$$
$$\leq \sum_{t=1}^{T} \frac{1}{\sqrt{T}}(\frac{5kG^2}{2\sqrt{T}} + (1 - \frac{2}{k})^t A_0) \tag{84}$$
$$\leq k\beta \tag{85}$$

where $\beta = \frac{5kG^2}{2} + \frac{kA_0}{2\sqrt{T}}$ and finally for update of S we get:

$$\frac{\sum_{t=1}^{T} \gamma_t(\bar{\phi}(\hat{\mathbf{x}}_t) - 2f(\hat{\mathbf{x}}_t, \hat{S}_t))}{\sum_{t=1}^{T} \gamma_t} \leq \frac{k\beta}{\sqrt{T}} \tag{86}$$

Adding up (86) and (73) we have:

$$\frac{\sum_{t=1}^{T} \gamma_t \left[0.5\bar{\phi}(\hat{\mathbf{x}}_t) - f(\mathbf{x}, \hat{S}_t)\right]}{\sum_{t=1}^{T} \gamma_t} \leq \frac{K_1}{\sqrt{T}} + \frac{k\beta}{\sqrt{T}} \leq \frac{K}{\sqrt{T}} \tag{87}$$

from this for every S

$$\frac{\sum_{t=1}^{T} \gamma_t \left[0.5f(\hat{\mathbf{x}}_t, S) - f(\mathbf{x}, \hat{S}_t)\right]}{\sum_{t=1}^{T} \gamma_t} \leq \frac{K}{\sqrt{T}} \tag{88}$$

Similar to (69) to (71), (88) results $\mathbf{x}_{sol} = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{x}}_t$, to be $(1/2, \epsilon)$-approximate minimax solution.

A.7 Maxmin Result

In this section, we introduce maxmin convex-submodular problem and discuss how we can exploit the algorithms described in the previous sections for the maxmin problem. Formally, consider the function $f : \mathbb{R}^d \times 2^V \to \mathbb{R}_+$, where $f(\mathbf{x}, .)$ is submodular for every $\mathbf{x}$ and $f(., S)$ is convex for every S. Then, the maxmin convex-submodular

problem is an optimization problem where the maximization is over continuous variable and minimization is over a discrete variable as

$$\text{OPT}_{maxmin} \triangleq \max_{S\in\mathcal{I}} \min_{x\in\mathcal{X}} f(\mathbf{x}, S), \tag{89}$$

Due to hardness of the max-min problem as we stated in Theorem 1 and Appendix A.1, we cannot drive the same result for the maxmin problem as we did for minimax problem. In general, finding an approximation solution for problem (89) is NP-hard. Our result as stated in theorem 8 proves that $\cup_{t=1}^{T} S_t$ is an approximate solution for (89) which has a larger cardinality than our cardinality constraint (at most Tk elements). Although, the set $\cup_{t=1}^{T} S_t$ is not feasible solution, our algorithm converges quickly, and we can use the small number of steps to solve such a problem which means even for small T the set $\cup_{t=1}^{T} S_t$ can solve maxmin problem approximately. This result is similar to the bi-criterion solutions for robust submodular maximization studied in (Krause et al., 2008), where the authors propose an approach that finds a set that violates the cardinality constraint, but it is within logarithmic factor of the constraint.

Theorem 8. *Consider all algorithms stated in Algorithms section, if the functions f is convex monotone submodular, and Assumption 1 holds (and Assumption 2 holds for Gradient Greedy(GG), and Gradient Replacement-greedy(GRG)), then the set $\cup_{t=1}^{T} S_t$ is $(\alpha, \epsilon)-$approximate solution for maxmin convex-submodular problem with cardinality constraint after $\mathcal{O}(1/\epsilon^2)$ iterations. Note that parameter α is $\alpha = (1-1/e)^{-1}$ for Gradient and Extra-gradient Greedy, $\alpha = 2$ for Gradient Replacement-greedy, $\alpha = 2 + \frac{k}{k-1}$ for Extra-gradient Replacement-greedy.*

A.7.1 Proof of Theorem 8 for Gradient Greedy(GG)

If we let $S^* = \arg\max_S \min_{\mathbf{x}} f(\mathbf{x}, S)$ we know that for every t we have $f(\mathbf{x}_{t-1}, S^*) \geq \min_{\mathbf{x}} f(\mathbf{x}, S^*) = \max_S \min_{\mathbf{x}} f(\mathbf{x}, S)$. Therefore, if we let $S = S^*$ in (27) we have:

$$(1-\frac{1}{e}) \max_{S} \min_{\mathbf{x}} f(\mathbf{x}, S) - \frac{\sum_{t=1}^{T} \alpha(f(\tilde{\mathbf{x}}, S_t))}{\sum_{t=1}^{T} \alpha} \leq \frac{3M^2\alpha T + \frac{H}{2\alpha}}{T} \tag{90}$$

Also, if we let $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x}, \cup_t S_t)$ and put $\tilde{\mathbf{x}} = \hat{\mathbf{x}}$ in (90) then because $f(\hat{\mathbf{x}}, S_t) \leq f(\hat{\mathbf{x}}, \cup_t S_t)$ we have:

$$(1-\frac{1}{e}) \max_{S} \min_{\mathbf{x}} f(\mathbf{x}, S) - \min_{\mathbf{x}} f(\mathbf{x}, \cup_t S_t) \leq \frac{3M^2\alpha T + \frac{H}{2\alpha}}{T} \tag{91}$$

and by using $\alpha = \frac{1}{\sqrt{T}}$:

$$(1-\frac{1}{e}) \max_{S} \min_{\mathbf{x}} f(\mathbf{x}, S) - \min_{\mathbf{x}} f(\mathbf{x}, \cup_t S_t) \leq \frac{3M^2\alpha T + \frac{H}{2\alpha}}{\sqrt{T}} \tag{92}$$

Now, using specific choices $K = 2M^2 + \frac{H}{2}$ and let $T = \frac{K^2}{\epsilon^2}$; we obtain that $S_{sol} = \cup_t S_t$ is a $((1-1/e)^{-1}, \epsilon)$-approximate maxmin solution.

A.7.2 Proof of Theorem 8 for Gradient Replacement-greedy(GRG)

If we let $S^* = \arg\max_S \min_{\mathbf{x}} f(\mathbf{x}, S)$ we know that for every t we have $f(\mathbf{x}_{t-1}, S^*) \geq min_{\mathbf{x}} f(\mathbf{x}, S^*) = \max_S \min_{\mathbf{x}} f(\mathbf{x}, S)$. Therefore, if we let $S = S^*$ in (45) we have :

$$\frac{1}{2} \max_{S} \min_{\mathbf{x}} f(\mathbf{x}, S) - \frac{1}{T} \sum_{t=1}^{T} f(\tilde{\mathbf{x}}, S_t) \leq \frac{K}{\sqrt{T}} \tag{93}$$

Let $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x}, \cup_t S_t)$ and put $\tilde{\mathbf{x}} = \hat{\mathbf{x}}$ in (94) then because $f(\hat{\mathbf{x}}, S_t) \leq f(\hat{\mathbf{x}}, \cup_t S_t)$ we have:

$$\frac{1}{2} \max_{S} \min_{\mathbf{x}} f(\mathbf{x}, S) - \min_{\mathbf{x}} f(\mathbf{x}, \cup_t S_t) \leq \frac{1}{2} \max_{S} \min_{\mathbf{x}} f(\mathbf{x}, S) - \frac{1}{T} \sum_{t=1}^{T} f(\hat{\mathbf{x}}, S_t) \leq \frac{K}{\sqrt{T}} \tag{94}$$

let $T = \frac{K^2}{\epsilon^2}$; then $S_{sol} = \cup_t S_t$ is a $(2, \epsilon)$-approximate maxmin solution.

A.7.3 Proof of Theorem 8 for Extra-gradient Greedy(EGG)

If we let $S^* = \arg\max_S \min_{\mathbf{x}} f(\mathbf{x}, S)$ we know that for every t we have $f(\hat{\mathbf{x}}_{t-1}, S^*) \geq min_{\mathbf{x}} f(\mathbf{x}, S^*) = \max_S \min_{\mathbf{x}} f(\mathbf{x}, S)$. Therefore, if we let $S = S^*$ in (66) we have :

$$(1-1/e)\max_S \min_{\mathbf{x}} f(\mathbf{x}, S) - \frac{1}{T}\sum_{t=1}^{T} f(\tilde{\mathbf{x}}, S_t) \leq \frac{K}{\sqrt{T}} \tag{95}$$

Let $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x}, \cup_t S_t)$ and put $\tilde{\mathbf{x}} = \hat{\mathbf{x}}$ in (94) then because $f(\hat{\mathbf{x}}, S_t) \leq f(\hat{\mathbf{x}}, \cup_t S_t)$ we have:

$$(1-1/e)\max_S \min_{\mathbf{x}} f(\mathbf{x}, S) - \min_{\mathbf{x}} f(\mathbf{x}, \cup_t S_t) \tag{96}$$

$$\leq (1-1/e)\max_S \min_{\mathbf{x}} f(\mathbf{x}, S) - \frac{1}{T}\sum_{t=1}^{T} f(\hat{\mathbf{x}}, S_t) \leq \frac{K}{\sqrt{T}} \tag{97}$$

Let $T = \frac{H^2}{\epsilon^2}$; then, $S_{sol} = \cup_t S_t$ is $((1-1/e)^{-1}, \epsilon)$ approximate maxmin solution.

A.7.4 Proof of Theorem 8 for Extra-gradient Replacement-greedy(EGRG)

Similar to (95), and (96), (88) results $S_{sol} = \cup_t S_t$ to be $((2 + \frac{k}{k-1}), \epsilon)$-approximate maxmin solution.

A.8 Proof of Theorem 7: Extra Gradient on Continuous Extension Convergence

In this section, we will focus on convergence analysis of Extra Gradient on continuous extension. We first provide two propositions and matroid definition that will help us in the proof.

Definition 7. *Let $\mathcal{I}$ be a nonempty family of allowable subsets of the ground set V, then the tuple $(V, \mathcal{I})$ is a matroid if and only if the following conditions hold:*

1. *For any $A \subset B \subset V$, if $B \in \mathcal{I}$, then $A \in \mathcal{I}$*
2. *For all $A, B \in \mathcal{I}$, if $|A| < |B|$, then there is an $e \in B \backslash A$ such that $A \cup \{e\} \in \mathcal{I}$.*

Proposition 1. *we have that*

$$\text{OPT} \triangleq \min_{\mathbf{x}\in\mathcal{C}} \max_{S\in\mathcal{I}} f(\mathbf{x}, S) = \min_{\mathbf{x}\in\mathcal{C}} \max_{\mathbf{y}\in\mathcal{K}} F(\mathbf{x}, \mathbf{y}). \tag{98}$$

Furthermore, the function F has the following properties (assuming differentibility):

Proposition 2. *we have for function F ([Hassani et al., 2017])):*

$$\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d : F(\mathbf{x}_1, \mathbf{y}) - F(\mathbf{x}_2, \mathbf{y}) \leq \langle \nabla_{\mathbf{x}} F(\mathbf{x}_1, \mathbf{y}), \mathbf{x}_1 - \mathbf{x}_2 \rangle,$$
$$\forall \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^d : F(\mathbf{x}, \mathbf{y}_2) - 2F(\mathbf{x}, \mathbf{y}_1) \leq \langle \nabla_{\mathbf{y}} F(\mathbf{x}, \mathbf{y}_1), \mathbf{y}_2 - \mathbf{y}_1 \rangle.$$

using same procedure as Extra-gradient Greedy we drive following equations similar to (53):

$$\begin{aligned}
&\|\hat{\mathbf{x}}_t - \mathbf{x}\|^2 \\
&\leq \|\mathbf{x}_t - \mathbf{x} - \gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, y_t)^\top (\mathbf{x}_t - \mathbf{x}) + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_t) \\
&\leq \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + 2(\mathbf{x}_t - \hat{\mathbf{x}}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}_t) \\
&= \|\mathbf{x}_t - \mathbf{x}\|^2 - 2\gamma_t \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)^\top (\hat{\mathbf{x}}_t - \mathbf{x}) + \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 - 2\|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2
\end{aligned}$$

and similarly we have:

$$
\begin{aligned}
2\langle -\gamma_t \nabla_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{y}}_t - \mathbf{y}\rangle &\leq \|\mathbf{y} - \mathbf{y}_t\|^2 - \|\mathbf{y} - \hat{\mathbf{y}}_t\|^2 - \|\hat{\mathbf{y}}_t - \mathbf{y}_t\|^2 \\
2\langle \gamma_t \nabla_{\mathbf{x}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{x}}_t - \mathbf{x}\rangle &\leq \|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \hat{\mathbf{x}}_t\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 \\
2\langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathbf{y}_{t+1} - \mathbf{y}\rangle &\leq \|\mathbf{y} - \mathbf{y}_t\|^2 - \|\mathbf{y} - \mathbf{y}_{t+1}\|^2 - \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\
2\langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathbf{x}_{t+1} - \mathbf{x}\rangle &\leq \|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
2\langle -\gamma_t \nabla_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{y}}_t - \mathbf{y}_{t+1}\rangle &\leq \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 - \|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t\|^2 - \|\mathbf{y}_t - \hat{\mathbf{y}}_t\|^2 \\
2\langle \gamma_t \nabla_{\mathbf{x}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\rangle &\leq \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 - \|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2
\end{aligned}
$$

combing the above equations we have:

$$
\begin{aligned}
\langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \hat{\mathbf{y}}_t - \mathbf{y}\rangle &= \langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \hat{\mathbf{y}}_t - \mathbf{y}_{t+1}\rangle + \langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathbf{y}_{t+1} - \mathbf{y}\rangle \\
&= \langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \gamma_t \nabla_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{y}}_t - \mathbf{y}_{t+1}\rangle \\
&\quad + \langle -\gamma_t \nabla_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{y}}_t - \mathbf{y}_{t+1}\rangle \\
&\quad + \langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathbf{y}_{t+1} - \mathbf{y}\rangle \\
&\leq \langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \gamma_t \nabla_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{y}}_t - \mathbf{y}_{t+1}\rangle \\
&\quad + 0.5(-\|\hat{\mathbf{y}}_t - \mathbf{y}_{t+1}\|^2 - \|\mathbf{y}_t - \hat{\mathbf{y}}_t\|^2 \\
&\quad + \|\mathbf{y} - \mathbf{y}_t\|^2 - \|\mathbf{y} - \mathbf{y}_{t+1}\|^2)
\end{aligned}
$$

and

$$
\begin{aligned}
\langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \hat{\mathbf{x}}_t - \mathbf{x}\rangle &= \langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\rangle + \langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathbf{x}_{t+1} - \mathbf{x}\rangle \\
&= \langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - \gamma_t \nabla_{\mathbf{x}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\rangle \\
&\quad + \langle \gamma_t \nabla_{\mathbf{x}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\rangle + \langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \mathbf{x}_{t+1} - \mathbf{x}\rangle \\
&\leq \langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - \gamma_t \nabla_{\mathbf{x}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\rangle + 0.5(-\|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2 \\
&\quad + \|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2)
\end{aligned}
$$

let

$$
\sigma_t^{\mathbf{x}} = \langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - \gamma_t \nabla_{\mathbf{x}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\rangle + 0.5(-\|\hat{\mathbf{x}}_t - \mathbf{x}_{t+1}\|^2 - \|\mathbf{x}_t - \hat{\mathbf{x}}_t\|^2)
$$

and

$$
\sigma_t^{\mathbf{y}} = \langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \gamma_t \nabla_{\mathbf{y}} F(\mathbf{x}_t, \mathbf{y}_t), \hat{\mathbf{y}}_t - \mathbf{y}_{t+1}\rangle + 0.5(-\|\hat{\mathbf{y}}_t - \mathbf{y}_{t+1}\|^2 - \|\mathbf{y}_t - \hat{\mathbf{y}}_t\|^2)
$$

then $2\sigma_t^{\mathbf{x}} + \sigma_t^{\mathbf{y}} \leq 0$ if $\gamma_t \leq \frac{1}{3\max\{L_{\mathbf{x}}, L_{\mathbf{y}}\}}$(check (Nemirovski, 2004) for more details) ; which results in:

$$
2\langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{y}}_t, \hat{\mathbf{y}}_t), \hat{\mathbf{y}}_t - \mathbf{y}\rangle \leq \|\mathbf{y} - \mathbf{y}_t\|^2 - \|\mathbf{y} - \mathbf{y}_{t+1}\|^2 \tag{99}
$$

$$
2\langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \hat{\mathbf{x}}_t - \mathbf{x}\rangle \leq \|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2 \tag{100}
$$

combing above equations with proposition 2 we have:

$$
2\gamma_t F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) - 2\gamma_t F(\mathbf{x}, \hat{\mathbf{y}}_t) \leq 2\langle \gamma_t \nabla_{\mathbf{x}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \hat{\mathbf{x}}_t - x\rangle \leq \|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2 \tag{101}
$$

$$
-2\gamma_t F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t) + \gamma_t F(\hat{\mathbf{x}}_t, \mathbf{y}) \leq \langle -\gamma_t \nabla_{\mathbf{y}} F(\hat{\mathbf{x}}_t, \hat{\mathbf{y}}_t), \hat{\mathbf{y}}_t - \mathbf{y}\rangle \leq 0.5\|\mathbf{y} - \mathbf{y}_t\|^2 - 0.5\|\mathbf{y} - \mathbf{y}_{t+1}\|^2 \tag{102}
$$

$$
-2\gamma_t F(\mathbf{x}, \hat{\mathbf{y}}_t) + \gamma_t F(\hat{\mathbf{x}}_t, \mathbf{y}) \leq 0.5\|\mathbf{y} - \mathbf{y}_t\|^2 - 0.5\|\mathbf{y} - \mathbf{y}_{t+1}\|^2 + \|\mathbf{x} - \mathbf{x}_t\|^2 - \|\mathbf{x} - \mathbf{x}_{t+1}\|^2 \tag{103}
$$

summing over t in (103) and divide both side by $\sum_{t=1}^{T}\gamma_t$ (set of variable $\mathbf{x}$ and $\mathbf{y}$ is bounded i.e. $\|\mathbf{y}\|^2 \leq H, \|\mathbf{x}\|^2 \leq H$):

$$\frac{\sum_{t=1}^{T}\gamma_t\left[-2F(\mathbf{x},\hat{\mathbf{y}}_t)+F(\hat{\mathbf{x}}_t,\mathbf{y})\right]}{\sum_{t=1}^{T}\gamma_t} \leq \frac{0.5\|\mathbf{y}-\mathbf{y}_1\|^2+\|\mathbf{x}-\mathbf{x}_1\|^2}{\sum_{t=1}^{T}\gamma_t} \leq \frac{1.5H}{\gamma T} \tag{104}$$

which means same as before let $T=\frac{\sqrt{1.5H}}{\gamma\epsilon}$ and constant step size $\gamma_t=\gamma$, and $\mathbf{x}^*=\arg\min\max f(\mathbf{x},\mathbf{y})$ we have:

$$\frac{1}{2}f(\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{x}}_t,\mathbf{y})-\min_{\mathbf{x}}\max_{\mathbf{y}}F(\mathbf{x},\mathbf{y}) \leq \frac{1}{2}F(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t,\mathbf{y})-\frac{1}{T}\sum_{t=1}^{T}f(\mathbf{x}^*,\mathbf{y}^t) \leq \epsilon \tag{105}$$

then using proposition 1, $\mathbf{x}_{sol}=\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{x}}_t$ is $(0.5,\epsilon)$-approximate minimax solution.