Non-smooth and Hölder-smooth Submodular Maximization

Duksang Lee [1 2] Nam Ho-Nguyen [3] Dabeen Lee [4 *]

October 12, 2022

Abstract

We study the problem of maximizing a continuous DR-submodular function that is not necessarily smooth. We prove that the continuous greedy algorithm achieves an $[(1-1/e)\text{OPT}-\epsilon]$ guarantee when the function is monotone and Hölder-smooth, meaning that it admits a Hölder-continuous gradient. For functions that are non-differentiable or non-smooth, we propose a variant of the mirror-prox algorithm that attains an $[(1/2)\text{OPT}-\epsilon]$ guarantee. We apply our algorithmic frameworks to robust submodular maximization and distrbituionally robust submodular maximization under Wasserstein ambiguity. In particular, the mirror-prox method applies to robust submodular maximization to obtain a single feasible solution whose value is at least $(1/2)\text{OPT}-\epsilon$. For distributionally robust maximization under Wasserstein ambiguity, we deduce and work over a submodular-convex maximin reformulation whose objective function is Hölder-smooth, for which we may apply both the continuous greedy method and the mirror-prox method.

1 Introduction

Submodularity is a structural property that is exploited in both discrete and continuous domains and has numerous applications in optimization. Submodular set functions are often associated with the problem of selecting the optimal group of items from a large pool of discrete candidates. Optimizing submodular set functions has applications in sensor placement [15], influence maximization [20], feature and variable selection [21], dictionary learning [11], document summarization [24, 25], image summarization [13, 28, 43], and active set selection in non-parametric learning [27]. Continuous submodularity also naturally arises in a wide range of application domains where the objective exhibits the diminishing returns property, such as MAP inference for determinantal point processes and mean-field inference of probabilistic graphical models [6], and robust budget allocation [35, 36].

An important line of research regarding submodularity is robust submodular maximization. In many application scenarios, objective submodular functions are often non-stationary or has underlying randomness, e.g., sensor

[1]Department of Mathematical Sciences, KAIST, Daejeon 34126, Republic of Korea
[2]Discrete Mathematics Group, Institute for Basic Science (IBS), Daejeon 34126, Republic of Korea
[3]Discipline of Business Analytics, The University of Sydney, Sydney, NSW 2006, Australia
[4]Department of Industrial and Systems Engineering, KAIST, Daejeon 34126, Republic of Korea

*Correspondence to <dabeenl@kaist.ac.kr>

placement [22] and influence maximization [10, 18, 26]. Such functions are typically estimated by data and thus subject to estimation noise. The robust optimization framework considers a family of submodular functions and attempts to maximize them simultaneously by taking the point-wise minimum of the functions [9, 22, 42, 46]. The distributionally robust submodular maximization framework extends robust submodular maximization. It considers a set of distributions over a family of submodular functions and attempts to maximize the minimum expectation over the distributions [37].

The celebrated continuous greedy algorithm [8, 44] is used for maximizing a submodular set function over a matroid constraint and is based on optimizing the multilinear extension, which is a continuous relaxation of the given submodular set function. A modification of the continuous greedy algorithm extends to solve continuous DR-submodular maximization [5], and there are other first-order methods for the continuous case [17, 31]. The multilinear extension of a submodular set function is differentiable and smooth, meaning that it has a Lipschitz continuous gradient, and the existing algorithms for continuous submodular maximization require smoothness as well. However, taking the point-wise minimum for robust maximization does not preserve differentiability nor smoothness, and similarly, the minimum of the expectation of a family of submodular functions is non-differentiable and non-smooth in general. For this reason, the aforementioned methods cannot directly apply to robust submodular maximization, and the existing methods for the robust problem cannot avoid violating constraints [3, 9, 22, 42, 46]. The existing methods for distributionally robust problem require an expensive smoothing step [46] or a high sample variance assumption [37] to guarantee smoothness.

Our contributions Motivated by this, we study the problem of maximizing a continuous submodular function that does not admit a Lipschitz continuous gradient, and may not be differentiable. We develop solution methods that do not require a smoothing step nor a smoothness assumption. We focus on deterministic continuous DR-submodular functions that are monotone, while our framework extends to the stochastic case.

- We prove that the continuous greedy algorithm works for a differentiable continuous DR-submodular function as long as it is differentiable and has a Hölder continuous gradient, i.e., Hölder-smooth. Here, Hölder continuity is a relaxed notion of Lipschitz continuity. We show that the algorithm returns a solution whose value is at least $(1 - 1/e)\mathrm{OPT} - \epsilon$ after $O(1/\epsilon^{\frac{1}{\sigma}})$ iterations, where OPT is the optimal value and σ is the Hölder exponent.
- For functions that are non-differentiable or differentiable but non-smooth, we propose a variant of the mirror-prox method that achieves a value at least $\mathrm{OPT}/2 - \epsilon$ after $O(1/\epsilon^{\frac{2}{1+\sigma}})$ iterations. In particular, we have $\sigma = 0$ for non-differentiable functions, in which case, the number of required iterations is $O(1/\epsilon^2)$.
- We show that the mirror-prox method can be used to solve robust submodular maximization, providing algorithms that finitely converges to a solution that achieves value at least $\mathrm{OPT}/2 - \epsilon$ after $O(1/\epsilon^2)$ iterations and does not violate constraints.
- We consider a distributionlly robust formulation of submodular maximization under Wasserstein ambiguity and provide a reformulation where the objective function is not smooth but admits a Hölder continuous gradient with parameter $\sigma = 1/2$. We then show that the continuous greedy algorithm returns a value at

least $(1-1/e)\mathrm{OPT}-\epsilon$ after $O(1/\epsilon^2)$ iterations and that the mirror-prox method achieves a value at least $\mathrm{OPT}/2-\epsilon$ after $O(1/\epsilon^{\frac{4}{3}})$ iterations.

Our framework also applies to the case of robust and distributionally robust maximization of submodular set functions via their multilinear extensions. Recently, the problem of maximizing a "submodular + concave" function, which is the sum of a submodular function and a concave function, is introduced [29]. Submodular + concave functions are not necessarily submodular but belong to the class of up-concave functions, which includes submodular functions. In fact, our framework works for any up-concave functions that are not necessarily differentiable nor smooth.

2 Related work

Starting with the work of Nemhauser et al. [32], algorithms for maximizing a submodular set function under various types of constraints have been extensively studied. One popular example is the continuous greedy algorithm [8, 44] that was first introduced to solve the problem under a matroid constraint, and the algorithm applies to other complicated constraints such as a system of linear constraints and the intersection of multiple matroid constraints [8, 23]. Its main idea is to solve the continuous relaxation obtained by the so-called multilinear extension of a given submodular set function and then apply rounding procedures to the solution from the continuous relaxation. Since then, the multilinear extension has gained attention [8, 14, 23, 41, 44, 45]. The multilinear extension of a submodular set function has nice properties such as smoothness and DR-submodularity, which is an extension of the diminishing returns property of set functions to continuous functions.

Continuous DR-submodular functions are a continuous analogue of submodular set functions, generalizing the notion of diminishing returns property. Bian et al. [5] proposed a variant of the continuous greedy algorithm for the problem of maximizing a continuous DR-submodular function. The algorithm is often referred to as the Frank-Wolfe algorithm or the conditional gradient method, and if the constraint set is down-closed and convex, it returns a solution whose value is at least $(1-1/e)\mathrm{OPT}-\epsilon$ after $O(d/\epsilon)$ iterations where d is the ambient dimension. Hassani et al. [17] showed that the stochastic gradient ascent achieves a value at least $\mathrm{OPT}/2-\epsilon$ after $O(1/\epsilon)$ iterations under an arbitrary convex constraint set. Moreover, they considered a restricted setting where not the exact gradients but their unbiased stochastic estimates with a bounded variance are available, in which case, they showed that the same asymptotic guarantee can be achieved after $O(1/\epsilon^2)$ iterations. Shortly after this, Mokhtari et al. [31] proposed the stochastic continuous greedy algorithm for the stochastic case that guarantees a value at least $(1-1/e)\mathrm{OPT}-\epsilon$ after $O(1/\epsilon^3)$ iterations. Later, Karbasi et al. [19] and Zhang et al. [47] developed variants of the stochastic continuous greedy algorithm that attain the same guarantee after $O(1/\epsilon^2)$ iterations. We remark that all these works focus on functions that are differentiable and smooth as the multilinear extension.

A related area is robust submodular maximization, in which we aim to make decisions that are robust against multiple submodular objective functions by maximizing the point-wise minimum of the functions [22]. One of its major application domains is robust influence maximization [10, 18, 26], where the goal is to maximize the worst-case influence spread characterized by multiple information diffusion processes. Krause et al. [22] showed that there is no polynomial time constant approximation algorithm for robust submodular maximization unless

$P = NP$. They considered the cardinality constrained case, for which they developed a bi-criteria approximation algorithm which sacrifices feasibility to achieve an approximation guarantee. Later, Anari et al. [3], Chen et al. [9], Torrico et al. [42], Wilder [46] generalized the idea of bi-criteria approximation for general constraints such as matroid and knapsack constraints. Unfortunately, the continuous greedy method and other gradient based methods [5, 17, 19, 31, 47] do not directly apply to the robust problem, as the robust submodular objective is not smooth in general. The mirror-prox algorithm was applied to smooth submodular problems by Adibi et al. [1]. Our work expands its applicability to Hölder-smooth and non-smooth submodular problems by extending the results of Nemirovski [33], Stonyakin et al. [40].

Distributionally robust submodular maximization is a framework to find a solution that is robust against a family of probability distributions over multiple submodular objective functions. More precisely, the objective is to maximize the minimum expected function value where the minimum is taken over the family of distributions. Therefore, it is a generalization of both robust and stochastic submodular maximization. Its formulation depends on how we construct a family of distributions, and a popular way is to consider distributions that are close to a reference distribution, typically set to the empirical distribution, based on a metric measuring the distance between two probability distributions. Staib et al. [37] considered the distributionally robust submodular maximization framework defined by the χ^2-divergence. They showed that under a high sample variance condition, the distributionally robust objective becomes smooth. However, such a smooth assumption does not hold in general, and instead we can apply the method of Wilder [46], which uses a randomized smoothing technique by Duchi et al. [12].

An ambiguity set based on the χ^2-divergence contains only distributions whose support is the same as that of the reference distribution. In particular, when we use the empirical distribution on a finite sample as a reference, we cannot consider any data points outside the sample set. Using a Wasserstein ambiguity set alleviates this since it considers optimal transport of data points, subject to a transportation budget. Blanchet and Murthy [7], Mohajerin Esfahani and Kuhn [30] showed that how various distributionally robust problems with Wasserstein ambiguity can be reformulated using duality theory. In the context of DR-submodular maximization, a reformulation of the problem with Wasserstein ambiguity leads to a Hölder-smooth objective. Our work analyzing the continuous greedy and mirror-prox algorithms show that they can be applied to this application.

3 Preliminaries

We say that a function $F : \mathbb{R}^d \to \mathbb{R}$ is (continuous) *DR-submodular* [5, 34] if F satisfies

$$F(\boldsymbol{x} + z\boldsymbol{e_i}) - F(\boldsymbol{x}) \geq F(\boldsymbol{y} + z\boldsymbol{e_i}) - F(\boldsymbol{y})$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ such that $\boldsymbol{x} \leq \boldsymbol{y}$, standard basic vector $\boldsymbol{e_i} \in \mathbb{R}^d$ and $z \in \mathbb{R}_+$. It is known that when F is differentiable, $\nabla F(\boldsymbol{x}) \geq \nabla F(\boldsymbol{y})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ such that $\boldsymbol{x} \leq \boldsymbol{y}$ [5]. Moreover, when F is twice-differentiable, F is DR-submodular if and only if its Hessian $\nabla^2 F(\boldsymbol{x})$ at any $\boldsymbol{x} \in \mathbb{R}^d$ has all components non-positive [5]. Hence, DR-submodular functions are neither convex nor concave in general. Furthermore, we assume that F is monotone, i.e., $F(\boldsymbol{x}) \leq F(\boldsymbol{y})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ such that $\boldsymbol{x} \leq \boldsymbol{y}$, and we assume that F is nonnegative, i.e., $F(\boldsymbol{x}) \geq 0$ for for any $\boldsymbol{x} \in \mathbb{R}^d$.

DR-submodularity naturally arises from *submodular set functions*. A set function $f : 2^V \to \mathbb{R}_+$ over a finite ground set V is submodular if $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$ for all $S, T \subseteq V$. Then the so-called *multilinear extension* of f, given by

$$F(\boldsymbol{x}) = \sum_{S \subseteq V} f(S) \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j), \quad \boldsymbol{x} \in [0, 1]^V$$

which provides a continuous relaxation of f, is DR-submodular [44]. Moreover, if f is monotone, i.e., $f(S) \leq f(T)$ for any $S \subseteq T$, then its multilinear extension is monotone as well.

There are many other classes of DR-submodular functions [4]. A quadratic function $F(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}/2 + \boldsymbol{b}^\top \boldsymbol{x} + c$ is DR-submodular if and only if all entries of A are nonpositive. Functions involving $\varphi_{ij}(x_i - x_j)$ where $\varphi_{ij} : \mathbb{R} \to \mathbb{R}$ for $i, j \in [m]$ are convex, $g(\sum_{i \in [m]} \lambda_i x_i)$ where g is concave and $\boldsymbol{\lambda} \geq \mathbf{0}$, and $\log \det(\sum_{i \in [m]} x_i \boldsymbol{A_i})$ where $\boldsymbol{A_i}$'s are positive definite and $\boldsymbol{x} \geq 0$ are all examples of DR-submodular functions.

While many applications involve a DR-submodular objective F, it turns out that there are important settings where DR-submodularity is not necessarily satisfied. On the other hand, the important property of DR-submodular functions that most algorithmic frameworks rely on is *up-concavity*, or equivalently *coordinate-wise concave* [5, 46]. A function $F : \mathbb{R}^d \to \mathbb{R}$ is *up-concave* if for any $\boldsymbol{x} \in \mathbb{R}^d$ and a nonnegative direction $\boldsymbol{v} \geq \mathbf{0}$, function $F(\boldsymbol{x} + t\boldsymbol{v})$ is concave with respect to $t \in \mathbb{R}$. Therefore, we will study the following problem:

$$\sup_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x}) \tag{SFM}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a constraint set that is convex and F is up-concave, monotone, and nonnegative, but not necessarily DR-submodular. We next provide a list of applications where such an objective is relevant.

Robust DR-submodular maximization. Let $F_1, \ldots, F_n : \mathbb{R}^d \to \mathbb{R}_+$ be nonnegative DR-submodular functions, and consider $\sup_{\boldsymbol{x} \in \mathcal{X}} \min_{i \in [n]} F_i(\boldsymbol{x})$. This problem is often referred to as *robust DR-submodular maximization*. Here, the function $F := \min_{i \in [n]} F_i$ can be shown to be up-concave but not necessarily DR-submodular. We consider robust DR-submodular maximization in Section 6.

Distributionally robust DR-submodular maximization. Let Q be a probability distribution over a sample space $\Xi \subseteq \mathbb{R}^p$, and let $\{F_{\boldsymbol{\xi}}\}_{\boldsymbol{\xi} \in \Xi}$ be a collection of functions from $\mathbb{R}^d$ to $\mathbb{R}_+$. Then we consider a family $\mathcal{B}$ of distributions over Ξ containing Q and the following distributionally robust optimization problem $\sup_{\boldsymbol{x} \in \mathcal{X}} \inf_{P \in \mathcal{B}} \mathbb{E}_{\boldsymbol{\xi} \sim P} [F_{\boldsymbol{\xi}}(\boldsymbol{x})]$ where $\mathcal{X}$ is the constraint set for variables $\boldsymbol{x}$. Assuming that $F_{\boldsymbol{\xi}}$ is DR-submodular for any $\boldsymbol{\xi} \in \Xi$, we can show that $F := \inf_{P \in \mathcal{B}} \mathbb{E}_{\boldsymbol{\xi} \sim P} [F_{\boldsymbol{\xi}}]$ is up-concave. We consider a family of distributions constructed based on the so-called Wasserstein distance. We provide algorithms for solving the distributionally robust formulation of DR-submodular maximization based on the Wasserstein distance in Section 7.

Another application of our framework is "Submodular + Concave" maximization. Mitra et al. [29] study the problem of maximizing $F + C$ where F is a smooth DR-submodular function and C is a smooth concave function. As a concave function is automatically up-concave, it follows that $F + C$ is up-concave.

We say that a function F is smooth with respect to a norm $\|\cdot\|$ if there exists some constant $\beta > 0$ such that

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\|_* \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|$$

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. However, the up-concave functions that arise from robust and distributionally robust DR-submodular maximization problems are not necessarily smooth. Moreover, when considering $F + C$ where F is DR-submodular and C is concave, $F + C$ is not smooth if F or C is not smooth.

In this paper, we relax the smoothness assumption on the given up-concave function and develop algorithms for maximizing a non-smooth up-concave function. We introduce the notion of *Hölder-smoothness*, defined as follows.

Definition 1. *We say that a function $F : \mathbb{R}^d \to \mathbb{R}$ is Hölder-smooth with respect to a norm $\|\cdot\|$ and a monotone increasing function $h : \mathbb{R}_+ \to \mathbb{R}_+$ with $\lim_{x\to 0} h(x) = 0$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$\|\nabla F(\boldsymbol{x}) - \nabla F(\boldsymbol{y})\|_* \leq h(\|\boldsymbol{x} - \boldsymbol{y}\|). \tag{1}$$

Here, if h is given by $h(z) = \beta \cdot z$ for some constant $\beta > 0$, then Hölder-smoothness with respect to h reduces to smoothness. Throughout the paper, the particular form of h we focus on is

$$h(z) = \sum_{i=1}^{k} \beta_i \cdot z^{\sigma_i} \tag{2}$$

for some integer $k \geq 1$ and constants $\beta_1, \ldots, \beta_k > 0$ and $0 < \sigma_1, \ldots, \sigma_k \leq 1$. Although we allow h to have multiple terms, i.e., $k > 1$, a more common definition of Hölder smoothness is by a function h with a single term [39]. If there is no function h of the form (2) where at least one of $\sigma_1, \ldots, \sigma_k$ being positive with which F satisfies condition (1), then we say that F is *non-smooth*.

The existing literature focuses on the case where the objective function F is differentiable, but in this paper, we also consider the non-differentiable case as well. However, when F is not differentiable, it is not clear whether a first-order method for maximizing F has to use subgradients or supergradients as F is neither convex nor concave. Instead, we introduce the notion of up-super-gradients defined as follows.

Definition 2. *We say that $\boldsymbol{g}_{\boldsymbol{x}}$ is an* up-super-gradient *at $\boldsymbol{x} \in \mathbb{R}^d$ if*

$$F(\boldsymbol{y}) \leq F(\boldsymbol{x}) + \boldsymbol{g}_{\boldsymbol{x}}^\top (\boldsymbol{y} - \boldsymbol{x}) \quad \textit{for any } \boldsymbol{y} \in \mathbb{R}^d \textit{ such that } \boldsymbol{y} \geq \boldsymbol{x} \textit{ or } \boldsymbol{y} \leq \boldsymbol{x}. \tag{3}$$

Then we define $\partial^\uparrow F(\boldsymbol{x})$ as the set of up-super-gradients of F at $\boldsymbol{x}$.

In fact, if F is differentiable, $\partial^\uparrow F(\boldsymbol{x})$ reduces to the gradient of F (see Section 5).

Definition 3. *We say that a solution $\bar{\boldsymbol{x}} \in \mathcal{X}$ is an (α, ϵ)-approximate solution to* (SFM) *for some $0 \leq \alpha < 1$ and $\epsilon > 0$ if*

$$F(\bar{\boldsymbol{x}}) \geq \alpha \cdot \sup_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x}) - \epsilon.$$

In this paper, we develop first-order iterative algorithms that guarantee an (α, ϵ)-approximate solution to (SFM) where the number of required iterations is bounded by $O(1/\epsilon^c)$ for some constant c.

4 Conditional gradient method

In this section, we consider (SFM), the problem of maximizing a monotone up-concave function, for the differentiable case. We present Algorithm 1, which is a variant of the continuous greedy algorithm [5] for monotone up-concave function maximization.

Algorithm 1 Continuous greedy algorithm for maximizing a differentiable monotone up-concave functions

Initialize $\boldsymbol{x_0} \leftarrow \mathbf{0}$.

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

Fetch $\boldsymbol{g_t}$.

$\boldsymbol{v_t} = \arg\max_{\boldsymbol{v} \in \mathcal{X}} \langle \boldsymbol{g_t}, \boldsymbol{v} \rangle$.

Update $\boldsymbol{x_t} = \left(1 - \frac{1}{t}\right) \boldsymbol{x_{t-1}} + \frac{1}{t} \boldsymbol{v_t} = \frac{1}{t}(\boldsymbol{v_1} + \cdots + \boldsymbol{v_t})$.

end for

Return $\boldsymbol{x_T} = \frac{1}{T}(\boldsymbol{v_1} + \dots \boldsymbol{v_T})$.

In Algorithm 1, we take $x_t = (\boldsymbol{v_1} + \cdots + \boldsymbol{v_t})/t$ at each iteration t, in which case, x_t is a convex combination of points in $\mathcal{X}$. It is more common to take $\boldsymbol{x_t} = (\boldsymbol{v_1} + \cdots + \boldsymbol{v_t})/T$ for the continuous greedy algorithm [5], but the point is not necessarily a point in $\mathcal{X}$ when $t < T$. Since F is monotone, scaling up $(\boldsymbol{v_1} + \cdots + \boldsymbol{v_t})/T$ to obtain $(\boldsymbol{v_1} + \cdots + \boldsymbol{v_t})/t$ always improves the objective, so we take the latter. Another important point of Algorithm 1 is that we allow errors in computing the gradient at each iteration. We will show that as long as $\|\nabla F((t/T)\boldsymbol{x_t}) - \boldsymbol{g_t}\|_* \leq \delta$ over all t for some $\delta > 0$, the total accumulated error is bounded by $O(\delta)$. Furthermore, the results given by [5] assume that the objective function F is smooth, but in our analysis, we show that Hölder-smoothness is sufficient to guarantee a finite convergence to an approximate solution.

4.1 Convergence of continuous greedy for Hölder-smooth functions

The following is the main lemma to show finite convergence of Algorithm 1.

Lemma 4.1. *Let $F : \mathbb{R}^d \to \mathbb{R}_+$ be a differentiable, monotone, and up-concave function. Assume that $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\| \leq R$ for some $R > 0$ and there exists a function $\tilde{h} : \mathbb{R}_+ \to \mathbb{R}_+$ such that for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$F(\boldsymbol{x}) \geq F(\boldsymbol{y}) + \langle \nabla F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle - \tilde{h}(\|\boldsymbol{x} - \boldsymbol{y}\|). \tag{4}$$

Let $\boldsymbol{x_0}, \boldsymbol{v_1}, \boldsymbol{x_1}, \dots, \boldsymbol{x_{T-1}}, \boldsymbol{v_T}, \boldsymbol{x_T}$ be the sequence generated by Algorithm 1. Then

$$F(\boldsymbol{x_T}) \geq \left(1 - \frac{1}{e}\right) \sup_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x}) - \sum_{s=0}^{T-1} \left(1 - \frac{1}{T}\right)^s \kappa_{T-s} + \frac{1}{e} F(\mathbf{0}) \tag{5}$$

where for $1 \leq t \leq T - 1$,

$$\kappa_t = \frac{2R}{T} \left\| F\left(\frac{t-1}{T} \boldsymbol{x_{t-1}}\right) - \boldsymbol{g_t} \right\|_* + \tilde{h}\left(\left\| \frac{1}{T} \boldsymbol{v_t} \right\|\right).$$

We defer the proof of this lemma to the next subsection. We now show that if F is Hölder-smooth with respect to a norm $\|\cdot\|$ and a function $h : \mathbb{R}_+ \to \mathbb{R}_+$ of the form (2), then there exists a function $\tilde{h} : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying (4).

Lemma 4.2. *If a differentiable function $F : \mathbb{R}^d \to \mathbb{R}_+$ is Hölder-smooth with respect to a norm $\|\cdot\|$ and h, then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,*

$$F(\boldsymbol{x}) \geq F(\boldsymbol{y}) + \langle \nabla F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle - \sum_{i=1}^{k} \frac{\beta_i}{1+\sigma_i} \|\boldsymbol{x} - \boldsymbol{y}\|^{1+\sigma_i}. \tag{6}$$

Proof. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$. Since F is differentiable, we have that

$$F(\boldsymbol{x}) - F(\boldsymbol{y}) = \int_0^1 \nabla F(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))^\top (\boldsymbol{x} - \boldsymbol{y}) dt. \tag{7}$$

Next we show that the following relations hold.

$$\begin{aligned}
&\left| F(\boldsymbol{x}) - F(\boldsymbol{y}) - \nabla F(\boldsymbol{y})^\top (\boldsymbol{x} - \boldsymbol{y}) \right| \\
&= \left| \int_0^1 \langle \nabla F(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \, dt \right| \\
&\leq \int_0^1 \left| \langle \nabla F(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla F(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \right| dt \\
&\leq \int_0^1 \|\nabla F(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla F(\boldsymbol{y})\|_* \cdot \|\boldsymbol{x} - \boldsymbol{y}\| dt \\
&\leq \int_0^1 h(t\|\boldsymbol{x} - \boldsymbol{y}\|) \cdot \|\boldsymbol{x} - \boldsymbol{y}\| dt \\
&= \int_0^1 \left(\sum_{i=1}^{k} \beta_i t^{\sigma_i} \|\boldsymbol{x} - \boldsymbol{y}\|^{1+\sigma_i} \right) dt \\
&= \sum_{i=1}^{k} \frac{\beta_i}{1+\sigma_i} \|\boldsymbol{x} - \boldsymbol{y}\|^{1+\sigma_i}
\end{aligned} \tag{8}$$

where the first equality comes from (7), the second inequality is by Hölder's inequality. Then (6) follows from (8). □

Based on Lemmas 4.1 and 4.2, we obtain the following convergence result on Algorithm 1.

Theorem 4.3. *Let $F : \mathbb{R}^d \to \mathbb{R}_+$ be a differentiable, monotone, and up-concave function that is Hölder-smooth with respect to a norm $\|\cdot\|$ and h. Assume that for each iteration of Algorithm 1, $\boldsymbol{g_t}$ is chosen so that $\|\nabla F((t/T)\boldsymbol{x_t}) - \boldsymbol{g_t}\|_* \leq \delta$ and $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\| \leq R$ for some $R > 0$. Then*

$$F(\boldsymbol{x_T}) - F(\boldsymbol{0}) \geq \left(1 - \frac{1}{e}\right) \sup_{\boldsymbol{x} \in \mathcal{X}} \{F(\boldsymbol{x}) - F(\boldsymbol{0})\} - \left(2\delta R + \sum_{i=1}^{k} \frac{\beta_i R^{1+\sigma_i}}{T^{\sigma_i}} \right). \tag{9}$$

Proof. For $1 \leq t \leq T$, we have

$$\kappa_t = \frac{2R}{T} \|\nabla F(t\boldsymbol{x_t}/T) - \boldsymbol{g_t}\|_* + \sum_{i=1}^{k} \frac{\beta_i}{1+\sigma_i} \|\boldsymbol{v_t}/T\|^{1+\sigma_i} \leq \frac{2\delta R}{T} + \sum_{i=1}^{k} \frac{\beta_i}{1+\sigma_i} \left(\frac{R}{T}\right)^{1+\sigma_i}$$

because $\|\nabla F(t\boldsymbol{x_t}/T) - \boldsymbol{g_t}\|_* \leq \delta$ and $\|\boldsymbol{v_t}\| \leq R$. Therefore, it follows that

$$\begin{aligned}\sum_{s=0}^{T-1}\left(1-\frac{1}{T}\right)^s \kappa_{T-s} &\leq \frac{1}{T}\sum_{s=0}^{T-1}\left(1-\frac{1}{T}\right)^s\left(2\delta R + \sum_{i=1}^{k}\beta_i\left(\frac{R^{1+\sigma_i}}{T^{\sigma_i}}\right)\right)\\ &= \left(1-\left(1-\frac{1}{T}\right)^T\right)\left(2\delta R + \sum_{i=1}^{k}\beta_i\left(\frac{R^{1+\sigma_i}}{T^{\sigma_i}}\right)\right)\\ &\leq \left(2\delta R + \sum_{i=1}^{k}\beta_i\left(\frac{R^{1+\sigma_i}}{T^{\sigma_i}}\right)\right).\end{aligned}$$

Plugging this to (5) in Lemma 4.1, we deduce that

$$\left(1-\frac{1}{e}\right)(F(\boldsymbol{x^*}) - F(\boldsymbol{x_0})) \leq F(\boldsymbol{x_T}) - F(\boldsymbol{x_0}) + \left(2\delta R + \sum_{i=1}^{k}\beta_i\left(\frac{R^{1+\sigma_i}}{T^{\sigma_i}}\right)\right)$$

as required. □

As an immediate corollary, we obtain the following convergence result.

Corollary 4.4. *Let $F : \mathbb{R}^d \to \mathbb{R}_+$ be a differentiable, monotone, and up-concave function that is Hölder-smooth with respect to a norm $\|\cdot\|$ and h. Let $\sigma := \min_{i\in[k]}\sigma_i$. If $F(\mathbf{0}) = 0$ and $\delta = O(\epsilon)$, then Algorithm 1 returns an $(1-1/e, \epsilon)$-approximate solution after $O(1/\epsilon^{1/\sigma})$ iterations.*

By Corollary 4.4, as long as $\min_{i\in[k]}\sigma_i > 0$, Algorithm 1 converges to an $(1-1/e,\epsilon)$ after a finite number of iterations. However, we have $\min_{i\in[k]}\sigma_i = 0$ for the non-smooth case, in which case, the performance bound (9) given in Theorem 4.3 does not guarantee that Algorithm 1 converges to an $(1-1/e,\epsilon)$-approximate solution.

4.2 Proof of Lemma 4.1

Take $\boldsymbol{y_t} = t\boldsymbol{x_t}/T$ for $0 \leq t \leq T$ and $\boldsymbol{x^*} \in \arg\max_{\boldsymbol{x}\in\mathcal{X}} F(\boldsymbol{x})$. Let $\boldsymbol{v_t^*} = \max\{\boldsymbol{x^*} - \boldsymbol{y_{t-1}}, \mathbf{0}\}$ for $1 \leq t \leq 1$. Then for each $1 \leq t \leq T$,

$$\begin{aligned}F(\boldsymbol{y_t}) &\geq F(\boldsymbol{y_{t-1}}) + \langle\nabla F(\boldsymbol{y_{t-1}}), \boldsymbol{y_t} - \boldsymbol{y_{t-1}}\rangle - \tilde{h}(\|\boldsymbol{y_t} - \boldsymbol{y_{t-1}}\|)\\ &= F(\boldsymbol{y_{t-1}}) + \frac{1}{T}\langle\nabla F(\boldsymbol{y_{t-1}}), \boldsymbol{v_t}\rangle - \tilde{h}\left(\left\|\frac{\boldsymbol{v_t}}{T}\right\|\right).\end{aligned} \tag{10}$$

Next we define $u : \mathbb{R}_+ \to \mathbb{R}_+$ as $u(\gamma) = F(\boldsymbol{y_{t-1}} + \gamma\boldsymbol{v_t^*})$ for $\gamma \geq 0$. Since F is up-concave and $\boldsymbol{v_t^*} \geq 0$, u is concave and differentiable. In particular, we have $F(\boldsymbol{y_{t-1}}+\gamma\boldsymbol{v_t^*}) - F(\boldsymbol{y_{t-1}}) = u(\gamma) - u(0) \leq u'(0)\gamma = \gamma\langle\nabla F(\boldsymbol{y_{t-1}}), \boldsymbol{v_t^*}\rangle$. Consequently, setting $\gamma = 1$ results in

$$F(\boldsymbol{y_{t-1}} + \boldsymbol{v_t^*}) - F(\boldsymbol{y_{t-1}}) \leq \langle\nabla F(\boldsymbol{y_{t-1}}), \boldsymbol{v_t^*}\rangle. \tag{11}$$

Furthermore, $F(\boldsymbol{y_{t-1}} + \boldsymbol{v_t^*}) \geq F(\boldsymbol{x^*})$ since F is monotone and $\boldsymbol{x^*} \leq \boldsymbol{y_{t-1}} + \boldsymbol{v_t^*}$. Then it follows that

$$\begin{aligned} F(\boldsymbol{x^*}) - F(\boldsymbol{y_{t-1}}) &\leq F(\boldsymbol{y_{t-1}} + \boldsymbol{v_t^*}) - F(\boldsymbol{y_{t-1}}) \\ &\leq \langle \nabla F(\boldsymbol{y_{t-1}}), \boldsymbol{v_t^*} \rangle \\ &= \langle \boldsymbol{g_t}, \boldsymbol{v_t^*} \rangle + \langle \nabla F(\boldsymbol{y_{t-1}}) - \boldsymbol{g_t}, \boldsymbol{v_t^*} \rangle \\ &\leq \langle \boldsymbol{g_t}, \boldsymbol{v_t} \rangle + \langle \nabla F(\boldsymbol{y_{t-1}}) - \boldsymbol{g_t}, \boldsymbol{v_t^*} \rangle \\ &= \langle \nabla F(\boldsymbol{y_{t-1}}), \boldsymbol{v_t} \rangle + \langle \nabla F(\boldsymbol{y_{t-1}}) - \boldsymbol{g_t}, \boldsymbol{v_t^*} - \boldsymbol{v_t} \rangle \end{aligned} \tag{12}$$

where the second inequality is from (11) and the third inequality is due to our choice of $\boldsymbol{v_t}$. Combining (10) and (12), we have

$$\begin{aligned} F(\boldsymbol{y_t}) &\geq F(\boldsymbol{y_{t-1}}) + \frac{1}{T} \langle \nabla F(\boldsymbol{y_{t-1}}), \boldsymbol{v_t} \rangle - \tilde{h}\left(\left\| \frac{\boldsymbol{v_t}}{T} \right\| \right) \\ &\geq F(\boldsymbol{y_{t-1}}) + \frac{1}{T} \left(F(\boldsymbol{x^*}) - F(\boldsymbol{y_{t-1}}) \right) \\ &\quad - \frac{1}{T} \langle \nabla F(\boldsymbol{y_{t-1}}) - \boldsymbol{g_t}, \boldsymbol{v_t^*} - \boldsymbol{v_t} \rangle - \tilde{h}\left(\left\| \frac{\boldsymbol{v_t}}{T} \right\| \right). \end{aligned}$$

Define $\eta_t = F(\boldsymbol{x^*}) - F(\boldsymbol{y_{t-1}})$. Then

$$\begin{aligned} \eta_t - \eta_{t+1} &= F(\boldsymbol{y_t}) - F(\boldsymbol{y_{t-1}}) \\ &\geq \frac{1}{T} \left(F(\boldsymbol{x^*}) - F(\boldsymbol{y_{t-1}}) \right) - \frac{1}{T} \langle \nabla F(\boldsymbol{y_{t-1}}) - \boldsymbol{g_t}, \boldsymbol{v_t^*} - \boldsymbol{v_t} \rangle - \tilde{h}\left(\left\| \frac{\boldsymbol{v_t}}{T} \right\| \right) \\ &\geq \frac{1}{T} \left(F(\boldsymbol{x^*}) - F(\boldsymbol{y_{t-1}}) \right) - \frac{1}{T} \| F(\boldsymbol{y_{t-1}}) - \boldsymbol{g_t} \|_* \| \boldsymbol{v_t^*} - \boldsymbol{v_t} \| - \tilde{h}\left(\left\| \frac{\boldsymbol{v_t}}{T} \right\| \right) \\ &\geq \frac{1}{T} \eta_t - \frac{2R}{T} \| F(\boldsymbol{y_{t-1}}) - \boldsymbol{g_t} \|_* - \tilde{h}\left(\left\| \frac{\boldsymbol{v_t}}{T} \right\| \right) \end{aligned} \tag{13}$$

where the second inequality is by Hölder's inequality and the third inequality is from our assumption that $\sup_{\boldsymbol{x} \in \mathcal{X}} \|x\| \leq R$. Then (13) implies that $\eta_{t+1} \leq (1 - 1/T)\, \eta_t + \kappa_t$. Unwinding this recursion, we get

$$\eta_{t+1} \leq \left(1 - \frac{1}{T}\right)^t \eta_1 + \sum_{s=0}^{t-1} \left(1 - \frac{1}{T}\right)^s \kappa_{t-s}.$$

In particular, when $t = T$, we obtain

$$F(\boldsymbol{x^*}) - F(\boldsymbol{y_T}) \leq \left(1 - \frac{1}{T}\right)^T \left(F(\boldsymbol{x^*}) - F(\boldsymbol{y_0}) \right) + \sum_{s=0}^{T-1} \left(1 - \frac{1}{T}\right)^s \kappa_{T-s}.$$

Since $\boldsymbol{y_0} = \boldsymbol{0}$, $\boldsymbol{y_T} = \boldsymbol{x_T}$, and $(1 - 1/T)^T \leq 1/e$, it follows that

$$\left(1 - \frac{1}{e}\right) \left(F(\boldsymbol{x^*}) - F(\boldsymbol{0}) \right) \leq F(\boldsymbol{x_T}) - F(\boldsymbol{0}) + \sum_{s=0}^{T-1} \left(1 - \frac{1}{T}\right)^s \kappa_{T-s}$$

as required.

5 Mirror-prox method

For the continuous greedy algorithm, given by Algorithm 1, to guarantee finite convergence, we need the objective function F to be differentiable and Hölder-smooth. If F is non-differentiable or non-smooth, then the continuous

greedy algorithm does not necessarily converge to a deisred approximate solution. Motivated by this, we develop an algorithm that guarantees finite convergence to an approximate solution even when F is non-differentiable or non-smooth.

We assume that F satisfies the following condition with a norm $\|\cdot\|$ and a function h of the form (2) given by $h(z) = \sum_{i=1}^{k} \beta_i \cdot z^{\sigma_i}$ with $k \geq 1$, $\beta_1, \ldots, \beta_k > 0$, and $0 \leq \sigma_1, \ldots, \sigma_k \leq 1$: for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, there exist $\boldsymbol{g_x} \in \partial^{\uparrow} F(\boldsymbol{x})$ and $\boldsymbol{g_y} \in \partial^{\uparrow} F(\boldsymbol{y})$ such that

$$\|\boldsymbol{g_x} - \boldsymbol{g_y}\|_* \leq h(\|\boldsymbol{x} - \boldsymbol{y}\|). \tag{14}$$

Lemma 5.1. *If $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable and up-concave, $\partial^{\uparrow} F(\boldsymbol{x}) = \{\nabla F(\boldsymbol{x})\}$ for any $\boldsymbol{x} \in \mathbb{R}^d$.*

Therefore, when F is differentiable, (14) reduces to (1) except that we may have $\min_{i \in [k]} \sigma_i = 0$ in (14). The proof of Lemma 5.1 is given in Appendix A.

We propose a variant of the mirror-prox method, given in Algorithm 2, which achieves a finite convergence to a constant approximation. In particular, Algorithm 2 admits the case when $\min_{i \in [k]} \sigma_i = 0$ even if Algorithm 1 only works only when $\min_{i \in [k]} \sigma_i > 0$. Moreover, Algorithm 2 does not require F to be differentiable unlike Algorithm 1.

Let Φ be a mirror map that is 1-strongly convex on $\mathbb{R}^d$ with respect to the norm $\|\cdot\|$. We define

$$V_{\boldsymbol{x}}(\boldsymbol{z}) := \Phi(\boldsymbol{z}) - \Phi(\boldsymbol{x}) - \langle \nabla \Phi(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x} \rangle$$

which is often referred to as the *Bregman divergence* associated with Φ. Let Prox be a proximal operator given by

$$\mathrm{Prox}_{\boldsymbol{x}}(\boldsymbol{\xi}) := \arg\min_{\boldsymbol{z} \in \mathcal{X}} \left\{ \langle \boldsymbol{\xi}, \boldsymbol{z} \rangle + V_{\boldsymbol{x}}(\boldsymbol{z}) \right\}.$$

Now we provide a pseudocode of the mirror-prox method for solving (SFM). At each iteration t, we obtain an

Algorithm 2 Mirror-prox algorithm for (SFM)

Initialize $\boldsymbol{x_1} \in \mathrm{argmin}_{x \in \mathcal{X}} \Phi(\boldsymbol{x})$.
for $t = 1, \ldots, T-1$ **do**
  Fetch $\boldsymbol{g_t}$.
  $\boldsymbol{x_{t+1/2}} = \mathrm{Prox}_{\boldsymbol{x_t}}(-\gamma_t \boldsymbol{g_t})$.
  Fetch $\boldsymbol{g_{t+1/2}}$.
  $\boldsymbol{x_{t+1}} = \mathrm{Prox}_{\boldsymbol{x_t}}(-\gamma_t \boldsymbol{g_{t+1/2}})$.
end for
Return $x^* \in \mathrm{argmax} \left\{ F(\boldsymbol{x}) : \boldsymbol{x} \in \{\boldsymbol{x_{t+1/2}} : t = \lfloor (T-2)/3 \rfloor + 1, \ldots, T-1\} \right\}$.

estimate $\boldsymbol{g_t}$ of an up-super-gradient of F at $\boldsymbol{x_t}$ and an estimate $\boldsymbol{g_{t+1/2}}$ of an up-super-gradient of F at $\boldsymbol{x_{t+1/2}}$. As long as the estimation error, measured by the norm of the gap, is bounded, we obtain a finite convergence.

5.1 Convergence of the mirror-prox algorithm

The following lemma is important for providing a constant approximation guarantee on Algorithm 2 when maximizing an up-concave function. It is similar in spirit to the first-order characterization of concave functions. Hassani et al. [17] have already considered the differentiable case, while our lemma extends the result to the non-differentiable case.

Lemma 5.2. *If F is monotone and up-concave, then for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $\boldsymbol{g} \in \partial^{\uparrow}(F(\boldsymbol{x}))$, we have*

$$F(\boldsymbol{y}) - 2F(\boldsymbol{x}) \leq \boldsymbol{g}^{\top}(\boldsymbol{y} - \boldsymbol{x}).$$

Proof. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. Since $\boldsymbol{x} \wedge \boldsymbol{y} \leq \boldsymbol{x} \leq \boldsymbol{x} \vee \boldsymbol{y}$, it follows from the definition of $\partial^{\uparrow} F(\boldsymbol{x})$ that for any $\boldsymbol{g} \in \partial^{\uparrow} F(\boldsymbol{x})$, we have

$$F(\boldsymbol{x} \wedge \boldsymbol{y}) - F(\boldsymbol{x}) \leq \boldsymbol{g}^{\top}(\boldsymbol{x} \wedge \boldsymbol{y} - \boldsymbol{x})$$
$$F(\boldsymbol{x} \vee \boldsymbol{y}) - F(\boldsymbol{x}) \leq \boldsymbol{g}^{\top}(\boldsymbol{x} \vee \boldsymbol{y} - \boldsymbol{x})$$

for every $\boldsymbol{g} \in \partial^{\uparrow}(F(\boldsymbol{x}))$. Adding up the two inequalities, we obtain

$$-2F(\boldsymbol{x}) + F(\boldsymbol{x} \vee \boldsymbol{y}) + F(\boldsymbol{x} \wedge \boldsymbol{y}) \leq \boldsymbol{g}^{\top}(\boldsymbol{x} \vee \boldsymbol{y} + \boldsymbol{x} \wedge \boldsymbol{y} - 2\boldsymbol{x}) = \boldsymbol{g}^{\top}(\boldsymbol{y} - \boldsymbol{x}).$$

Since F is nonnegative and monotone, it follows that $F(\boldsymbol{y}) - 2F(\boldsymbol{x}) \leq F(\boldsymbol{x} \vee \boldsymbol{y}) - 2F(\boldsymbol{x}) \leq \boldsymbol{g}^{\top}(\boldsymbol{y} - \boldsymbol{x})$. □

Based on Lemma 5.2, we can prove the following theorem that provides a convergence guarantee of Algorithm 2.

Theorem 5.3. *Let $F : \mathbb{R}^d \to \mathbb{R}_+$ be a monotone and up-concave function that satisfies (14) with a norm $\|\cdot\|$ and function h. In particular, there exist $\boldsymbol{g}(\boldsymbol{x_t}) \in \partial^{\uparrow} F(\boldsymbol{x_t})$ and $\boldsymbol{g}(\boldsymbol{x_{t+1/2}}) \in \partial^{\uparrow} F(\boldsymbol{x_{t+1/2}})$ such that $\|\boldsymbol{g}(\boldsymbol{x_t}) - \boldsymbol{g}(\boldsymbol{x_t})\|_* \leq h(\|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|)$ for $t \geq 1$. Assume that for each $t \geq 1$, $\boldsymbol{g_t}$ and $\boldsymbol{g_{t+1/2}}$ are chosen so that $\|\boldsymbol{g_t} - \boldsymbol{g}(\boldsymbol{x_t})\|_*, \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g}(\boldsymbol{x_{t+1/2}})\|_* \leq \delta$. Assume further that $\sup_{\boldsymbol{x},\boldsymbol{y} \in \mathcal{X}} V_{\boldsymbol{y}}(\boldsymbol{x}) \leq D$ for some $D > 0$ and γ_t is set to*

$$\gamma_t = \left(2t^{\frac{1-\sigma}{2}} \sum_{i=1}^{k} \beta_i\right)^{-1}$$

where $\sigma = \min_{i \in [k]} \sigma_i$. If Let $\hat{\boldsymbol{x}}$ be the solution returned by Algorithm 2, then

$$F(\hat{\boldsymbol{x}}) \geq \frac{1}{2} \sup_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x}) - \delta\sqrt{2D} - \frac{12(D+1)\sum_{i=1}^{k} \beta_i}{T^{\frac{1+\delta}{2}}} - \frac{T^{\frac{1+\sigma}{2}}}{\sum_{i=1}^{k} \beta_i} \delta^2.$$

We defer the proof of this theorem to the next subsection. By Theorem 5.3, we obtain the following convergence result for Algorithm 2.

Corollary 5.4. *Let $F : \mathbb{R}^d \to \mathbb{R}_+$ be a monotone, and up-concave function that satisfies (14) with a norm $\|\cdot\|$ and h. Let $\sigma := \min_{i \in [k]} \sigma_i$. If $\delta = O(\epsilon)$, then Algorithm 2 returns an $(1/2, \epsilon)$-approximate solution after $O(1/\epsilon^{2/(1+\sigma)})$ iterations.*

By Corollary 5.4, Algorithm 2 converges to an $(1/2, \epsilon)$-approximate solution after a finite number of iterations although $\min_{i \in [k]} \sigma_i = 0$.

5.2 Proof of Theorem 5.3

By Lemma 5.2, for each $x \in \mathcal{X}$, we obtain

$$\begin{aligned}
&F(\boldsymbol{x}) - 2F(\boldsymbol{x_{t+1/2}}) \\
&\leq \langle \boldsymbol{g(x_{t+1/2})}, \boldsymbol{x} - \boldsymbol{x_{t+1/2}} \rangle \\
&= -\langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x} \rangle + \langle \boldsymbol{g_{t+1/2}} - \boldsymbol{g(x_{t+1/2})}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x} \rangle \\
&\leq -\langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x} \rangle + \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g(x_{t+1/2})}\|_* \|\boldsymbol{x_{t+1/2}} - \boldsymbol{x}\| \\
&\leq -\langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x} \rangle + \delta\sqrt{2D}
\end{aligned} \tag{15}$$

where the second inequality is due to Hölder's inequality and the last inequality is because $\|\boldsymbol{x_{t+1/2}} - \boldsymbol{x}\|^2 \leq 2V_{\boldsymbol{x}}(\boldsymbol{x_{t+1/2}})$. We have that

$$-\langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x} \rangle \tag{16}$$

$$= -\langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1}} - \boldsymbol{x} \rangle - \langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}} \rangle \tag{17}$$

$$= -\langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1}} - \boldsymbol{x} \rangle - \langle \boldsymbol{g_t}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}} \rangle - \langle \boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}} \rangle. \tag{18}$$

Observe that $\boldsymbol{x_{t+1/2}} = \arg\min_{\boldsymbol{x} \in \mathcal{X}} \{\langle -\gamma_t \boldsymbol{g_t}, \boldsymbol{x} \rangle + V_{\boldsymbol{x_t}}(\boldsymbol{x})\}$. Then, by the first-order optimality condition of convex minimization, we have

$$\langle -\gamma_t \boldsymbol{g_t} + \nabla\Phi(\boldsymbol{x_{t+1/2}}) - \nabla\Phi(\boldsymbol{x_t}), \boldsymbol{x_{t+1/2}} - \boldsymbol{x} \rangle \leq 0 \tag{19}$$

for every $\boldsymbol{x} \in \mathcal{X}$. Hence, for any $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned}
-\gamma_t \langle \boldsymbol{g_t}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x} \rangle &\leq \langle \nabla\Phi\left(\boldsymbol{x_t}\right) - \nabla\Phi\left(\boldsymbol{x_{t+1/2}}\right), \boldsymbol{x_{t+1/2}} - \boldsymbol{x} \rangle \\
&= V_{\boldsymbol{x_t}}(\boldsymbol{x}) - V_{\boldsymbol{x_{t+1/2}}}(\boldsymbol{x}) - V_{\boldsymbol{x_t}}(\boldsymbol{x_{t+1/2}}).
\end{aligned} \tag{20}$$

where the inequality is due to (19) and the equality is from a well-known fact about the Bregman divergence. Similarly, for any $\boldsymbol{x} \in \mathcal{X}$,

$$\begin{aligned}
-\gamma_t \langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1}} - \boldsymbol{x} \rangle &\leq \langle \nabla\Phi\left(\boldsymbol{x_t}\right) - \nabla\Phi\left(\boldsymbol{x_{t+1}}\right), \boldsymbol{x_{t+1}} - \boldsymbol{x} \rangle \\
&= V_{\boldsymbol{x_t}}(\boldsymbol{x}) - V_{\boldsymbol{x_{t+1}}}(\boldsymbol{x}) - V_{\boldsymbol{x_t}}(\boldsymbol{x_{t+1}}).
\end{aligned} \tag{21}$$

Summing up (20) with $\boldsymbol{x} = \boldsymbol{x_{t+1}}$ and (21) with $\boldsymbol{x} = \boldsymbol{x_{t+1/2}}$, we obtain

$$-\gamma_t \langle \boldsymbol{g_t} - \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}} \rangle \leq \langle \nabla\Phi(\boldsymbol{x_{t+1}}) - \nabla\Phi(\boldsymbol{x_{t+1/2}}), \boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}} \rangle. \tag{22}$$

Note that the right-hand side of (22) is bounded above by $-\|\boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}}\|^2$ as Φ is 1-strongly convex with respect to the norm $\|\cdot\|$. Furthermore, we can apply Hölder's inequality to obtain a lower bound on the left-hand side of (22). Then we have

$$\begin{aligned}
-\gamma_t \|\boldsymbol{g_t} - \boldsymbol{g_{t+1/2}}\|_* \|\boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}}\| &\leq -\gamma_t \langle \boldsymbol{g_t} - \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}} \rangle \\
&\leq -\|\boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}}\|^2.
\end{aligned} \tag{23}$$

(23) implies that $\|\boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}}\| \leq \gamma_t \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}\|_*$, by which we get

$$\begin{aligned} -\gamma_t \langle \boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}} \rangle &\leq \gamma_t \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}\|_* \|\boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}}\| \\ &\leq \gamma_t^2 \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}\|_*^2. \end{aligned} \tag{24}$$

Next we deduce the following inequalities. Note that

$$\begin{aligned} &- \gamma_t \langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x} \rangle \\ &= -\gamma_t \langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1}} - \boldsymbol{x} \rangle - \gamma_t \langle \boldsymbol{g_t}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}} \rangle \\ &\quad - \gamma_t \langle \boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x_{t+1}} \rangle \\ &\leq \left(V_{\boldsymbol{x_t}}(\boldsymbol{x}) - V_{\boldsymbol{x_{t+1}}}(\boldsymbol{x}) - V_{\boldsymbol{x_t}}(\boldsymbol{x_{t+1}}) \right) \\ &\quad + \left(V_{\boldsymbol{x_t}}(\boldsymbol{x_{t+1}}) - V_{\boldsymbol{x_{t+1/2}}}(\boldsymbol{x_{t+1}}) - V_{\boldsymbol{x_t}}(\boldsymbol{x_{t+1/2}}) \right) + \gamma_t^2 \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}\|_*^2 \\ &= V_{\boldsymbol{x_t}}(\boldsymbol{x}) - V_{\boldsymbol{x_{t+1}}}(\boldsymbol{x}) + \gamma_t^2 \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}\|_*^2 - V_{\boldsymbol{x_{t+1/2}}}(\boldsymbol{x_{t+1}}) - V_{\boldsymbol{x_t}}(\boldsymbol{x_{t+1/2}}) \\ &\leq V_{\boldsymbol{x_t}}(\boldsymbol{x}) - V_{\boldsymbol{x_{t+1}}}(\boldsymbol{x}) + \gamma_t^2 \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}\|_*^2 \\ &\quad - \frac{1}{2} [\|\boldsymbol{x_{t+1}} - \boldsymbol{x_{t+1/2}}\|^2 + \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^2] \\ &\leq V_{\boldsymbol{x_t}}(\boldsymbol{x}) - V_{\boldsymbol{x_{t+1}}}(\boldsymbol{x}) + \gamma_t^2 \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}\|_*^2 - \frac{1}{2} \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^2 \end{aligned} \tag{25}$$

where the first inequality is from (20) with $\boldsymbol{x} = \boldsymbol{x_{t+1}}$, (21), and (24) while the second inequality is due to the 1-strong convexity of Φ with respect to the norm $\|\cdot\|$. To bound the last part of the inequalities, we consider the following terms. Notice that

$$\begin{aligned} &\gamma_t^2 \|\boldsymbol{g_{t+1/2}} - \boldsymbol{g_t}\|_*^2 - \frac{1}{2} \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^2 \\ &\leq \gamma_t^2 (\|\boldsymbol{g_{t+1/2}} - \boldsymbol{g}(\boldsymbol{x_{t+1/2}})\|_* + \|\boldsymbol{g}(\boldsymbol{x_{t+1/2}}) - \boldsymbol{g}(\boldsymbol{x_t})\|_* + \|\boldsymbol{g}(\boldsymbol{x_t}) - \boldsymbol{g_t}\|_*)^2 \\ &\quad - \frac{1}{2} \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^2 \\ &\leq \gamma_t^2 \left(2\delta + \sum_{i=1}^k \beta_i \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^{\sigma_i} \right)^2 - \frac{1}{2} \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^2 \\ &\leq 2\gamma_t^2 \left(4\delta^2 + \left(\sum_{i=1}^k \beta_i \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^{\sigma_i} \right)^2 \right) - \frac{1}{2} \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^2 \\ &= 8\gamma_t^2 \delta^2 + 2\gamma_t^2 \left(\sum_{i=1}^k \beta_i \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^{\sigma_i} \right)^2 - \frac{1}{2} \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^2 \end{aligned} \tag{26}$$

where the first inequality is by the subadditivity of norms, the second inequality is due to our assumption that $\|\boldsymbol{g_{t+1/2}} - \boldsymbol{g}(\boldsymbol{x_{t+1/2}})\|_*, \|\boldsymbol{g}(\boldsymbol{x_t}) - \boldsymbol{g_t}\|_* \leq \delta$, and the third inequality is because $\|\boldsymbol{a} + \boldsymbol{b}\|^2 \leq 2(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2)$ for any $\boldsymbol{a}, \boldsymbol{b}$. The following lemma provides an upper bound on the last part of (26). Recall that we set $\gamma_t = 1/\left(2t^{\frac{1-\sigma}{2}} \sum_{i=1}^k \beta_i\right)$. Next we need the following lemma whose proof is given in Appendix A.

Lemma 5.5. *For $1 \leq t \leq T-1$, we have*

$$2\gamma_t^2 \left(\sum_{i=1}^k \beta_i \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^{\sigma_i} \right)^2 - \frac{1}{2} \|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\|^2 \leq \frac{1}{2t}. \tag{27}$$

By (25), (26), and Lemma 5.5,

$$-\gamma_t\langle \boldsymbol{g_{t+1/2}}, \boldsymbol{x_{t+1/2}} - \boldsymbol{x}\rangle \le V_{\boldsymbol{x_t}}(\boldsymbol{x}) - V_{\boldsymbol{x_{t+1}}}(\boldsymbol{x}) + 8\gamma_t^2\delta^2 + \frac{1}{2t}. \tag{28}$$

Together with (15), (28) implies that

$$\gamma_t F(\boldsymbol{x}) - 2\gamma_t F(\boldsymbol{x_{t+1/2}}) \le V_{\boldsymbol{x_t}}(\boldsymbol{x}) - V_{\boldsymbol{x_{t+1}}}(\boldsymbol{x}) + \gamma_t\delta\sqrt{2D} + 8\delta^2\gamma_t^2 + \frac{1}{2t}. \tag{29}$$

Summing (29) up for $t = T_0, \dots, T-1$ where $T_0 = \lfloor \frac{T-2}{3} \rfloor + 1$, we have

$$\begin{aligned}
&\left(\sum_{t=T_0}^{T-1}\gamma_t\right)F(\boldsymbol{x}) - 2\sum_{t=T_0}^{T-1}\gamma_t F(\boldsymbol{x_{t+1/2}}) \\
&\le \left(\sum_{t=T_0}^{T-1}\gamma_t\right)\delta\sqrt{2D} + V_{\boldsymbol{x_{T_0}}}(\boldsymbol{x}) - V_{\boldsymbol{x_T}}(\boldsymbol{x}) + \sum_{t=T_0}^{T-1}\frac{1}{2t} + 8\delta^2\left(\sum_{t=T_0}^{T-1}\gamma_t^2\right) \\
&\le \left(\sum_{t=T_0}^{T-1}\gamma_t\right)\delta\sqrt{2D} + V_{\boldsymbol{x_{T_0}}}(\boldsymbol{x}) + \sum_{t=T_0}^{T-1}\frac{1}{2t} + 8\delta^2\left(\sum_{t=T_0}^{T-1}\gamma_t\right)^2 \\
&\le \left(\sum_{t=T_0}^{T-1}\gamma_t\right)\delta\sqrt{2D} + \left(D + \frac{3}{2(T-1)}\cdot(T-1)\right) + 8\delta^2\left(\sum_{t=T_0}^{T-1}\gamma_t\right)^2 \\
&\le \left(\sum_{t=T_0}^{T-1}\gamma_t\right)\delta\sqrt{2D} + (D+2) + 8\delta^2\left(\sum_{t=T_0}^{T-1}\gamma_t\right)^2
\end{aligned}$$

where the third inequality is because $D \ge \sup_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}} V_{\boldsymbol{y}}(\boldsymbol{x})$ and $T_0 \ge (T-2)/3 - 2/3 + 1 = (T-1)/3$. Dividing each side by $\sum_{t=T_0}^{T-1}\gamma_t$, we obtain

$$F(\boldsymbol{x}) - 2\frac{\sum_{t=T_0}^{T-1}\gamma_t F(\boldsymbol{x_{t+1/2}})}{\sum_{t=T_0}^{T-1}\gamma_t} \le \delta\sqrt{2D} + \frac{D+2}{\sum_{t=T_0}^{T-1}\gamma_t} + 8\delta^2\sum_{t=T_0}^{T-1}\gamma_t. \tag{30}$$

The following lemma provide upper and lower bounds on the sum $\sum_{t=T_0}^{T-1}\gamma_t$, by which and (30), we can complete the proof of Theorem 5.3. We need the following lemma whose proof is given in Appendix A.

Lemma 5.6. *Let* $T_0 = \lfloor (T-2)/3 \rfloor + 1$. *Then*

$$\frac{1}{12\sum_{i=1}^k\beta_i}T^{\frac{1+\sigma}{2}} \le \sum_{t=T_0}^{T-1}\gamma_t \le \frac{1}{\sum_{i=1}^k\beta_i}T^{\frac{1+\sigma}{2}}$$

By Lemma 5.6 and (30), we get that for any $\boldsymbol{x}\in\mathcal{X}$,

$$\frac{\sum_{t=T_0}^{T-1}\gamma_t F(\boldsymbol{x_{t+1/2}})}{\sum_{t=T_0}^{T-1}\gamma_t} \ge \frac{1}{2}F(\boldsymbol{x}) - \delta\sqrt{2D} - \frac{12(D+2)\sum_{i=1}^k\beta_i}{T^{\frac{1+\delta}{2}}} - \frac{T^{\frac{1+\sigma}{2}}}{\sum_{i=1}^k\beta_i}\delta^2. \tag{31}$$

Let $\hat{\boldsymbol{x}} = \operatorname{argmax}\{F(\boldsymbol{x}) : \boldsymbol{x}\in\{\boldsymbol{x_{T_0+1/2}}, \dots, \boldsymbol{x_{(T-1)+1/2}}\}\}$. As the left-hand side of (31) is a convex combination of $F(\boldsymbol{x_{T_0+1/2}}), \dots, F(\boldsymbol{x_{(T-1)+1/2}})$, it is bounded above by $F(\hat{\boldsymbol{x}})$. Furthermore, taking the supremum of the right-hand side over $\boldsymbol{x}\in\mathcal{X}$, it follows that

$$F(\hat{\boldsymbol{x}}) \ge \frac{1}{2}\sup_{\boldsymbol{x}\in\mathcal{X}}F(\boldsymbol{x}) - \delta\sqrt{2D} - \frac{12(D+2)\sum_{i=1}^k\beta_i}{T^{\frac{1+\delta}{2}}} - \frac{T^{\frac{1+\sigma}{2}}}{\sum_{i=1}^k\beta_i}\delta^2, \tag{32}$$

as required.

6 Robust submodular maximization

In this section, we consider robust maximization of nonnegative monotone DR-submodular functions $F_1, \dots, F_n : \mathbb{R}^d \to \mathbb{R}_+$ formulated by $\sup_{\boldsymbol{x} \in \mathcal{X}} \min_{i \in [n]} F_i(\boldsymbol{x})$. We further assume that $F_1, \dots, F_n$ are differentiable. To apply our framework, we show that the problem is an instance of monotone up-concave function maximization.

Lemma 6.1. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be defined as $F = \min_{i \in [n]} F_i$ where $F_1, \dots, F_n : \mathbb{R}^d \to \mathbb{R}$ are nonnegative, monotone, and up-concave. Then F is also nonnegative, monotone, and up-concave.*

Although $F_1, \dots, F_n$ are differentiable, $F = \min_{i \in [n]} F_i$, the point-wise minimum of $F_1, \dots, F_n$, is not necessarily differentiable. Nevertheless, we may apply the mirror-prox algorithm given by Algorithm 2 as long as F satisfies (14) for some h. To argue that there exists a function h of the form (2) with which $F = \min_{i \in [n]} F_i$ satisfies (14), we need the following two lemmas.

Lemma 6.2. $\operatorname{conv}\{\nabla F_i(\boldsymbol{x}) : i \in \arg\min_{i \in [n]} F_i(\boldsymbol{x})\} \subseteq \partial^{\uparrow} F(\boldsymbol{x})$ *where F is defined as $F = \min_{i \in [n]} F_i$.*

We assume that $F_1, \dots, F_n$ are L-Lipschitz continuous with respect to the norm $\|\cdot\|$ for some $L > 0$, i.e., $|F_i(\boldsymbol{x}) - F_i(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and each $i \in [n]$.

Lemma 6.3. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be defined as $F = \min_{i \in [n]} F_i$ where $F_1, \dots, F_n : \mathbb{R}^d \to \mathbb{R}$ are L-Lipschitz continuous. Then F is L-Lipschitz continuous and $\|\boldsymbol{g}\|_* \leq L$ for any $\boldsymbol{g} \in \operatorname{conv}\{\nabla F_i(\boldsymbol{x}) : i \in \arg\min_{i \in [n]} F_i(\boldsymbol{x})\}$.*

Based on Lemmas 6.1–6.3, $F = \min_{i \in [n]} F_i$ is nonnegative, monotone, and up-concave and satisfies (14) with norm $\|\cdot\|$ and function $h = 2L$, a constant function. We provide the proofs of the lemmas in Appendix B. Hence, we may apply the mirror-prox method, presented as Algorithm 2 for solving the robust submodular maximization given by $\sup_{\boldsymbol{x} \in \mathcal{X}} \min_{i \in [n]} F_i(\boldsymbol{x})$. Algorithm 3 is the mirror-prox method tailored for robust submodular maximization.

Algorithm 3 Mirror-prox algorithm for robust submodular maximization

Initialize $\boldsymbol{x_1} \in \operatorname{argmin}_{x \in \mathcal{X}} \Phi(\boldsymbol{x})$.
for $t = 1, \dots, T-1$ **do**
  $\boldsymbol{x_{t+1/2}} = \operatorname{Prox}_{\boldsymbol{x_t}}(-\gamma_t \nabla F_i(\boldsymbol{x_t}))$ for some $i \in \arg\min_{i \in [n]} F_i(\boldsymbol{x_t})$.
  $\boldsymbol{x_{t+1}} = \operatorname{Prox}_{\boldsymbol{x_t}}(-\gamma_t \nabla F_i(\boldsymbol{x_{t+1/2}}))$ for some $i \in \arg\min_{i \in [n]} F_i(\boldsymbol{x_{t+1/2}})$.
end for
Return $x^* \in \operatorname{argmax}\left\{F(\boldsymbol{x}) : \boldsymbol{x} \in \{\boldsymbol{x_{t+1/2}} : t = \lfloor (T-2)/3 \rfloor + 1, \dots, T-1\}\right\}$.

Based on Theorem 5.3 and Corollary 5.4, we deduce the following convergence result for Algorithm 3.

Theorem 6.4. *Let $F_1, \dots, F_n : \mathbb{R}^d \to \mathbb{R}_+$ be monotone, up-concave, differentiable, and L-Lipschitz continuous for some $L > 0$. Then Algorithm 3 returns a solution $\hat{\boldsymbol{x}} \in \mathcal{X}$ such that*

$$\min_{i \in [n]} F_i(\hat{\boldsymbol{x}}) \geq \frac{1}{2} \sup_{\boldsymbol{x} \in \mathcal{X}} \min_{i \in [n]} F_i(\boldsymbol{x}) - \frac{24(D+1)L}{\sqrt{T}}$$

where $\sup_{\boldsymbol{x}, \boldsymbol{y}} V_{\boldsymbol{y}}(\boldsymbol{x}) \leq D$. Hence, Algorithm 3 returns an $(1/2, \epsilon)$-approximate solution to $\sup_{\boldsymbol{x} \in \mathcal{X}} \min_{i \in [n]} F_i(\boldsymbol{x})$ after $O(1/\epsilon^2)$ iterations.

For robust submodular set function maximization, we may use the known fact that the multilinear extension of a monotone submodular set function is nonnegative, monotone, differentiable, up-concave, and L-Lipschitz continuous for some $L > 0$ (see [46]).

7 Distributionally robust formulation under Wasserstein ambiguity

Let Q be a probability distribution over a sample space $\Xi \subseteq \mathbb{R}^m$. We consider the *2-Wasserstein ball* of radius θ, based on the ℓ_2 norm, centered at distribution Q:

$$\mathcal{B}(Q, \theta) = \{P : \ d_W(P, Q) \leq \theta\}$$

where $d_W(P, Q)$ denotes the *2-Wasserstein distance* between two probability distributions P, Q over sample space Ξ, which is defined as

$$d_W(P, Q) = \inf_{\Pi} \left\{ \left(\mathbb{E}_{(\boldsymbol{\xi}, \boldsymbol{\zeta}) \sim \Pi} \left[\|\boldsymbol{\xi} - \boldsymbol{\zeta}\|_2^2\right]\right)^{1/2} : \ \Pi \text{ has marginal distributions } P, Q \right\}.$$

Basically, $\mathcal{B}(Q, \theta)$ is the family of probability distributions that are within a 2-Wassertein distance of θ from Q. In the distributionally robust optimization (DRO) literature, it is often assumed that Q has a finite support, e.g., Q is an empirical distribution over N samples $\boldsymbol{\xi^1}, \ldots, \boldsymbol{\xi^N}$, that is, $Q = \sum_{i \in [N]} \frac{1}{N} \delta_{\boldsymbol{\xi^i}}$ where $\delta_{\boldsymbol{\xi^i}}$ is the Dirac measure on sample $\boldsymbol{\xi^i}$. Throughout this paper, we assume that $Q = \sum_{i \in [N]} p_i \delta_{\boldsymbol{\xi^i}}$ for some $p \in \{p' \in \mathbb{R}_+^N : \sum_{i \in [N]} p'_i = 1\}$ and that $p_i > 0$ for each $i \in [N]$.

Given a nonnegative functions $f : \mathbb{R}^d \times \Xi \to \mathbb{R}_+$, we consider the following distributionally robust maximization problem:

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \inf_{P \in \mathcal{B}(Q, \theta)} \mathbb{E}_{\boldsymbol{\xi} \sim P} \left[f(\boldsymbol{x}, \boldsymbol{\xi})\right] \tag{DR-SFM}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is some compact feasible region for variables $\boldsymbol{x}$. In particular, we consider settings where f has the following properties. For any $\boldsymbol{x} \in \mathcal{X}$, $f(\boldsymbol{x}, \boldsymbol{\xi})$ is convex with respect to $\boldsymbol{\xi} \in \Xi$. Moreover, for any $\boldsymbol{\xi} \in \Xi$, $f(\boldsymbol{x}, \boldsymbol{\xi})$ is DR-submodular with respect to $\boldsymbol{x} \in \mathcal{X}$. Furthermore, we assume that for any $\boldsymbol{\xi} \in \Xi$, $f(\boldsymbol{x}, \boldsymbol{\xi})$ is monotone with respect to $\boldsymbol{x} \in \mathcal{X}$, i.e., $f(\boldsymbol{x}, \boldsymbol{\xi}) \leq f(\boldsymbol{y}, \boldsymbol{\xi})$ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ such that $\boldsymbol{x} \leq \boldsymbol{y}$.

7.1 Submodular-convex reformulation

To solve the distributionally robust submodular function maximization problem defined in (DR-SFM), we provide a reformulation of the inner infimum of (DR-SFM).

$$F(\boldsymbol{x}) = \inf_{P \in \mathcal{B}(Q, \theta)} \mathbb{E}_{\boldsymbol{\xi} \sim P} \left[f(\boldsymbol{x}, \boldsymbol{\xi})\right].$$

Here, $\mathcal{B}(Q, \theta)$ contains infinitely many distributions, so computing a worst-case distribution directly involves solving an infinite-dimensional problem. Instead, we provide an equivalent description of F by the strong duality result of Blanchet and Murthy [7] for the Wasserstein ball $\mathcal{B}(Q, \theta)$.

Proposition 7.1.

$$\inf_{P \in \mathcal{B}(Q, \theta)} \mathbb{E}_{\boldsymbol{\xi} \sim P} \left[f(\boldsymbol{x}, \boldsymbol{\xi})\right] = \inf_{(\boldsymbol{\zeta^1}; \cdots; \boldsymbol{\zeta^N}) \in Z} \sum_{i \in [N]} p_i f(\boldsymbol{x}, \boldsymbol{\zeta^i}).$$

where Z *is defined as*

$$\mathrm{Z} = \left\{ (\boldsymbol{\zeta^1}; \cdots; \boldsymbol{\zeta^N}) : \sum_{i\in[N]} p_i \|\boldsymbol{\xi^i} - \boldsymbol{\zeta^i}\|_2^2 \le \theta^2,\ \boldsymbol{\zeta^i} \in \Xi\ \ \forall i \in [N] \right\}.$$

Proof. As $\mathcal{B}(Q,\theta)$ is a 2-Wasserstein ball, the strong duality result of Blanchet and Murthy [7, Eq. (11)] states that

$$\begin{aligned} F(\boldsymbol{x}) &= - \sup_{P\in\mathcal{B}(Q,\theta)} \mathbb{E}_{\boldsymbol{\xi}\sim P}\left[-f(\boldsymbol{x},\boldsymbol{\xi})\right] \\ &= - \inf_{\lambda\ge 0} \left\{ \theta^2\lambda + \mathbb{E}_{\boldsymbol{\xi}\sim Q}\left[\sup_{\boldsymbol{\zeta}\in\Xi} \left\{ -f(\boldsymbol{x},\boldsymbol{\zeta}) - \lambda\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|_2^2 \right\} \right] \right\} \\ &= \sup_{\lambda\ge 0} \left\{ \mathbb{E}_{\boldsymbol{\xi}\sim Q}\left[\inf_{\boldsymbol{\zeta}\in\Xi} \left\{ f(\boldsymbol{x},\boldsymbol{\zeta}) + \lambda\|\boldsymbol{\xi}-\boldsymbol{\zeta}\|_2^2 \right\} \right] - \theta^2\lambda \right\}. \end{aligned} \tag{33}$$

Furthermore, by our assumption that $Q = \sum_{i\in[N]} p_i \delta_{\boldsymbol{\xi}^i}$ has a finite support, the last term of (33) can be rewritten as the following finite-dimensional max-min problem:

$$F(\boldsymbol{x}) = \sup_{\lambda\ge 0} \sum_{i\in[N]} p_i \left\{ \inf_{\boldsymbol{\zeta}\in\Xi} \left\{ f(\boldsymbol{x},\boldsymbol{\zeta}) + \lambda\|\boldsymbol{\xi^i}-\boldsymbol{\zeta}\|_2^2 \right\} - \theta^2\lambda \right\}. \tag{34}$$

As $f(\boldsymbol{x},\boldsymbol{\zeta})$ is convex in $\boldsymbol{\zeta}$, the function $\sum_{i\in[N]} p_i f(\boldsymbol{x},\boldsymbol{\zeta^i})$ is also convex. Moreover, $\sum_{i\in[N]} p_i\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 - \theta^2 \le 0$ is a convex constraint with respect to $\boldsymbol{\zeta^1},\ldots,\boldsymbol{\zeta^N}$. Notice that choosing $\boldsymbol{\zeta^i} = \boldsymbol{\xi^i}$ for $i\in[N]$ gives $\sum_{i\in[N]} p_i\|\boldsymbol{\xi^i} - \boldsymbol{\zeta^i}\|_2^2 - \theta^2 = -\theta^2 < 0$. Then strong Lagrangian duality holds to give

$$\begin{aligned} &\inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in\mathrm{Z}} \sum_{i\in[N]} p_i f(\boldsymbol{x},\boldsymbol{\zeta^i}) \\ &= \sup_{\lambda\ge 0} \inf_{\boldsymbol{\zeta^1},\ldots,\boldsymbol{\zeta^N}\in\Xi} \sum_{i\in[N]} p_i f(\boldsymbol{x},\boldsymbol{\zeta^i}) - \lambda\left(\sum_{i\in[N]} p_i\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 - \theta^2 \right), \end{aligned}$$

and it follows from (34) that

$$\inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in\mathrm{Z}} \sum_{i\in[N]} p_i f(\boldsymbol{x},\boldsymbol{\zeta^i}) = F(\boldsymbol{x}) = \inf_{P\in\mathcal{B}(Q,\theta)} \mathbb{E}_{\boldsymbol{\xi}\sim P}\left[f(\boldsymbol{x},\boldsymbol{\xi})\right],$$

as required. □

By Proposition 7.1, (DR-SFM) admits the following reformulation.

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in\mathrm{Z}} \sum_{i\in[N]} p_i f(\boldsymbol{x},\boldsymbol{\zeta^i}). \tag{DR-SFM'}$$

Here, Z is bounded due to the constraint $\sum_{i\in[N]} p_i\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 \le \theta^2$, although the sample space Ξ is not necessarily bounded. Moreover, if $(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N}) \in \mathrm{Z}$, then $\boldsymbol{\zeta^1},\ldots,\boldsymbol{\zeta^N}$ all belong to Γ where

$$\Gamma = \bigcup_{i\in[N]} \left\{ \boldsymbol{\zeta}\in\Xi :\ p_i\|\boldsymbol{\xi^i}-\boldsymbol{\zeta}\|_2^2 \le \theta^2 \right\}. \tag{35}$$

Note that Γ is bounded as well. Therefore, (DR-SFM') is a maximin submodular-convex saddle-point problem over a bounded domain. To leverage existing machineries in continuous optimization, we want to guarantee Lipschitz-continuity on the convex part of f, for which, we impose the following mild assumption.

Assumption 1. *$f(\boldsymbol{x}, \cdot)$ for every $\boldsymbol{x} \in \mathcal{X}$ is bounded over Γ, in which case, $f(\boldsymbol{x}, \cdot)$ for every $\boldsymbol{x} \in \mathcal{X}$ is L_2-Lipschitz continuous in the ℓ_2 norm over Γ where $L_2 = 2 \sup_{(\boldsymbol{x}, \boldsymbol{\xi}) \in \mathcal{X} \times \Gamma} |f(\boldsymbol{x}, \boldsymbol{\xi})|$.*

Since $\mathcal{X}$ and Γ are bounded, the Lipschitz constant in Assumption 1 is bounded. Furthermore, we also impose an additional assumption to have some smooth structures on the submodular part of f.

Assumption 2. *There exist some constants L_1, λ_1, and λ_2 such that f satisfies the following structures.*

- *$f(\cdot, \boldsymbol{\xi})$ for every $\boldsymbol{\xi} \in \Gamma$ is differentiable and L_1-Lipschitz continuous with respect to a norm $\|\cdot\|$ over $\mathcal{X}$.*
- *$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x^1}, \boldsymbol{\xi}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x^2}, \boldsymbol{\xi})\|_* \leq \lambda_1 \|\boldsymbol{x^1} - \boldsymbol{x^2}\|$ for any $\boldsymbol{x^1}, \boldsymbol{x^2} \in \mathcal{X}$ and $\boldsymbol{\xi} \in \Gamma$.*
- *$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{y}, \boldsymbol{\xi^1}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{y}, \boldsymbol{\xi^2})\|_* \leq \lambda_2 \|\boldsymbol{\xi^1} - \boldsymbol{\xi^2}\|_2$ for any $\boldsymbol{y} \in \mathcal{X}$ and $\boldsymbol{\xi^1}, \boldsymbol{\xi^2} \in \Gamma$.*

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$.

We show in Appendix D that if F is the multilinear extension of some submodular-convex function, then F satisfies Assumption 2.

7.2 Preprocessing

We develop a conditional gradient method and a mirror prox method for solving (DR-SFM) based on the reformulation (DR-SFM'). However, there are a few issues when applying Algorithms 1 and 2 to the formulation $\sup_{\boldsymbol{x} \in \mathcal{X}} F(\boldsymbol{x})$. The first issue is that the function F may be non-differentiable, although $f(\cdot, \boldsymbol{\xi})$ is smooth for every $\boldsymbol{\xi} \in \Gamma$ by Assumption 2. The second issue is that we may not have an access to the gradient of F and obtaining even an approximate gradient can be difficult. To access the gradient or a subgradient of F at $\boldsymbol{x}$, we solve $\inf_{(\boldsymbol{\zeta^1}; \cdots; \boldsymbol{\zeta^N}) \in Z} \sum_{i \in [N]} p_i f(\boldsymbol{x}, \boldsymbol{\zeta^i})$, for which we obtain an approximate solution, not an exact optimal solution. Here, if $f(\boldsymbol{x}, \cdot)$ is not strongly convex, the distance between the approximate solution and an optimal solution can be large, which potentially incurs a large error in the gradient computation.

To remedy the challenges, we construct a function that has favorable structural properties, and at the same time, well approximates the original function F. We construct the following function by adding a regularization term to F.

$$\begin{aligned} H(\boldsymbol{x}) = & \inf_{(\boldsymbol{\zeta^1}; \cdots; \boldsymbol{\zeta^N}) \in Z} R(\boldsymbol{x}, \boldsymbol{\zeta^1}, \ldots, \boldsymbol{\zeta^N}) \\ & \text{where } R(\boldsymbol{x}, \boldsymbol{\zeta^1}, \ldots, \boldsymbol{\zeta^N}) = \sum_{i \in [N]} p_i \left(f(\boldsymbol{x}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2} \left\| \boldsymbol{\xi^i} - \boldsymbol{\zeta^i} \right\|_2^2 \right) \end{aligned} \tag{36}$$

where θ is the radius of the Wasserstein ball and ϵ is the accuracy parameter. Note that the function R for any fixed $\boldsymbol{x}$ is strongly convex thanks to the regularization term. With H defined as in (36), we consider the following approximation of (DR-SFM'):

$$\sup_{\boldsymbol{x} \in \mathcal{X}} H(\boldsymbol{x}). \tag{DR-SFM''}$$

In fact, solving (DR-SFM'') is equivalent to solving (DR-SFM) up to some additive error.

Lemma 7.2. *Let $\hat{\boldsymbol{x}}$ be an (α, ϵ')-approximate solution to (DR-SFM''). Then*

$$F(\hat{\boldsymbol{x}}) \geq \alpha \cdot OPT - \epsilon' - \frac{\epsilon}{2}$$

where OPT is the value of (DR-SFM)*, i.e., $\hat{\boldsymbol{x}}$ is an $(\alpha, \epsilon' + \epsilon/2)$-approximate solution to* (DR-SFM).

Proof. We first show that for any $\boldsymbol{x} \in \mathcal{X}$, $F(\boldsymbol{x}) \le H(\boldsymbol{x}) \le F(\boldsymbol{x}) + \epsilon/2$. Note that

$$\begin{aligned} H(\boldsymbol{x}) &= \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in Z} \sum_{i\in[N]} p_i \left(f(\boldsymbol{x}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2} \left\| \boldsymbol{\xi^i} - \boldsymbol{\zeta^i} \right\|_2^2 \right) \\ &\ge \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in Z} \sum_{i\in[N]} p_i f(\boldsymbol{x}, \boldsymbol{\zeta^i}) + \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in Z} \frac{\epsilon}{2\theta^2} \sum_{i\in[N]} p_i \left\| \boldsymbol{\xi^i} - \boldsymbol{\zeta^i} \right\|_2^2 \\ &= F(\boldsymbol{x}) + \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in Z} \frac{\epsilon}{2\theta^2} \sum_{i\in[N]} p_i \left\| \boldsymbol{\xi^i} - \boldsymbol{\zeta^i} \right\|_2^2. \end{aligned}$$

The last term of this inequality is at least $F(\boldsymbol{x})$, so $H(\boldsymbol{x}) \ge F(\boldsymbol{x})$, as required. To show $H(\boldsymbol{x}) \le F(\boldsymbol{x}) + \epsilon/2$, take $(\boldsymbol{\zeta^{1*}};\cdots;\boldsymbol{\zeta^{N*}}) \in Z$ such that $F(\boldsymbol{x}) = \sum_{i\in[N]} p_i f(\boldsymbol{x}, \boldsymbol{\zeta^{i*}})$. Then

$$\begin{aligned} H(\boldsymbol{x}) &= \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in Z} \sum_{i\in[N]} p_i \left(f(\boldsymbol{x}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2} \left\| \boldsymbol{\xi^i} - \boldsymbol{\zeta^i} \right\|_2^2 \right) \\ &\le \sum_{i\in[N]} p_i f(\boldsymbol{x}, \boldsymbol{\zeta^{i*}}) + \frac{\epsilon}{2\theta^2} \sum_{i\in[N]} p_i \left\| \boldsymbol{\xi^i} - \boldsymbol{\zeta^{i*}} \right\|_2^2 \\ &= F(\boldsymbol{x}) + \frac{\epsilon}{2\theta^2} \sum_{i\in[N]} p_i \left\| \boldsymbol{\xi^i} - \boldsymbol{\zeta^{i*}} \right\|_2^2 \\ &\le F(\boldsymbol{x}) + \frac{\epsilon}{2} \end{aligned}$$

where the first and second inequalities hold because $(\boldsymbol{\zeta^{1*}};\cdots;\boldsymbol{\zeta^{N*}}) \in Z$. Therefore, $H(\boldsymbol{x}) \le F(\boldsymbol{x}) + \epsilon/2$ holds, as required.

Let $\hat{\boldsymbol{x}}$ be an (α, ϵ')-approximate solution to (DR-SFM"). Let $\boldsymbol{x}^*$ be an optimal solution to (DR-SFM), i.e., $F(\boldsymbol{x}^*) = \text{OPT}$. Note that

$$F(\hat{\boldsymbol{x}}) \ge H(\hat{\boldsymbol{x}}) - \frac{\epsilon}{2} \ge \alpha \cdot \sup_{\boldsymbol{x}\in\mathcal{X}} H(\boldsymbol{x}) - \epsilon' - \frac{\epsilon}{2}$$

where the second inequality is from the assumption. Therefore, $\hat{\boldsymbol{x}}$ is an $(\alpha, \epsilon' + \epsilon/2)$-approximate solution to (DR-SFM). □

Next we prove several structural properties of H. We first show that H is Hölder-smooth.

Lemma 7.3. *H is differentiable, and for any $\boldsymbol{x^1}, \boldsymbol{x^2} \in \mathcal{X}$,*

$$\|\nabla H(\boldsymbol{x^1}) - \nabla H(\boldsymbol{x^2})\|_* \le \lambda_1 \|\boldsymbol{x^1} - \boldsymbol{x^2}\| + 2\lambda_2 \theta \sqrt{\frac{L_1}{\epsilon}} \|\boldsymbol{x^1} - \boldsymbol{x^2}\|^{1/2}.$$

We defer the proof of this lemma to Section 7.4

Due to the term $\|\boldsymbol{x^1} - \boldsymbol{x^2}\|^{1/2}$, the function H is not uniformly smooth. Nevertheless, as H satisfies (1) with $h(z) = \beta_1 z + \beta_2 z^{1/2}$ for some $\beta_1, \beta_2 > 0$, we may apply Algorithms 1 and 2.

Next, we show the strongly convex structure of H results in the following lemma whose proof is given in Appendix C.

Lemma 7.4. *Let $\boldsymbol{x} \in \mathcal{X}$, and let $(\hat{\boldsymbol{\zeta}}^{\mathbf{1}}; \cdots; \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) \in \mathrm{Z}$ be such that $R(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\mathbf{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) \leq H(\boldsymbol{x}) + \Delta$. Then*

$$\left\| \nabla H(\boldsymbol{x}) - \sum_{i \in [N]} p_i \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{i}}) \right\|_* \leq \lambda_2 \theta \sqrt{\frac{2\Delta}{\epsilon}}.$$

By Lemma 7.4, if we solve the inner minimization problem for H up to some small additive error Δ, then we obtain an approximation of the gradient of H. Lastly, we show that H preserves some concave structure present in DR-submodular functions, although H is not necessarily DR-submodular.

Lemma 7.5. *H is monotone and up-concave.*

In summary, H is nonnegative, monotone, up-concave, differentiable, and Hölder-smooth with respect to the norm $\|\cdot\|$, as required.

7.3 Algorithms for distributionally robust submodular maximization

In this section, we present Algorithms 4 and 5 for solving (DR-SFM), which are the applications of the conditional gradient method (Algorithm 1) and the mirror-prox method (Algorithm 2) to the distributionally robust submodular maximization, respectively.

Algorithm 4 Continuous greedy algorithm for (DR-SFM)

Initialize $\boldsymbol{x_0} \leftarrow \mathbf{0}$.
for $t = 1, \ldots, T$ **do**
  Find $(\hat{\boldsymbol{\zeta}}_{\boldsymbol{t}}^{\mathbf{1}}, \ldots, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t}}^{\boldsymbol{N}}) \in Z$ such that $R(\boldsymbol{x_t}, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t}}^{\mathbf{1}}, \ldots, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t}}^{\boldsymbol{N}}) - H(\boldsymbol{x_t}) \leq \Delta$.
  $\boldsymbol{v_t} = \mathrm{argmax}_{\boldsymbol{v} \in \mathcal{X}} \left\{ \sum_{i \in [N]} p_i \left\langle \nabla_{\boldsymbol{x}} f(\boldsymbol{x_t}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{i}}), \boldsymbol{v} \right\rangle \right\}$.
  Update $\boldsymbol{x_t} \leftarrow \left(1 - \frac{1}{t}\right) \boldsymbol{x_{t-1}} + \frac{1}{t} \boldsymbol{v_t} = \frac{1}{t}(\boldsymbol{v_1} + \cdots + \boldsymbol{v_t})$.
end for
Return $\boldsymbol{x_T} = \frac{1}{T}(\boldsymbol{v_1} + \cdots + \boldsymbol{v_T})$.

At each iteration, the algorithm solves the inner minimization problem, which is convex, and obtains an Δ-close solution. As H is strongly convex, we obtain an approximation of the gradient by Lemma 7.4. Then, based on Corollary 4.4, we deduce the following convergence result for Algorithm 4. Recall that H is a differentiable, monotone, and up-concave function that is Hölder-smooth with respect to a function $h : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$h(z) = \lambda_1 \|z\| + 2\lambda_2 \theta \sqrt{\frac{L_1}{\epsilon}} \|z\|^{1/2}. \tag{37}$$

Here, h has $\sigma = \{1, 1/2\} = 1/2$.

Theorem 7.6. *If $\Delta = O(\epsilon^3)$, Algorithm 4 returns an $(1 - 1/e, \epsilon)$-approximate solution to (DR-SFM) after $O(1/\epsilon^2)$ iterations.*

Next we present Algorithm 5 which is the application of our mirror-prox method to distributionally robust submodular maximization.

Then we deduce the following convergence result for Algorithm 5 based on Corollary 5.4.

Algorithm 5 Mirror-prox algorithm for (DR-SFM)

Initialize $\boldsymbol{x_1} \in \operatorname{argmin}_{x\in\mathcal{X}} \Phi(\boldsymbol{x})$.

for $t = 1, \ldots, T-1$ **do**

Find $(\hat{\boldsymbol{\zeta}}_{\boldsymbol{t}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t}}^{\boldsymbol{N}}) \in Z$ such that $R(\boldsymbol{x_t}, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t}}^{\boldsymbol{N}}) - H(\boldsymbol{x_t}) \le \Delta$.

$\boldsymbol{x_{t+1/2}} = \operatorname{Prox}_{\boldsymbol{x_t}}(-\gamma_t \sum_{i\in[N]} p_i \nabla_{\boldsymbol{x}} f(\boldsymbol{x_t}, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t}}^{\boldsymbol{i}}))$.

Find $(\hat{\boldsymbol{\zeta}}_{\boldsymbol{t+1/2}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t+1/2}}^{\boldsymbol{N}}) \in Z$ such that $R(\boldsymbol{x_{t+1/2}}, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t+1/2}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t+1/2}}^{\boldsymbol{N}}) - H(\boldsymbol{x_{t+1/2}}) \le \Delta$.

$\boldsymbol{x_{t+1}} = \operatorname{Prox}_{\boldsymbol{x_t}}(-\gamma_t \sum_{i\in[N]} p_i \nabla_{\boldsymbol{x}} f(\boldsymbol{x_{t+1/2}}, \hat{\boldsymbol{\zeta}}_{\boldsymbol{t+1/2}}^{\boldsymbol{i}}))$.

end for

Return $x^* \in \operatorname{argmax}\left\{H(\boldsymbol{x}) : \ \boldsymbol{x} \in \{\boldsymbol{x_{t+1/2}} : t = \lfloor (T-2)/3 \rfloor + 1, \ldots, T-1\}\right\}$.

Theorem 7.7. *If $\Delta = O(\epsilon^3)$, Algorithm 5 returns an $(1/2, \epsilon)$-approximate solution to (DR-SFM) after $O(1/\epsilon^{4/3})$ iterations.*

7.4 Proof of Lemma 7.3 on the Hölder-smoothness of H

First, to show that H is differentiable, we argue that $\inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})} R(\boldsymbol{x}, \boldsymbol{\zeta^1}, \ldots, \boldsymbol{\zeta^N})$ has a unique solution for any $\boldsymbol{x} \in \mathcal{X}$. This is essentially due to the strongly convex structure of R. To be more precise, we state the following lemma whose proof is given in Appendix C.

Lemma 7.8. *Let $\boldsymbol{x} \in \mathcal{X}$, and let $(\boldsymbol{\zeta^{1*}};\cdots;\boldsymbol{\zeta^{N*}}) \in \mathrm{Z}$ be such that $H(\boldsymbol{x}) = R(\boldsymbol{x}, \boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}})$. Then, for any $(\hat{\boldsymbol{\zeta}}^{\boldsymbol{1}};\cdots;\hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) \in \mathrm{Z}$, we have*

$$R(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - R(\boldsymbol{x}, \boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}}) \ge \frac{\epsilon}{2\theta^2} \sum_{i\in[N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} - \boldsymbol{\zeta^{i*}}\|_2^2. \tag{38}$$

By Lemma 7.8, if $H(\boldsymbol{x}) = R(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) = R(\boldsymbol{x}, \boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}})$, then $\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} = \boldsymbol{\zeta^{i*}}$ for all $i \in [N]$. Hence, H is differentiable because $f(\cdot, \boldsymbol{\xi})$ is differentiable for any $\boldsymbol{\xi} \in \Gamma$. Moreover,

$$\partial H(\boldsymbol{x}) = \left\{\nabla_{\boldsymbol{x}} R(\boldsymbol{x}, \boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}})\right\} = \left\{\sum_{i\in[N]} p_i \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{\zeta^{i*}})\right\}$$

where $(\boldsymbol{\zeta^{1*}};\cdots;\boldsymbol{\zeta^{N*}})$ is the unique solution satisfying $H(\boldsymbol{x}) = R(\boldsymbol{x}, \boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}})$.

Next, to prove that H satisfies the desired smooth property, we take $\hat{\boldsymbol{x}}, \tilde{\boldsymbol{x}} \in \mathcal{X}$. Then

$$H(\hat{\boldsymbol{x}}) = R(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}), \quad H(\tilde{\boldsymbol{x}}) = R(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{N}})$$

for some $(\hat{\boldsymbol{\zeta}}^{\boldsymbol{1}};\cdots;\hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}), (\tilde{\boldsymbol{\zeta}}^{\boldsymbol{1}};\cdots;\tilde{\boldsymbol{\zeta}}^{\boldsymbol{N}}) \in \mathrm{Z}$. Note that

$$\begin{aligned}
\|\nabla H(\hat{\boldsymbol{x}}) - \nabla H(\tilde{\boldsymbol{x}})\|_* &= \|\nabla_{\boldsymbol{x}} R(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - \nabla_{\boldsymbol{x}} R(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{N}})\|_* \\
&\le \|\nabla_{\boldsymbol{x}} R(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - \nabla_{\boldsymbol{x}} R(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}})\|_* \\
&\quad + \|\nabla_{\boldsymbol{x}} R(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - \nabla_{\boldsymbol{x}} R(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{N}})\|_* \\
&\le \lambda_1 \|\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\| + \lambda_2 \sum_{i\in[N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} - \tilde{\boldsymbol{\zeta}}^{\boldsymbol{i}}\|_2
\end{aligned} \tag{39}$$

where the second inequality is due to Assumption 2. Next, to bound $\|\nabla H(\hat{\boldsymbol{x}}) - \nabla H(\tilde{\boldsymbol{x}})\|_*$ based on (39), we consider the term $\sum_{i\in[N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} - \tilde{\boldsymbol{\zeta}}^{\boldsymbol{i}}\|_2$. By the Cauchy-Schwarz inequality and Lemma 7.8, we obtain the following.

$$\begin{aligned}\sum_{i\in[N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} - \tilde{\boldsymbol{\zeta}}^{\boldsymbol{i}}\|_2 &\leq (\sum_{i\in[N]} p_i)^{1/2} (\sum_{i\in[N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} - \tilde{\boldsymbol{\zeta}}^{\boldsymbol{i}}\|_2^2)^{1/2} \\ &\leq \sqrt{\frac{2\theta^2}{\epsilon}} (R(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - R(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{N}}))^{1/2}.\end{aligned} \tag{40}$$

Moreover, we need the following lemma whose proof is given in Appendix C.

Lemma 7.9. *H is L_1-Lipschitz continuous in the norm $\|\cdot\|$ over $\mathcal{X}$.*

Then it follows that

$$\begin{aligned}&R(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - R(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \tilde{\boldsymbol{\zeta}}^{\boldsymbol{N}}) \\ &\leq |R(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - R(\hat{\boldsymbol{x}}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}})| + |H(\hat{\boldsymbol{x}}) - H(\tilde{\boldsymbol{x}})| \\ &\leq 2L_1 \|\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\|\end{aligned} \tag{41}$$

holds because of Assumption 2 and Lemma 7.9. Then it follows from (39), (40), and (41) that

$$\|\nabla H(\hat{\boldsymbol{x}}) - \nabla H(\tilde{\boldsymbol{x}})\|_* \leq \lambda_1 \|\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\| + 2\lambda_2 \theta \sqrt{\frac{L_1}{\epsilon}} \|\hat{\boldsymbol{x}} - \tilde{\boldsymbol{x}}\|^{1/2},$$

as required.

8 Numerical experiments

In this section, we present our experimental results to test the numerical performance of our algorithmic framework for the problem of maximizing a continuous DR-submodular function that is non-smooth or Hölder-smooth. We evaluated the efficacy of our algorithms, the continuous greedy method (Algorithm 1) and the mirror-prox method (Algorithm 2), on two sets of test instances: (1) the multi-resolution data summary problem with a non-differentiable utility function and (2) the distributionally robust movie recommendation problem. We used instances generated by synthetic simulated data for the multi-resolution data summary problem, which is described in Section 8.1. For the distributionally robust movie recommendation problem, we used the MovieLens 1M Dataset [16] to obtain the users' movie rating data. We explain the experimental setup for the multi-resolution data summary problem with a non-differentiable utility function in Section 8.1 and that for the distributionally robust movie recommendation problem in Section 8.2. We report and summarize numerical results in Section 8.3.

8.1 Multi-resolution data summarization

Our first experiment is on the multi-resolution summary problem [5] with a non-differentiable utility function. Given a collection of data or items $E = \{e_1, \ldots, e_k\}$, we assign each item e_i a nonnnegative score x_i to measure its importance, by which we may recommend a subset of items. We set a threshold τ so that we report the set $S_\tau = \{e_i : x_i \geq \tau\}$ of items whose scores exceed the given threshold. By adjusting the value of τ, we can decide

the level of details or resolution of the summary. The scores of items basically represent the relative importance of items, and they are determined so that a utility function is maximized. The utility function is given by

$$F(\boldsymbol{x}) = \sum_{i=1}^{k}\sum_{j=1}^{k}\phi(x_j)s_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{k}x_i x_j s_{ij}$$

where $s_{ij} \geq 0$ is the similarity index between two items e_i and e_j. The first sum consists of the terms $\phi(x_j)s_{ij}$, which captures how much item e_j contributes to item e_i when e_j has weight x_j. Here, Φ is a monotone concave function to model the diminishing returns property. The second sum consists of terms $x_i x_j s_{ij}$, which has a high value when two similar items e_i and e_j with a high similarity index s_{ij} get large weights at the same time. Hence, taking away the second sum from the first encourages that if two items are similar, at most one of them gets a large score.

Note that F is up-concave because the first sum is a concave function in $\boldsymbol{x}$ and $-\sum_{i=1}^{k}\sum_{j=1}^{k}x_i x_j s_{ij}$ is DR-submodular with respect to $\boldsymbol{x}$. We consider the case when ϕ is the following piece-wise linear function defined on $[0, 1]$.

$$\phi(x) = \begin{cases} 7x & \text{if } x \in [0, \frac{1}{2}], \\ 6x + \frac{1}{2} & \text{if } x \in [\frac{1}{2}, \frac{3}{4}], \\ 5x + \frac{5}{4} & \text{if } x \in [\frac{3}{4}, 1]. \end{cases}$$

In our experiments, we used randomly generated instances with $k = 50$ items and similarity indices s_{ij} sampled from the uniform distribution on $[0, 1]$. For each instance, we determined a score vector by solving $\max_{\boldsymbol{x}\in\mathcal{P}} F(\boldsymbol{x})$ where $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^{\cdot}0 \leq x_i \leq 1, \sum_{i=1}^{k} x_i = 5\}$. Note that the objective F is not differentiable as ϕ is not differentiable. We tested the mirror-prox method, which works for non-differentiable objective functions, with the step size set to $\gamma_t = 1/(2\sqrt{T})$. Although we do not have a theoretical ground for the continuous greedy method for the case of non-differentiable functions, we also ran the method to test its empirical effectiveness. We tested 30 randomly generated instances, and for each of the instances, we ran the algorithms for 50 iterations.

8.2 Distributionally robust movie recommendation

Our second problem is the distributionally robust formulation of the movie recommendation problem considered in [31, 38]. We have a collection of n users, labelled by $\{1, \ldots, n\}$, and m movies. Each user i has a submodular set function $f(S, i)$ for evaluating a set S of movies. We consider the setting where $f(S, i)$ is given by

$$f(S, i) = \max_{j\in S} r_{ij}$$

where r_{ij} is user i's preference index on movie j. Basically, $f(S, i)$ takes the maximum preference value of user i among the movies in S. Here, we treat an user's random preference indices over the m movies as a sample $\boldsymbol{\xi} \in \mathbb{R}^m$. Hence, the sample space Ξ consists of n samples that correspond to the n users. Moreover, we can rewrite $f(S, i)$ as $f(S, \boldsymbol{\xi})$. Let P^* be the distribution of the preference index vectors of the entire population of the n users. Then we can recommend a set of k movies for the population of n users based on solving

$$\max_{|S|=k} \mathbb{E}_{\boldsymbol{\xi}\sim P^*}\left[f(S, \boldsymbol{\xi})\right].$$

When P^* is unknown to us, we may obtain a few sample preference vectors from some subset of users. Let Q be the distribution of the preference vectors among the subset of users. Then we may obtain a list of movies to recommend by solving the following distributionally robust formulation.

$$\max_{|S|=k} \min_{P \in \mathcal{B}(Q,\theta)} \mathbb{E}_{\boldsymbol{\xi} \sim P} \left[f(S, \boldsymbol{\xi}) \right]$$

where $\mathcal{B}(Q, \theta)$ is the Wasserstein ambiguity set of radius θ around the nominal distribution Q. We can solve the discrete problem by solving its continuous relaxation obtained by the multilinear extension and applying a proper rounding scheme, e.g., [2, 45], where the multilinear extension of $f(\cdot, \boldsymbol{\xi})$ is given by

$$F_{\boldsymbol{\xi}}(\boldsymbol{x}) = \sum_{S \subseteq [m]} f(S, \boldsymbol{\xi}) \prod_{i \in S} x_i \prod_{i \notin S} (1 - x_i).$$

For our experiments, we used the MovieLens 1M Dataset from [16] that consists of $n = 6041$ users and $m = 4000$ movies. For each instance, we choose $N = 10$ samples from the $n = 6041$ data uniformly at random and consider the empirical distribution Q on the samples. Then we consider the continuous relaxation

$$\max_{\boldsymbol{x} \in \mathcal{X}} \min_{P \in \mathcal{B}(Q,\theta)} \mathbb{E}_{\boldsymbol{\xi} \sim P} \left[F_{\boldsymbol{\xi}}(\boldsymbol{x}) \right]$$

where $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^m : \sum_{i=1}^m x_i = 5\}$. We reformulate this based on our framework described in Section 7, which is a submodular-convex maximin problem. Lastly, we set the Wasserstein radius θ to $\theta = 0.2$. We observed that each rating value r_{ij} is from 1 to 5 in the dataset, in which case, Assumptions 1 and 2 are satisfied with $L_1 = \lambda_1 = L_2 = 5$ and $\lambda_2 = 10$. Lastly, to determine the accuracy parameter for inner minimization, we set $\epsilon = 0.01$.

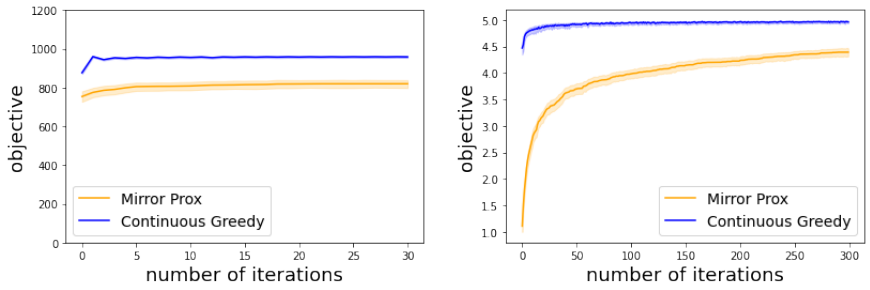
For each iteration of the continuous greedy algorithm (Algorithm 4) and the mirror prox method (Algorithm 5), we need to be able to compute the gradient $\nabla_{\boldsymbol{x}} F_{\boldsymbol{\xi}}(\boldsymbol{x})$ of the multilinear extension. However, as there are exponentially many movies, computing the exact gradient is expensive. Instead, we obtained an unbiased estimator of the gradient $\nabla_{\boldsymbol{x}} F_{\boldsymbol{\xi}}(\boldsymbol{x})$ based on the approach described in [31, Section 9.3].

Moreover, at each iteration, we need to solve the inner minimization problem. For that, we used the projected gradient descent method where the projection step is made onto $\mathcal{X}$ that is the base polytope of a uniform matroid. An issue with this approach is that it may be expensive to compute the values of R, H, and ∇R (see Section 7), because they involve exponentially many multilinear terms as in the multilinear extension. To remedy this, we used the standard sampling approach to obtain an unbiased estimator of each of R, H, and ∇R. To elaborate, note that $F_{\boldsymbol{\xi}}(\boldsymbol{x}) = \mathbb{E}_{S \sim \boldsymbol{x}} \left[f(S, \boldsymbol{\xi}) \right]$ where $S \sim \boldsymbol{x}$ means that a set S can be sampled by picking each item $i \in [m]$ with probability x_i. Then obtaining sets $S_1, \ldots, S_B$ where $B = 10$ and taking $(1/B) \sum_{i=1}^B f(S_i, \boldsymbol{\xi})$, we obtain an unbiased estimator of $F_{\boldsymbol{\xi}}(\boldsymbol{x})$. Similarly, we obtain an unbiased estimator of $\nabla_{\boldsymbol{\xi}} F_{\boldsymbol{\xi}}(\boldsymbol{x})$.

For each problem instance, we ran both the conditional gradient descent algorithm and the mirror-prox algorithm for 300 iterations.

8.3 Experimental results

Figure 1(a) shows the numerical results for the multi-resolution summarization problem. We may observe that both the continuous greedy method and the mirror-prox algorithm converge after a few iterations. However, the



(a) (b)

Figure 1: Results of (left) the multi-resolution summary problem and (right) the distributionally robust movie recommendation.

continuous greedy algorithm converges to a higher objective value than the mirror-prox algorithm. This is interesting because the objective function for the multi-resolution summary problem is non-differentiable, in which case, the continuous greedy algorithm does not necessarily converge. A possible explanation would be that although the objective is non-differentiable, it is differentiable almost everywhere, and in particular, the function might have been differentiable at every iteration of the continuous greedy algorithm. The mirror-prox algorithm converges to a value that is about 20% less than the value to which the continuous greedy algorithm, and as our convergence result suggests, the value should be at least $1/2$ times the optimal value.

Figure 1(b) shows the experimental results from the distributionally robust movie recommendation problem. We observe that the continuous greedy algorithm converges fast after less than 10 iterations. In contrast, the mirror-prox method exhibits a slower convergence pattern. Moreover, the continuous greedy algorithm converges to a higher value than the mirror-prox method, as expected from our theoretical results.

Acknowledgements This research is supported, in part, by the Institute for Basic Science (IBS-R029-C1, Y2).

References

[1] Arman Adibi, Aryan Mokhtari, and Hamed Hassani. Minimax optimization: The case of convex-submodular. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 3556–3580. PMLR, 28–30 Mar 2022.

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

[3] Nima Anari, Nika Haghtalab, Seffi Naor, Sebastian Pokutta, Mohit Singh, and Alfredo Torrico. Structured robust submodular maximization: Offline and online algorithms. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 3128–3137. PMLR, 16–18 Apr 2019.

[4] Fancis Bach. Submodular functions: from discrete to continuous domains. *Mathematical Programming*, 175: 419–459, 2019.

[5] Andrew An Bian, Baharan Mirzasoleiman, Joachim Buhmann, and Andreas Krause. Guaranteed Non-convex Optimization: Submodular Maximization over Domains. In Aarti Singh and Jerry Zhu, editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54 of *Proceedings of Machine Learning Research*, pages 111–120, 20–22 Apr 2017.

[6] Yatao Bian, Joachim M. Buhmann, and Andreas Krause. Continuous submodular function maximization. *arXiv preprint arXiv:2006.13474*, 2020.

[7] Jose Blanchet and Karthyek Murthy. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019. doi: 10.1287/moor.2018.0936.

[8] Gruia Calinescu, Chandra Chekuri, Martin Pál, and Jan Vondrák. Maximizing a monotone submodular function subject to a matroid constraint. 40(6):783–792, 2011.

[9] Robert Chen, Brendan Lucier, Yaron Singer, and Vasilis Syrgkanis. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, page 4708–4717, 2017.

[10] Wei Chen, Tian Lin, Zihan Tan, Mingfei Zhao, and Xuren Zhou. Robust influence maximization. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 795–804, 2016.

[11] Abhimanyu Das and David Kempe. Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection. In *International Conference on International Conference on Machine Learning (ICML)*, pages 1057–1064, 2011.

[12] John C. Duchi, Peter L. Bartlett, and Martin J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

[13] Moran Feldman, Karbasi Karbasi, and Ehsan Kazemi. Do less, get more: Streaming submodular maximization with subsampling. In *Advances in Neural Information Processing Systems*, pages 730–740, 2018.

[14] Shayan Oveis Gharan and Jan Vondrák. Submodular maximization by simulated annealing. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '11, page 1098–1116, 2011.

[15] Carlos Guestrin, Andreas Krause, and Ajit Paul Singh. Near-optimal sensor placements in gaussian processes. In *International Conference on Machine Learning (ICML)*, pages 265–272, 2005.

[16] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015. URL https://doi.org/10.1145/2827872.

[17] Hamed Hassani, Mahdi Soltanolkotabi, and Amin Karbasi. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems*, page 5843–5853, 2017.

[18] Xinran He and David Kempe. Robust influence maximization. In *Proceedings of the 22nd ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, KDD '16, page 885–894, 2016.

[19] Amin Karbasi, Hamed Hassani, Aryan Mokhtari, and Zebang Shen. Stochastic continuous greedy ++: When upper and lower bounds match. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/456048afb7253926e1fbb7486e699180-Paper.pdf.

[20] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 137–146, 2003.

[21] Andreas Krause and Carlos Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 324–331, 2005.

[22] Andreas Krause, H. Brendan McMahan, Carlos Guestrin, and Anupam Gupta. Robust submodular observation selection. *Journal of Machine Learning Research*, 9(93):2761–2801, 2008. URL http://jmlr.org/papers/v9/krause08b.html.

[23] Ariel Kulik, Hadas Shachnai, and Tami Tamir. Maximizing submodular set functions subject to multiple linear constraints. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '09, page 545–554, 2009.

[24] Hui Lin and Jeff Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT)*, pages 912–920, 2010.

[25] Hui Lin and Jeff Bilmes. A class of submodular functions for document summarization. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT)*, pages 510–520, 2011.

[26] Meghna Lowalekar, Pradeep Varakantham, and Akshat Kumar. Robust influence maximization: (extended abstract). In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, AAMAS '16, page 1395–1396, 2016.

[27] Baharan Mirzasoleiman, Amin Karbasi, Rik Sarkar, and Andreas Krause. Distributed submodular maximization. *Journal of Machine Learning Research*, 17(1):1–44, 2016.

[28] Baharan Mirzasoleiman, Stefanie Jegelka, and Andreas Krause. Streaming non-monotone submodular maximization: Personalized video summarization on the fly. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1379–1386, 2018.

[29] Siddharth Mitra, Moran Feldman, and Amin Karbasi. Submodular + concave. In *Advances in Neural Information Processing Systems*, volume 34, pages 11577–11591, 2021.

[30] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*, 171

(1):115–166, 2018.

[31] Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. Conditional gradient method for stochastic submodular maximization: Closing the gap. In Amos Storkey and Fernando Perez-Cruz, editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 84 of *Proceedings of Machine Learning Research*, pages 1886–1895, 09–11 Apr 2018.

[32] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions - i. *Mathematical Programming*, 14(1):265–294, 1978.

[33] Arkadi 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.

[34] Tasuku Soma and Yuichi Yoshida. A generalization of submodular cover via the diminishing return property on the integer lattice. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28. Curran Associates, Inc., 2015.

[35] Tasuku Soma, Naonori Kakimura, Kazuhiro Inaba, and Ken-ichi Kawarabayashi. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In Eric P. Xing and Tony Jebara, editors, *International Conference on Machine Learning (ICML)*, volume 32 of *Proceedings of Machine Learning Research*, pages 351–359, Bejing, China, 22–24 Jun 2014. PMLR.

[36] Matthew Staib and Stefanie Jegelka. Robust budget allocation via continuous submodular functions. In Doina Precup and Yee Whye Teh, editors, *International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 3230–3240, 06–11 Aug 2017.

[37] Matthew Staib, Bryan Wilder, and Stefanie Jegelka. Distributionally robust submodular maximization. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *Proceedings of Machine Learning Research*, pages 506–516, 16–18 Apr 2019.

[38] Serban Stan, Morteza Zadimoghaddam, Andreas Krause, and Amin Karbasi. Probabilistic submodular maximization in sub-linear time. In Doina Precup and Yee Whye 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 2017.

[39] Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10(4):1040 – 1053, 1982.

[40] Fedor Stonyakin, Alexander Gasnikov, Pavel Dvurechensky, Alexander Titov, and Mohammad Alkousa. Generalized mirror prox algorithm for monotone variational inequalities: Universality and inexact oracle. *Journal of Optimization Theory and Applications*, 194(3):988–1013, 2022.

[41] Maxim Sviridenko, Jan Vondrák, and Justin Ward. Optimal approximation for submodular and supermodular optimization with bounded curvature. *Mathematics of Operations Research*, 42(4):1197–1218, 2017.

[42] Alfredo Torrico, Mohit Singh, Sebastian Pokutta, Nika Haghtalab, Joseph (Seffi) Naor, and Nima Anari. Structured robust submodular maximization: Offline and online algorithms. *INFORMS Journal on Computing*, 33(4):1590–1607, 2021.

[43] Sebastian Tschiatschek, Rishabh Iyer, Haochen Wei, and Jeff Bilmes. Learning mixtures of submodular functions for image collection summarization. In *Advances in Neural Information Processing Systems*, pages 1413–1421, 2014.

[44] Jan Vondrak. Optimal approximation for the submodular welfare problem in the value oracle model. In *ACM Symposium on Theory of Computing (STOC)*, page 67–74, 2008.

[45] Jan Vondrák, Chandra Chekuri, and Rico Zenklusen. Submodular function maximization via the multilinear relaxation and contention resolution schemes. In *ACM Symposium on Theory of Computing (STOC)*, pages 783–792, 2011.

[46] Bryan Wilder. Equilibrium computation and robust optimization in zero sum games with submodular structure. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1274–1281, 2018.

[47] Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic frank-wolfe. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4012–4023. PMLR, 26–28 Aug 2020.

A Proofs for results in Section 5 on the mirror prox algorithm

A.1 Proof of Lemma 5.1: The up-super-differential reduces to the gradient for differentiable functions

Take $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{g} \in \partial^\uparrow F(\boldsymbol{x})$. Let $j \in [d]$, and let $\boldsymbol{e_j}$ be the jth standard basis vector, i.e., $(\boldsymbol{e_j})_\ell = 1$ if $\ell = j$ and $(\boldsymbol{e_j})_\ell = 0$ if $\ell \neq j$. Since F is up-concave, we have $F(\boldsymbol{x} + t\boldsymbol{e_j}) - F(\boldsymbol{x}) \leq \boldsymbol{g}^\top (t\boldsymbol{e_j})$ for any $t \in \mathbb{R}_+$. Then it follows that

$$\boldsymbol{e_j}^\top \nabla F(\boldsymbol{x}) = \lim_{t \to 0+} \frac{F(\boldsymbol{x} + t\boldsymbol{e_j}) - F(\boldsymbol{x})}{t} \leq \boldsymbol{e_j}^\top \boldsymbol{g},$$

and therefore $\boldsymbol{e_j}^\top (\nabla F(\boldsymbol{x}) - \boldsymbol{g}) \leq 0$. Similarly, we have $F(\boldsymbol{x} - t\boldsymbol{e_j}) - F(\boldsymbol{x}) \leq \boldsymbol{g}^\top (-t\boldsymbol{e_j})$ for any $t \in \mathbb{R}_+$, which implies that

$$-\boldsymbol{e_j}^\top \nabla F(\boldsymbol{x}) = \lim_{t \to 0+} \frac{F(\boldsymbol{x} - t\boldsymbol{e_j}) - F(\boldsymbol{x})}{t} \leq -\boldsymbol{e_j}^\top \boldsymbol{g}.$$

Thus we obtain $\boldsymbol{e_j}^\top (\nabla F(\boldsymbol{x}) - \boldsymbol{g}) \geq 0$. We have just proved that $\boldsymbol{e_j}^\top (\nabla F(\boldsymbol{x}) - \boldsymbol{g}) = 0$ for every $j \in [d]$, implying in turn that $\boldsymbol{g} = \nabla F(\boldsymbol{x})$. Therefore, $\partial^\uparrow F(\boldsymbol{x}) = \{\nabla F(\boldsymbol{x})\}$ for $i \in [n]$.

A.2 Proof of Lemma 5.5

As $\|\boldsymbol{x_t} - \boldsymbol{x_{t+1/2}}\| \geq 0$, the left-hand side of (27) is bounded above by $\sup_{d \geq 0} f_t(d)$ where f_t is defined as

$$f_t(d) = 2\gamma_t^2 \left(\sum_{i=1}^k \beta_i d^{\sigma_i} \right)^2 - \frac{1}{2} d^2 = \frac{1}{2} \left(\frac{1}{t^{1-\sigma}} \left(\frac{\sum_{i=1}^k \beta_i d^{\sigma_i}}{\sum_{i=1}^k \beta_i} \right)^2 - d^2 \right).$$

Recall that $0 \leq \sigma_1, \ldots, \sigma_k \leq 1$. Assuming that $\sigma_1 \leq \cdots \leq \sigma_k$, we have $\sigma = \sigma_1$, $f_t(d) \leq f_{t,1}(d)$ for $0 \leq d \leq 1$ and $f_t(d) \leq f_{t,2}(d)$ for $d \geq 1$ where

$$f_{t,1}(d) = \frac{1}{2} \left(\frac{1}{t^{1-\sigma}} d^{2\sigma_1} - d^2 \right), \quad f_{t,2}(d) = \frac{1}{2} \left(\frac{1}{t^{1-\sigma}} d^{2\sigma_k} - d^2 \right).$$

Then $f_t(d) \leq \sup_{d \geq 0} f_{t,1}(d)$ for $0 \leq d \leq 1$ and $f_t(d) \leq \sup_{d \geq 0} f_{t,2}(d)$ for $d \geq 1$. This implies that $f_t(d) \leq \max\left\{\sup_{d \geq 0} f_{t,1}(d), \sup_{d \geq 0} f_{t,2}(d)\right\}$, taking the supremum of the left-hand side of which over $d \geq 0$, we obtain

$$\sup_{d \geq 0} f_t(d) \leq \max\left\{ \sup_{d \geq 0} f_{t,1}(d), \sup_{d \geq 0} f_{t,2}(d) \right\}.$$

To show that (27) holds, it suffices to argue that

$$\sup_{d \geq 0} f_{t,1}(d) \leq 1/t \quad \text{and} \quad \sup_{d \geq 0} f_{t,2}(d) \leq 1/t.$$

If $\sigma_1 = 1$, as $t^{1-\sigma} \geq 1$, we have $f_{t,1}(d) \leq 0$ for all $d \geq 0$. Likewise, if $\sigma_k = 1$, then $f_{t,2}(d) \leq 0$ for any $d \geq 0$. Thus, we may assume that $\sigma_1, \sigma_k < 1$. Moreover, as $\sigma = \min_{i \in [k]} \sigma_i$, we have $\sigma \leq \sigma_1, \sigma_k < 1$.

Let $\sigma' \in [\sigma, 1)$ and $g_t(d)$ be defined as

$$g_t(d) = \frac{1}{2} \left(\frac{1}{t^{1-\sigma}} d^{2\sigma'} - d^2 \right).$$

Then it is sufficient to argue that $\sup_{d\geq 0} g_t(d) \leq 1/t$. The derivative of $g_t(d)$ is given by

$$g_t'(d) = \frac{\sigma'}{t^{1-\sigma}} d^{2\sigma'-1} - d = d\left(\frac{\sigma'}{t^{1-\sigma}} d^{2(\sigma'-1)} - 1\right).$$

Note that $g_t'(d) \geq 0$ when $0 \leq d \leq \left(t^{1-\sigma}/\sigma'\right)^{1/2(\sigma'-1)}$ and $g_t'(d) \leq 0$ when $d \geq \left(t^{1-\sigma}/\sigma'\right)^{1/2(\sigma'-1)}$. As a consequence, the maximum of g_t over $d \geq 0$ is attained at $d = \left(t^{1-\sigma}/\sigma'\right)^{1/2(\sigma'-1)}$. Then it follows that

$$\sup_{d\geq 0} g_t(d) = \frac{1}{2t^{\frac{1-\sigma}{1-\sigma'}}} \left((\sigma')^{\frac{\sigma'}{1-\sigma'}} - (\sigma')^{\frac{1}{1-\sigma'}}\right) \leq \frac{1}{2t}$$

where the inequality is because $t \geq 1$, $\sigma' < 1$, and $(1-\sigma)/(1-\sigma') \geq 1$.

A.3 Proof of Lemma 5.6

Observe that

$$\sum_{t=T_0}^{T-1} \gamma_t = \frac{1}{2\sum_{i=1}^k \beta_i} \sum_{t=T_0}^{T-1} t^{\frac{-1+\sigma}{2}}. \tag{42}$$

To bound the sum $\sum_{t=T_0}^{T-1} \gamma_t$, we consider $\sum_{t=T_0}^{T-1} t^{\frac{-1+\sigma}{2}}$. Since $1-\sigma \geq 0$, we know that $1/t^{\frac{1-\sigma}{2}}$ is a decreasing function. This implies that

$$\int_{T_0}^{T} t^{\frac{-1+\sigma}{2}} dt \leq \sum_{t=T_0}^{T-1} t^{\frac{-1+\sigma}{2}} \leq \int_{T_0-1}^{T-1} t^{\frac{-1+\sigma}{2}} dt. \tag{43}$$

Note that $T_0 \leq (T+1)/3$ and $(T+1)/3 \leq 2T/3$. Then

$$\begin{aligned}\int_{T_0}^{T} t^{\frac{-1+\sigma}{2}} dt &= \frac{2}{1+\sigma}\left(T^{\frac{1+\sigma}{2}} - T_0^{\frac{1+\sigma}{2}}\right) \geq \frac{2}{1+\sigma}\left(T^{\frac{1+\sigma}{2}} - \left(\frac{2T}{3}\right)^{\frac{1+\sigma}{2}}\right) \\ &\geq (1-\sqrt{2/3})T^{\frac{1+\sigma}{2}} \geq \frac{T^{\frac{1+\sigma}{2}}}{6}\end{aligned}$$

where the first inequality is from $T_0 \leq 2T/3$, the second inequality holds because $0 \leq \sigma \leq 1$, and the third inequality is due to $1 - \sqrt{2/3} \geq 1/6$. Moreover,

$$\int_{T_0-1}^{T-1} t^{\frac{-1+\sigma}{2}} dt = \frac{2}{1+\sigma}\left((T-1)^{\frac{1+\sigma}{2}} - (T_0-1)^{\frac{1+\sigma}{2}}\right) \leq 2T^{\frac{1+\sigma}{2}} \tag{44}$$

where the inequality holds because $\sigma \geq 0$.

B Proofs for the results in Section 6 on robust submodular maximization

B.1 Proof of Lemma 6.1

We first argue that F is up-concave. Let $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{v} \in \mathbb{R}^d_+$. To show that $F(\boldsymbol{x} + \boldsymbol{v}t)$ is concave with respect to t for each $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{v} \geq 0$, take $t_1, t_2 \geq 0$ and $\lambda \in [0, 1]$. Note that

$$\begin{aligned}
F(\boldsymbol{x} + (\lambda t_1 + (1-\lambda)t_2)\boldsymbol{v}) &= \min_{i \in [n]} F_i(\boldsymbol{x} + (\lambda t_1 + (1-\lambda)t_2)\boldsymbol{v}) \\
&\geq \min_{i \in [n]} \{\lambda F_i(\boldsymbol{x} + t_1\boldsymbol{v}) + (1-\lambda)F_i(\boldsymbol{x} + t_2\boldsymbol{v})\} \\
&\geq \lambda \min_{i \in [n]} \{F_i(\boldsymbol{x} + t_1\boldsymbol{v})\} + (1-\lambda) \min_{i \in [n]} \{F_i(\boldsymbol{x} + t_2\boldsymbol{v})\} \\
&= \lambda F(\boldsymbol{x} + t_1\boldsymbol{v}) + (1-\lambda)F(\boldsymbol{x} + t_2\boldsymbol{v})
\end{aligned}$$

where the first inequality is because each F_i is up-concave. Hence, F is up-concave.

Next we show that F is monotone, let us take $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ such that $\boldsymbol{x} \leq \boldsymbol{y}$. Since F_i is monotone for each $i \in [n]$, we have $F_i(\boldsymbol{x}) \leq F_i(\boldsymbol{y})$. Taking the minimum of the left-hand side of this inequality over $i \in [n]$, it follows that $\min_{i \in [n]} F_i(\boldsymbol{x}) \leq F_i(\boldsymbol{y})$, implying in turn that $F(\boldsymbol{x}) \leq F_i(\boldsymbol{y})$. Then taking the minimum of the right-hand side, we obtain $F(\boldsymbol{x}) \leq \min_{i \in [n]} F_i(\boldsymbol{y}) = F(\boldsymbol{y})$. Thus F is monotone.

Lastly, it is straightforward that F is nonnegative as all of $F_1, \ldots, F_n$ are nonnegative, as required.

B.2 Proof of Lemma 6.2

Since $F_1, \ldots, F_n$ are differentiable, $\partial^\uparrow F_i(\boldsymbol{x}) = \{\nabla F_i(\boldsymbol{x})\}$ for $i \in [n]$ by Lemma 5.1. Then we argue that $\operatorname{conv}\{\nabla F_i(\boldsymbol{x}) : i \in \arg\min_{i \in [n]} F_i(\boldsymbol{x})\} \subseteq \partial^\uparrow F(\boldsymbol{x})$. Let $i \in \arg\min_{i \in [n]} F_i(\boldsymbol{x})$. Then

$$F(\boldsymbol{y}) \leq F_i(\boldsymbol{y}) \leq F_i(\boldsymbol{x}) + \nabla F_i(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) = F(\boldsymbol{x}) + \nabla F_i(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}),$$

implying in turn that $\nabla F_i(\boldsymbol{x}) \in \partial^\uparrow F(\boldsymbol{x})$. Furthermore, as $F(\boldsymbol{y}) - F(\boldsymbol{x}) \leq \nabla F_i(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x})$ for any $i \in \arg\min_{i \in [n]} F_i(\boldsymbol{x})$, it follows that $F(\boldsymbol{y}) - F(\boldsymbol{x}) \leq \boldsymbol{g}^\top (\boldsymbol{y} - \boldsymbol{x})$ for any $\boldsymbol{g} \in \operatorname{conv}\{\nabla F_i(\boldsymbol{x}) : i \in \arg\min_{i \in [n]} F_i(\boldsymbol{x})\}$.

B.3 Proof of Lemma 6.3

Let us first argue that F is L-Lipschitz continuous with respect to the norm $\|\cdot\|$. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ and $i \in [n]$, we have

$$F_i(\boldsymbol{x}) = F_i(\boldsymbol{x}) - F_i(\boldsymbol{y}) + F_i(\boldsymbol{y}) \leq L\|\boldsymbol{x} - \boldsymbol{y}\| + F_i(\boldsymbol{y}).$$

Therefore, we have $F_i(\boldsymbol{x}) \leq L\|\boldsymbol{x} - \boldsymbol{y}\| + F_i(\boldsymbol{y})$. Taking the minimum of the left-hand side over $i \in [n]$, we deduce that $F(\boldsymbol{x}) \leq L\|\boldsymbol{x} - \boldsymbol{y}\| + F_i(\boldsymbol{y})$. Then we take the minimum of its right-hand side over $i \in [n]$, which results in $F(\boldsymbol{x}) \leq L\|\boldsymbol{x} - \boldsymbol{y}\| + F(\boldsymbol{y})$. Similarly, we can show that $F(\boldsymbol{y}) \leq L\|\boldsymbol{y} - \boldsymbol{x}\| + F(\boldsymbol{x})$. Therefore, it follows that $|F(\boldsymbol{x}) - F(\boldsymbol{y})| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$, and therefore, F is L-Lipschitz continuous in the norm $\|\cdot\|$.

Let $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{g} \in \operatorname{conv}\{\nabla F_i(\boldsymbol{x}) : i \in \arg\min_{i \in [n]} F_i(\boldsymbol{x})\}$. Then for some $i_1, \ldots, i_k \in [n]$, $\boldsymbol{g}$ is a convex combination of $\nabla F_{i_1}(\boldsymbol{x}), \ldots, \nabla F_{i_k}(\boldsymbol{x})$. Hence, $\boldsymbol{g} = \sum_{\ell=1}^k \lambda_\ell \nabla F_{i_\ell}(\boldsymbol{x})$ for some $\lambda_1, \ldots, \lambda_k \geq 0$ that add up to 1.

Therefore,

$$\|\boldsymbol{g}\|_* \leq \sum_{\ell=1}^{k} \lambda_\ell \|\nabla F_{i_\ell}(\boldsymbol{x})\|_* \leq L$$

where the first and second inequalities are the triangle inequality and the absolute homogeneity of the dual norm $\|\cdot\|_*$ and the last inequality is because $F_1, \ldots, F_n$ are differentiable and L-Lipschitz continuous. Therefore, $\|\boldsymbol{g}\|_* \leq L$.

C Proofs for results in Section 7

C.1 Proof of Lemma 7.8

For $(\boldsymbol{\zeta^1}; \cdots; \boldsymbol{\zeta^N}) \in \mathrm{Z}$, let $R'(\boldsymbol{\zeta^1}, \ldots, \boldsymbol{\zeta^N})$ be defined as

$$R'(\boldsymbol{x}, \boldsymbol{\zeta^1}, \ldots, \boldsymbol{\zeta^N}) := R(\boldsymbol{x}, \boldsymbol{\zeta^1}, \ldots, \boldsymbol{\zeta^N}) - \frac{\epsilon}{2\theta^2} \sum_{i \in [N]} p_i \|\boldsymbol{\zeta^i}\|_2^2.$$

Here, $R'(\boldsymbol{\zeta^1}, \ldots, \boldsymbol{\zeta^N})$ is convex with respect to $(\boldsymbol{\zeta^1}; \cdots; \boldsymbol{\zeta^N})$. Since $(\boldsymbol{\zeta^{1*}}; \cdots; \boldsymbol{\zeta^{N*}})$ minimizes R over Z, there exists a subgradient $g \in \partial R_{(\boldsymbol{\zeta^1}; \cdots; \boldsymbol{\zeta^N})}(\boldsymbol{x}, \boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}})$ such that

$$g^\top (\hat{\boldsymbol{\zeta}}^{\boldsymbol{1}} - \boldsymbol{\zeta^{1*}}; \cdots; \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}} - \boldsymbol{\zeta^{N*}}) \geq 0. \tag{45}$$

Moreover,

$$g - \frac{\epsilon}{\theta^2}(p_1 \boldsymbol{\zeta^{1*}}; \cdots; p_N \boldsymbol{\zeta^{N*}}) \in \partial R'(\boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}}).$$

Then, since R' is convex, it follows that

$$\begin{aligned}
&R'(\hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) \\
&\geq R'(\boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}}) + \left(g - \frac{\epsilon}{\theta^2}(p_1 \boldsymbol{\zeta^{1*}}; \cdots; p_N \boldsymbol{\zeta^{N*}})\right)^\top (\hat{\boldsymbol{\zeta}}^{\boldsymbol{1}} - \boldsymbol{\zeta^{1*}}; \cdots; \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}} - \boldsymbol{\zeta^{N*}}).
\end{aligned}$$

From this, we deduce that

$$\begin{aligned}
&R(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - R(\boldsymbol{x}, \boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}}) - g^\top (\hat{\boldsymbol{\zeta}}^{\boldsymbol{1}} - \boldsymbol{\zeta^{1*}}; \cdots; \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}} - \boldsymbol{\zeta^{N*}}) \\
&\geq \frac{\epsilon}{2\theta^2} \sum_{i \in [N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}}\|_2^2 - \frac{\epsilon}{2\theta^2} \sum_{i \in [N]} p_i \|\boldsymbol{\zeta^{i*}}\|_2^2 - \frac{\epsilon}{\theta^2} \sum_{i \in [N]} p_i ((\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}})^\top \boldsymbol{\zeta^{i*}} - \|\boldsymbol{\zeta^{i*}}\|_2^2) \\
&= \frac{\epsilon}{2\theta^2} \sum_{i \in [N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} - \boldsymbol{\zeta^{i*}}\|_2^2.
\end{aligned} \tag{46}$$

Combining (45) and (46), it follows that (38) holds, as required.

C.2 Proof of Lemma 7.9

For each $(\boldsymbol{\zeta^1}; \cdots; \boldsymbol{\zeta^N}) \in Z$, we know that $\boldsymbol{\zeta^i} \in \Gamma$ for each $i \in [N]$. For $\boldsymbol{x^1}, \boldsymbol{x^2} \in \mathcal{X}$ and $\boldsymbol{\zeta^i} \in \Gamma$, we have

$$f(\boldsymbol{x^1}, \boldsymbol{\zeta^i}) = f(\boldsymbol{x^1}, \boldsymbol{\zeta^i}) - f(\boldsymbol{x^2}, \boldsymbol{\zeta^i}) + f(\boldsymbol{x^2}, \boldsymbol{\zeta^i}) \leq L_1 \|\boldsymbol{x^1} - \boldsymbol{x^2}\| + f(\boldsymbol{x^2}, \boldsymbol{\zeta^i})$$

where the inequality holds due to Assumption 2. Since $\sum_{i\in[N]} p_i = 1$,

$$\sum_{i\in[N]} p_i f(\boldsymbol{x^1}, \boldsymbol{\zeta^i}) \leq L_1 \|\boldsymbol{x^1} - \boldsymbol{x^2}\| + \sum_{i\in[N]} p_i f(\boldsymbol{x^2}, \boldsymbol{\zeta^i}).$$

By adding the term $\frac{\epsilon}{2\theta^2} \sum_{i\in[N]} p_i \left\|\boldsymbol{\xi^i} - \boldsymbol{\zeta^i}\right\|_2^2$ to both sides, we obtain

$$\begin{aligned}&\sum_{i\in[N]} p_i \left(f(\boldsymbol{x^1}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2} \left\|\boldsymbol{\xi^i} - \boldsymbol{\zeta^i}\right\|_2^2\right)\\ &\leq L_1 \|\boldsymbol{x^1} - \boldsymbol{x^2}\| + \sum_{i\in[N]} p_i \left(f(\boldsymbol{x^2}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2} \left\|\boldsymbol{\xi^i} - \boldsymbol{\zeta^i}\right\|_2^2\right).\end{aligned}$$

By taking the infimum of its left-hand side over $(\boldsymbol{\zeta^1}, \ldots, \boldsymbol{\zeta^N}) \in Z$ and that of the right-hand side next, we obtain

$$\begin{aligned}&\inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in Z} \left\{\sum_{i\in[N]} p_i \left(f(\boldsymbol{x^1}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2} \left\|\boldsymbol{\xi^i} - \boldsymbol{\zeta^i}\right\|_2^2\right)\right\}\\ &\leq L_1 \|\boldsymbol{x^1} - \boldsymbol{x^2}\| + \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in Z} \left\{\sum_{i\in[N]} p_i \left(f(\boldsymbol{x^2}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2} \left\|\boldsymbol{\xi^i} - \boldsymbol{\zeta^i}\right\|_2^2\right)\right\},\end{aligned}$$

and therefore, $H(\boldsymbol{x^1}) - H(\boldsymbol{x^2}) \leq L_1 \|\boldsymbol{x^1} - \boldsymbol{x^2}\|$. Similarly, we deduce that $-H(\boldsymbol{x^1}) + H(\boldsymbol{x^2}) \leq L_1 \|\boldsymbol{x^1} - \boldsymbol{x^2}\|$, so we have just proved that H is L_1-Lipschitz continuous in the norm $\|\cdot\|$ over $\mathcal{X}$, as required.

C.3 Proof of Lemma 7.4

Let $(\boldsymbol{\zeta^{1*}}; \cdots; \boldsymbol{\zeta^{N*}}) \in$ Z be such that $H(\boldsymbol{x}) = R(\boldsymbol{x}, \boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}})$. Then

$$R(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - R(\boldsymbol{x}, \boldsymbol{\zeta^{1*}}, \ldots, \boldsymbol{\zeta^{N*}}) = R(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{1}}, \ldots, \hat{\boldsymbol{\zeta}}^{\boldsymbol{N}}) - H(\boldsymbol{x}) \leq \delta. \tag{47}$$

By Lemma 7.8, we obtain

$$\sum_{i\in[N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} - \boldsymbol{\zeta^{i*}}\|_2^2 \leq \frac{2\theta^2\delta}{\epsilon}. \tag{48}$$

Next, based on the last statement of Assumption 2, we deduce the following.

$$\begin{aligned}\left\|\nabla H(\boldsymbol{x}) - \sum_{i\in[N]} p_i \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{i}})\right\|_* &\leq \sum_{i\in[N]} p_i \left\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{\zeta^{i*}}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{i}})\right\|_*\\ &\leq \lambda_2 \sum_{i\in[N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} - \boldsymbol{\zeta^{i*}}\|_2\\ &\leq \lambda_2 \Big(\sum_{i\in[N]} p_i\Big)^{1/2} \cdot \Big(\sum_{i\in[N]} p_i \|\hat{\boldsymbol{\zeta}}^{\boldsymbol{i}} - \boldsymbol{\zeta^{i*}}\|_2^2\Big)^{1/2}\\ &\leq \lambda_2 \theta \sqrt{\frac{2\delta}{\epsilon}}\end{aligned} \tag{49}$$

where the first inequality is by the triangle inequality, the second inequality follows from the last statement of Assumption 2, the third inequality is due to the Cauchy-Schwarz inequality, and the last inequality comes from (48). Then (49) implies $\left\|\nabla H(\boldsymbol{x}) - \sum_{i\in[N]} p_i \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{\zeta}}^{\boldsymbol{i}})\right\|_* \leq \lambda_2 \theta \sqrt{2\delta/\epsilon}$, as required.

C.4 Proof of Lemma 7.5

We aim to show that $H(\boldsymbol{x}+\boldsymbol{v}t)$ is concave with respect to t for each $\boldsymbol{x} \in \mathcal{X} + B(\mathbf{0}, r)$ and $\boldsymbol{v} \geq 0$. For each $t_1, t_2 \geq 0$ and $\lambda \in [0,1]$, we have

$$\begin{aligned}
&H(\boldsymbol{x}+(\lambda t_1+(1-\lambda)t_2)\boldsymbol{v}) \\
&= \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in \mathrm{Z}} \sum_{i\in[N]} p_i \left(f(\boldsymbol{x}+(\lambda t_1+(1-\lambda)t_2)\boldsymbol{v}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2}\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 \right) \\
&\geq \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in \mathrm{Z}} \left\{ \lambda \sum_{i\in[N]} p_i \left(f(\boldsymbol{x}+t_1\boldsymbol{v}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2}\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 \right) \right. \\
&\qquad \left. +(1-\lambda) \sum_{i\in[N]} p_i \left(f(\boldsymbol{x}+t_2\boldsymbol{v}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2}\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 \right) \right\} \\
&\geq \lambda H(\boldsymbol{x}+t_1\boldsymbol{v}) + (1-\lambda) H(\boldsymbol{x}+t_2\boldsymbol{v}).
\end{aligned}$$

Hence, H is up-concave.

To show that H is monotone, let us take $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ such that $\boldsymbol{x} \leq \boldsymbol{y}$. Since $f(\cdot, \boldsymbol{\xi})$ is monotone for any $\boldsymbol{\xi}$, we have

$$\sum_{i\in[N]} p_i \left(f(\boldsymbol{x}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2}\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 \right) \leq \sum_{i\in[N]} p_i \left(f(\boldsymbol{y}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2}\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 \right).$$

Take the infimum of it left-hand side over $(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N}) \in \mathrm{Z}$ and that of the right-hand side next, we obtain

$$\begin{aligned}
&\inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in \mathrm{Z}} \sum_{i\in[N]} p_i \left(f(\boldsymbol{x}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2}\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 \right) \\
&\leq \inf_{(\boldsymbol{\zeta^1};\cdots;\boldsymbol{\zeta^N})\in \mathrm{Z}} \sum_{i\in[N]} p_i \left(f(\boldsymbol{y}, \boldsymbol{\zeta^i}) + \frac{\epsilon}{2\theta^2}\|\boldsymbol{\xi^i}-\boldsymbol{\zeta^i}\|_2^2 \right),
\end{aligned}$$

implying in turn that $H(\boldsymbol{x}) \leq H(\boldsymbol{y})$. Therefore, H is monotone, as required.

D Discrete submodular functions

Lemma D.1. *Let $f : 2^V \times \Xi \to \mathbb{R}_+$ be a nonnegative function such that $f(S, \boldsymbol{\xi})$ is monotone and submodular with respect to $S \subseteq V$ and convex with respect to $\boldsymbol{\xi} \in \Xi$. We further assume that $f(\emptyset, \boldsymbol{\xi}) = 0$ for $\boldsymbol{\xi} \in \cup_{i\in[N]}\Xi_i$. If $F(\boldsymbol{x}, \boldsymbol{\xi})$ for $\xi \in \Xi$ is the multilinear extension of $f(\cdot, \boldsymbol{\xi})$, then the statements in Assumption 2 hold with $L_1 = \lambda_1 = \sup_{(j,\boldsymbol{\xi})\in V\times\cup_{i\in[N]}\Xi_i} f(\{j\}, \boldsymbol{\xi})$ and $\lambda_2 = 2L_2$.*

Proof. For any $\boldsymbol{x} \in [0,1]^V$ and $\boldsymbol{\xi} \in \Xi$, let $F_{\boldsymbol{\xi}}(\boldsymbol{x})$ denote the multilinear extension of $f(\boldsymbol{x}, \boldsymbol{\xi})$. By Calinescu et al. [8], we have for $\boldsymbol{y} \in [0,1]^V$ and $\boldsymbol{\zeta} \in \cup_{i\in[N]}\Xi_i$,

$$0 \leq \frac{\partial F_{\boldsymbol{\zeta}}(\boldsymbol{y})}{\partial x_i} = F_{\boldsymbol{\zeta}}(\boldsymbol{y})|_{y_i=1} - F_{\boldsymbol{\zeta}}(\boldsymbol{y})|_{y_i=0} \leq f(\{i\}, \boldsymbol{\zeta})$$

where the last inequality is by submodularity of f. Therefore, $\|\nabla_{\boldsymbol{x}} F_{\boldsymbol{\zeta}}(\boldsymbol{y})\|_\infty \leq \max_{i\in V} f(\{i\}, \boldsymbol{\zeta})$, which implies that $F_{\boldsymbol{\xi}}$ for every $\boldsymbol{\xi} \in \cup_{i\in[N]}\Xi_i$ is L_1-Lipschitz continuous in the ℓ_1 norm over $\mathcal{X}$ where $L_1 = \sup_{(j,\boldsymbol{\xi})\in V\times\cup_{i\in[N]}\Xi_i} f(\{j\}, \boldsymbol{\xi})$.

Next we argue that the second statement of Assumption 2 holds. By Hassani et al. [17, Lemma C.1], for $\boldsymbol{y} \in [0,1]^V$ and $\boldsymbol{\zeta} \in \cup_{i\in[N]}\Xi_i$,

$$-\max_{j\in V} f(\{j\},\boldsymbol{\zeta}) \le \frac{\partial^2 F_{\boldsymbol{\zeta}}(\boldsymbol{y})}{\partial x_i \partial x_j} \le 0$$

Let $\boldsymbol{z} \in \mathbb{R}^V$. Note that

$$\left|\boldsymbol{z}^\top \nabla^2_{\boldsymbol{x}} F_{\boldsymbol{\zeta}}(\boldsymbol{y})\boldsymbol{z}\right| \le \sum_{i,j\in V} \left|\frac{\partial^2 F_{\boldsymbol{\zeta}}(\boldsymbol{y})}{\partial x_i \partial x_j}\right| \cdot |z_i z_j| \le \max_{j\in V} f(\{j\},\boldsymbol{\zeta})\|\boldsymbol{z}\|_1^2.$$

Therefore, $F(\cdot,\boldsymbol{\xi})$ for a fixed $\boldsymbol{\xi}$ is $(\max_{j\in V} f(\{j\},\boldsymbol{\xi}))$-smooth, so for any $\boldsymbol{y^1},\boldsymbol{y^2} \in [0,1]^V$, we have

$$\|\nabla_{\boldsymbol{x}} F_{\boldsymbol{\zeta}}(\boldsymbol{y^1}) - \nabla_{\boldsymbol{x}} F_{\boldsymbol{\zeta}}(\boldsymbol{y^2})\|_\infty \le \sup_{(j,\boldsymbol{\xi})\in V\times\cup_{i\in[N]}\Xi_i} f(\{j\},\boldsymbol{\xi})\|\boldsymbol{y^1}-\boldsymbol{y^2}\|_1.$$

Therefore, the second statement of Assumption 2 holds. Lastly, we consider the last statement of Assumption 2. It is known [8] that

$$\frac{\partial}{\partial x_j} F_{\boldsymbol{\zeta}}(\boldsymbol{y}) = \sum_{S\subseteq V\setminus\{j\}} (f(S\cup\{j\},\boldsymbol{\zeta}) - f(S,\boldsymbol{\zeta})) \prod_{\ell\in S} y_j \prod_{\ell\in V\setminus\{j\}} (1-y_j).$$

Then it follows that

$$\begin{aligned}
&\frac{\partial}{\partial x_j} F_{\boldsymbol{\zeta^1}}(\boldsymbol{y}) - \frac{\partial}{\partial x_j} F_{\boldsymbol{\zeta^2}}(\boldsymbol{y}) \\
&= \sum_{S\subseteq V\setminus\{j\}} \left(\left(f(S\cup\{j\},\boldsymbol{\zeta^1}) - f(S,\boldsymbol{\zeta^1})\right) - \left(f(S\cup\{j\},\boldsymbol{\zeta^2}) - f(S,\boldsymbol{\zeta^2})\right)\right) \prod_{\ell\in S} y_j \prod_{\ell\in V\setminus\{j\}} (1-y_j) \\
&= \sum_{S\subseteq V\setminus\{j\}} \left(\left(f(S\cup\{j\},\boldsymbol{\zeta^1}) - f(S\cup\{j\},\boldsymbol{\zeta^2})\right) - \left(f(S,\boldsymbol{\zeta^1}) - f(S,\boldsymbol{\zeta^2})\right)\right) \prod_{\ell\in S} y_j \prod_{\ell\in V\setminus\{j\}} (1-y_j) \\
&\le \sum_{S\subseteq V\setminus\{j\}} \left(L_2\|\boldsymbol{\zeta^1}-\boldsymbol{\zeta^2}\| + L_2\|\boldsymbol{\zeta^1}-\boldsymbol{\zeta^2}\|\right) \prod_{\ell\in S} y_j \prod_{\ell\in V\setminus\{j\}} (1-y_j) \\
&= 2L_2\|\boldsymbol{\zeta^1}-\boldsymbol{\zeta^2}\| \sum_{S\subseteq V\setminus\{j\}} \prod_{\ell\in S} y_j \prod_{\ell\in V\setminus\{j\}} (1-y_j) \\
&= 2L_2\|\boldsymbol{\zeta^1}-\boldsymbol{\zeta^2}\|
\end{aligned} \tag{50}$$

where the inequality holds because $f(S,\cdot)$ is M-Lipschitz continuous for any $S \subseteq V$. Hence the last statement of Assumption 2 holds true with $\lambda_2 = 2L_2$. □