Continuous DR-submodular Maximization: Structure and Algorithms

An Bian[*]
ETH Zurich
ybian@inf.ethz.ch

Kfir Y. Levy
ETH Zurich
yehuda.levy@inf.ethz.ch

Andreas Krause
ETH Zurich
krausea@ethz.ch

Joachim M. Buhmann
ETH Zurich
jbuhmann@inf.ethz.ch

Abstract

DR-submodular continuous functions are important objectives with wide real-world applications spanning MAP inference in determinantal point processes (DPPs), and mean-field inference for probabilistic submodular models, amongst others. DR-submodularity captures a subclass of non-convex functions that enables both exact minimization and approximate maximization in polynomial time.

In this work we study the problem of maximizing *non-monotone* continuous DR-submodular functions under general down-closed convex constraints. We start by investigating geometric properties that underlie such objectives, e.g., a strong relation between (approximately) stationary points and global optimum is proved. These properties are then used to devise two optimization algorithms with provable guarantees. Concretely, we first devise a "two-phase" algorithm with $1/4$ approximation guarantee. This algorithm allows the use of existing methods for finding (approximately) stationary points as a subroutine, thus, harnessing recent progress in non-convex optimization. Then we present a non-monotone FRANK-WOLFE variant with $1/e$ approximation guarantee and sublinear convergence rate. Finally, we extend our approach to a broader class of generalized DR-submodular continuous functions, which captures a wider spectrum of applications. Our theoretical findings are validated on synthetic and real-world problem instances.

1 Introduction

Submodularity is classically most well known for set function optimization, where it enables efficient minimization [23] and approximate maximization [31; 25] in polynomial time. Submodularity has recently been studied on the integer lattice [34; 33] and on continuous domains [3; 4; 36; 21], with significant theoretical results and practical applications. For set functions, it is well known that submodularity is equivalent to the diminishing returns (DR) property. However, this does not hold for integer-lattice functions or continuous functions, where the DR property defines a subclass of submodular functions, called DR-submodular functions.

In continuous domains, applying convex optimization techniques enables efficient minimization of submodular continuous functions [3; 36] (despite the non-convex nature of such objectives). In [4] it is further shown that continuous submodularity enables constant-factor approximation schemes for constrained monotone DR-submodular maximization and "box" constrained non-monotone submodular maximization problems.

---

[*]Now known as Yatao A. Bian. ORCID: orcid.org/0000-0002-2368-4084

Many real-world non-convex problems, such as maximizing the softmax extension of DPPs, require maximizing a *non-monotone* DR-submodular function over a *general* down-closed convex constraint. Yet, current theory [3; 4; 36] does not apply to this general problem setting, which motivates us to develop guaranteed and efficient algorithms for such problems.

Exploring the structure that underlies DR-submodularity is crucial to deriving guaranteed algorithms. Combined with a notion of non-stationarity for constrained optimization problems and a new notion of "strong DR-submodularity", we find a rich structure in the problem of continuous DR-submodular maximization. This in turn gives rise to two approximation algorithms with provable guarantees. Specifically, we make the following contributions:

- We bound the difference between objective values of stationary points and the global optimum. Our analysis shows that the bound is even tighter if the objective is strongly DR-submodular (see Definition 3).

- Based on the geometric properties, we present two algorithms: (i) A two-phase FRANK-WOLFE-style algorithm with $1/4$ approximation guarantee converges with a $1/\sqrt{k}$ rate; (ii) a non-monotone FRANK-WOLFE variant exhibits a $1/e$ approximation guarantee and converges sublinearly. Even though the worst-case guarantee of the first one is worse than the second, it yields several practical advantages, which we discuss in Section 4.2.

- We investigate a generalized class of submodular functions on "conic" lattices. This allows us to model a larger class of non-trivial applications. These include logistic regression with a non-convex separable regularizer, non-negative PCA, etc. To optimize them, we provide a reduction that enables to invoke algorithms for continuous submodular optimization problems.

- We experimentally demonstrate the applicability of our methods on both synthetic and real-world problem instances.

1.1 Problem Statement

Notation. We use boldface letters, e.g., x to represent a vector, boldface capital letters, e.g., $\mathbf{A}$ to denote a matrix. x_i is the ith entry of x, A_{ij} is the $(ij)^{\text{th}}$ entry of $\mathbf{A}$. We use e_i to denote the standard i^{th} basis vector. $f(\cdot)$ is used to denote a continuous function, and $F(\cdot)$ to represent a set function. $[n] := \{1, ..., n\}$ for an integer $n \geq 1$. $\|\cdot\|$ means the Euclidean norm by default. Given two vectors x, y, $x \leq y$ means $x_i \leq y_i, \forall i$. $x \vee y$ and $x \wedge y$ denote coordinate-wise maximum and coordinate-wise minimum, respectively.

The general setup of constrained non-monotone DR-submodular (see Definition 1 below) maximization is,

$$\max_{\boldsymbol{x} \in \mathcal{P}} f(\boldsymbol{x}), \tag{P}$$

where $f : \mathcal{X} \to \mathbb{R}$ is continuous DR-submodular, $\mathcal{X} = \prod_{i=1}^{n} \mathcal{X}_i$, each $\mathcal{X}_i$ is an interval [3; 4]. Wlog[2], we assume that the lower bound $\underline{u}$ of $\mathcal{X}$ is 0, i.e., $\mathcal{X} = [0, \bar{u}]$. The set $\mathcal{P} \subseteq [0, \bar{u}]$ is assumed to be a down-closed convex set, where down-closedness means: $x \in \mathcal{P}$ and $0 \leq y \leq x$ implies that $y \in \mathcal{P}$. The diameter of $\mathcal{P}$ is $D := \max_{\boldsymbol{x}, \boldsymbol{y} \in \mathcal{P}} \|\boldsymbol{x} - \boldsymbol{y}\|$, and it holds that $D \leq \|\bar{u}\|$. We use x^* to denote the global maximum of (P). One can assume f is non-negative over $\mathcal{X}$, since otherwise one just needs to find a lower bound for the minimum function value of f over $[0, \bar{u}]$ (and box-constrained submodular minimization can be solved to arbitrary precision in polynomial time [3]). Over continuous domains, a DR-submodular function [4] is a submodular function with the diminishing returns (DR) property,

Definition 1 (DR-submodular & DR property). *A function* $f : \mathcal{X} \mapsto \mathbb{R}$ *is DR-submodular (has the DR property) if* $\forall \boldsymbol{a} \leq \boldsymbol{b} \in \mathcal{X}, \forall i \in [n], \forall k \in \mathbb{R}_+$ *s.t.* $(k\boldsymbol{e}_i + \boldsymbol{a})$ *and* $(k\boldsymbol{e}_i + \boldsymbol{b})$ *are still in* $\mathcal{X}$, *it holds,*

$$f(k\boldsymbol{e}_i + \boldsymbol{a}) - f(\boldsymbol{a}) \geq f(k\boldsymbol{e}_i + \boldsymbol{b}) - f(\boldsymbol{b}). \tag{1}$$

If f is differentiable, one can show that Definition 1 is equivalent to ∇f being an antitone mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$. Furthermore, if f is twice-differentiable, the DR property is equivalent to all of

---

[2]Since otherwise one can work on a new function $g(\boldsymbol{x}) := f(\boldsymbol{x} + \underline{u})$ that has 0 as the lower bound of its domain, and all properties of the function are still preserved.

the entries of its Hessian being non-positive, i.e., $\nabla^2_{ij} f(\boldsymbol{x}) \leq 0, \forall \boldsymbol{x} \in \mathcal{X}, i, j \in [n]$. A function $f : \mathcal{X} \mapsto \mathbb{R}$ is DR-supermodular iff $-f$ is DR-submodular. We also assume that f has Lipschitz gradients,

Definition 2. *A function f has L-Lipschitz gradients if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ it holds that,*

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|. \tag{2}$$

A brief summary of related work appears in Section 6.

2 Motivating Real-world Examples

Many continuous objectives in practice turn out to be DR-submodular. Here we list several of them. More can be found in Appendix B.

Softmax extension. Determinantal point processes (DPPs) are probabilistic models of repulsion, that have been used to model diversity in machine learning [26]. The constrained MAP (maximum a posteriori) inference problem of a DPP is an NP-hard combinatorial problem in general. Currently, the methods with the best approximation guarantees are based on either maximizing the multilinear extension [6] or the softmax extension [20], both of which are DR-submodular functions (details in Appendix F.1). The multilinear extension is given as an expectation over the original set function values, thus evaluating the objective of this extension requires expensive sampling. In constast, the softmax extension has a closed form expression, which is much more appealing from a computational perspective. Let $\mathbf{L}$ be the positive semidefinite kernel matrix of a DPP, its softmax extension is:

$$f(\boldsymbol{x}) = \log\det\left(\text{diag}(\boldsymbol{x})(\mathbf{L} - \mathbf{I}) + \mathbf{I}\right), \boldsymbol{x} \in [0,1]^n, \tag{3}$$

where $\mathbf{I}$ is the identity matrix, $\text{diag}(\boldsymbol{x})$ is the diagonal matrix with diagonal elements set as $\boldsymbol{x}$. The problem of MAP inference in DPPs corresponds to the problem $\max_{\boldsymbol{x} \in \mathcal{P}} f(\boldsymbol{x})$, where $\mathcal{P}$ is a down-closed convex constraint, e.g., a matroid polytope or a matching polytope.

Mean-field inference for log-submodular models. Log-submodular models [9] are a class of probabilistic models over subsets of a ground set $\mathcal{V} = [n]$, where the log-densities are submodular set functions $F(S)$: $p(S) = \frac{1}{Z}\exp(F(S))$. The partition function $Z = \sum_{S \subseteq \mathcal{V}} \exp(F(S))$ is typically hard to evaluate. One can use mean-field inference to approximate $p(S)$ by some factorized distribution $q_{\boldsymbol{x}}(S) := \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j), \boldsymbol{x} \in [0,1]^n$, by minimizing the distance measured w.r.t. the Kullback-Leibler divergence between $q_{\boldsymbol{x}}$ and p, i.e., $\sum_{S \subseteq \mathcal{V}} q_{\boldsymbol{x}}(S) \log \frac{q_{\boldsymbol{x}}(S)}{p(S)}$. It is,

$$\text{KL}(\boldsymbol{x}) = -\sum_{S \subseteq \mathcal{V}} \prod_{i \in S} x_i \prod_{j \notin S} (1 - x_j) F(S) + \sum_{i=1}^{n} [x_i \log x_i + (1 - x_i)\log(1 - x_i)] + \log Z.$$

$\text{KL}(\boldsymbol{x})$ is DR-supermodular w.r.t. $\boldsymbol{x}$ (details in Appendix F.1). Minimizing the Kullback-Leibler divergence $\text{KL}(\boldsymbol{x})$ amounts to maximizing a DR-submodular function.

2.1 Motivating Example Captured by Generalized Submodularity on Conic Lattices

Submodular continuous functions can already model many scenarios. Yet, there are several interesting cases which are in general not (DR-)Submodular, but can still be captured by a generalized notion. This generalization enables to develop polynomial algorithms with guarantees by using ideas from continuous submodular optimization. We present one representative objective here (more in Appendix B). In Appendix A we show the technical details on how they are covered by a class of submodular continuous functions over conic lattices.

Consider the logistic regression model with a *non-convex* separable regularizer. This flexibility may result in better statistical performance (e.g., in recovering discontinuities, [2]) compared to classical models with convex regularizers. Let $\boldsymbol{z}^1, ..., \boldsymbol{z}^m$ in $\mathbb{R}^n$ be m training points with corresponding binary labels $\boldsymbol{y} \in \{\pm 1\}^m$. Assume that the following mild assumption is satisfied: For any fixed dimension i, all the data points have the same sign, i.e., $\text{sign}(z_i^j)$ is the same for all $j \in [m]$ (which can be achieved by easily scaling if not). The task is to solve the following non-convex optimization problem,

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}) := m^{-1} \sum_{j=1}^{m} f_j(\boldsymbol{x}) + \lambda r(\boldsymbol{x}), \tag{4}$$

3

where $f_j(\boldsymbol{x}) = \log(1 + \exp(-y_j \boldsymbol{x}^\top \boldsymbol{z}^j))$ is the logistic loss; $\lambda > 0$ is the regularization parameter, and $r(\boldsymbol{x})$ is some non-convex separable regularizer. Such separable regularizers are popular in statistics, and two notable choices are $r(\boldsymbol{x}) = \sum_{i=1}^{n} \frac{\gamma x_i^2}{1+\gamma x_i^2}$, and $r(\boldsymbol{x}) = \sum_{i=1}^{n} \min\{\gamma x_i^2, 1\}$ (see [2]). Let us define a vector $\boldsymbol{\alpha} \in \{\pm 1\}^n$ as $\alpha_i = \text{sign}(z_i^j), i \in [n]$ and $l(\boldsymbol{x}) := \frac{1}{m} \sum_{j=1}^{m} f_j(\boldsymbol{x})$. One can show that $l(\boldsymbol{x})$ is not DR-submodular or DR-supermodular. Yet, in Appendix A we will show that $l(\boldsymbol{x})$ is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-supermodular, where the latter generalizes DR-supermodularity. Usually, one can assume the optimal solution $\boldsymbol{x}^*$ lies in some box $[\underline{\boldsymbol{u}}, \bar{\boldsymbol{u}}]$. Then the problem is an instance of constrained non-monotone $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular maximization.

3 Underlying Properties of Constrained DR-submodular Maximization

In this section we present several properties arising in DR-submodular function maximization. First we show properties related to concavity of the objective along certain directions, then we establish the relation between locally stationary points and the global optimum (thus called "local-global relation"). These properties will be used to derive guarantees for the algorithms in Section 4. All omitted proofs are in Appendix D.

3.1 Properties Along Non-negative/Non-positive Directions

A DR-submodular function f is *concave* along any non-negative/non-positive direction [4]. Notice that DR-submodularity is a stronger condition than concavity along directions $\boldsymbol{v} \in \pm\mathbb{R}_+^n$: for instance, a concave function is concave along any direction, but it may not be a DR-submodular function.

For a DR-submodular function with L-Lipschitz gradients, one can get the following quadratic lower bound using standard techniques by combing the concavity and Lipschitz gradients in (2).
Quadratic lower bound. If f is DR-submodular with a L-Lipschitz gradient, then for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{v} \in \pm\mathbb{R}_+^n$, it holds,

$$f(\boldsymbol{x} + \boldsymbol{v}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle - \frac{L}{2}\|\boldsymbol{v}\|^2. \tag{5}$$

It will be used in Section 4.2 for analyzing the non-monotone FRANK-WOLFE variant (Algorithm 2).

Strong DR-submodularity. DR-submodular objectives may be strongly concave along directions $\boldsymbol{v} \in \pm\mathbb{R}_+^n$, e.g., for DR-submodular quadratic functions. We will show that such additional structure may be exploited to obtain stronger guarantees for the local-global relation.

Definition 3 (Strongly DR-submodular). *A function f is μ-strongly DR-submodular ($\mu \geq 0$) if for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{v} \in \pm\mathbb{R}_+^n$, it holds that,*

$$f(\boldsymbol{x} + \boldsymbol{v}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle - \frac{\mu}{2}\|\boldsymbol{v}\|^2. \tag{6}$$

3.2 Relation Between Approximately Stationary Points and Global Optimum

First of all, we present the following Lemma, which will motivate us to consider a non-stationarity measure for general constrained optimization problems.

Lemma 1. *If f is μ-strongly DR-submodular, then for any two points $\boldsymbol{x}$, $\boldsymbol{y}$ in $\mathcal{X}$, it holds:*

$$(\boldsymbol{y} - \boldsymbol{x})^\top \nabla f(\boldsymbol{x}) \geq f(\boldsymbol{x} \vee \boldsymbol{y}) + f(\boldsymbol{x} \wedge \boldsymbol{y}) - 2f(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{7}$$

Lemma 1 implies that if $\boldsymbol{x}$ is stationary (i.e., $\nabla f(\boldsymbol{x}) = 0$), then $2f(\boldsymbol{x}) \geq f(\boldsymbol{x} \vee \boldsymbol{y}) + f(\boldsymbol{x} \wedge \boldsymbol{y}) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$, which gives an implicit relation between $\boldsymbol{x}$ and $\boldsymbol{y}$. While in practice finding an exact stationary point is not easy, usually non-convex solvers will arrive at an approximately stationary point, thus requiring a proper measure of non-stationarity for the constrained optimization problem.

Non-stationarity measure. Looking at the LHS of (7), it naturally suggests to use $\max_{\boldsymbol{y} \in \mathcal{P}}(\boldsymbol{y} - \boldsymbol{x})^\top \nabla f(\boldsymbol{x})$ as the non-stationarity measure, which happens to coincide with the measure proposed by recent work of [27], and it can be calculated for free for Frank-Wolfe-style algorithms (e.g.,

4

Algorithm 3). In order to adapt it to the local-global relation, we give a slightly more general definition here: For any constraint set $\mathcal{Q} \subseteq \mathcal{X}$, the non-stationarity of a point $\boldsymbol{x} \in \mathcal{Q}$ is,

$$g_{\mathcal{Q}}(\boldsymbol{x}) := \max_{\boldsymbol{v} \in \mathcal{Q}} \langle \boldsymbol{v} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle \qquad \text{(non-stationarity)}. \qquad (8)$$

It always holds that $g_{\mathcal{Q}}(\boldsymbol{x}) \geq 0$, and $\boldsymbol{x}$ is a stationary point in $\mathcal{Q}$ iff $g_{\mathcal{Q}}(\boldsymbol{x}) = 0$, so (8) is a natural generalization of the non-stationarity measure $\|\nabla f(\boldsymbol{x})\|$ for unconstrained optimization. As the next statement shows, $g_{\mathcal{Q}}(\boldsymbol{x})$ plays an important role in characterizing the local-global relation.

Proposition 1 (Local-Global Relation). *Let $\boldsymbol{x}$ be a point in $\mathcal{P}$ with non-stationarity $g_{\mathcal{P}}(\boldsymbol{x})$, and $\mathcal{Q} := \{\boldsymbol{y} \in \mathcal{P} \mid \boldsymbol{y} \leq \bar{\boldsymbol{u}} - \boldsymbol{x}\}$. Let $\boldsymbol{z}$ be a point in $\mathcal{Q}$ with non-stationarity $g_{\mathcal{Q}}(\boldsymbol{z})$. It holds that,*

$$\max\{f(\boldsymbol{x}), f(\boldsymbol{z})\} \geq \frac{1}{4} \left[f(\boldsymbol{x}^*) - g_{\mathcal{P}}(\boldsymbol{x}) - g_{\mathcal{Q}}(\boldsymbol{z}) \right] + \frac{\mu}{8} \left(\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 + \|\boldsymbol{z} - \boldsymbol{z}^*\|^2 \right), \qquad (9)$$

where $\boldsymbol{z}^* := \boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{x}$.

Proof sketch of Proposition 1: The proof uses Lemma 1, the non-stationarity in (8) and a key observation in the following Claim. The detailed proof is in Appendix D.2.

Claim 1. *It holds that $f(\boldsymbol{x} \vee \boldsymbol{x}^*) + f(\boldsymbol{x} \wedge \boldsymbol{x}^*) + f(\boldsymbol{z} \vee \boldsymbol{z}^*) + f(\boldsymbol{z} \wedge \boldsymbol{z}^*) \geq f(\boldsymbol{x}^*)$.*

Note that [7; 20] propose a similar relation for the special cases of multilinear/softmax extensions by mainly proving the same conclusion as in Claim 1. Their relation does not incorporate the properties of non-stationarity or strong DR-submodularity. They both use the proof idea of constructing a complicated auxiliary set function tailored to specific DR-submodular functions. We present a different proof method by directly utilizing the DR property on carefully constructed auxiliary points (e.g., $(\boldsymbol{x} + \boldsymbol{z}) \vee \boldsymbol{x}^*$ in the proof of Claim 1), this is arguably more succint and straightforward than that of [7; 20].

4 Algorithms for Constrained DR-submodular Maximization

Based on the properties, we present two algorithms for solving (P). The first is based on the local-global relation, and the second is a FRANK-WOLFE variant adapted for the non-monotone setting. All the omitted proofs are deferred to Appendix E.

4.1 An Algorithm Based on the Local-Global Relation

Algorithm 1: TWO-PHASE FRANK-WOLFE for non-monotone DR-submodular maximization

---
Input: $\max_{\boldsymbol{x} \in \mathcal{P}} f(\boldsymbol{x})$, stopping tolerance ϵ_1, ϵ_2, #iterations K_1, K_2
1 $\boldsymbol{x} \leftarrow$ NON-CONVEX FRANK-WOLFE$(f, \mathcal{P}, K_1, \epsilon_1, \boldsymbol{x}^{(0)})$;　　　　　　　 // $\boldsymbol{x}^{(0)} \in \mathcal{P}$
2 $\mathcal{Q} \leftarrow \mathcal{P} \cap \{\boldsymbol{y} \in \mathbb{R}_+^n \mid \boldsymbol{y} \leq \bar{\boldsymbol{u}} - \boldsymbol{x}\}$;
3 $\boldsymbol{z} \leftarrow$ NON-CONVEX FRANK-WOLFE$(f, \mathcal{Q}, K_2, \epsilon_2, \boldsymbol{z}^{(0)})$;　　　　　　　 // $\boldsymbol{z}^{(0)} \in \mathcal{Q}$
Output: $\arg\max\{f(\boldsymbol{x}), f(\boldsymbol{z})\}$;

We summarize the TWO-PHASE algorithm in Algorithm 1. It is generalized from the "two-phase" method in [7; 20]. It invokes some non-convex solver (we use the NON-CONVEX FRANK-WOLFE by [27]; pseudocode is included in Algorithm 3 of Appendix C) to find approximately stationary points in $\mathcal{P}$ and $\mathcal{Q}$, respectively, then returns the solution with the larger function value. Though we use NON-CONVEX FRANK-WOLFE as the subroutine here, it is worth noting that any algorithm that is guaranteed to find an approximately stationary point can be plugged into Algorithm 1 as the subroutine. We give an improved approximation bound by considering more properties of DR-submodular functions. Borrowing the results from [27] for the NON-CONVEX FRANK-WOLFE subroutine, we get the following,

Theorem 1. *The output of Algorithm 1 satisfies,*

$$\max\{f(\boldsymbol{x}), f(\boldsymbol{z})\} \geq \frac{\mu}{8} \left(\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 + \|\boldsymbol{z} - \boldsymbol{z}^*\|^2 \right) \qquad (10)$$

$$+ \frac{1}{4}\left[f(\boldsymbol{x}^*) - \min\left\{ \frac{\max\{2h_1, C_f(\mathcal{P})\}}{\sqrt{K_1+1}}, \epsilon_1 \right\} - \min\left\{ \frac{\max\{2h_2, C_f(\mathcal{Q})\}}{\sqrt{K_2+1}}, \epsilon_2 \right\} \right],$$

where $h_1 := \max_{\boldsymbol{x} \in \mathcal{P}} f(\boldsymbol{x}) - f(\boldsymbol{x}^{(0)})$, $h_2 := \max_{\boldsymbol{z} \in \mathcal{Q}} f(\boldsymbol{z}) - f(\boldsymbol{z}^{(0)})$ *are the initial suboptimalities,* $C_f(\mathcal{P}) := \sup_{\boldsymbol{x}, \boldsymbol{v} \in \mathcal{P}, \gamma \in [0,1], \boldsymbol{y} = \boldsymbol{x} + \gamma(\boldsymbol{v} - \boldsymbol{x})} \frac{2}{\gamma^2}(f(\boldsymbol{y}) - f(\boldsymbol{x}) - (\boldsymbol{y} - \boldsymbol{x})^\top \nabla f(\boldsymbol{x}))$ *is the curvature of* f *w.r.t.* $\mathcal{P}$, *and* $\boldsymbol{z}^* = \boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{x}$.

Theorem 1 indicates that Algorithm 1 has a $1/4$ approximation guarantee and $1/\sqrt{k}$ rate. However, it has good empirical performance as demonstrated by the experiments in Section 5. Informally, this can be partially explained by the term $\frac{\mu}{8}\left(\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 + \|\boldsymbol{z} - \boldsymbol{z}^*\|^2\right)$ in (10): if $\boldsymbol{x}$ is away from $\boldsymbol{x}^*$, this term will augment the bound; if $\boldsymbol{x}$ is close to $\boldsymbol{x}^*$, by the smoothness of f, it should be close to optimal.

4.2 The Non-monotone FRANK-WOLFE Variant

Algorithm 2: Non-monotone FRANK-WOLFE variant for DR-submodular maximization

---
Input: $\max_{\boldsymbol{x} \in \mathcal{P}} f(\boldsymbol{x})$, prespecified step size $\gamma \in (0, 1]$
1   $\boldsymbol{x}^{(0)} \leftarrow 0$, $t^{(0)} \leftarrow 0$, $k \leftarrow 0$;   // k : *iteration index,* $t^{(k)}$: *cumulative step size*
2   **while** $t^{(k)} < 1$ **do**
3      $\boldsymbol{v}^{(k)} \leftarrow \arg\max_{\boldsymbol{v} \in \mathcal{P}, \boldsymbol{v} \le \bar{\boldsymbol{u}} - \boldsymbol{x}^{(k)}} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}^{(k)}) \rangle$;   // shrunken LMO
4   use uniform step size $\gamma_k = \gamma$; set $\gamma_k \leftarrow \min\{\gamma_k, 1 - t^{(k)}\}$;
5      $\boldsymbol{x}^{(k+1)} \leftarrow \boldsymbol{x}^{(k)} + \gamma_k \boldsymbol{v}^{(k)}$, $t^{(k+1)} \leftarrow t^{(k)} + \gamma_k$, $k \leftarrow k+1$;
     Output: $\boldsymbol{x}^{(K)}$;   // *assuming there are* K *iterations in total*

Algorithm 2 summarizes the non-monotone FRANK-WOLFE variant, which is inspired by the unified continuous greedy algorithm in [13] for maximizing the multilinear extension of a submodular set function. It initializes the solution $\boldsymbol{x}^{(0)}$ to be 0, and maintains $t^{(k)}$ as the cumulative step size. At iteration k, it maximizes the linearization of f over a "shrunken" constraint set: $\{\boldsymbol{v} | \boldsymbol{v} \in \mathcal{P}, \boldsymbol{v} \le \bar{\boldsymbol{u}} - \boldsymbol{x}^{(k)}\}$, which is different from the classical LMO of Frank-Wolfe-style algorithms (hence we refer to it as the "shrunken LMO"). Then it employs an update step in the direction $\boldsymbol{v}^{(k)}$ chosen by the LMO with a uniform step size $\gamma_k = \gamma$. The cumulative step size $t^{(k)}$ is used to ensure that the overall step sizes sum to one, thus the output solution $\boldsymbol{x}^{(K)}$ is a convex combination of the LMO outputs, hence also lies in $\mathcal{P}$.

The shrunken LMO (Step 3) is the key difference compared to the monotone FRANK-WOLFE variant in [4]. The extra constraint $\boldsymbol{v} \le \bar{\boldsymbol{u}} - \boldsymbol{x}^{(k)}$ is added to prevent too aggressive growth of the solution, since in the non-monotone setting such aggressive growth may hurt the overall performance. The next theorem states the guarantees of Algorithm 2.

Theorem 2. *Consider Algorithm 2 with uniform step size* γ. *For* $k = 1, ..., K$ *it holds that,*

$$f(\boldsymbol{x}^{(k)}) \ge t^{(k)} e^{-t^{(k)}} f(\boldsymbol{x}^*) - \frac{LD^2}{2} k \gamma^2 - O(\gamma^2) f(\boldsymbol{x}^*). \tag{11}$$

By observing that $t^{(K)} = 1$ and applying Theorem 2, we get the following Corollary:

Corollary 1. *The output of Algorithm 2 satisfies* $f(\boldsymbol{x}^{(K)}) \ge e^{-1} f(\boldsymbol{x}^*) - \frac{LD^2}{2K} - O\left(\frac{1}{K^2}\right) f(\boldsymbol{x}^*)$.

Corollary 1 shows that Algorithm 2 enjoys a sublinear convergence rate towards some point $\boldsymbol{x}^{(K)}$ inside $\mathcal{P}$, with a $1/e$ approximation guarantee.

Proof sketch of Theorem 2: The proof is by induction. To prepare the building blocks, we first of all show that the growth of $\boldsymbol{x}^{(k)}$ is indeed bounded,

Lemma 2. *Assume* $\boldsymbol{x}^{(0)} = 0$. *For* $k = 0, ..., K-1$, *it holds* $x_i^{(k)} \le \bar{u}_i[1 - (1-\gamma)^{t^{(k)}/\gamma}], \forall i \in [n]$.

Then the following Lemma provides a lower bound, which gets the global optimum involved,

Lemma 3 (Generalized from Lemma 7 in [8]). *Given* $\boldsymbol{\theta} \in (0, \bar{\boldsymbol{u}}]$, *let* $\lambda' = \min_{i \in [n]} \frac{\bar{u}_i}{\theta_i}$. *Then for all* $\boldsymbol{x} \in [0, \boldsymbol{\theta}]$, *it holds* $f(\boldsymbol{x} \vee \boldsymbol{x}^*) \ge (1 - \frac{1}{\lambda'}) f(\boldsymbol{x}^*)$.

Then the key ingredient for induction is the relation between $f(\boldsymbol{x}^{(k+1)})$ and $f(\boldsymbol{x}^{(k)})$ indicated by:

Claim 2. *For $k = 0, ..., K-1$ it holds $f(\boldsymbol{x}^{(k+1)}) \geq (1-\gamma)f(\boldsymbol{x}^{(k)}) + \gamma(1-\gamma)^{t^{(k)}/\gamma}f(\boldsymbol{x}^*) - \frac{LD^2}{2}\gamma^2,$*

which is derived by a combination of the quadratic lower bound in (5), Lemma 2 and Lemma 3.

Remarks on the two algorithms. Notice that though the TWO-PHASE algorithm has a worse guarantee than the non-monotone FRANK-WOLFE variant, it is still of interest: i) It allows flexibility in using a wide range of existing solvers for finding an (approximately) stationary point. ii) The guarantees that we present rely on a worst-case analysis. The empirical performance of the TWO-PHASE algorithm is often comparable or better than that of the FRANK-WOLFE variant. This suggests to explore more properties in concrete problems that may favor the TWO-PHASE algorithm, which we leave for future work.

5 Experimental Results

We test the performance of the analyzed algorithms, while considering the following baselines: 1) QUADPROGIP [39], which is a global solver for non-convex quadratic programming; 2) Projected gradient ascent (PROJGRAD) with diminishing step sizes ($\frac{1}{k+1}$, k starts from 0). We run all the algorithms for 100 iterations. For the subroutine (Algorithm 3) of TWO-PHASE FRANK-WOLFE, we set $\epsilon_1 = \epsilon_2 = 10^{-6}, K_1 = K_2 = 100$. All the synthetic results are the average of 20 repeated experiments. All experiments were implemented using MATLAB. Source code can be found at: https://github.com/bianan/non-monotone-dr-submodular.

5.1 DR-submodular Quadratic Programming

As a state-of-the-art global solver, QUADPROGIP[3] [39] can find the global optimum (possibly in exponential time), which were used to calculate the approximation ratios. Our problem instances are synthetic DR-submodular quadratic objectives with down-closed polytope constraints, i.e., $f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top\mathbf{H}\boldsymbol{x} + \boldsymbol{h}^\top\boldsymbol{x} + c$ and $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}_+^n \mid \mathbf{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \leq \bar{\boldsymbol{u}}, \mathbf{A} \in \mathbb{R}_{++}^{m \times n}, \boldsymbol{b} \in \mathbb{R}_+^m\}$. Both objective and constraints were randomly generated, in the following two manners:

1) Uniform distribution. $\mathbf{H} \in \mathbb{R}^{n \times n}$ is a symmetric matrix with uniformly distributed entries in $[-1, 0]$; $\mathbf{A} \in \mathbb{R}^{m \times n}$ has uniformly distributed entries in $[\nu, \nu + 1]$, where $\nu = 0.01$ is a small positive constant in order to make entries of $\mathbf{A}$ strictly positive.
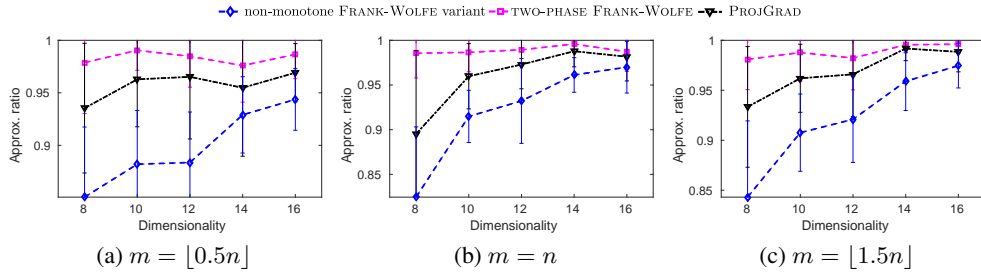

Figure 1: Results on DR-submodular quadratic instances with uniform distribution.

2) Exponential distribution. The entries of $-\mathbf{H}$ and $\mathbf{A}$ were sampled from exponential distributions $\text{Exp}(\lambda)$ (For a random variable $y \geq 0$, its probability density function is $\lambda e^{-\lambda y}$, and for $y < 0$, its density is 0). Specifically, each entry of $-\mathbf{H}$ was sampled from $\text{Exp}(1)$, then the matrix $-\mathbf{H}$ was made to be symmetric. Each entry of $\mathbf{A}$ was sampled from $\text{Exp}(0.25) + \nu$, where $\nu = 0.01$ is a small positive constant.

In both the above two cases, we set $\boldsymbol{b} = \mathbf{1}^m$, and $\bar{\boldsymbol{u}}$ to be the tightest upper bound of $\mathcal{P}$ by $\bar{u}_j = \min_{i \in [m]} \frac{b_i}{A_{ij}}, \forall j \in [n]$. In order to make f non-monotone, we set $\boldsymbol{h} = -0.2 * \mathbf{H}^\top\bar{\boldsymbol{u}}$. To make sure that f is non-negative, we first of all solve the problem $\min_{\boldsymbol{x} \in \mathcal{P}} \frac{1}{2}\boldsymbol{x}^\top\mathbf{H}\boldsymbol{x} + \boldsymbol{h}^\top\boldsymbol{x}$ using QUADPROGIP, let the solution to be $\hat{\boldsymbol{x}}$, then set $c = -f(\hat{\boldsymbol{x}}) + 0.1 * |f(\hat{\boldsymbol{x}})|$.

---

[3]We used the open source code provided by [39], and the IBM CPLEX optimization studio https://www.ibm.com/jm-en/marketplace/ibm-ilog-cplex as the subroutine.
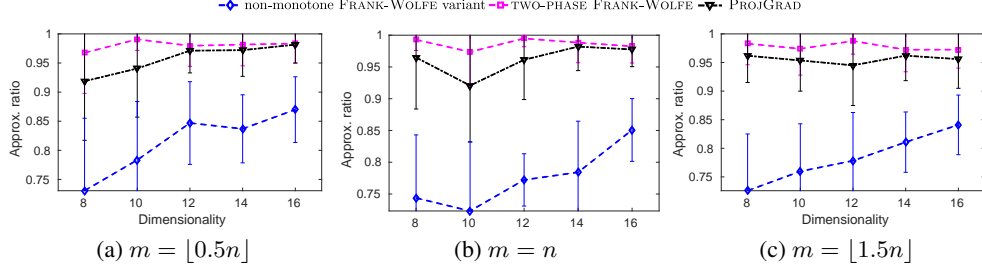
Figure 2: Results on quadratic instances with exponential distribution.

The approximation ratios w.r.t. dimensionalities (n) are plotted in Figures 1 and 2, for the two manners of data generation. We set the number of constraints to be $m = \lfloor 0.5n \rfloor$, $m = n$ and $m = \lfloor 1.5n \rfloor$ in Figures 1a to 1c (and Figures 2a to 2c), respectively.

One can see that TWO-PHASE FRANK-WOLFE usually performs the best, PROJGRAD follows, and non-monotone FRANK-WOLFE variant is the last. The good performance of TWO-PHASE FRANK-WOLFE can be partially explained by the strong DR-submodularity of quadratic functions according to Theorem 1. Performance of the two analyzed algorithms is consistent with the theoretical bounds: the approximation ratios of FRANK-WOLFE variant are always much higher than $1/e$.

5.2 Maximizing Softmax Extensions

With some derivation, one can see the derivative of the softmax extension in (3) is: $\nabla_i f(\boldsymbol{x}) = \mathrm{tr}((\mathrm{diag}(\boldsymbol{x})(\mathbf{L} - \mathbf{I}) + \mathbf{I})^{-1}(\mathbf{L}-\mathbf{I})_i), \forall i \in [n]$, where $(\mathbf{L}-\mathbf{I})_i$ denotes the matrix obtained by zeroing all entries except the ith row of $(\mathbf{L} - \mathbf{I})$. Let $\mathbf{C} := (\mathrm{diag}(\boldsymbol{x})(\mathbf{L} - \mathbf{I}) + \mathbf{I})^{-1}, \mathbf{D} := (\mathbf{L} - \mathbf{I})$, one can see that $\nabla_i f(\boldsymbol{x}) = \mathbf{D}_{i\cdot}^{\top}\mathbf{C}_{\cdot i}$, which gives an efficient way to calculate the gradient $\nabla f(\boldsymbol{x})$.
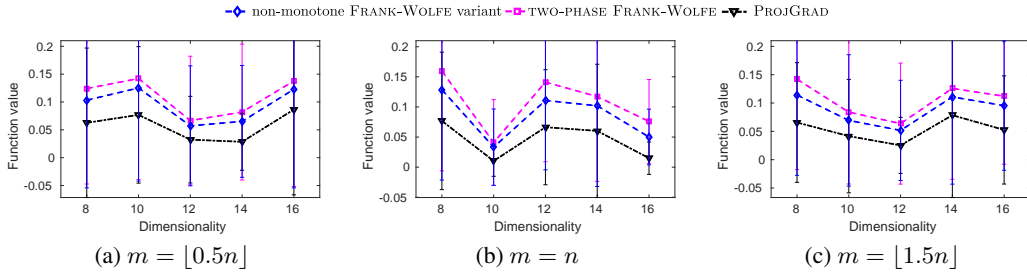

Figure 3: Results on softmax instances with polytope constraints generated from uniform distribution.

Results on synthetic data. We generate the softmax objectives (see (3)) in the following way: first generate the n eigenvalues $\boldsymbol{d} \in \mathbb{R}_+^n$, each randomly distributed in $[0, 1.5]$, and set $\mathbf{D} = \mathrm{diag}(\boldsymbol{d})$. After generating a random unitary matrix $\mathbf{U}$, we set $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{U}^{\top}$. One can verify that $\mathbf{L}$ is positive semidefinite and has eigenvalues as the entries of $\boldsymbol{d}$.

We generate the down-closed polytope constraints in the same form and same way as that for DR-submodular quadratic functions, except for setting $\boldsymbol{b} = 2 * \mathbf{1}^m$. Function values returned by different solvers w.r.t. n are shown in Figure 3, for which the random polytope constraints were generated with uniform distribution (results for which the random polytope constraints were generated with exponential distribution are deferred to Appendix G). The number of constraints was set to be $m = \lfloor 0.5n \rfloor$, $m = n$ and $m = \lfloor 1.5n \rfloor$ in Figures 3a to 3c, respectively. One can observe that TWO-PHASE FRANK-WOLFE still has the best performance, the non-monotone FRANK-WOLFE variant follows, and PROJGRAD has the worst performance.

Real-world results on matched summarization. The task of "matched summarization" is to select a set of document *pairs* out of a corpus of documents, such that the two documents within a pair are similar, and the overall set of pairs is as diverse as possible. The motivation for this task is very practical: it could be, for example, to compare the opinions of various politicians on a range of representative topics.

8

In our experiments, we used a similar setting to the one in [20]. We experimented on the 2012 US Republican debates data, which consists of 8 candidates: Bachman, Gingrich, Huntsman, Paul, Perry, Romney and Santorum. Each task involves one pair of candidates, so in total there are $28 = 7 * 8/2$ tasks. Figure 4a plots the averaged function values returned by the three solvers over 28 tasks, w.r.t. different values of a hyperparameter reflecting the matching quality (details see [20]).

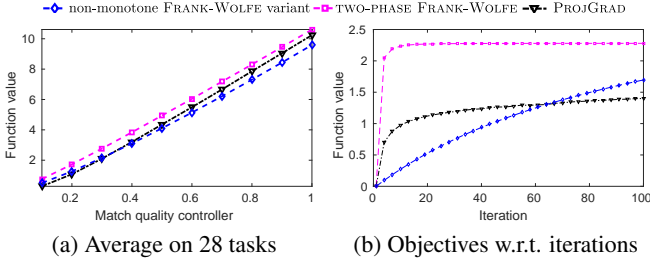(a) Average on 28 tasks (b) Objectives w.r.t. iterations

Figure 4: Results on 2012 US Republican debates data.

Figure 4b traces the objectives w.r.t. iterations for a specific candidate pair (Bachman, Romney). For TWO-PHASE FRANK-WOLFE, the objectives of the selected phase were plotted. One can see that TWO-PHASE FRANK-WOLFE also achieves the best performance, while the performance of non-monotone FRANK-WOLFE variant and PROJGRAD is comparable.

6 Related Work

Submodular optimization and, more broadly, non-convex optimization are extensively studied in the literature, which renders it very difficult comprehensively surveying all previous work. Here we only briefly summarize some of the most related papers.

Submodular optimization over integer-lattice and continuous domains. Many results from submodular set function optimization have been generalized to the integer-lattice case [34; 33; 12; 24]. Of particular interest is the reduction [12] from an integer-lattice DR-submodular maximization problem to a submodular set function maximization problem. Submodular optimization over continuous domains has attracted considerable attention recently [3; 4; 36]. Two classes of functions that are covered by continuous submodularity are the Lovasz extensions [28] and multilinear extensions [6] of submodular set functions. Particularly, multilinear extensions of submodular set functions are also continuous DR-submodular [3], but with the special property that they are coordinate-wise linear. Combined with the rounding technique of contention resolution [7], maximizing multilinear extensions [38; 19; 13; 8; 11] has become the state-of-the-art method for submodular set function maximization. Some of the techniques in maximizing multilinear extensions [13; 7; 8] have inspired this work. However, we are the first to explore the rich properties and devise algorithms for the general constrained DR-submodular maximization problem over continuous domains.

Non-convex optimization. Non-convex optimization receives a surge of attention in the past years. One active research topic is to reach a stationary point for unconstrained optimization [35; 32; 1] or constrained optimization [18; 27]. However, without proper assumptions, a stationary point may not lead to any global approximation guarantee. The local-global relation (in Proposition 1) provides a strong relation between (approximately) stationary points and global optimum, thus making it flexible to incorporate progress in this area.

7 Conclusion

We have studied the problem of constrained non-monotone DR-submodular continuous maximization. We explored the structural properties of such problems, and established a local-global relation. Based on these properties, we presented a TWO-PHASE algorithm with a $1/4$ approximation guarantee, and a non-monotone FRANK-WOLFE variant with a $1/e$ approximation guarantee. We further generalized submodular continuous function over conic lattices, which enabled us to model a larger class of applications. Lastly, our theoretical findings were verified by synthetic and real-world experiments.

References

[1] Allen-Zhu, Zeyuan and Hazan, Elad. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning (ICML)*, pp. 699–707, 2016.

[2] Antoniadis, Anestis, Gijbels, Irène, and Nikolova, Mila. Penalized likelihood regression for generalized linear models with non-quadratic penalties. *Annals of the Institute of Statistical Mathematics*, 63(3):585–615, 2011.

[3] Bach, Francis. Submodular functions: from discrete to continous domains. *arXiv preprint arXiv:1511.00394*, 2015.

[4] Bian, Andrew An, Mirzasoleiman, Baharan, Buhmann, Joachim M., and Krause, Andreas. Guaranteed non-convex optimization: Submodular maximization over continuous domains. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 111–120, 2017.

[5] Boyd, Stephen and Vandenberghe, Lieven. *Convex optimization*. Cambridge university press, 2004.

[6] Calinescu, Gruia, Chekuri, Chandra, Pál, Martin, and Vondrák, Jan. Maximizing a submodular set function subject to a matroid constraint. In *Integer programming and combinatorial optimization*, pp. 182–196. Springer, 2007.

[7] Chekuri, Chandra, Vondrák, Jan, and Zenklusen, Rico. Submodular function maximization via the multilinear relaxation and contention resolution schemes. *SIAM Journal on Computing*, 43 (6):1831–1879, 2014.

[8] Chekuri, Chandra, Jayram, TS, and Vondrák, Jan. On multiplicative weight updates for concave and submodular function maximization. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pp. 201–210. ACM, 2015.

[9] Djolonga, Josip and Krause, Andreas. From map to marginals: Variational inference in bayesian submodular models. In *Neural Information Processing Systems (NIPS)*, pp. 244–252, 2014.

[10] Eghbali, Reza and Fazel, Maryam. Designing smoothing functions for improved worst-case competitive ratio in online optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3279–3287. 2016.

[11] Ene, Alina and Nguyen, Huy L. Constrained submodular maximization: Beyond 1/e. In *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*, pp. 248–257, 2016.

[12] Ene, Alina and Nguyen, Huy L. A reduction for optimizing lattice submodular functions with diminishing returns. *arXiv preprint arXiv:1606.08362*, 2016.

[13] Feldman, Moran, Naor, Joseph, and Schwartz, Roy. A unified continuous greedy algorithm for submodular maximization. In *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*, pp. 570–579. IEEE, 2011.

[14] Friedland, S and Gaubert, S. Submodular spectral functions of principal submatrices of a hermitian matrix, extensions and applications. *Linear Algebra and its Applications*, 438(10): 3872–3884, 2013.

[15] Fuchssteiner, Benno and Lusky, Wolfgang. *Convex cones*, volume 56. Elsevier, 2011.

[16] Fujishige, Satoru. *Submodular functions and optimization*, volume 58. Elsevier, 2005.

[17] Garg, Vijay K. *Introduction to lattice theory with computer science applications*. John Wiley & Sons, 2015.

[18] Ghadimi, Saeed, Lan, Guanghui, and Zhang, Hongchao. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

[19] Gharan, Shayan Oveis and Vondrák, Jan. Submodular maximization by simulated annealing. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pp. 1098–1116. Society for Industrial and Applied Mathematics, 2011.

[20] Gillenwater, Jennifer, Kulesza, Alex, and Taskar, Ben. Near-optimal map inference for determinantal point processes. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 2735–2743, 2012.

[21] Hassani, Hamed, Soltanolkotabi, Mahdi, and Karbasi, Amin. Gradient methods for submodular maximization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 5837–5847, 2017.

[22] Ito, Shinji and Fujimaki, Ryohei. Large-scale price optimization via network flow. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 3855–3863, 2016.

[23] Iwata, Satoru, Fleischer, Lisa, and Fujishige, Satoru. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.

[24] Khodabakhsh, Ali and Nikolova, Evdokia. Maximizing non-monotone dr-submodular functions with cardinality constraints. *arXiv preprint arXiv:1611.09474*, 2016.

[25] Krause, Andreas and Golovin, Daniel. Submodular function maximization. *Tractability: Practical Approaches to Hard Problems*, 3:19, 2012.

[26] Kulesza, Alex, Taskar, Ben, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012.

[27] Lacoste-Julien, Simon. Convergence rate of frank-wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

[28] Lovász, László. Submodular functions and convexity. In *Mathematical Programming The State of the Art*, pp. 235–257. Springer, 1983.

[29] Montanari, Andrea and Richard, Emile. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *IEEE Transactions on Information Theory*, 62(3): 1458–1484, 2016.

[30] Motzkin, Theodore S and Straus, Ernst G. Maxima for graphs and a new proof of a theorem of turán. *Canad. J. Math*, 17(4):533–540, 1965.

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

[32] Reddi, Sashank J., Sra, Suvrit, Poczos, Barnabas, and Smola, Alexander J. Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1145–1153. 2016.

[33] Soma, Tasuku and Yoshida, Yuichi. A generalization of submodular cover via the diminishing return property on the integer lattice. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 847–855, 2015.

[34] Soma, Tasuku, Kakimura, Naonori, Inaba, Kazuhiro, and Kawarabayashi, Ken-ichi. Optimal budget allocation: Theoretical guarantee and efficient algorithm. In *International Conference on Machine Learning (ICML)*, pp. 351–359, 2014.

[35] Sra, Suvrit. Scalable nonconvex inexact proximal splitting. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 530–538, 2012.

[36] Staib, Matthew and Jegelka, Stefanie. Robust budget allocation via continuous submodular functions. In *International Conference on Machine Learning (ICML)*, pp. 3230–3240, 2017.

[37] Topkis, Donald M. Minimizing a submodular function on a lattice. *Operations research*, 26 (2):305–321, 1978.

[38] Vondrák, Jan. Optimal approximation for the submodular welfare problem in the value oracle model. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 67–74, 2008.

[39] Xia, Wei, Vera, Juan, and Zuluaga, Luis F. Globally solving non-convex quadratic programs via linear integer programming techniques. *arXiv preprint arXiv:1511.02423*, 2015.

[40] Zass, Ron and Shashua, Amnon. Nonnegative sparse pca. *Advances in Neural Information Processing Systems (NIPS)*, pp. 1561–1568, 2007.

Appendix

A Continuous Submodular Functions on Conic Lattices and the Reduction

Motivated by the objectives that can not be modeled by continuous submodular functions, we consider the more general submodular continuous functions over lattices induced by conic inequalities. Furthermore we provide a reduction to the original (DR-)submodular optimization problem.

A.1 Definitions and Properties

Let us look at the proper cone that will be used to define a conic inequality firstly. A cone $\mathcal{K} \subseteq \mathbb{R}^n$ is a *proper cone* if it is convex, closed, solid (having nonempty interior) and pointed (contains no line, i.e., $\boldsymbol{x} \in \mathcal{K}, -\boldsymbol{x} \in \mathcal{K}$ implies $\boldsymbol{x} = 0$). A proper cone $\mathcal{K}$ can be used to define a conic inequality (a.k.a. generalized inequality [5, Chapter 2.4]): $\boldsymbol{a} \preceq_{\mathcal{K}} \boldsymbol{b}$ iff $\boldsymbol{b} - \boldsymbol{a} \in \mathcal{K}$, which also defines a partial ordering since the binary relation $\preceq_{\mathcal{K}}$ is reflexive, antisymmetric and transitive. Then it is easy to see that $(\mathcal{X}, \preceq_{\mathcal{K}})$ is a partial ordered set (poset).

If two elements $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{X}$ have a least upper bound (greatest lower bound), it is denoted as the "join": $\boldsymbol{a} \vee \boldsymbol{b}$ (the "meet": $\boldsymbol{a} \wedge \boldsymbol{b}$). A lattice is a poset that contains the join and meet of each pair of its elements [17]. A "lattice cone" [15] is the proper cone that can be used to define a lattice. Note that not all conic inequalities can be used to define a lattice. For example, the positive semidefine cone $\mathcal{K}_{\mathrm{PSD}} = \{\mathbf{A} \in \mathbb{R}^{n \times n} | \mathbf{A} \text{ is symmetric}, \mathbf{A} \succeq 0\}$ is a proper cone, but its induced ordering can not be used to define a lattice. There is a simple counter example to show this in Appendix F.2.

Specifically, we name the lattice that can be defined through a conic inequality as "conic lattice", since it is of particular interest for modeling the real-world applications in this paper.

Definition 4 (Conic Lattice). *Given a poset $(\mathcal{X}, \preceq_{\mathcal{K}})$ induced by the conic inequality $\preceq_{\mathcal{K}}$, if there exist joint and meet operations for every pair of elements $(\boldsymbol{a}, \boldsymbol{b})$ in $\mathcal{X} \times \mathcal{X}$, s.t. $\boldsymbol{a} \vee \boldsymbol{b}$ and $\boldsymbol{a} \wedge \boldsymbol{b}$ are still in $\mathcal{X}$, then we call $(\mathcal{X}, \preceq_{\mathcal{K}})$ a conic lattice.*

In one word, a conic lattice $(\mathcal{X}, \preceq_{\mathcal{K}})$ is a lattice induced by a conic inequality $\preceq_{\mathcal{K}}$. In the following we introduce a class of conic lattices to model the applications in this work. We further provide a general characterization about submodularity on this conic lattice.

Orthant conic lattice. Given a sign vector $\boldsymbol{\alpha} \in \{\pm 1\}^n$, the orthant cone is defined as $\mathcal{K}_{\boldsymbol{\alpha}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid x_i \alpha_i \geq 0, \forall i \in [n]\}$, one can see that $\mathcal{K}_{\boldsymbol{\alpha}}$ is a proper cone. For any two points $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{X}$, one can further define the join and meet operations: $(\boldsymbol{a} \vee \boldsymbol{b})_i := \alpha_i \max\{\alpha_i a_i, \alpha_i b_i\}$, $(\boldsymbol{a} \wedge \boldsymbol{b})_i := \alpha_i \min\{\alpha_i a_i, \alpha_i b_i\}, \forall i \in [n]$. Then one can show that the poset $(\mathcal{X}, \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}})$ is a conic lattice.

A function $f : \mathcal{X} \mapsto \mathbb{R}$ is submodular on a lattice [37; 16] if for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{X}$, it holds that,

$$f(\boldsymbol{x}) + f(\boldsymbol{y}) \geq f(\boldsymbol{x} \vee \boldsymbol{y}) + f(\boldsymbol{x} \wedge \boldsymbol{y}). \tag{12}$$

One can establish the characterizations of submodularity on the orthant conic lattice $(\mathcal{X}, \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}})$ similarly as that in [4]:

Proposition 2 (Characterizations of Submodularity on Orthant Conic Lattice $(\mathcal{X}, \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}})$). *If a function f is is submodular on the lattice $(\mathcal{X}, \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}})$ (called $\mathcal{K}_{\boldsymbol{\alpha}}$-submodular), then we have the following two equivalent characterizations:*
a) $\forall \boldsymbol{a}, \boldsymbol{b} \in \mathcal{X}$ s.t. $\boldsymbol{a} \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{b}$, $\forall i$ s.t. $a_i = b_i$, $\forall k \in \mathbb{R}_+$ s.t. $(k\boldsymbol{e}_i + \boldsymbol{a})$ and $(k\boldsymbol{e}_i + \boldsymbol{b})$ are still in $\mathcal{X}$, it holds that,

$$\alpha_i[f(k\boldsymbol{e}_i + \boldsymbol{a}) - f(\boldsymbol{a})] \geq \alpha_i[f(k\boldsymbol{e}_i + \boldsymbol{b}) - f(\boldsymbol{b})]. \quad (\texttt{weak DR}) \tag{13}$$

b) If f is twice differentiable, then $\forall \boldsymbol{x} \in \mathcal{X}$ it holds,

$$\alpha_i \alpha_j \nabla_{ij}^2 f(\boldsymbol{x}) \leq 0, \ \forall i, j \in [n], i \neq j. \tag{14}$$

Proposition 2 can be proved by directly generalizing the proof of Proposition 1 in [4], proof is omitted here due to the high similarity. Next, we generalize the definition of DR-submodularity to the conic lattice $(\mathcal{X}, \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}})$:

Definition 5 ($\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular). *A function $f : \mathcal{X} \mapsto \mathbb{R}$ is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular if $\forall \boldsymbol{a}, \boldsymbol{b} \in \mathcal{X}$ s.t. $\boldsymbol{a} \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{b}$, $\forall i \in [n], \forall k \in \mathbb{R}_+$ s.t. $(k\boldsymbol{e}_i + \boldsymbol{a})$ and $(k\boldsymbol{e}_i + \boldsymbol{b})$ are still in $\mathcal{X}$, it holds that,*

$$\alpha_i[f(k\boldsymbol{e}_i + \boldsymbol{a}) - f(\boldsymbol{a})] \geq \alpha_i[f(k\boldsymbol{e}_i + \boldsymbol{b}) - f(\boldsymbol{b})]. \tag{15}$$

In correspondence to the relation between DR-submodularity and submodularity over continuous domains (Proposition 2 in [4]), one can easily get the similar relation (with highly similar proof) in bellow:

Proposition 3 ($\mathcal{K}_{\boldsymbol{\alpha}}$-submodular + coordinate-wise concave $\Leftrightarrow$ $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular). *A function f is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular iff it is $\mathcal{K}_{\boldsymbol{\alpha}}$-submodular and coordinate-wise concave.*

Combining (14) and Proposition 3, one can show that if f is twice differentiable and $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular, then $\forall \boldsymbol{x} \in \mathcal{X}$ it holds that,

$$\alpha_i \alpha_j \nabla_{ij}^2 f(\boldsymbol{x}) \leq 0, \ \forall i, j \in [n]. \tag{16}$$

Similarly, a function f is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-supermodular iff $-f$ is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular.

Remark: We only consider the orthant conic lattice $(\mathcal{X}, \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}})$ here, since it can already model the applications in this paper. However, it is noteworthy that the framework can be generalized to arbitrary conic lattices, which may be of interest to model more complicated applications. We left this as future exploration.

A.2 A Reduction to Optimizing Submodular Functions over Continuous Domains

To be succint, in this section we only discuss the reduction for the $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular maximization problems. However, it is easy to see that the reduction works for all kinds of $\mathcal{K}_{\boldsymbol{\alpha}}$-submodular optimization problems, e.g., $\mathcal{K}_{\boldsymbol{\alpha}}$-submodular minimization problem.

Suppose g is a $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular function, and the $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular maximization problem is $\max_{\boldsymbol{y} \in \mathcal{P}'} g(\boldsymbol{y})$, where $\mathcal{P}' = \{\boldsymbol{y} \in \mathbb{R}^n | h_i(\boldsymbol{y}) \leq b_i, \forall i \in [m], \boldsymbol{y} \succeq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{0}\}$ is down-closed w.r.t. the conic inequality $\preceq_{\mathcal{K}_{\boldsymbol{\alpha}}}$. The down-closedness here means if $\boldsymbol{a} \in \mathcal{P}'$ and $\boldsymbol{0} \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{b} \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{a}$, then $\boldsymbol{b} \in \mathcal{P}'$ as well.

Let $\mathbf{A} := \text{diag}(\boldsymbol{\alpha})$, and a function $f(\boldsymbol{x}) := g(\mathbf{A}\boldsymbol{x})$. One can see that if g is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular, then f is DR-submodular: assume wlog.[4] that g is twice differentiable, then $\nabla^2 f(\boldsymbol{x}) = \mathbf{A}^\top \nabla^2 g \mathbf{A}$, and $\nabla_{ij}^2 f(\boldsymbol{x}) = \alpha_i \alpha_j \nabla_{ij}^2 g \leq 0$, so f is DR-submodular.

By the affine transformation $\boldsymbol{y} := \mathbf{A}\boldsymbol{x}$, one can transform the $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular maximization problem to be a DR-submodular maximization problem $\max_{\boldsymbol{x} \in \mathcal{P}} g(\mathbf{A}\boldsymbol{x})$, where $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^n | h_i(\mathbf{A}\boldsymbol{x}) \leq b_i, \forall i \in [m], \mathbf{A}\boldsymbol{x} \succeq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{0}\}$ is down-closed w.r.t. the ordinary component-wise inequality $\leq$. To verify the down-closedness of $\mathcal{P}$ w.r.t. to the ordinary inequality $\leq$ here, let $\boldsymbol{y}_1 = \mathbf{A}\boldsymbol{x}_1 \in \mathcal{P}'$ (so $\boldsymbol{x}_1 \in \mathcal{P}$). Suppose there is a point $\boldsymbol{y}_2 = \mathbf{A}\boldsymbol{x}_2$ s.t. $\boldsymbol{0} \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{y}_2 \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{y}_1$. From the down-closedness of $\mathcal{P}'$, we know that $\boldsymbol{y}_2 \in \mathcal{P}'$, thus $\boldsymbol{x}_2 \in \mathcal{P}$. Looking at $\boldsymbol{0} \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{y}_2 \preceq_{\mathcal{K}_{\boldsymbol{\alpha}}} \boldsymbol{y}_1$, it is equivalent to that $0 \leq \boldsymbol{x}_2 \leq \boldsymbol{x}_1$. Thus we establish the down-closedness of $\mathcal{P}$.

Given the reduction, we can reuse the algorithms for the original DR-submodular maximization problem (P).

A.3 Proof for the Logistic Loss in Section 2.1

Remember that the logistic loss is:

$$l(\boldsymbol{x}) = \frac{1}{m} \sum_{j=1}^{m} f_j(\boldsymbol{x}) = \frac{1}{m} \sum_{j=1}^{m} \log(1 + \exp(-y_j \boldsymbol{x}^\top \boldsymbol{z}^j)) \tag{17}$$

Claim 3. *$l(\boldsymbol{x})$ in (17) is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-supermodular.*

Proof of Claim 3. To show that $l(\boldsymbol{x})$ is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-supermodular, we can check the second-order condition in (16), that is, whether it holds that $\alpha_p \alpha_q \nabla_{pq}^2 l(\boldsymbol{x}) \geq 0, \ \forall p, q \in [n]$. One can easily see

---

[4]If not one can still use other equivalent characterizations, for instance, the characterization in (12) or in (13) to formulate this.

that,

$$\frac{\partial l(\boldsymbol{x})}{\partial x_p} = \frac{1}{m} \sum\nolimits_{j=1}^{m} \frac{-y_j z_p^j}{\exp\left(y_j \boldsymbol{x}^\top \boldsymbol{z}^j\right) + 1}$$

$$\frac{\partial^2 l(\boldsymbol{x})}{\partial x_p \partial x_q} = \frac{1}{m} \sum\nolimits_{j=1}^{m} \frac{\exp\left(y_j \boldsymbol{x}^\top \boldsymbol{z}^j\right)}{[\exp\left(y_j \boldsymbol{x}^\top \boldsymbol{z}^j\right) + 1]^2} z_p^j z_q^j.$$

Since $\alpha_p = \text{sign}(z_p^j)$, so $\alpha_p \alpha_q \nabla_{pq}^2 l(\boldsymbol{x}) \geq 0$, $\forall p, q \in [n]$. Thus $l(\boldsymbol{x})$ in (17) is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-supermodular according to (16). ∎

B More Applications

We present more applications that fall into submodular or $\mathcal{K}_{\boldsymbol{\alpha}}$-submodular optimization problems. One class of notable examples are the objectives studied in [10] in the online setting. These objectives are captured by the DR-submodular property over continuous domains. One can also refer to Section 2.2 in [3] to see more examples.

DR-submodular quadratic functions. Price optimization with continuous prices is a DR-submodular quadratic optimization problem [22]. Another representative class of DR-submodular quadratic objectives arises when computing the stability number $s(G)$ of a graph $G = (V, E)$ [30], $s(G)^{-1} = \min_{\boldsymbol{x} \in \Delta} \boldsymbol{x}^\top (\mathbf{A} + \mathbf{I}) \boldsymbol{x}$, where $\mathbf{A}$ is the adjacency matrix of the graph G, Δ is the standard simplex. This is an instance of a convex-constrained DR-submodular maximization problem.

Non-negative PCA (NN-PCA). NN-PCA [40; 29] is widely used as alternative models of PCA for dimension reduction, since its projection involves only non-negative weights—a required property in fields like economics, bioinformatics and computer vision. For a given set of m data points $\boldsymbol{z}^j \in \mathbb{R}^n, j \in [m]$, NN-PCA aims to solve the following non-convex optimization problem:

$$\min_{\|\boldsymbol{x}\|_2 \leq 1, \boldsymbol{x} \geq 0} f(\boldsymbol{x}) := -\frac{1}{2} \boldsymbol{x}^\top \left(\sum\nolimits_{j=1}^{m} \boldsymbol{z}^j \boldsymbol{z}^{j\top} \right) \boldsymbol{x}. \tag{18}$$

Let $\mathbf{A} = \sum_{j=1}^{m} \boldsymbol{z}^j \boldsymbol{z}^{j\top}$, one can see that,

$$A_{pp} = \sum\nolimits_{j=1}^{m} (z_p^j)^2 \geq 0, \ A_{pq} = \sum\nolimits_{j=1}^{m} z_p^j z_q^j = A_{qp}.$$

Let us make the following weak assumption: For one dimension/feature i, all the data points have the same sign, i.e., $\text{sign}(z_i^j)$ is the same for all $j \in [m]$ (which can be achieved by easily scaling if not). Now, by choosing the sign vector $\boldsymbol{\alpha} \in \{\pm 1\}^n$ to be $\alpha_p = \text{sign}(z_p^j), \forall p \in [n]$, one can easily verify that $A_{pq} \alpha_p \alpha_q \geq 0, \forall p, q \in [n]$. Notice that $\nabla^2 f$ in (18) is $-\mathbf{A}$, so it holds that $\alpha_p \alpha_q \nabla_{pq}^2 f \leq 0, \forall p, q \in [n]$, thus $f(\boldsymbol{x})$ is $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular according to (16). Thus we can treat (18) as a constrained $\mathcal{K}_{\boldsymbol{\alpha}}$-DR-submodular minimization problem.

Submodular spectral functions. As discussed by [3], submodular spectral functions [14] in the following form are DR-submodular,

$$f(\boldsymbol{x}) = \log \det \left(\sum\nolimits_{i=1}^{n} x_i \mathbf{A}_i \right), \boldsymbol{x} \in \mathbb{R}_+^n, \tag{19}$$

where $\mathbf{A}_i$ are positive definite matrices. One can check the DR-submodularity of $f(\boldsymbol{x})$ by checking its second-order-derivatives.

C The Subroutine Algorithm

Algorithm 3 is taken from [27], the only difference lies in the output: we output the solution $\boldsymbol{x}^{(k')}$ with the minimum non-stationarity, which is needed to apply the local-global relation. While [27] outputs the solution in the last step. Since $C_f(\mathcal{P})$ is generally hard to evaluate, we used the classical FRANK-WOLFE step size $\frac{2}{k+2}$ in the experiments.

Algorithm 3: NON-CONVEX FRANK-WOLFE $(f, \mathcal{P}, K, \epsilon, \boldsymbol{x}^{(0)})$[27]

Input: $\max_{\boldsymbol{x} \in \mathcal{P}} f(\boldsymbol{x})$, $\mathcal{P}$: convex set, K: number of iterations, ϵ: stopping tolerance

1 **for** $k = 0, ..., K$ **do**

2   find $\boldsymbol{v}^{(k)}$ s.t. $\langle \boldsymbol{v}^{(k)}, \nabla f(\boldsymbol{x}^{(k)}) \rangle \geq \max_{\boldsymbol{v} \in \mathcal{P}} \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}^{(k)}) \rangle$;   *// LMO*

3    $\boldsymbol{d}_k \leftarrow \boldsymbol{v}^{(k)} - \boldsymbol{x}^{(k)}$, $g_k := \langle \boldsymbol{d}_k, \nabla f(\boldsymbol{x}^{(k)}) \rangle$;   *// g_k: non-stationarity measure*

4   **if** $g_k \leq \epsilon$ **then return** $\boldsymbol{x}^{(k)}$;

5   Option I: $\gamma_k \in \arg\min_{\gamma \in [0,1]} f(\boldsymbol{x}^{(k)} + \gamma \boldsymbol{d}^{(k)})$, Option II: $\gamma_k \leftarrow \min\{\frac{g_k}{C}, 1\}$ for
   $C \geq C_f(\mathcal{P})$;

6    $\boldsymbol{x}^{(k+1)} \leftarrow \boldsymbol{x}^{(k)} + \gamma_k \boldsymbol{d}^{(k)}$;

Output: $\boldsymbol{x}^{(k')}$ and $g_{k'} = \min_{0 \leq k \leq K} g_k$;   *// modified output solution compared to [27]*

D   Proofs for Properties

D.1   Proof of Lemma 1

Proof of Lemma 1. Since f is DR-submodular, so it is concave along any direction $\boldsymbol{v} \in \pm\mathbb{R}_+^n$. We know that $\boldsymbol{x} \vee \boldsymbol{y} - \boldsymbol{x} \geq 0$ and $\boldsymbol{x} \wedge \boldsymbol{y} - \boldsymbol{x} \leq 0$, so from the strong DR-submodularity in (6),

$$f(\boldsymbol{x} \vee \boldsymbol{y}) - f(\boldsymbol{x}) \leq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} \vee \boldsymbol{y} - \boldsymbol{x} \rangle - \frac{\mu}{2} \|\boldsymbol{x} \vee \boldsymbol{y} - \boldsymbol{x}\|^2,$$

$$f(\boldsymbol{x} \wedge \boldsymbol{y}) - f(\boldsymbol{x}) \leq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} \wedge \boldsymbol{y} - \boldsymbol{x} \rangle - \frac{\mu}{2} \|\boldsymbol{x} \wedge \boldsymbol{y} - \boldsymbol{x}\|^2.$$

Summing the above two inequalities and notice that $\boldsymbol{x} \vee \boldsymbol{y} + \boldsymbol{x} \wedge \boldsymbol{y} = \boldsymbol{x} + \boldsymbol{y}$, we arrive,

$$(\boldsymbol{y} - \boldsymbol{x})^\top \nabla f(\boldsymbol{x}) \geq f(\boldsymbol{x} \vee \boldsymbol{y}) + f(\boldsymbol{x} \wedge \boldsymbol{y}) - 2f(\boldsymbol{x}) + \frac{\mu}{2}(\|\boldsymbol{x} \vee \boldsymbol{y} - \boldsymbol{x}\|^2 + \|\boldsymbol{x} \wedge \boldsymbol{y} - \boldsymbol{x}\|^2)$$

$$= f(\boldsymbol{x} \vee \boldsymbol{y}) + f(\boldsymbol{x} \wedge \boldsymbol{y}) - 2f(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2,$$

the last equality holds since $\|\boldsymbol{x} \vee \boldsymbol{y} - \boldsymbol{x}\|^2 + \|\boldsymbol{x} \wedge \boldsymbol{y} - \boldsymbol{x}\|^2 = \|\boldsymbol{y} - \boldsymbol{x}\|^2$.   ■

D.2   Proof of Proposition 1

Proof of Proposition 1. Consider the point $\boldsymbol{z}^* := \boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{x} = (\boldsymbol{x}^* - \boldsymbol{x}) \vee 0$. One can see that: 1) $0 \leq \boldsymbol{z}^* \leq \boldsymbol{x}^*$; 2) $\boldsymbol{z}^* \in \mathcal{P}$ (down-closedness); 3) $\boldsymbol{z}^* \in \mathcal{Q}$ (because of $\boldsymbol{z}^* \leq \bar{\boldsymbol{u}} - \boldsymbol{x}$). From Lemma 1,

$$\langle \boldsymbol{x}^* - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + 2f(\boldsymbol{x}) \geq f(\boldsymbol{x} \vee \boldsymbol{x}^*) + f(\boldsymbol{x} \wedge \boldsymbol{x}^*) + \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|^2, \tag{20}$$

$$\langle \boldsymbol{z}^* - \boldsymbol{z}, \nabla f(\boldsymbol{z}) \rangle + 2f(\boldsymbol{z}) \geq f(\boldsymbol{z} \vee \boldsymbol{z}^*) + f(\boldsymbol{z} \wedge \boldsymbol{z}^*) + \frac{\mu}{2}\|\boldsymbol{z} - \boldsymbol{z}^*\|^2. \tag{21}$$

Let us first of all prove the following key Claim.

Claim 1. *It holds that* $f(\boldsymbol{x} \vee \boldsymbol{x}^*) + f(\boldsymbol{x} \wedge \boldsymbol{x}^*) + f(\boldsymbol{z} \vee \boldsymbol{z}^*) + f(\boldsymbol{z} \wedge \boldsymbol{z}^*) \geq f(\boldsymbol{x}^*)$.

Proof of Claim 1. Firstly, we are going to prove that

$$f(\boldsymbol{x} \vee \boldsymbol{x}^*) + f(\boldsymbol{z} \vee \boldsymbol{z}^*) \geq f(\boldsymbol{z}^*) + f((\boldsymbol{x} + \boldsymbol{z}) \vee \boldsymbol{x}^*), \tag{22}$$

which is equivalent to $f(\boldsymbol{x} \vee \boldsymbol{x}^*) - f(\boldsymbol{z}^*) \geq f((\boldsymbol{x} + \boldsymbol{z}) \vee \boldsymbol{x}^*) - f(\boldsymbol{z} \vee \boldsymbol{z}^*)$. It can be shown that $\boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{z}^* = (\boldsymbol{x} + \boldsymbol{z}) \vee \boldsymbol{x}^* - \boldsymbol{z} \vee \boldsymbol{z}^*$. Combining this with the fact that $\boldsymbol{z}^* \leq \boldsymbol{z} \vee \boldsymbol{z}^*$, and using the DR property (see (1)) implies (22). Then we establish,

$$\boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{z}^* = (\boldsymbol{x} + \boldsymbol{z}) \vee \boldsymbol{x}^* - \boldsymbol{z} \vee \boldsymbol{z}^* . \tag{23}$$

We will show that both the RHS and LHS of the above equation are equal to $\boldsymbol{x}$: for the LHS of (23) we can write $\boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{z}^* = \boldsymbol{x} \vee \boldsymbol{x}^* - (\boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{x}) = \boldsymbol{x}$. For the RHS of (23) let us consider any coordinate $i \in [n]$,

$$(x_i + z_i) \vee x_i^* - z_i \vee z_i^* = (x_i + z_i) \vee x_i^* - ((x_i + z_i) - x_i) \vee ((x_i \vee x_i^*) - x_i) = x_i,$$

where the last equality holds easily for the two situations: $(x_i + z_i) \geq x_i^*$ and $(x_i + z_i) < x_i^*$.

Next, we are going to prove that,

$$f(\boldsymbol{z}^*) + f(\boldsymbol{x} \wedge \boldsymbol{x}^*) \geq f(\boldsymbol{x}^*) + f(0) \tag{24}$$

it is equivalent to $f(\boldsymbol{z}^*) - f(0) \geq f(\boldsymbol{x}^*) - f(\boldsymbol{x} \wedge \boldsymbol{x}^*)$, which can be done similarly by the DR property: Notice that

$$\boldsymbol{x}^* - \boldsymbol{x} \wedge \boldsymbol{x}^* = \boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{x} = \boldsymbol{z}^* - 0 \text{ and } 0 \leq \boldsymbol{x} \wedge \boldsymbol{x}^*$$

thus (24) holds from the DR property. Combining (22) and (24) one can get,

$$f(\boldsymbol{x} \vee \boldsymbol{x}^*) + f(\boldsymbol{z} \vee \boldsymbol{z}^*) + f(\boldsymbol{x} \wedge \boldsymbol{x}^*) + f(\boldsymbol{z} \wedge \boldsymbol{z}^*) \geq f(\boldsymbol{x}^*) + f(0) + f((\boldsymbol{x} + \boldsymbol{z}) \vee \boldsymbol{x}^*) + f(\boldsymbol{z} \wedge \boldsymbol{z}^*)$$
$$\geq f(\boldsymbol{x}^*) \quad \text{(non-negativity of } f) \,.$$

∎

Combining (20) and (21) and Claim 1 it reads,

$$\langle \boldsymbol{x}^* - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle + \langle \boldsymbol{z}^* - \boldsymbol{z}, \nabla f(\boldsymbol{z}) \rangle + 2(f(\boldsymbol{x}) + f(\boldsymbol{z})) \geq f(\boldsymbol{x}^*) + \frac{\mu}{2}(\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 + \|\boldsymbol{z} - \boldsymbol{z}^*\|^2) \tag{25}$$

From the definition of non-stationarity in (8) one can get,

$$g_{\mathcal{P}}(\boldsymbol{x}) := \max_{\boldsymbol{v} \in \mathcal{P}} \langle \boldsymbol{v} - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle \overset{\boldsymbol{x}^* \in \mathcal{P}}{\geq} \langle \boldsymbol{x}^* - \boldsymbol{x}, \nabla f(\boldsymbol{x}) \rangle \tag{26}$$

$$g_{\mathcal{Q}}(\boldsymbol{z}) := \max_{\boldsymbol{v} \in \mathcal{Q}} \langle \boldsymbol{v} - \boldsymbol{z}, \nabla f(\boldsymbol{z}) \rangle \overset{\boldsymbol{z}^* \in \mathcal{Q}}{\geq} \langle \boldsymbol{z}^* - \boldsymbol{z}, \nabla f(\boldsymbol{z}) \rangle \tag{27}$$

Putting together (25) to (27) we can get,

$$2(f(\boldsymbol{x}) + f(\boldsymbol{z})) \geq f(\boldsymbol{x}^*) - g_{\mathcal{P}}(\boldsymbol{x}) - g_{\mathcal{Q}}(\boldsymbol{z}) + \frac{\mu}{2}(\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 + \|\boldsymbol{z} - \boldsymbol{z}^*\|^2).$$

so it arrives $\max\{f(\boldsymbol{x}), f(\boldsymbol{z})\} \geq \frac{1}{4}[f(\boldsymbol{x}^*) - g_{\mathcal{P}}(\boldsymbol{x}) - g_{\mathcal{Q}}(\boldsymbol{z})] + \frac{\mu}{8}(\|\boldsymbol{x} - \boldsymbol{x}^*\|^2 + \|\boldsymbol{z} - \boldsymbol{z}^*\|^2).$ ∎

E Proofs for Algorithms

E.1 Proof of Theorem 1

Proof of Theorem 1. Let $g_{\mathcal{P}}(\boldsymbol{x}), g_{\mathcal{Q}}(\boldsymbol{z})$ to be the non-stationarity of $\boldsymbol{x}$ and $\boldsymbol{z}$, respectively. Since we are using the NON-CONVEX FRANK-WOLFE (Algorithm 3) as subroutine, according to [27, Theorem 1], one can get,

$$g_{\mathcal{P}}(\boldsymbol{x}) \leq \min\left\{\frac{\max\{2h_1, C_f(\mathcal{P})\}}{\sqrt{K_1 + 1}}, \epsilon_1\right\}$$

$$g_{\mathcal{Q}}(\boldsymbol{z}) \leq \min\left\{\frac{\max\{2h_2, C_f(\mathcal{Q})\}}{\sqrt{K_2 + 1}}, \epsilon_2\right\},$$

Plugging the above into Proposition 1 we reach the conclusion in (10). ∎

E.2 Proof of Lemma 2

Lemma 2. *Assume $\boldsymbol{x}^{(0)} = 0$. For $k = 0, ..., K - 1$, it holds $x_i^{(k)} \leq \bar{u}_i[1 - (1 - \gamma)^{t^{(k)}/\gamma}], \forall i \in [n]$.*

Proof of Lemma 2. We prove by induction. First of all, it holds when $k = 0$, since $x_i^{(0)} = 0$, and $t^{(0)} = 0$ as well. Assume it holds for k. Then for $k + 1$, we have

$$x_i^{(k+1)} = x_i^{(k)} + \gamma v_i^{(k)}$$
$$\leq x_i^{(k)} + \gamma(\bar{u}_i - x_i^{(k)}) \quad \text{(constraint of shrunken LMO)} \tag{28}$$
$$= (1 - \gamma)x_i^{(k)} + \gamma\bar{u}_i$$
$$\leq (1 - \gamma)\bar{u}_i[1 - (1 - \gamma)^{t^{(k)}/\gamma}] + \gamma\bar{u}_i \quad \text{(induction)} \tag{29}$$
$$= \bar{u}_i[1 - (1 - \gamma)^{t^{(k+1)}/\gamma}].$$

∎

E.3 Proof of Lemma 3

Lemma 3 (Generalized from Lemma 7 in [8]). *Given $\boldsymbol{\theta} \in (0, \bar{\boldsymbol{u}}]$, let $\lambda' = \min_{i \in [n]} \frac{\bar{u}_i}{\theta_i}$. Then for all $\boldsymbol{x} \in [0, \boldsymbol{\theta}]$, it holds $f(\boldsymbol{x} \vee \boldsymbol{x}^*) \geq (1 - \frac{1}{\lambda'})f(\boldsymbol{x}^*)$.*

Proof of Lemma 3. Consider $r(\lambda) = \boldsymbol{x}^* + \lambda(\boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{x}^*)$, it is easy to see that $r(\lambda) \geq 0, \forall \lambda \geq 0$.

Notice that $\lambda' \geq 1$. Let $\boldsymbol{y} = r(\lambda') = \boldsymbol{x}^* + \lambda'(\boldsymbol{x} \vee \boldsymbol{x}^* - \boldsymbol{x}^*)$, it is easy to see that $\boldsymbol{y} \geq 0$, it also hold that $\boldsymbol{y} \leq \bar{\boldsymbol{u}}$: Consider one coordinate i, 1) if $x_i \geq x_i^*$, then $y_i = x_i^* + \lambda'(x_i - x_i^*) \leq \lambda' x_i \leq \lambda' \theta_i \leq \bar{u}_i$; 2) if $x_i < x_i^*$, then $y_i = x_i^* \leq \bar{u}_i$. So $f(\boldsymbol{y}) \geq 0$.

Note that

$$\boldsymbol{x} \vee \boldsymbol{x}^* = (1 - \frac{1}{\lambda'})\boldsymbol{x}^* + \frac{1}{\lambda'}\boldsymbol{y} = (1 - \frac{1}{\lambda'})r(0) + \frac{1}{\lambda'}r(\lambda'),$$

since f is concave along $r(\lambda)$, so it holds that,

$$f(\boldsymbol{x} \vee \boldsymbol{x}^*) \geq (1 - \frac{1}{\lambda'})f(\boldsymbol{x}^*) + \frac{1}{\lambda'}f(\boldsymbol{y}) \geq (1 - \frac{1}{\lambda'})f(\boldsymbol{x}^*).$$

∎

E.4 Proof of Theorem 2

Proof of Theorem 2. First of all, let us prove the Claim:

Claim 2. *For $k = 0, ..., K-1$ it holds $f(\boldsymbol{x}^{(k+1)}) \geq (1-\gamma)f(\boldsymbol{x}^{(k)}) + \gamma(1-\gamma)^{t^{(k)}/\gamma}f(\boldsymbol{x}^*) - \frac{LD^2}{2}\gamma^2$,*

Proof of Claim 2. Consider a point $\boldsymbol{z}^{(k)} := \boldsymbol{x}^{(k)} \vee \boldsymbol{x}^* - \boldsymbol{x}^{(k)}$, one can observe that: 1) $\boldsymbol{z}^{(k)} \leq \bar{\boldsymbol{u}} - \boldsymbol{x}^{(k)}$; 2) since $\boldsymbol{x}^{(k)} \geq 0, \boldsymbol{x}^* \geq 0$, so $\boldsymbol{z}^{(k)} \leq \boldsymbol{x}^*$, which implies that $\boldsymbol{z}^{(k)} \in \mathcal{P}$ (from down-closedness of $\mathcal{P}$). So $\boldsymbol{z}^{(k)}$ is a candidate solution for the new LMO (Step 3). We have,

$$f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)}) \geq \gamma\langle\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{v}^{(k)}\rangle - \frac{L}{2}\gamma^2\|\boldsymbol{v}^{(k)}\|^2 \quad \text{(Quadratic lower bound from (5))}$$

$$\geq \gamma\langle\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{v}^{(k)}\rangle - \frac{L}{2}\gamma^2 D^2 \quad \text{(diameter of } \mathcal{P)}$$

$$\geq \gamma\langle\nabla f(\boldsymbol{x}^{(k)}), \boldsymbol{z}^{(k)}\rangle - \frac{L}{2}\gamma^2 D^2 \quad \text{(shrunken LMO)}$$

$$\geq \gamma(f(\boldsymbol{x}^{(k)} + \boldsymbol{z}^{(k)}) - f(\boldsymbol{x}^{(k)})) - \frac{L}{2}\gamma^2 D^2 \quad \text{(concave along } \boldsymbol{z}^{(k)})$$

$$= \gamma[f(\boldsymbol{x}^{(k)} \vee \boldsymbol{x}^*) - f(\boldsymbol{x}^{(k)})] - \frac{L}{2}\gamma^2 D^2$$

$$\geq \gamma[(1 - \frac{1}{\lambda'})f(\boldsymbol{x}^*) - f(\boldsymbol{x}^{(k)})] - \frac{L}{2}\gamma^2 D^2 \quad \text{(Lemma 3)}$$

$$= \gamma[(1 - \gamma)^{t^{(k)}/\gamma}f(\boldsymbol{x}^*) - f(\boldsymbol{x}^{(k)})] - \frac{L}{2}\gamma^2 D^2$$

where the last equality comes from setting $\boldsymbol{\theta} := \bar{\boldsymbol{u}}(1 - (1-\gamma)^{t^{(k)}/\gamma})$ according to Lemma 2, thus $\lambda' = \min_i \frac{\bar{u}_i}{\theta_i} = (1 - (1-\gamma)^{t^{(k)}/\gamma})^{-1}$.

After rearrangement, we reach the claim. ∎

Then, let us prove Theorem 2 by *induction*.

First of all, it holds when $k = 0$ (notice that $t^{(0)} = 0$). Assume that it holds for k. Then for $k + 1$, considering the fact $e^{-t} - O(\gamma) \leq (1-\gamma)^{t/\gamma}$ when $0 < \gamma \leq t \leq 1$ and Claim 2 we get,

$$f(\boldsymbol{x}^{(k+1)})$$

$$\geq (1-\gamma)f(\boldsymbol{x}^{(k)}) + \gamma(1-\gamma)^{t^{(k)}/\gamma}f(\boldsymbol{x}^*) - \frac{LD^2}{2}\gamma^2$$

$$\geq (1-\gamma)f(\boldsymbol{x}^{(k)}) + \gamma[e^{-t^{(k)}} - O(\gamma)]f(\boldsymbol{x}^*) - \frac{LD^2}{2}\gamma^2 \tag{30}$$

$$\geq (1-\gamma)[t^{(k)}e^{-t^{(k)}}f(\boldsymbol{x}^*) - \frac{LD^2}{2}k\gamma^2 - O(\gamma^2)f(\boldsymbol{x}^*)] + \gamma[e^{-t^{(k)}} - O(\gamma)]f(\boldsymbol{x}^*) - \frac{LD^2}{2}\gamma^2$$

$$= [(1-\gamma)t^{(k)}e^{-t^{(k)}} + \gamma e^{-t^{(k)}}]f(\boldsymbol{x}^*) - \frac{LD^2}{2}\gamma^2[(1-\gamma)k + 1] - [(1-\gamma)O(\gamma^2) + \gamma O(\gamma)]f(\boldsymbol{x}^*)$$

$$\geq [(1-\gamma)t^{(k)}e^{-t^{(k)}} + \gamma e^{-t^{(k)}}]f(\boldsymbol{x}^*) - \frac{LD^2}{2}\gamma^2(k+1) - O(\gamma^2)f(\boldsymbol{x}^*). \tag{31}$$

Let us consider the term $[(1-\gamma)t^{(k)}e^{-t^{(k)}} + \gamma e^{-t^{(k)}}]f(\boldsymbol{x}^*)$. We know that the function $g(t) = te^{-t}$ is concave in $[0, 2]$, so $g(t^{(k)} + \gamma) - g(t^{(k)}) \leq \gamma g'(t^{(k)})$, which amounts to

$$[(1-\gamma)t^{(k)}e^{-t^{(k)}} + \gamma e^{-t^{(k)}}]f(\boldsymbol{x}^*) \geq (t^{(k)} + \gamma)e^{-(t^{(k)}+\gamma)}f(\boldsymbol{x}^*) = t^{(k+1)}e^{-t^{(k+1)}}f(\boldsymbol{x}^*) \tag{32}$$

Plugging (32) into (31) we get,

$$f(\boldsymbol{x}^{(k+1)}) \geq t^{(k+1)}e^{-t^{(k+1)}}f(\boldsymbol{x}^*) - \frac{LD^2}{2}\gamma^2(k+1) - O(\gamma^2)f(\boldsymbol{x}^*).$$

Thus proving the induction, and proving the theorem as well. ∎

F Miscellaneous Results

F.1 Verifying DR-submodularity of the Objectives

Softmax extension. For softmax extension, the objective is,

$$f(\boldsymbol{x}) = \log\det\left(\text{diag}(\boldsymbol{x})(\mathbf{L} - \mathbf{I}) + \mathbf{I}\right), \boldsymbol{x} \in [0, 1]^n.$$

Its DR-submodularity can be established by directly applying Lemma 3 in [20]: [20, Lemma 3] immediately implies that all entries of $\nabla^2 f$ are non-positive, so $f(\boldsymbol{x})$ is DR-submodular.

Multilinear extension. The DR-submodularity of multilinear extension can be directly recognized by considering the conclusion in Appendix A.2 of [3] and the fact that multilinear extension is coordinate-wise linear.

KL($\boldsymbol{x}$). The Kullback-Leibler divergence between $q_{\boldsymbol{x}}$ and p, i.e., $\sum_{S \subseteq \mathcal{V}} q_{\boldsymbol{x}}(S) \log\frac{q_{\boldsymbol{x}}(S)}{p(S)}$ is,

$$\text{KL}(\boldsymbol{x}) = -\sum_{S \subseteq \mathcal{V}}\prod_{i \in S}x_i\prod_{j \notin S}(1 - x_j)F(S) + \sum_{i=1}^{n}[x_i\log x_i + (1 - x_i)\log(1 - x_i)] + \log Z.$$

The first term is the negative of a multilinear extension, so it is DR-supermodular. The second term is separable, and coordinate-wise convex, so it will not affect the off-diagonal entries of $\nabla^2\text{KL}(\boldsymbol{x})$, it will only contribute to the diagonal entries. Now, one can see that all entries of $\nabla^2\text{KL}(\boldsymbol{x})$ are non-negative, so $\text{KL}(\boldsymbol{x})$ is DR-supermodular w.r.t. $\boldsymbol{x}$.

F.2 A Counter Example to Show PSD Cone is not Lattice

The positive semidefine cone $\mathcal{K}_{\text{PSD}} = \{\mathbf{A} \in \mathbb{R}^{n \times n} | \mathbf{A} \text{ is symmetric}, \mathbf{A} \succeq 0\}$ is a proper cone, but not a lattice cone. That is, it can not be used to define a lattice over the space of symmetric matrices.

Let us consider the two dimensional symmetric matrix space S^2. Specifically, the following two symmetric matrices,

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{Y} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

For the conic inequality $\preceq_{\mathcal{K}_{\text{PSD}}}$, assume that there exists a least upper bound, i.e., the join of $\mathbf{X}, \mathbf{Y}$: $\mathbf{Z} := \mathbf{X} \vee \mathbf{Y}$. From the definition of least upper bound, $\forall \mathbf{W} \in S^2$ it should hold that,

$$\mathbf{W} \succeq_{\mathcal{K}_{\text{PSD}}} \mathbf{X} \text{ and } \mathbf{W} \succeq_{\mathcal{K}_{\text{PSD}}} \mathbf{Y} \text{ iff } \mathbf{W} \succeq_{\mathcal{K}_{\text{PSD}}} \mathbf{Z}. \tag{33}$$

Suppose $\mathbf{Z} = \begin{bmatrix} b & a \\ a & c \end{bmatrix}$. Firstly, consider $\mathbf{W}$ to be diagonal matrices, one can verify that $\mathbf{Z}$ must be in the form of $\begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix}$, then considering $\mathbf{W} = \mathbf{I}$ forcing $\mathbf{Z}$ to be $\mathbf{I}$.

Now let $\mathbf{W} = \frac{2}{3} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$, which is $\succeq_{\mathcal{K}_{\text{PSD}}} \mathbf{X}$ and $\succeq_{\mathcal{K}_{\text{PSD}}} \mathbf{Y}$. However, $\mathbf{W} - \mathbf{I} = \frac{1}{3} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \notin \mathcal{K}_{\text{PSD}}$, thus contradicting (33).

G Additional Experimental Results

We generate the down-closed polytope constraints in the same form and same way as that for DR-submodular quadratic functions.

Figure 5 shows the function values returned by different solvers w.r.t. n, for which the random polytope constraints were generated with exponential distribution. Specifically, the random polytope is in the form of $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^n_+ \mid \mathbf{A}\boldsymbol{x} \leq \boldsymbol{b}, \boldsymbol{x} \leq \bar{\boldsymbol{u}}, \mathbf{A} \in \mathbb{R}^{m \times n}_{++}, \boldsymbol{b} \in \mathbb{R}^m_+\}$. Each entry of $\mathbf{A}$ was sampled from $\text{Exp}(1) + \nu$, where $\nu = 0.01$ is a small positive constant. We set $\boldsymbol{b} = 2 * \mathbf{1}^m$, and set $\bar{\boldsymbol{u}}$ to be the tightest upper bound of $\mathcal{P}$ by $\bar{u}_j = \min_{i \in [m]} \frac{b_i}{A_{ij}}, \forall j \in [n]$. One can see that the TWO-PHASE FRANK-WOLFE has the best performance, while non-monotone FRANK-WOLFE and PROJGRAD have comparable performance.
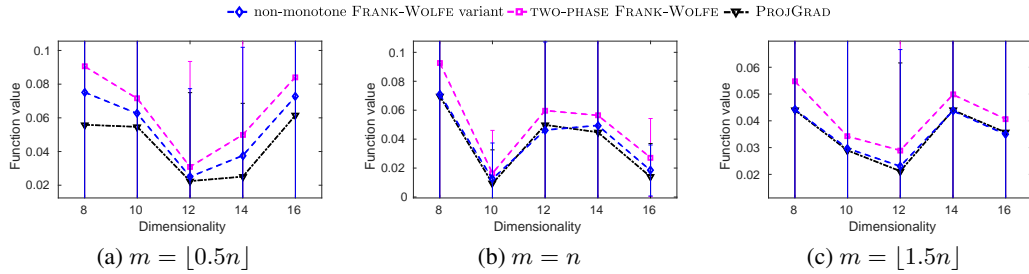

Figure 5: Results on softmax instances with random polytope constraints generated from exponential distribution.