On Maximization of Weakly Modular Functions: Guarantees of Multi-stage Algorithms, Tractability, and Hardness

Shinsaku Sakaue
NTT Communication Science Laboratories

Abstract

Maximization of *non-submodular* functions appears in various scenarios, and many previous works studied it based on some measures that quantify the closeness to being submodular. On the other hand, some practical non-submodular functions are actually close to being *modular*, which has been utilized in few studies. In this paper, we study cardinality-constrained maximization of *weakly modular* functions, whose closeness to being modular is measured by *submodularity* and *supermodularity ratios*, and reveal what we can and cannot do by using the weak modularity. We first show that guarantees of multi-stage algorithms can be proved with the weak modularity, which generalize and improve some existing results, and experiments confirm their effectiveness. We then show that weakly modular maximization is *fixed-parameter tractable* under certain conditions; as a byproduct, we provide a new time–accuracy trade-off for ℓ_0-constrained minimization. We finally prove that, even if objective functions are weakly modular, no polynomial-time algorithms can improve the existing approximation guarantee achieved by the greedy algorithm in general.

1 INTRODUCTION

We consider the following set function maximization with a cardinality constraint:

$$\underset{\mathsf{S}\subseteq[d]}{\text{maximize}}\ F(\mathsf{S}) \quad \text{subject to } |\mathsf{S}| \leq k, \tag{1}$$

where $d, k \in \mathbb{Z}_{>0}$, $[d] \coloneqq \{1, \dots, d\}$, and $F : 2^{[d]} \to \mathbb{R}$. We assume F to be monotone, normalized, and *weakly modular* (WM), where the closeness to being modular is represented with *submodularity ratio* (SBR) $\gamma \in [0, 1]$ and *supermodularity ratio* (SPR) $\beta \in [0, 1]$; (see, Section 1.2 for precise definitions). We say F is *weakly submodular* (*weakly supermodular*) if its SBR (SPR) is lower bounded. The larger SBR and SPR are, the closer F is to being submodular and supermodular, respectively, and F is modular if $\gamma = \beta = 1$.

Many previous studies on non-submodular maximization are based on some measures that quantify the deviation from being submodular (Elenberg et al., 2018; Qian and Singer, 2019), and SBR is one of the most prevalent among such measures. As regards weakly submodular maximization, Das and Kempe (2018) proved a well-known $(1 - \mathrm{e}^{-\gamma})$-approximation guarantee of the greedy algorithm (`Greedy`).

When it comes to practical non-submodular maximization instances, it can be effective to employ additional measures other than those quantifying the distance to being submodular. Bian et al. (2017) considered a class such that F has bounded SBR and *curvature* $\alpha \in [0, 1]$, and they proved a $\frac{1}{\alpha}(1 - \mathrm{e}^{-\alpha\gamma})$-approximation guarantee of `Greedy`. Namely, an improved approximation guarantee is possible if $\alpha < 1$. Unfortunately, however, $\alpha = 1$ occurs quite naturally in many applications as discussed in Section 2 (see also (Soma and Yoshida, 2018)), which motivates us to consider a wider class of non-submodular maximization that can capture the structures of various practical problems.

Weakly modular maximization (WMM) forms a wider class than that of (Bian et al., 2017). In fact, SPR β and curvature α always satisfy $\beta \geq 1 - \alpha$ (Bogunovic et al., 2018); i.e., SPR β can be bounded even if $\alpha = 1$. As shown in Section 2, various problems including feature selection (Das and Kempe, 2018) and production planning (Bian et al., 2017) strictly belong to WMM; that is, F has bounded SPR β even though $\alpha = 1$ in general. This fact suggests the importance of studying WMM. However, few previous works have studied problem (1) by utilizing the weak modularity, and so WMM remains to be studied.

Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

1.1 Our Contribution

Our first contribution provides guarantees of efficient algorithms for WMM. As shown in Section 2, WMM can model various continuous optimization problems including ℓ_0-constrained minimization and linear programming (LP) with a cardinality constraint. Given such WMM instances, the evaluation of objective functions involves solving optimization subproblems, which is often so costly that even standard `Greedy` becomes impractical. To overcome this hardship, we consider using *multi-stage* algorithms for WMM.

Guarantees of Multi-stage Algorithms In Section 3, we show that the multi-stage approach is effective for WMM instances; with this approach, we accelerate greedy-style algorithms by adding multiple elements in each iteration, instead of a single element. The only existing study that proved guarantees of multi-stage algorithms is (Wei et al., 2014); their result requires the submodularity, and the approximation ratio is expressed as $\frac{1}{\alpha}(1-\mathrm{e}^{-\alpha(1-\alpha)})$ in general, which becomes 0 if curvature α is equal to 1. Our guarantee of the multi-stage greedy algorithm (`Multi-Greedy`) for WMM is advantageous relative to the previous result in two aspects: it can be applied to WM functions, which are generally non-submodular, and it can yield positive approximation ratios even if $\alpha = 1$ as long as SBR and SPR are bounded. Our result also includes the $(1-\mathrm{e}^{-\gamma})$-approximation guarantee of (Das and Kempe, 2018) as a special case. We then focus on ℓ_0-constrained minimization and prove a guarantee of the multi-stage orthogonal matching pursuit (`Multi-OMP`), which can achieve a better approximation ratio than `Multi-Greedy`. Surprisingly, our result matches that of standard `OMP` (Elenberg et al., 2018), while `Multi-OMP` can run faster than `OMP`. Our result also improves that of the latest feature selection algorithm (Qian and Singer, 2019). Experiments show that the multi-stage approach successfully accelerates `Greedy` and `OMP` at the cost of a slight decline in solution quality.

Our second and their contributions, presented in Section 4, are related to theoretical properties of WMM. These contributions are important for revealing what we can and cannot do with the weak modularity.

Fixed-parameter Tractability In Section 4.1, we show that ϵ-error solutions for WMM can be obtained with a randomized *fixed-parameter tractable* (FPT) algorithm, whose computation cost depends arbitrarily on certain inputs including SBR γ, SPR β, sparsity k, and ϵ, but it is polynomial in d. The FPT algorithm was first proposed by Skowron (2017), but its guarantee was proved only for a special case of monotone submodular maximization. As a byproduct, we provide a time–accuracy trade-off for ℓ_0-constrained minimization, which is contrasted with the existing sparsity–accuracy trade-off (Shalev-Shwartz et al., 2010).

Hardness of Improving Approximation Ratio As mentioned before, if curvature α is bounded by a constant smaller than 1, the $\frac{1}{\alpha}(1-\mathrm{e}^{-\alpha\gamma})$-approximation guarantee of (Bian et al., 2017) improves the approximation ratio, $1-\mathrm{e}^{-\gamma}$, of (Das and Kempe, 2018). When it comes to WMM, not curvature α but SPR β ($\geq 1-\alpha$) is bounded. Given this background, the following question arises: Can we improve the approximation ratio, $1-\mathrm{e}^{-\gamma}$, if SPR β is bounded by a constant, instead of curvature α. In Section 4.2, we show that it is generally impossible. More precisely, we prove that, even if $\gamma = 1$ and $\beta \geq 1/2$, no polynomial-time algorithms can improve the $(1-\mathrm{e}^{-1})$-approximation guarantee in general in the value oracle model; i.e., bounded SPR β does not always help us to improve the ratio, $1-\mathrm{e}^{-\gamma}$. This result clarifies the theoretical gap between SPR β and curvature α.

1.2 Notation and Definitions

Given any $F : 2^{[d]} \to \mathbb{R}$, we define $F(\mathsf{T} \mid \mathsf{S}) \coloneqq F(\mathsf{S} \cup \mathsf{T}) - F(\mathsf{S})$ for any $\mathsf{S}, \mathsf{T} \subseteq [d]$. All the set functions considered in this paper are monotone ($F(\mathsf{T} \mid \mathsf{S}) \geq 0$, $\forall \mathsf{S}, \mathsf{T} \subseteq [d]$) and normalized ($F(\emptyset) = 0$). We say F is submodular (supermodular) if $F(j \mid \mathsf{S}) \geq F(j \mid \mathsf{T})$ ($F(j \mid \mathsf{S}) \leq F(j \mid \mathsf{T})$) holds for any $\mathsf{S} \subseteq \mathsf{T}$ and $j \notin \mathsf{T}$. We assume that F can be evaluated in polynomial time w.r.t. d (or poly(d) time). Given any $\mathsf{S} \subseteq [d]$ and $\boldsymbol{x} \in \mathbb{R}^{[d]}$, whose j-th entry $\boldsymbol{x}_j$ is associated with $j \in [d]$, $\boldsymbol{x}_\mathsf{S} \in \mathbb{R}^\mathsf{S}$ denotes the restriction of $\boldsymbol{x}$ to S. We define the support of $\boldsymbol{x}$ as $\mathrm{supp}(\boldsymbol{x}) \coloneqq \{j \in [d] \mid \boldsymbol{x}_j \neq 0\}$.

SBR and SPR Given any monotone $F : 2^{[d]} \to \mathbb{R}$, $\mathsf{U} \subseteq [d]$, and $s \in \mathbb{Z}_{>0}$, we define SBR $\gamma_{\mathsf{U},s}$ and SPR $\beta_{\mathsf{U},s}$ as the largest scalars that satisfy

$$\gamma_{\mathsf{U},s} F(\mathsf{S} \mid \mathsf{L}) \leq \sum\nolimits_{j \in \mathsf{S}} F(j \mid \mathsf{L}) \quad \text{and}$$
$$\sum\nolimits_{j \in \mathsf{S}} F(j \mid \mathsf{L}) \leq \beta_{\mathsf{U},s}^{-1} F(\mathsf{S} \mid \mathsf{L}),$$

respectively, for any disjoint $\mathsf{L}, \mathsf{S} \subseteq [d]$ such that $\mathsf{L} \subseteq \mathsf{U}$ and $|\mathsf{S}| \leq s$. We say F is $(\gamma_{\mathsf{U}_1,s_1}, \beta_{\mathsf{U}_2,s_2})$-WM if F has bounded $\gamma_{\mathsf{U}_1,s_1}$ and β_{U_2,s_2}. Note that $\gamma_{\mathsf{U}',s'} \geq \gamma_{\mathsf{U},s}$ and $\beta_{\mathsf{U}',s'} \geq \beta_{\mathsf{U},s}$ hold for any $\mathsf{U}' \subseteq \mathsf{U}$ and $s' \leq s$. We can confirm that $\gamma_{\mathsf{U},s} \in [0,1]$ and $\beta_{\mathsf{U},s} \in [1/s, 1]$ hold for any U and s. We define $\gamma_{s',s} \coloneqq \min_{|\mathsf{U}| \leq s'} \gamma_{\mathsf{U},s}$ and $\beta_{s',s} \coloneqq \min_{|\mathsf{U}| \leq s'} \beta_{\mathsf{U},s}$; we sometimes use $\gamma_s \coloneqq \gamma_{s,s}$ and $\beta_s \coloneqq \beta_{s,s}$. We have $\gamma_d = 1$ ($\beta_d = 1$) iff F is submodular (supermodular).

Curvature Given monotone $F : 2^{[d]} \to \mathbb{R}$, its curvature $\alpha \in [0,1]$ is defined as the smallest scalar that

satisfies

$$F(j \mid \mathsf{S}\backslash\{j\} \cup \mathsf{M}) \geq (1-\alpha)F(j \mid \mathsf{S}\backslash\{j\})$$

for any $\mathsf{S}, \mathsf{M} \subseteq [d]$ and $j \in \mathsf{S}\backslash\mathsf{M}$. We have $\beta_{\mathsf{U},s} \geq 1-\alpha$ for any U and s (see, (Bogunovic et al., 2018)).

Restricted Strong Convexity and Restricted Smoothness When studying ℓ_0-constrained minimization algorithms, the restricted strong convexity (RSC) and restricted smoothness (RSM) of loss function $l : \mathbb{R}^d \to \mathbb{R}$ is often used (Jain et al., 2014; Elenberg et al., 2018; Yuan et al., 2018). We assume l to be differentiable. Given any fixed $s_1, s_2 \in \mathbb{Z}_{>0}$, we say l is μ_{s_1,s_2}-RSC and ν_{s_1,s_2}-RSM if it satisfies

$$l(\boldsymbol{y}) \geq l(\boldsymbol{x}) + \langle \nabla l(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x} \rangle + \frac{\mu_{s_1,s_2}}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \quad \text{and}$$
$$l(\boldsymbol{y}) \leq l(\boldsymbol{x}) + \langle \nabla l(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x} \rangle + \frac{\nu_{s_1,s_2}}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2,$$

respectively, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ such that $\|\boldsymbol{x}\|_0 \leq s_1$, $\|\boldsymbol{y}\|_0 \leq s_1$, and $\|\boldsymbol{x}-\boldsymbol{y}\|_0 \leq s_2$. If l is quadratic, the above inequalities reduce to those of the well-known restricted isometric property (RIP) condition (Candès et al., 2006). We let $\mu_s \coloneqq \mu_{s,s}$ and $\nu_s \coloneqq \nu_{s,s}$. We define the restricted condition number as $\kappa_s \coloneqq \nu_s/\mu_s$. Typically, l with a smaller κ_s value is easier to deal with. If l is μ_d-RSC and ν_d-RSM, we abbreviate the subscript and say l is μ-strongly convex (μ-SC) and ν-smooth (ν-SM); we call $\kappa \coloneqq \nu/\mu$ a condition number.

1.3 Related Work

For the case where F is submodular, Nemhauser et al. (1978) proved the $(1-\mathrm{e}^{-1})$-approximation guarantee of `Greedy`. Nemhauser and Wolsey (1978) proved that no polynomial-time algorithms can improve this guarantee in the value oracle model, and Feige (1998) proved the NP-hardness for the case of *Max k-cover*. As regards tractability, Skowron (2017) developed a randomized FPT approximation algorithm for maximization of monotone submodular functions with a special property called *p-separability*. Unlike our results, those results hold only for monotone submodular maximization.

When it comes to non-submodular maximization, various notions have been introduced to obtain theoretical guarantees (Krause and Cevher, 2010; Feige and Izsak, 2013; Horel and Singer, 2016; Wang et al., 2016; Zhou and Spanos, 2016). Das and Kempe (2018) proposed SBR, one of the most prevalent notion used in many studies (Hu et al., 2016; Elenberg et al., 2017; Khanna et al., 2017a,b; Chen et al., 2018; Qian and Singer, 2019), and they proved that `Greedy` outputs solution $\mathsf{S} \subseteq [d]$ with a $(1-\mathrm{e}^{-\gamma_{\mathsf{S},k}})$-approximation guarantee. Harshaw et al. (2019) proved that no polynomial-time algorithms can improve the $(1-\mathrm{e}^{\gamma_d})$-approximation guarantee in general for every $\gamma_d \in (0,1]$ value. This result is different from our hardness result since they do not assume SPR to be bounded by a constant; this difference is critical since bounded SPR could make the problem easier.

The definition of SPR that we use was introduced by Bogunovic et al. (2018). Other SPR-like notions have been used in the context of minimization problems (Takeda et al., 2013; Liberty and Sviridenko, 2017), but those are different from SPR, which quantifies the deviation from being supermodular in the context of maximization problems.

Curvature α (Conforti and Cornuéjols, 1984; Bian et al., 2017) is also used in many studies (Iyer et al., 2013; Sviridenko et al., 2015; Bai and Bilmes, 2018). Its value is, however, often pessimistic (i.e., $\alpha \approx 1$) as pointed out by Soma and Yoshida (2018), and to bound the curvature value is more demanding than to bound SPR. Hence our results obtained with SPR are different from existing guarantees that use curvature; although those results can sometimes be improved by using *greedy curvature* $\alpha_G \leq \alpha$ (Bian et al., 2017), no lower bounds of α_G for WM functions have been proved.

We remark that our work is different from some previous studies on set functions that are close to being modular. As mentioned before, Bian et al. (2017) studied the case where the curvature and SBR are bounded, and they proved that `Greedy` finds solution S with a $\frac{1}{\alpha}(1-\mathrm{e}^{-\alpha\gamma_{\mathsf{S},k}})$-approximation guarantee. They also proved that `Greedy` cannot improve this guarantee. Unlike this result, our hardness result considers every polynomial-time algorithm. Bogunovic et al. (2018) considered the case where SBR and SPR are bounded. However, they are interested in obtaining guarantees for robust maximization, not for the standard cardinality-constrained maximization (1), which is of our interest. Chierichetti et al. (2015) defined the approximate modularity as the ℓ_∞-distance to being modular, which is different from the weak modularity.

Wei et al. (2014) provided the curvature-dependent approximation guarantees of multi-stage algorithms for submodular maximization. Marsousi et al. (2013) applied `Multi-OMP` to a special case of ℓ_0-constrained minimization where the loss function l is quadratic, but its theoretical guarantee has not been proved.

The idea of adding multiple elements in each round is also considered in the context of parallel algorithms (Balkanski and Singer, 2018). Qian and Singer (2019) recently developed a parallel approximation algorithm that runs in $O(\ln d)$ rounds for ℓ_0-constrained minimization. Surprisingly, thanks to the use of the weak modularity, we can show that `Multi-OMP` with only one round achieves a better approximation ratio.

2 APPLICATIONS

We motivate to study WMM by presenting its applications. For each application, we present lower bounds of SBR and SPR. We also provide an example such that $\alpha = 1$ for one of the applications, and such examples for the other applications are presented in Appendix A.

ℓ_0-constrained Minimization Given a differentiable loss function $l : \mathbb{R}^d \to \mathbb{R}$, we consider the ℓ_0-constrained minimization problem: $\min_{\|\boldsymbol{x}\|_0 \leq k} l(\boldsymbol{x})$. It is generally NP-hard (Natarajan, 1995) and appears in many practical scenarios: feature selection (Das and Kempe, 2018) and M-estimation (Jain et al., 2014). The problem can be rewritten as in (1) with $F(\mathsf{S}) = l(0) - \min_{\mathrm{supp}(\boldsymbol{x}) \subseteq \mathsf{S}} l(\boldsymbol{x})$, which has SBR $\gamma_{\mathsf{U},s} \geq \mu_{|\mathsf{U}|+s}/\nu_{|\mathsf{U}|+1,1} \geq 1/\kappa_{|\mathsf{U}|+s}$ (Elenberg et al., 2018) and SPR $\beta_{\mathsf{U},s} \geq \mu_{|\mathsf{U}|+1}/\nu_{|\mathsf{U}|+s,s} \geq 1/\kappa_{|\mathsf{U}|+s}$ (Appendix A.1); the later bound improves an existing result, $\beta_{\mathsf{U},s} \geq \mu/\nu = 1/\kappa$, of (Bogunovic et al., 2018). The evaluation of $F(\mathsf{S})$ involves solving $\min_{\mathrm{supp}(\boldsymbol{x}) \subseteq \mathsf{S}} l(\boldsymbol{x})$. If l is quadratic, we can solve it by computing a pseudo-inverse matrix. Given a more general l, we can use iterative methods (e.g., (Shalev-Shwartz and Zhang, 2016)) to solve the minimization problem.

LP with a Cardinality Constraint We consider the following constrained LP that models optimal production planning problem (Bian et al., 2017). Given a set of d items and k production lines, we design a production plan so that the total profit is maximized; i.e., we aim to solve $\max_{\boldsymbol{x} \in \mathcal{P}, \|\boldsymbol{x}\|_0 \leq k} \boldsymbol{c}^\top \boldsymbol{x}$, where $\boldsymbol{c} \in \mathbb{R}^d$ and $\mathcal{P} \subseteq \mathbb{R}^d$ represent the profit of each item and a polytope specified by continuous constraints (e.g., upper bounds on the total quantities of materials), respectively. This problem can be reformulated as in (1) with $F(\mathsf{S}) \coloneqq \max_{\boldsymbol{x} \in \mathcal{P}} \boldsymbol{c}_{\mathsf{S}}^\top \boldsymbol{x}_{\mathsf{S}}$. As in (Bian et al., 2017), SBR $\gamma_{\mathsf{U},s}$ of F is lower bounded by some $\gamma_0 > 0$ for any U and s under the *non-degeneracy* assumption. Furthermore, from the definition of SPR, we have $\beta_{\mathsf{U},s} \geq 1/s$; although the lower bound, $1/s$, can be small if $s \approx d$, this is not always the case. For example, in the guarantee of `Multi-Greedy` (Theorem 1), s is a controllable parameter, $b_{\max}$; i.e., $\beta_{\mathsf{U},s} \geq 1/b_{\max}$ holds.

Coverage Maximization Submodular functions sometimes have bounded SPR $\beta_{\mathsf{U},s}$, and such functions can be seen as special WM functions such that $\gamma_{\mathsf{U},s} = 1$ for any U and s. One such example is the coverage function. Let V be a finite set and $w_v \geq 0$ ($v \in V$). We define d groups $I_1, \dots, I_d \subseteq V$, and we let $I_{\mathsf{S}} \coloneqq \bigcup_{j \in \mathsf{S}} I_j$ for any $\mathsf{S} \subseteq [d]$. The coverage function is defined as $F(\mathsf{S}) \coloneqq \sum_{v \in I_{\mathsf{S}}} w_v$, which is submodular and used in many scenarios including document summarization (Lin and Bilmes, 2011) and itemset mining (Kumar et al., 2015). Given $s \in \mathbb{Z}_{>0}$, we assume that any collection of up to s groups covers every $v \in V$ at most c_s times; i.e., $c_s \coloneqq \max_{v \in V, |\mathsf{S}| \leq s} |\{j \in \mathsf{S} \mid v \in I_j\}|$. Note that $c_s \leq s$ always holds. In this case, SPR $\beta_{\mathsf{U},s}$ of F is lower bounded by $1/c_s$ as proved in Appendix A.3.

Example with Unbounded Curvature We provide an example of LP with a cardinality constraint such that $\alpha = 1$. Let $d = 2$, $\boldsymbol{x} = (x_1, x_2)^\top$, and $\epsilon \in (0, 1)$. We consider a set function defined as $F(\mathsf{S}) = \max_{\mathrm{supp}(\boldsymbol{x}) \subseteq \mathsf{S}} \{x_1 + \epsilon x_2 \mid x_1 + x_2 \leq 1, x_1 \geq 0, x_2 \geq 0\}$ for any $\mathsf{S} \subseteq [d]$; i.e., $\boldsymbol{c} = (1, \epsilon)^\top$ and $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^2 \mid x_1 + x_2 \leq 1, x_1 \geq 0, x_2 \geq 0\}$. From the definitions of SBR and SPR, we can confirm that $\gamma_{\mathsf{U},s} = 1$ and $\beta_{\mathsf{U},s} \geq \frac{1}{1+\epsilon}$ hold for any U and s. On the other hand, we have $\alpha = 1$ since $F(\{2\} \mid \{1\}) \geq (1 - \alpha) F(\{1\})$ must hold for $F(\{2\} \mid \{1\}) = 0$ and $F(\{1\}) = 1$.

3 MULTI-STAGE ALGORITHMS

We study multi-stage algorithms for WMM. Let S^* and $\boldsymbol{x}^*$ be target solutions for WMM and ℓ_0-constrained minimization, respectively, and $k^* \coloneqq |\mathsf{S}^*| = \|\boldsymbol{x}^*\|_0$; note that we allow k^* to be different from k. As a warm-up, we first discuss two simple algorithms:

Single-stage Algorithm We compute $F(j)$ for $j \in [d]$ and let $\mathsf{S} = \mathrm{argmax}_{\mathsf{S}' : |\mathsf{S}'| \leq k} \sum_{j \in \mathsf{S}'} F(j)$. The algorithm requires to evaluate F only d times, and it can find optimal solutions if F is modular. However, its approximation ratio becomes poor if F lacks the modularity. We consider a coverage maximization instance with $d = 2k$ and $V = \{v_1, \dots, v_{2k}\}$. Let $w_{v_j} = 1$ and $I_j = \{v_j\}$ for $j = 1, \dots, k$, and let $w_{v_j} = \epsilon \ll 1$ and $I_j = \{v_1, v_j\}$ for $j = k+1, \dots 2k$. In this case, if $\epsilon > 0$ is sufficiently small, the approximation ratio achieved by the single-stage algorithm is $\frac{1+k\epsilon}{k+\epsilon} = O(1/d)$.

Greedy Algorithm Starting from $\mathsf{S} = \emptyset$, `Greedy` iteratively adds $\mathrm{argmax}_{j \notin \mathsf{S}} F(j \mid \mathsf{S})$ to S and outputs S after k iterations. Given F with SBR γ_{S,k^*}, `Greedy` achieves a $(1 - \exp(-\gamma_{\mathsf{S},k^*}))$-approximation guarantee. `Greedy` is, however, often costly due to the sequential evaluation of F, particularly when the evaluation of F involves solving optimization problems. For example, in the case of ℓ_0-constrained minimization, `Greedy` solves convex minimization problems $\Theta(dk)$ times.

Namely, while the single-stage algorithm can efficiently find optimal solutions if F is modular, `Greedy` can achieve better guarantees for non-modular F at the cost of more computational effort. In the case of WMM, since F is close to being modular, we can expect that an intermediate of the above two algorithms works well. The multi-stage approach provides such an intermediate. As in Algorithm 1, we perform m ($\leq k$) iterations

Algorithm 1 Multi-stage algorithm

$\mathsf{U} \leftarrow [d]$, $\mathsf{S} \leftarrow \emptyset$
for $i = 1, \dots, m$ **do**
  $\mathsf{B}_i \leftarrow \operatorname{argmax}_{\mathsf{B} \subseteq \mathsf{U} : |\mathsf{B}| \leq b_i} G_{\mathsf{S}}(\mathsf{B})$
  $\mathsf{S} \leftarrow \mathsf{S} \cup \mathsf{B}_i$
  $\mathsf{U} \leftarrow \mathsf{U} \backslash \mathsf{B}_i$
return S

to obtain a solution. In each i-th iteration, we choose subset $\mathsf{B}_i \subseteq [d]$ of size at most b_i so that it maximizes a *surrogate function*, G_{S}, where S is the current solution. To obtain fast multi-stage algorithms, G_{S} should be efficiently evaluated and maximized. Below we design G_{S} for `Multi-Greedy` and `Multi-OMP`, and we present their theoretical guarantees. We then experimentally evaluate the multi-stage algorithms.

3.1 Theoretical Guarantees

Let $\mathsf{S}_i = \mathsf{B}_1 \cup \dots \cup \mathsf{B}_i$ for $i \in [m]$ and $\mathsf{S}_0 = \emptyset$. We first present a guarantee of `Multi-Greedy` for WMM. We then focus on ℓ_0-constrained minimization and prove a guarantee of `Multi-OMP`. As detailed below, our results generalize and improve some existing results, which emphasizes that to utilize the weak modularity is effective for obtaining strong theoretical results. The proofs of the theorems are presented in Appendix B.1.

3.1.1 Multi-Greedy

`Multi-Greedy` uses $G_{\mathsf{S}}(\mathsf{B}) = \sum_{j \in \mathsf{B}} F(j \mid \mathsf{S})$ as a surrogate function. Therefore, `Multi-Greedy` evaluates F $\Theta(dm)$ times. We can show that `Multi-Greedy` enjoys the following approximation guarantee:

Theorem 1. *Let $b_{\max}$ be an integer satisfying $1 \leq b_{\max} \leq k^*$. Set $b_1, \dots, b_m$ so as to satisfy $b_i \in [b_{\max}]$ for $i \in [m]$ and $\sum_{i \in [m]} b_i = k$. If S is the solution obtained with* `Multi-Greedy` *and F is $(\gamma_{\mathsf{S},k^*}, \beta_{\mathsf{S},b_{\max}})$-WM, we have*

$$F(\mathsf{S}) \geq \left(1 - \prod_{i=1}^{m} \left(1 - \gamma_{\mathsf{S}_{i-1},k^*} \beta_{\mathsf{S}_{i-1},b_i} \frac{b_i}{k^*}\right)\right) F(\mathsf{S}^*) \geq \left(1 - \exp\left(-\gamma_{\mathsf{S},k^*} \beta_{\mathsf{S},b_{\max}} \frac{k}{k^*}\right)\right) F(\mathsf{S}^*).$$

Note that, if we set $b_{\max} = 1$, this result recovers the $(1 - \mathrm{e}^{-\gamma_{\mathsf{S},k}})$-approximation of `Greedy` (Das and Kempe, 2018) since $\beta_{\mathsf{S},1} = 1$. `Multi-Greedy` with $m = 1$ is the single-stage algorithm, which is studied as the oblivious algorithm in the field of ℓ_0-constrained minimization. Elenberg et al. (2018) proved that its approximation ratio is at least $\max\left\{\frac{1}{k}\kappa_k^{-1}, \frac{3}{4}\kappa_k^{-2}, \kappa_k^{-3}\right\}$. Note that Theorem 1 improves this result since, if $b_1 = k = k^*$, the approximation ratio becomes $\max\{\frac{1}{k}\kappa_k^{-1}, \kappa_k^{-2}\}$ from the lower bounds of SBR and SPR (Section 2) and $\beta_{\emptyset,k} \geq 1/k$. More generally, for $m \geq 1$, `Multi-Greedy` achieves a $1 - \exp(\kappa_{2k}^{-1}\kappa_{k+b_{\max}}^{-1})$-approximation. Below we show that a stronger guarantee for ℓ_0-constrained minimization can be obtained by using `Multi-OMP`.

3.1.2 Multi-OMP

We then focus on ℓ_0-constrained minimization; i.e., we assume $F(\mathsf{S}) = l(0) - \min_{\mathrm{supp}(\boldsymbol{x}) \subseteq \mathsf{S}} l(\boldsymbol{x})$ $(\forall \mathsf{S} \subseteq [d])$. We let $\mathbf{b}^{(\mathsf{S})} := \operatorname{argmin}_{\mathrm{supp}(\boldsymbol{x}') \subseteq \mathsf{S}} l(\boldsymbol{x}')$ for any $\mathsf{S} \subseteq [d]$. `Multi-OMP` uses $G_{\mathsf{S}}(\mathsf{B}) = \sum_{j \in \mathsf{B}} |\nabla l(\mathbf{b}^{(\mathsf{S})})_j|^2$ as a surrogate function. Thus, it requires to compute the gradient and to solve convex minimization problems m times. To prove the guarantee of `Multi-OMP`, we use the following lemma, which, roughly speaking, connects the decrease in l to the increase in F.

Lemma 1. *For any disjoint $\mathsf{A}, \mathsf{B} \subseteq [d]$, if $l(\cdot)$ is $\mu_{|\mathsf{A} \cup \mathsf{B}|}$-RSC and $\nu_{|\mathsf{B}|,|\mathsf{B} \backslash \mathsf{A}|}$-RSM, we have*

$$\frac{\|\nabla l(\mathbf{b}^{(\mathsf{A})})_{\mathsf{B}}\|_2^2}{2\nu_{|\mathsf{B}|,|\mathsf{B} \backslash \mathsf{A}|}} \leq F(\mathsf{B} \mid \mathsf{A}) \leq \frac{\|\nabla l(\mathbf{b}^{(\mathsf{A})})_{\mathsf{B}}\|_2^2}{2\mu_{|\mathsf{A} \cup \mathsf{B}|}}.$$

A special case of the lemma is implicitly used in (Elenberg et al., 2018). In Appendix A.1, we provide a slightly stronger version of the lemma, which we use for proving the guarantee of `Multi-OMP`. By using the lemma, we can employ the technique used when proving Theorem 1, which leads to the following result:

Theorem 2. *Set $b_1, \dots, b_m$ as in Theorem 1. If l is μ_{k+k^*}-RSC and $\nu_{k,b_{\max}}$-RSM, then* `Multi-OMP` *outputs solution S such that $\boldsymbol{x} = \operatorname{argmin}_{\mathrm{supp}(\boldsymbol{x}') \subseteq \mathsf{S}} l(\boldsymbol{x}')$ satisfies*

$$l(\boldsymbol{x}) \leq l(\boldsymbol{x}^*) + \prod_{i=1}^{m} \left(1 - \frac{\mu_{|\mathsf{S}_{i-1} \cup \mathsf{S}^*|}}{\nu_{|\mathsf{S}_i|,|\mathsf{B}_i|}} \frac{b_i}{k^*}\right) (l(0) - l(\boldsymbol{x}^*)) \leq l(\boldsymbol{x}^*) + \exp\left(-\frac{\mu_{k+k^*}}{\nu_{k,b_{\max}}} \frac{k}{k^*}\right) (l(0) - l(\boldsymbol{x}^*)) \leq l(\boldsymbol{x}^*) + \exp\left(-\frac{1}{\kappa_{k+k^*}} \frac{k}{k^*}\right) (l(0) - l(\boldsymbol{x}^*)).$$

Note that, if $k = k^*$, Theorem 2 gives a $(1 - \exp(\kappa_{2k}^{-1}))$-approximation guarantee, which improves the aforementioned guarantee of `Multi-Greedy`. Interestingly, the approximation ratio matches those of `OMP` and `Greedy` (Elenberg et al., 2018). Namely, the use of the multi-stage approach does not degrade the theoretical guarantee. If we let $b_1 = k = k^*$, `Multi-OMP` with only one round achieves a κ_{2k}^{-1}-approximation guarantee; this improves the existing $\left(1 - \exp\left(-\kappa_{2k}^{-4}\right)\right)$-approximation guarantee with $O(\ln d)$ rounds, recently proved by Qian and Singer (2019), in terms of both the approximation ratio and the computation complexity.

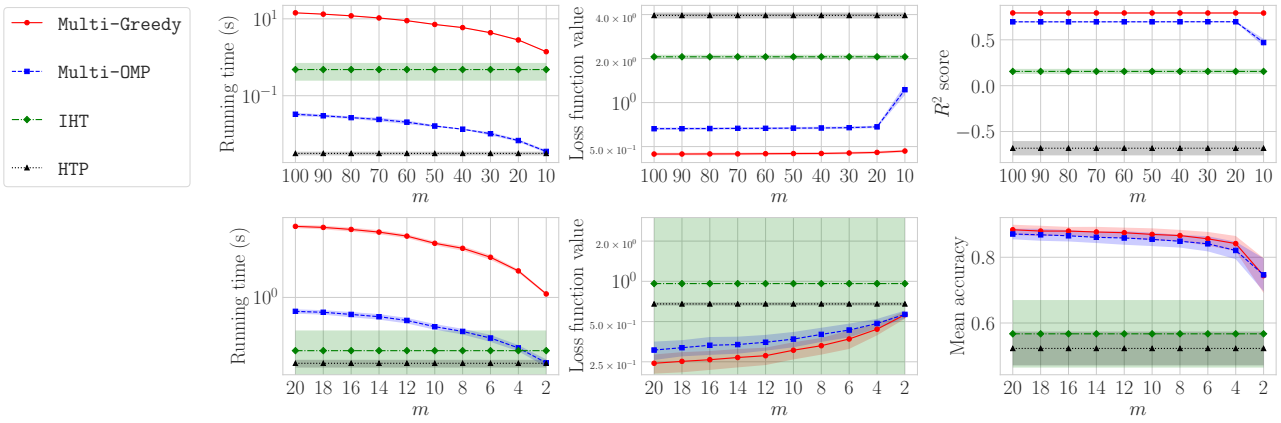


Figure 1: Results of ℓ_0-constrained minimization with various m values. Top (bottom) figures present those of regression (classification) instances. Running times and loss function values are shown with semi-log plots. Each curve and error band indicate the average and standard deviation, respectively, calculated over 100 instances.

3.2 Experiments

We evaluate the multi-stage algorithms via experiments with two kinds instances: ℓ_0-constrained minimization and LP with a cardinality constraint. We use Python3 to implement the algorithms, and we conduct experiments on a 64-bit macOS machine with 3.3GHz Intel Core i7 CPUs and 16 GB RAM. All the algorithms considered below can be accelerated via randomization (Li et al., 2016; Khanna et al., 2017b), but to simplify the comparisons we here do not employ such techniques.

3.2.1 ℓ_0-constrained Minimization

We use two instances with the real-world dataset available at PMLB (Olson et al., 2017). The first is a sparse regression instance with the square loss, $l(\boldsymbol{x}) = \frac{1}{2n}\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2$, where $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$ are obtained from "satellite_image" dataset. We use the 1st and 2nd order polynomial features; as a result, we have $d = 666$ features and a sample of size $N = 6435$. We set $k = 100$. The second is a sparse classification instance. We use the regularized logistic loss, $l(\boldsymbol{x}) = \frac{1}{n}\sum_{i\in[n]} \ln(1 + \exp(-\boldsymbol{y}_i(\mathbf{A}\boldsymbol{x})_i)) + \frac{\lambda}{2}\|\boldsymbol{x}\|_2^2$, where $\mathbf{A} \in \mathbb{R}^{n\times d}$ and $\boldsymbol{y} \in \mathbb{R}^n$ are obtained from "hill_valley_with_noise" dataset. The dataset has $d = 100$ features and a sample of size $N = 1212$. We let $\lambda = 0.01$ and $k = 20$. For each instance, we randomly split the sample into training and test data of sizes $\lceil N/2 \rceil$ and $\lfloor N/2 \rfloor$, respectively; we thus create 100 random instances. We consider multi-stage algorithms with various numbers of iterations, $m = k, 0.9k, \ldots, 0.1k$ ($m = k$ corresponds to standard `Greedy`/`OMP`). We set $b_1, \ldots, b_{k-m\lfloor k/m \rfloor}$ at $\lceil k/m \rceil$ and the rest at $\lfloor k/m \rfloor$. We use two baselines based on the projected gradient method: iterative hard thresholding (`IHT`) (Jain et al., 2014) and hard thresholding pursuit (`HTP`) (Yuan et al., 2018). We continue their iterations until the decrease in $l(\cdot)$ value becomes smaller than 10^{-5}. We evaluate the algorithms with running times, loss function values, R^2 scores (for regression), and mean accuracy (for classification); the last two are defined by the corresponding scikit-learn score functions.

Figure 1 summarizes the results. We see that the multi-stage algorithms speed up as m decreases; in particular, `Multi-OMP` becomes as fast as `HTP`. In the regression instances, multi-stage algorithms achieve better loss function values and R^2 scores than the baselines. Other than for `Multi-OMP` with $m = 10$, the decrease in m has negligible effects on loss function values and R^2 scores. In the classification instances, the loss function values of the multi-stage algorithms increase as m decreases, but they are smaller on average than those of `IHT` and `HTP`. The multi-stage algorithms also achieve better mean accuracy than the baselines. To conclude, by using the multi-stage approach, `Greedy` and `OMP` can become faster while outperforming the baselines. When addressing large-scale instances in practice, it would be effective to try multi-stage algorithms with a small m and increase it until an acceptable solution is obtained.

As regards solution quality, the gap between the greedy-style algorithms (`Multi-Greedy` and `Multi-OMP`) and the baselines (`IHT` and `HTP`) can partially be explained in terms of the restricted condition number. For example, `IHT` requires $k \geq \Omega(\kappa_{2k+k^*}^2 \ln \epsilon^{-1})$ to achieve ϵ-errors (Jain et al., 2014), while `Multi-OMP` requires $k \geq \Omega(\kappa_{k+k^*} \ln \epsilon^{-1})$ as implied in Theorem 2. This suggests that greedy-style algorithms can be more resistant to being ill-conditioned (or a large restricted condition number), which is often the case with real-world instances; hence the better performance of the greedy-style algorithms. Appendix B.2 presents further experiments with well- and ill-conditioned instances.

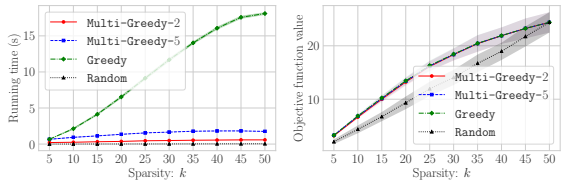


Figure 2: Results of LP with a cardinality constraint. Each curve (error band) indicates the average (standard deviation) calculated over 100 instances.

3.3 LP with a Cardinality Constraint

We use synthetic optimal production planning instances. We let $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^d \mid \mathbf{A}\boldsymbol{x} \leq \mathbf{b}, \mathbf{0} \leq \boldsymbol{x} \leq \mathbf{1}\}$. Each entry of $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\boldsymbol{c} \in \mathbb{R}^d$ is drawn from the uniform distribution on $[0, 1]$. We set $d = 50$, $m = 100$, and $\mathbf{b} = 0.5k \times \mathbf{1}$. We consider various sparsities $k = 5, 10, \ldots, 50$; for each k, we randomly generate 100 instances as above. We consider `Multi-Greedy` with $m = 2$ and $m = 5$, denoted by `Multi-Greedy-2` and `Multi-Greedy-5`, respectively. As baselines, we employ `Greedy` and `Random`, which chooses k elements from $[d]$ uniformly at random.

Figure 2 shows the results. We see that `Multi-Greedy` algorithms run far faster than `Greedy`, and they achieve almost the same objective values as those of `Greedy`. Namely, for optimal production planning instances, the multi-stage strategy can accelerate `Greedy` at a very slight sacrifice of solution quality.

4 THEORETICAL PROPERTIES

We study theoretical properties of WMM: In Section 4.1 we show that WMM is fixed-parameter tractable (FPT) under certain conditions, and in Section 4.2 we prove that no polynomial-time algorithms can improve the $(1 - \mathrm{e}^{-\gamma_{S,k}})$-approximation guarantee in general even if SBR and SPR are bounded by some constants.

4.1 Fixed-parameter Tractability

Here we discuss the computation cost of solving WMM almost optimally. If we are to find an optimal solution for WMM, a naive approach is exhaustive search; i.e., we examine $F(\mathsf{S})$ for all $\mathsf{S} \subseteq [d]$ of size k. This, however, incurs $\Omega(d^k)$ computation cost, which becomes too large as the instance size, d, increases. Taking this into account, the following question arises: Can we solve WMM (almost) optimally without requiring an $\Omega(d^k)$ computation cost? To answer this, we use the parameterized complexity framework (Cygan et al., 2015). We regard a part of the input as a fixed parameter(s), which is denoted by $\mathbf{p}$ and does not include the instance size, d. An algorithm is said to be FPT if it runs in $g(\mathbf{p}) \times \mathrm{poly}(d)$ time, where g is a computable function of $\mathbf{p}$. Note that, if k is a fixed parameter, algorithms that require $\Omega(d^k)$ time, including exhaustive search, are not FPT. Here, regarding k as a part of the fixed parameters, we show that ϵ-error solutions for WMM can be computed with a randomized FPT algorithm (Algorithm 2), which was originally developed by Skowron (2017) for a special case of monotone submodular maximization. Algorithm 2 performs `SingleRun`(), a randomized variant of `Greedy`, T times and returns the best solution. We can show that it enjoys the following guarantee for WMM:

Algorithm 2 Randomized FPT algorithm

Execute `SingleRun`() T times and return the best solution.
function `SingleRun`()
  $\mathsf{S}_0 \leftarrow \emptyset$
  for $i = 1, \ldots, k$ **do**
    Choose $j \in [d] \backslash \mathsf{S}_{i-1}$ randomly with probability $\propto F(j \mid \mathsf{S}_{i-1})$.
    $\mathsf{S}_i \leftarrow \mathsf{S}_{i-1} \cup \{j\}$
  return S_k

Theorem 3. *Assume F to be $(\gamma_k, \beta_{k,d})$-WM. Let S^* be an optimal solution for problem (1) and $\tilde{F} \coloneqq F([d]) - F(\mathsf{S}^*)$. For any $\epsilon > 0$, if*

$$T \geq \left\lceil \left(\frac{1}{\gamma_k \beta_{k,d}} \cdot \frac{\tilde{F} + \epsilon}{\epsilon} \right)^k \ln \delta^{-1} \right\rceil,$$

then Algorithm 2 returns solution S satisfying $F(\mathsf{S}) \geq F(\mathsf{S}^*) - \epsilon$ with a probability of at least $1 - \delta$.

The key to proving Theorem 3 is the fact that the probability of choosing $j \in \mathsf{S}^*$ in each iteration can be lower bounded thanks to the weak modularity. We present the proof in Appendix C.

Since F can be evaluated in $\mathrm{poly}(d)$ time as assumed in Section 1.2, Algorithm 2 is FPT if we regard $\mathbf{p} \coloneqq (k, \gamma_k, \beta_{k,d}, \tilde{F}, \epsilon, \delta)$ as fixed parameters. Note that, since $\tilde{F} \leq F([d])$, a sufficiently large T can be computed once we obtain lower bounds of SBR and SPR, which are available for various applications as in Section 2.

While Algorithm 2 is not so practical, Theorem 3 is beneficial for studying the tractability of WMM instances. In particular, we can obtain an interesting corollary related to ℓ_0-constrained minimization from the theorem. Let $\boldsymbol{x}^* \coloneqq \mathrm{argmin}_{\|\boldsymbol{x}\|_0 \leq k} l(\boldsymbol{x})$ be an optimal solution. As shown by Shalev-Shwartz et al. (2010), `Greedy` can find $\boldsymbol{x}$ such that $l(\boldsymbol{x}) \leq l(\boldsymbol{x}^*) + \epsilon$ if $\boldsymbol{x}$ is allowed to have $\Omega(\kappa \ln \epsilon^{-1})$ non-zeros; i.e., there is a trade-off between sparsity $\|\boldsymbol{x}\|_0$ and accuracy ϵ. In practice, however, $\boldsymbol{x}$ is not always allowed to have sufficiently

many non-zeros. For instance, when performing feature selection for medical analysis, the number of features used for predicting a patient's status is limited since to use many features requires the patient to undergo many medical tests, which is a considerable burden. Hence, to reveal whether we can solve ℓ_0-constrained minimization almost optimally with limited sparsity $k \geq \|\boldsymbol{x}\|_0$ is an important research subject. The following corollary, which is obtained from Theorem 3 and the lower bounds of SBR and SPR (Section 2), implies that it is possible at the cost of FPT computation time; i.e, there is a time–accuracy trade-off.

Corollary 3.a. *Let $F(\mathsf{S}) = l(0) - \min_{\mathrm{supp}(\boldsymbol{x}) \subseteq \mathsf{S}} l(\boldsymbol{x})$ for any $\mathsf{S} \subseteq [d]$ and assume l to be μ_{2k}-RSC, μ_{k+1}-RSC, $\nu_{k+1,1}$-RSM, and ν_d-RSM. Let $\tilde{l} \coloneqq l(\boldsymbol{x}^*) - \min_{\boldsymbol{x} \in \mathbb{R}^{[d]}} l(\boldsymbol{x})$. If Algorithm 2 runs with*

$$T \geq \left\lceil \left(\frac{\nu_{k+1,1}}{\mu_{2k}} \cdot \frac{\nu_d}{\mu_{k+1}} \cdot \frac{\tilde{l} + \epsilon}{\epsilon} \right)^k \ln \delta^{-1} \right\rceil$$

and outputs S, then $\boldsymbol{x} = \mathrm{argmin}_{\mathrm{supp}(\boldsymbol{x}') \subseteq \mathsf{S}} l(\boldsymbol{x}')$ satisfies $l(\boldsymbol{x}) \leq l(\boldsymbol{x}^*) + \epsilon$ with a probability of at least $1 - \delta$.

Namely, if we take $\mathbf{p} \coloneqq (k, \mu_{2k}, \mu_{k+1}, \nu_{k+1,1}, \nu_d, \tilde{l}, \epsilon, \delta)$ to be fixed parameters, ϵ-error solutions can be computed in FPT time with a high probability. Note that, unlike the aforementioned guarantee of `Greedy`, Corollary 3.a does not require $\|\boldsymbol{x}\|_0$ to be sufficiently large.

4.2 Hardness Result

We here prove the following hardness of improving the $(1 - \mathrm{e}^{-\gamma_{\mathsf{S},k}})$-approximation guarantee for WMM:

Theorem 4. *Even if F has SBR $\gamma_k = 1$ and SPR $\beta_k \geq 1/2 - o(1)$, no algorithms that evaluate F only on polynomially many subsets can achieve an approximation guarantee that exceeds $1 - \mathrm{e}^{-1} = 1 - \mathrm{e}^{-\gamma_k}$ for problem (1) in general.*

Note that the significance of Theorem 4 comes from SPR β_k that can be bounded by a universal constant: When curvature α, which satisfies $\beta_k \geq 1 - \alpha$, is upper bounded by a universal constant smaller than 1, then a strictly improved approximation ratio, $\frac{1}{\alpha}(1 - \mathrm{e}^{-\alpha\gamma_{\mathsf{S},k}})$, can be obtained thanks to (Bian et al., 2017). Namely, Theorem 4 reveals a non-trivial theoretical gap between SPR β_k and curvature α. Below we describe a proof sketch, and the full proof is presented in Appendix D.1.

Proof sketch. We make a WM function that is hard to maximize approximately. As with the proof of (Nemhauser and Wolsey, 1978), given unknown subset M of size k, we show that to achieve an approximation guarantee that exceeds $1 - \mathrm{e}^{-\gamma_k}$ is at least as hard as to find S such that $|\mathsf{S} \cap \mathsf{M}| > r$ and $|\mathsf{S}| \leq p_r^k \coloneqq 2k - r + 1$, where $r > 0$ is any fixed integer; this cannot be solved via polynomially many queries. To this end, we use F that satisfies the following conditions: $F(\mathsf{S})$ value depends on $|\mathsf{S}|$ and $|\mathsf{S} \cap \mathsf{M}|$ for any $\mathsf{S} \subseteq [d]$ and only on $|\mathsf{S}|$ if $|\mathsf{S} \cap \mathsf{M}| \leq r$ or $|\mathsf{S}| > p_r^k$, which, roughly speaking, means that the information about F values is useless. By using such function F, we can obtain the hardness result. The main difficulty remained in the above proof is to show that F is WM. In particular, obtaining $\beta_k \geq 1/2 - o(1)$ is the most challenging part. To prove this, we first rewrite SPR as $\beta_k = \min_{\mathsf{L},\mathsf{S} \subseteq [d]} \left\{ \frac{F(\mathsf{S}|\mathsf{L})}{\sum_{j \in \mathsf{S}} F(j|\mathsf{L})} \;\middle|\; \mathsf{L} \cap \mathsf{S} = \emptyset, |\mathsf{L}| \leq k, |\mathsf{S}| \leq k \right\}$, where we regard $0/0 = 1$. Then, by carefully designing F and using the fact that $F(\mathsf{S})$ depends only on $|\mathsf{S}|$ and $|\mathsf{S} \cap \mathsf{M}|$, we can lower bound β_k by the minimum value of some function with only three variables, and the minimum value can be proved to be at least $\frac{1}{2} - \frac{1}{2} \cdot \frac{r-1}{2k-r+1}$ (see, Lemma A.4 in Appendix D.1). By letting k increase with d and setting d at a sufficiently large value, we obtain the lower bound on β_k. □

Given solution S of `Greedy`, we always have $\gamma_{\mathsf{S},k} \geq \gamma_k$. Therefore, Theorem 4 implies that, even if β_k ($\leq \beta_{\mathsf{S},k}$) is lower bounded by a value that can be arbitrarily close to 1/2, no polynomial-time algorithms can improve the $(1 - \mathrm{e}^{-\gamma_{\mathsf{S},k}})$-approximation guarantee in general.

We remark that it may be possible to improve the approximation ratio for some easier subclasses of WMM; for example, if γ_d ($\leq \gamma_k$) and β_d ($\leq \beta_k$) are bounded, we may be able to obtain a better ratio than $1 - \mathrm{e}^{-\gamma_d}$ by using β_d. We discuss this topic in Appendix D.2. We also remark that Theorem 4 does not contradict the FPT result (Theorem 3) for the following reason: Theorem 4 is proved by using sparsity k that increases with d, and we cannot regard such a k as a fixed parameter.

5 CONCLUSION

We studied WMM, a class of non-submodular maximization that can model various practical problems. We proved guarantees of multi-stage algorithms, which generalize and improve some existing results, and confirmed their effectiveness via experiments. We then proved the fixed-parameter tractability of WMM, which yields the time–accuracy trade-off for ℓ_0-constrained minimization as a byproduct, and the hardness of improving the $(1 - \mathrm{e}^{-\gamma_{\mathsf{S},k}})$-approximation guarantee.

Recent studies (Khanna et al., 2017b; Qian et al., 2018) provided various techniques for accelerating greedy algorithms, and greedy-style methods for many different problem settings have also been studied (Bogunovic et al., 2018; Fujii and Soma, 2018). It will be interesting to study how to incorporate the multi-stage approach into those methods for further acceleration.

Acknowledgements

The author is grateful to Kaito Fujii and anonymous reviewers for providing valuable comments.

References

Bai, W. and Bilmes, J. (2018). Greed is still good: Maximizing monotone Submodular+Supermodular (BP) functions. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 304–313. PMLR.

Balkanski, E. and Singer, Y. (2018). The adaptive complexity of maximizing a submodular function. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1138–1151. ACM.

Bian, A. A., Buhmann, J. M., Krause, A., and Tschiatschek, S. (2017). Guarantees for greedy maximization of non-submodular functions with applications. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 498–507. PMLR.

Bogunovic, I., Zhao, J., and Cevher, V. (2018). Robust maximization of non-submodular objectives. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 890–899. PMLR.

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

Candès, E. J., Romberg, J. K., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223.

Chen, L., Feldman, M., and Karbasi, A. (2018). Weakly submodular maximization beyond cardinality constraints: Does randomization help greedy? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 804–813. PMLR.

Chierichetti, F., Das, A., Dasgupta, A., and Kumar, R. (2015). Approximate modularity. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1143–1162.

Conforti, M. and Cornuéjols, G. (1984). Submodular set functions, matroids and the greedy algorithm: Tight worst-case bounds and some generalizations of the Rado-Edmonds theorem. *Discrete Appl. Math.*, 7(3):251–274.

Cygan, M., Fomin, F. V., Kowalik, L., Lokshtanov, D., Marx, D., Pilipczuk, M., Pilipczuk, M., and Saurabh, S. (2015). *Parameterized Algorithms*. Springer Publishing Company, Incorporated, 1st edition.

Das, A. and Kempe, D. (2018). Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection. *J. Mach. Learn. Res.*, 19(3):1–34.

Elenberg, E. R., Dimakis, A. G., Feldman, M., and Karbasi, A. (2017). Streaming weak submodularity: Interpreting neural networks on the fly. In *Advances in Neural Information Processing Systems 30*, pages 4044–4054. Curran Associates, Inc.

Elenberg, E. R., Khanna, R., Dimakis, A. G., and Negahban, S. (2018). Restricted strong convexity implies weak submodularity. *Ann. Statist.*, 46(6B):3539–3568.

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

Feige, U. and Izsak, R. (2013). Welfare maximization and the supermodular degree. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, pages 247–256. ACM.

Fujii, K. and Soma, T. (2018). Fast greedy algorithms for dictionary selection with generalized sparsity constraints. In *Advances in Neural Information Processing Systems 31*, pages 4749–4758. Curran Associates, Inc.

Harshaw, C., Feldman, M., Ward, J., and Karbasi, A. (2019). Submodular maximization beyond non-negativity: Guarantees, fast algorithms, and applications. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2634–2643. PMLR.

Horel, T. and Singer, Y. (2016). Maximization of approximately submodular functions. In *Advances in Neural Information Processing Systems 29*, pages 3045–3053. Curran Associates, Inc.

Hu, H., Grubb, A., Bagnell, J. A., and Hebert, M. (2016). Efficient feature group sequencing for anytime linear prediction. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, pages 279–288. AUAI Press.

Iyer, R. K., Jegelka, S., and Bilmes, J. A. (2013). Curvature and optimal algorithms for learning and minimizing submodular functions. In *Advances in Neural Information Processing Systems 26*, pages 2742–2750. Curran Associates, Inc.

Jain, P., Tewari, A., and Kar, P. (2014). On iterative hard thresholding methods for high-dimensional M-estimation. In *Advances in Neural Information Processing Systems 27*, pages 685–693. Curran Associates, Inc.

Khanna, R., Elenberg, E. R., Dimakis, A. G., Ghosh, J., and Negahban, S. (2017a). On approximation guarantees for greedy low rank optimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1837–1846. PMLR.

Khanna, R., Elenberg, E. R., Dimakis, A. G., Negahban, S., and Ghosh, J. (2017b). Scalable greedy feature selection via weak submodularity. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 1560–1568. PMLR.

Krause, A. and Cevher, V. (2010). Submodular dictionary selection for sparse representation. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 567–574. Omnipress.

Kumar, R., Moseley, B., Vassilvitskii, S., and Vattani, A. (2015). Fast greedy algorithms in mapreduce and streaming. *ACM Trans. Parallel Comput.*, 2(3):14:1–14:22.

Li, X., Zhao, T., Arora, R., Liu, H., and Haupt, J. (2016). Stochastic variance reduced optimization for nonconvex sparse learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 917–925. PMLR.

Liberty, E. and Sviridenko, M. (2017). Greedy minimization of weakly supermodular set functions. In *Proceedings of Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 81, pages 19:1–19:11. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Lin, H. and Bilmes, J. A. (2011). A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 510–520. Association for Computational Linguistics.

Marsousi, M., Abhari, K., Babyn, P., and Alirezaie, J. (2013). MULTI-STAGE OMP sparse coding using local matching pursuit atoms selection. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1783–1787.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM J. Optim.*, 24(2):227–234.

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

Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. (1978). An analysis of approximations for maximizing submodular set functions-I. *Math. Program.*, 14(1):265–294.

Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., and Moore, J. H. (2017). PMLB: A large benchmark suite for machine learning evaluation and comparison. *BioData Min.*, 10(1):36.

Qian, C., Yu, Y., and Tang, K. (2018). Approximation guarantees of stochastic greedy algorithms for subset selection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1478–1484. International Joint Conferences on Artificial Intelligence Organization.

Qian, S. and Singer, Y. (2019). Fast parallel algorithms for statistical subset selection problems. In *Advances in Neural Information Processing Systems 32*, pages 5073–5082. Curran Associates, Inc.

Shalev-Shwartz, S., Srebro, N., and Zhang, T. (2010). Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM J. Optim.*, 20(6):2807–2832.

Shalev-Shwartz, S. and Zhang, T. (2016). Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Math. Program.*, 155(1):105–145.

Skowron, P. (2017). FPT approximation schemes for maximizing submodular functions. *Inform. Comput.*, 257:65–78.

Soma, T. and Yoshida, Y. (2018). A new approximation guarantee for monotone submodular function maximization via discrete convexity. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming*, volume 107, pages 99:1–99:14. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Sviridenko, M., Vondrák, J., and Ward, J. (2015). Optimal approximation for submodular and supermodular optimization with bounded curvature. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1134–1148. SIAM.

Takeda, A., Niranjan, M., Gotoh, J., and Kawahara, Y. (2013). Simultaneous pursuit of out-of-sample performance and sparsity in index tracking portfolios. *Comput. Manag. Sci.*, 10(1):21–49.

Wang, Z., Moran, B., Wang, X., and Pan, Q. (2016). Approximation for maximizing monotone non-decreasing set functions with a greedy method. *J. Comb. Optim.*, 31(1):29–43.

Wei, K., Iyer, R., and Bilmes, J. (2014). Fast multi-stage submodular maximization. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1494–1502. PMLR.

Yuan, X.-T., Li, P., and Zhang, T. (2018). Gradient hard thresholding pursuit. *J. Mach. Learn. Res.*, 18(166):1–43.

Zhou, Y. and Spanos, C. J. (2016). Causal meets submodular: Subset selection with directed information. In *Advances in Neural Information Processing Systems 29*, pages 2649–2657. Curran Associates, Inc.

Appendix

In Appendix A, we derive the lower bounds of SBR and SPR for each application presented in Section 2, and we also provide example instances such that $\alpha = 1$ holds even if SBR and SPR are bounded. In Appendix B, we prove the guarantees of the multi-stage algorithms, and we also present additional experimental results with well- and ill-conditioned synthetic ℓ_0-constrained minimization instances. In Appendix C, we prove the guarantee of the FPT algorithm. In Appendix D, we present the proof of the hardness result.

A APPLICATIONS

We show that SBR and SPR are lower bounded for each application presented in Section 2. We also present example instances such that $\alpha = 1$ holds even if SBR and SPR are lower bounded; regarding ℓ_0-constrained minimization, we show that *inverse curvature* $\check{\alpha} \in [0, 1]$ (Bogunovic et al., 2018) can also become equal to 1. Note that curvature α and inverse curvature $\check{\alpha}$ of F are defined as the smallest scalars that satisfy

$$F(j \mid \mathsf{S}\backslash\{j\} \cup \mathsf{M}) \geq (1-\alpha)F(j \mid \mathsf{S}\backslash\{j\}) \qquad \text{and} \qquad F(j \mid \mathsf{S}\backslash\{j\}) \geq (1-\check{\alpha})F(j \mid \mathsf{S}\backslash\{j\} \cup \mathsf{M}),$$

respectively, for any $\mathsf{S}, \mathsf{M} \subseteq [d]$ and $j \in \mathsf{S}\backslash\mathsf{M}$. Function F is submodular (supermodular) iff $\check{\alpha} = 0$ ($\alpha = 0$). As shown in (Bogunovic et al., 2018), for any $\mathsf{U} \subseteq [d]$ and $s \in \mathbb{Z}_{>0}$, we have

$$\gamma_{\mathsf{U},s} \geq 1 - \check{\alpha} \qquad \text{and} \qquad \beta_{\mathsf{U},s} \geq 1 - \alpha.$$

Namely, while the bounded curvature (inverse curvature) implies bounded SPR (SBR), the opposite is not always true.

A.1 ℓ_0-constrained Minimization

Lower Bounds of SBR and SPR We first introduce some definitions required in the following discussion. Given $\Omega \subseteq \mathbb{R}^{[d]} \times \mathbb{R}^{[d]}$, we say l is μ_Ω-RSC and ν_Ω-RSM if it satisfies

$$\frac{\mu_\Omega}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \leq l(\boldsymbol{y}) - l(\boldsymbol{x}) - \langle \nabla l(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle \leq \frac{\nu_\Omega}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2$$

for all $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega$. For convenience, we define $f(\boldsymbol{x}) \coloneqq l(0) - l(\boldsymbol{x})$. Note that we have

$$F(\mathsf{S}) = l(0) - \min_{\mathrm{supp}(\boldsymbol{x}) \subseteq \mathsf{S}} l(\boldsymbol{x}) = \max_{\mathrm{supp}(\boldsymbol{x}) \subseteq \mathsf{S}} f(\boldsymbol{x})$$

for any $\mathsf{S} \subseteq [d]$. If l is μ_Ω-RSC and ν_Ω-RSM, then f is μ_Ω-restricted strong concave (μ_Ω-RSC) and ν_Ω-restricted smooth (ν_Ω-RSM) as follows:

$$-\frac{\mu_\Omega}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \geq f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle \geq -\frac{\nu_\Omega}{2}\|\boldsymbol{y}-\boldsymbol{x}\|_2^2 \tag{A.1}$$

for any $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega$. We employ the following definitions for convenience:

- If (A.1) holds with

  $$\Omega = \Omega_{s_1,s_2} \coloneqq \{(\boldsymbol{x}, \boldsymbol{y}) \mid \|\boldsymbol{x}\|_0 \leq s_1, \|\boldsymbol{y}\|_0 \leq s_1, \text{ and } \|\boldsymbol{x}-\boldsymbol{y}\|_0 \leq s_2\},$$

  we say f is μ_{s_1,s_2}-RSC and ν_{s_1,s_2}-RSM. For simplicity, we define $\mu_s \coloneqq \mu_{s,s}$ and $\nu_s \coloneqq \nu_{s,s}$.
- Given $\mathsf{A}, \mathsf{B} \subseteq [d]$, if (A.1) holds with

  $$\Omega = \Omega_{\mathsf{A},\mathsf{B}} \coloneqq \{(\boldsymbol{x}, \boldsymbol{y}) \mid \mathrm{supp}(\boldsymbol{x}) \subseteq \mathsf{A}, \mathrm{supp}(\boldsymbol{y}) \subseteq \mathsf{B}\},$$

  we say f is $\mu_{\mathsf{A},\mathsf{B}}$-RSC and $\nu_{\mathsf{A},\mathsf{B}}$-RSM.
- Given $\mathsf{A} \subseteq \mathsf{B} \subseteq [d]$, if (A.1) holds with

  $$\Omega = \tilde{\Omega}_{\mathsf{A},\mathsf{B}} \coloneqq \{(\boldsymbol{x}, \boldsymbol{y}) \mid \mathrm{supp}(\boldsymbol{x}) \subseteq \mathsf{A}, \mathrm{supp}(\boldsymbol{y}) \subseteq \mathsf{B}, \text{ and } \mathrm{supp}(\boldsymbol{y}-\boldsymbol{x}) \subseteq \mathsf{B}\backslash\mathsf{A}\},$$

  we say f is $\tilde{\mu}_{\mathsf{A},\mathsf{B}}$-RSC and $\tilde{\nu}_{\mathsf{A},\mathsf{B}}$-RSM.

Given any Ω' and Ω satisfying $\Omega' \subseteq \Omega$, we can set $\mu_{\Omega'}$ and $\nu_{\Omega'}$ so that we have $\mu_{\Omega'} \geq \mu_{\Omega}$ and $\nu_{\Omega'} \leq \nu_{\Omega}$, respectively. In particular, we often use the following inequalities:

- For any $0 \leq s_1' \leq s_1$ and $0 \leq s_2' \leq s_2$, we have $\mu_{s_1,s_2} \leq \mu_{s_1',s_2'}$ and $\nu_{s_1,s_2} \geq \nu_{s_1',s_2'}$.
- For any $\mathsf{A}, \mathsf{B} \subseteq [d]$, we have $\mu_{|\mathsf{A}\cup\mathsf{B}|} \leq \mu_{\mathsf{A},\mathsf{B}}$ and $\nu_{|\mathsf{A}\cup\mathsf{B}|} \geq \nu_{\mathsf{A},\mathsf{B}}$.
- For any $\mathsf{A} \subseteq \mathsf{B} \subseteq [d]$, we have $\mu_{|\mathsf{B}|,|\mathsf{B}\setminus\mathsf{A}|} \leq \tilde{\mu}_{\mathsf{A},\mathsf{B}}$ and $\nu_{|\mathsf{B}|,|\mathsf{B}\setminus\mathsf{A}|} \geq \tilde{\nu}_{\mathsf{A},\mathsf{B}}$.

The following lemma is the key to obtaining the lower bounds of SBR and SPR, and we will also use it when proving the guarantee of `Multi-OMP`. A special case of the lemma is implicitly used in (Elenberg et al., 2018), but we here state and prove it clearly for completeness and convenience.

Lemma A.1. *For any $\mathsf{A} \subseteq [d]$, let $\mathbf{b}^{(\mathsf{A})} \coloneqq \operatorname{argmax}_{\operatorname{supp}(\boldsymbol{x})\subseteq\mathsf{A}} f(\boldsymbol{x})$. For any disjoint $\mathsf{A}, \mathsf{B} \subseteq [d]$, if f is $\mu_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}$-RSC and $\tilde{\nu}_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}$-RSM, we have*

$$\frac{1}{2\tilde{\nu}_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}\|\nabla f(\mathbf{b}^{(\mathsf{A})})_{\mathsf{B}}\|_2^2 \leq F(\mathsf{B} \mid \mathsf{A}) \leq \frac{1}{2\mu_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}\|\nabla f(\mathbf{b}^{(\mathsf{A})})_{\mathsf{B}}\|_2^2.$$

Proof. We show the first inequality. Since $\mathbf{b}^{(\mathsf{A}\cup\mathsf{B})}$ is the maximizer of f over $\{\boldsymbol{x} \in \mathbb{R}^{[d]} \mid \operatorname{supp}(\boldsymbol{x}) \subseteq \mathsf{A}\cup\mathsf{B}\}$, we have $f(\mathbf{b}^{(\mathsf{A}\cup\mathsf{B})}) \geq f(\boldsymbol{w} + \mathbf{b}^{(\mathsf{A})})$ for any $\operatorname{supp}(\boldsymbol{w}) \subseteq \mathsf{B}$. Therefore, from inequality (A.1), we obtain

$$F(\mathsf{B} \mid \mathsf{A}) = f(\mathbf{b}^{(\mathsf{A}\cup\mathsf{B})}) - f(\mathbf{b}^{(\mathsf{A})}) \geq f(\boldsymbol{w} + \mathbf{b}^{(\mathsf{A})}) - f(\mathbf{b}^{(\mathsf{A})}) \geq \langle \nabla f(\mathbf{b}^{(\mathsf{A})}), \boldsymbol{w}\rangle - \frac{\tilde{\nu}_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}{2}\|\boldsymbol{w}\|_2^2.$$

Setting $\boldsymbol{w}_{\mathsf{B}} = \frac{1}{\tilde{\nu}_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}\nabla f(\mathbf{b}^{(\mathsf{A})})_{\mathsf{B}}$ and $\boldsymbol{w}_{[d]\setminus\mathsf{B}} = 0$, we obtain the first inequality:

$$F(\mathsf{B} \mid \mathsf{A}) \geq \frac{1}{2\tilde{\nu}_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}\|\nabla f(\mathbf{b}^{(\mathsf{A})})_{\mathsf{B}}\|_2^2.$$

We then prove the second inequality. Thanks to inequality (A.1), we have

$$F(\mathsf{B} \mid \mathsf{A}) = f(\mathbf{b}^{(\mathsf{A}\cup\mathsf{B})}) - f(\mathbf{b}^{(\mathsf{A})}) \leq \langle \nabla f(\mathbf{b}^{(\mathsf{A})}), \mathbf{b}^{(\mathsf{A}\cup\mathsf{B})} - \mathbf{b}^{(\mathsf{A})}\rangle - \frac{\mu_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}{2}\|\mathbf{b}^{(\mathsf{A}\cup\mathsf{B})} - \mathbf{b}^{(\mathsf{A})}\|_2^2.$$

Let $\boldsymbol{w} \in \mathbb{R}^{[d]}$ be a vector such that $\operatorname{supp}(\boldsymbol{w}) \subseteq \mathsf{A}\cup\mathsf{B}$. We consider replacing $\mathbf{b}^{(\mathsf{A}\cup\mathsf{B})}$ in RHS with $\boldsymbol{w} + \mathbf{b}^{(\mathsf{A})}$ and maximizing RHS w.r.t. $\boldsymbol{w}$; we thus obtain an upper bound of $F(\mathsf{B} \mid \mathsf{A})$ as follows:

$$F(\mathsf{B} \mid \mathsf{A}) \leq \max_{\operatorname{supp}(\boldsymbol{w})\subseteq\mathsf{A}\cup\mathsf{B}} \langle \nabla f(\mathbf{b}^{(\mathsf{A})}), \boldsymbol{w}\rangle - \frac{\mu_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}{2}\|\boldsymbol{w}\|_2^2.$$

The maximum is attained with $\boldsymbol{w}_{\mathsf{A}\cup\mathsf{B}} = \frac{1}{\mu_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}\nabla f(\mathbf{b}^{(\mathsf{A})})_{\mathsf{A}\cup\mathsf{B}}$, and so we obtain

$$F(\mathsf{B} \mid \mathsf{A}) \leq \frac{1}{2\mu_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}\|\nabla f(\mathbf{b}^{(\mathsf{A})})_{\mathsf{A}\cup\mathsf{B}}\|_2^2 = \frac{1}{2\mu_{\mathsf{A},\mathsf{A}\cup\mathsf{B}}}\|\nabla f(\mathbf{b}^{(\mathsf{A})})_{\mathsf{B}}\|_2^2,$$

where the last equality comes from the first-order optimality condition (or the KKT condition with the linear independence constraint qualification) at $\mathbf{b}^{(\mathsf{A})}$: $\nabla f(\mathbf{b}^{(\mathsf{A})})_{\mathsf{A}} = 0$. □

By using this lemma, we can show that SBR and SPR can be lower bounded by ratios of RSC and RSM constants. The lower bound of SBR is adopted from (Elenberg et al., 2018), and that of SPR improves the existing one presented in (Bogunovic et al., 2018).

Proposition A.1. *For any $\mathsf{U} \subseteq [d]$ and $s \in \mathbb{Z}_{>0}$, SBR $\gamma_{\mathsf{U},s}$ and SPR $\beta_{\mathsf{U},s}$ of $F(\mathsf{S}) = l(0) - \min_{\operatorname{supp}(\boldsymbol{x})\subseteq\mathsf{S}} l(\boldsymbol{x})$ ($\forall \mathsf{S} \subseteq [d]$) are bounded with RSC and RSM constants of l as follows:*

$$\gamma_{\mathsf{U},s} \geq \frac{\mu_{|\mathsf{U}|+s}}{\nu_{|\mathsf{U}|+1,1}} \geq \frac{\mu_{|\mathsf{U}|+s}}{\nu_{|\mathsf{U}|+s}} = \frac{1}{\kappa_{|\mathsf{U}|+s}} \qquad \textit{and} \qquad \beta_{\mathsf{U},s} \geq \frac{\mu_{|\mathsf{U}|+1}}{\nu_{|\mathsf{U}|+s,s}} \geq \frac{\mu_{|\mathsf{U}|+s}}{\nu_{|\mathsf{U}|+s}} = \frac{1}{\kappa_{|\mathsf{U}|+s}}.$$

Proof. We refer readers to (Elenberg et al., 2018) for the proof of the lower bound of $\gamma_{\mathsf{U},s}$. Here, we show how to obtain the lower bound of $\beta_{\mathsf{U},s}$. From the definition of SPR, we have

$$\beta_{\mathsf{U},s} \coloneqq \min_{\substack{\mathsf{L},\mathsf{S}\,:\,\mathsf{L}\cap\mathsf{S}=\emptyset,\\ \mathsf{L}\subseteq\mathsf{U},|\mathsf{S}|\leq s}} \frac{F(\mathsf{S}\mid\mathsf{L})}{\sum_{j\in\mathsf{S}}F(j\mid\mathsf{L})},$$

where we regard $0/0=1$. Therefore, we obtain

$$\begin{aligned}
\beta_{\mathsf{U},s} &\geq \min_{\substack{\mathsf{L},\mathsf{S}\,:\,\mathsf{L}\cap\mathsf{S}=\emptyset,\\ \mathsf{L}\subseteq\mathsf{U},|\mathsf{S}|\leq s}} \frac{\|\nabla f(\mathbf{b}^{(\mathsf{L})})_{\mathsf{S}}\|_2^2}{2\tilde{\nu}_{\mathsf{L},\mathsf{L}\cup\mathsf{S}}}\left(\sum_{j\in\mathsf{S}}\frac{|\nabla f(\mathbf{b}^{(\mathsf{L})})_j|^2}{2\mu_{\mathsf{L},\mathsf{L}\cup\{j\}}}\right)^{-1} && \because \text{Lemma A.1}\\
&\geq \min_{\substack{\mathsf{L},\mathsf{S}\,:\,\mathsf{L}\cap\mathsf{S}=\emptyset,\\ \mathsf{L}\subseteq\mathsf{U},|\mathsf{S}|\leq s}} \frac{\mu_{|\mathsf{U}|+1}}{\tilde{\nu}_{\mathsf{L},\mathsf{L}\cup\mathsf{S}}}\cdot\frac{\|\nabla f(\mathbf{b}^{(\mathsf{L})})_{\mathsf{S}}\|_2^2}{\sum_{j\in\mathsf{S}}|\nabla f(\mathbf{b}^{(\mathsf{L})})_j|^2} && \because \mu_{\mathsf{L},\mathsf{L}\cup\{j\}}\geq\mu_{|\mathsf{U}|+1}\\
&= \min_{\substack{\mathsf{L},\mathsf{S}\,:\,\mathsf{L}\cap\mathsf{S}=\emptyset,\\ \mathsf{L}\subseteq\mathsf{U},|\mathsf{S}|\leq s}} \frac{\mu_{|\mathsf{U}|+1}}{\tilde{\nu}_{\mathsf{L},\mathsf{L}\cup\mathsf{S}}} && \because \|\nabla f(\mathbf{b}^{(\mathsf{L})})_{\mathsf{S}}\|_2^2=\sum_{j\in\mathsf{S}}|\nabla f(\mathbf{b}^{(\mathsf{L})})_j|^2\\
&\geq \frac{\mu_{|\mathsf{U}|+1}}{\nu_{|\mathsf{U}|+s,s}}. && \because \tilde{\nu}_{\mathsf{L},\mathsf{L}\cup\mathsf{S}}\leq\nu_{|\mathsf{U}|+s,s}
\end{aligned}$$

The proof is completed with $\nu_{|\mathsf{U}|+s,s}\leq\nu_{|\mathsf{U}|+s}$ and $\mu_{|\mathsf{U}|+1}\geq\mu_{|\mathsf{U}|+s}$. □

Example with Unbounded Curvature We show that there is an ℓ_0-constrained minimization instance that satisfies the following conditions: SBR and SPR of $F(\mathsf{S})=l(0)-\min_{\mathrm{supp}(\boldsymbol{x})\subseteq\mathsf{S}}l(\boldsymbol{x})$ are bounded by a constant, while its curvature α and inverse curvature $\breve{\alpha}$ are unbounded (i.e., $\alpha=\breve{\alpha}=1$). We define

$$\mathbf{B}\coloneqq\begin{bmatrix}1&1\\0&1\end{bmatrix},\qquad \boldsymbol{a}_1\coloneqq\begin{bmatrix}0\\1\end{bmatrix},\qquad \text{and}\qquad \boldsymbol{a}_2\coloneqq\begin{bmatrix}1\\1\end{bmatrix}.$$

Note that we have

$$\begin{aligned}
&\min_{x_1,x_2\in\mathbb{R}}\left\|\mathbf{B}\begin{bmatrix}x_1\\x_2\end{bmatrix}-\boldsymbol{a}_1\right\|_2^2=0, && \min_{x_1\in\mathbb{R}}\left\|\mathbf{B}\begin{bmatrix}x_1\\0\end{bmatrix}-\boldsymbol{a}_1\right\|_2^2=1, && \min_{x_2\in\mathbb{R}}\left\|\mathbf{B}\begin{bmatrix}0\\x_2\end{bmatrix}-\boldsymbol{a}_1\right\|_2^2=1/2,\\
&\min_{x_3,x_4\in\mathbb{R}}\left\|\mathbf{B}\begin{bmatrix}x_3\\x_4\end{bmatrix}-\boldsymbol{a}_2\right\|_2^2=0, && \min_{x_3\in\mathbb{R}}\left\|\mathbf{B}\begin{bmatrix}x_3\\0\end{bmatrix}-\boldsymbol{a}_2\right\|_2^2=1, && \min_{x_4\in\mathbb{R}}\left\|\mathbf{B}\begin{bmatrix}0\\x_4\end{bmatrix}-\boldsymbol{a}_2\right\|_2^2=0.
\end{aligned}$$

We define the loss function as $l(\boldsymbol{x})\coloneqq\|\mathbf{A}\boldsymbol{x}-\boldsymbol{y}\|_2^2$, where $\mathbf{A}\in\mathbb{R}^{[d]\times[d]}$ is a block-diagonal matrix and $\boldsymbol{y}\in\mathbb{R}^{[d]}$ is a vector defined as

$$\mathbf{A}\coloneqq\begin{bmatrix}\mathbf{B}&&&&\\&\mathbf{B}&&&\\&&1&&\\&&&\ddots&\\&&&&1\end{bmatrix}\qquad\text{and}\qquad \boldsymbol{y}\coloneqq\begin{bmatrix}\boldsymbol{a}_1\\\boldsymbol{a}_2\\0\\\vdots\\0\end{bmatrix},$$

respectively. We let $F(\mathsf{S})=l(0)-\min_{\mathrm{supp}(\boldsymbol{x})\subseteq\mathsf{S}}l(\boldsymbol{x})$ for any $\mathsf{S}\subseteq[d]$. Then we have

$$F(\{1\}\mid\{2\})=1/2,\qquad F(\{1\})=0,\qquad F(\{3\}\mid\{4\})=0,\qquad\text{and}\qquad F(\{3\})=1.$$

Since $\alpha,\breve{\alpha}\in[0,1]$ must satisfy

$$F(\{1\})\geq(1-\breve{\alpha})F(\{1\}\mid\{2\})\qquad\text{and}\qquad F(\{3\}\mid\{4\})\geq(1-\alpha)F(\{3\}),$$

we have $\alpha=\breve{\alpha}=1$. On the other hand, the condition number, κ, of l is bounded from above by the ratio of the largest and smallest eigenvalues of $\mathbf{A}^\top\mathbf{A}$, which are equal to $\frac{3+\sqrt{5}}{2}$ and $\frac{3-\sqrt{5}}{2}$, respectively; hence $\kappa\leq\frac{3+\sqrt{5}}{3-\sqrt{5}}$. Therefore, thanks to Proposition A.1, we have $\gamma_{\mathsf{U},s}\geq\frac{3-\sqrt{5}}{3+\sqrt{5}}$ and $\beta_{\mathsf{U},s}\geq\frac{3-\sqrt{5}}{3+\sqrt{5}}$ for any U and s.

A.2 LP with a Cardinality Constraint

Lower Bounds of SBR and SPR As in Section 2, SPR is lower bounded as $\beta_{\mathsf{U},s} \geq 1/s$. Furthermore, as shown in (Bian et al., 2017), SBR is lower bounded by some $\gamma_0 > 0$ under the non-degeneracy assumption: For any $\mathsf{S} \subseteq [d]$, any basic feasible solution of the corresponding LP in the standard form is non-degenerate.

Example with Unbounded Curvature An example with $\alpha = 1$ is provided in Section 2. Note that the example instance satisfies the non-degeneracy assumption.

A.3 Coverage Maximization

Lower Bounds of SBR and SPR Recall that the coverage function is defined as $F(\mathsf{S}) \coloneqq \sum_{v \in I_{\mathsf{S}}} w_v$, where $w_v \geq 0$ $(v \in V)$, $I_j \subseteq V$ $(j \in [d])$, and $I_{\mathsf{S}} \coloneqq \bigcup_{j \in \mathsf{S}} I_j$ for any $\mathsf{S} \subseteq [d]$. Since the function is submodular, we have $\gamma_{\mathsf{U},s} = 1$ for any U and s. As assumed in Section 2, any collection of up to s groups covers any $v \in V$ at most c_s times; i.e., $c_s \coloneqq \max_{v \in V, |\mathsf{S}| \leq s} |\{j \in \mathsf{S} \mid v \in I_j\}|$. Therefore,

$$\frac{F(\mathsf{S} \mid \mathsf{L})}{\sum_{j \in \mathsf{S}} F(j \mid \mathsf{L})} = \frac{\sum_{v \in I_{\mathsf{S} \cup \mathsf{L}} \setminus I_{\mathsf{L}}} w_v}{\sum_{j \in \mathsf{S}} \sum_{v \in I_{\mathsf{L} \cup j} \setminus I_{\mathsf{L}}} w_v} = \frac{\sum_{v \in I_{\mathsf{S} \cup \mathsf{L}} \setminus I_{\mathsf{L}}} w_v}{\sum_{v \in I_{\mathsf{S} \cup \mathsf{L}} \setminus I_{\mathsf{L}}} w_v |\{j \in \mathsf{S} \mid v \in I_j\}|} \geq \frac{1}{c_s}$$

holds for any disjoint $\mathsf{L}, \mathsf{S} \subseteq [d]$ such that $|\mathsf{S}| \leq s$, which implies $\beta_{\mathsf{U},s} \geq 1/c_s$ for any U and s.

Example with Unbounded Curvature We provide an example of a coverage function with bounded SPR $\beta_{\mathsf{U},s}$ and unbounded curvature $\alpha = 1$. Let $V = \{v_1, v_2, v_3\}$ and $w_v = 1$ $(v \in V)$; i.e., $F(\mathsf{S}) = |I_{\mathsf{S}}|$. We let $d = 3$ and define $I_1 = \{v_1, v_2\}$, $I_2 = \{v_2, v_3\}$, and $I_3 = \{v_1, v_3\}$. Since each $v \in V$ is covered by at most two groups, we have $c_s = 2$ for any s, which implies $\beta_{\mathsf{U},s} \geq 1/2$ for any U and s. On the other hand, we have $F(\{1\} \mid \{2, 3\}) = 0$ and $F(\{1\}) = 2$, which leads to $\alpha = 1$ since α must satisfy $F(j \mid \mathsf{S}) \geq (1 - \alpha)F(j)$ for any $\mathsf{S} \subseteq [d]$ and $j \notin \mathsf{S}$.

B MULTI-STAGE ALGORITHMS

We prove the theoretical guarantees of the multi-stage algorithms in Appendix B.1, and we present experimental results with well- and ill-conditioned synthetic ℓ_0-constrained instances in Appendix B.2.

Algorithm 1 Multi-stage algorithm

$\mathsf{U} \leftarrow [d]$, $\mathsf{S} \leftarrow \emptyset$
for $i = 1, \dots, m$ **do**
  $\mathsf{B}_i \leftarrow \operatorname{argmax}_{\mathsf{B} \subseteq \mathsf{U} : |\mathsf{B}| \leq b_i} G_{\mathsf{S}}(\mathsf{B})$
  $\mathsf{S} \leftarrow \mathsf{S} \cup \mathsf{B}_i$
  $\mathsf{U} \leftarrow \mathsf{U} \backslash \mathsf{B}_i$
return S

B.1 Theoretical Guarantees

Note that the surrogate functions, G_{S}, considered below are monotone, which means we have $|\mathsf{B}_i| = b_i$ in each i-th iteration. Let $\mathsf{S}_i = \mathsf{B}_1 \cup \dots \cup \mathsf{B}_i$ for $i \in [m]$ and $\mathsf{S}_0 = \emptyset$. We take S^* and $\boldsymbol{x}^*$, which satisfy $k^* = |\mathsf{S}^*| = \|\boldsymbol{x}^*\|_0$, to be target solutions for WMM and ℓ_0-constrained minimization, respectively. As is usual with the proof of greedy-style algorithms, we obtain approximation guarantees from a lower bound of the marginal gain in each iteration as in the following lemma:

Lemma A.2. *Given any $\theta_1, \dots, \theta_m$ such that $\theta_i \in [0, 1]$ ($i \in [m]$), if we can find $\mathsf{B}_i \subseteq [d]$ such that $b_i = |\mathsf{B}_i| \leq k^*$ and*

$$F(\mathsf{B}_i \mid \mathsf{S}_{i-1}) \geq \theta_i \frac{b_i}{k^*} (F(\mathsf{S}^*) - F(\mathsf{S}_{i-1})) \tag{A.2}$$

in each i-th iteration ($i \in [m]$), then the following inequality holds:

$$F(\mathsf{S}_m) \geq \left(1 - \prod_{i=1}^{m} \left(1 - \theta_i \frac{b_i}{k^*}\right)\right) F(\mathsf{S}^*) \geq \left(1 - \exp\left(-\frac{1}{k^*} \sum_{i=1}^{m} \theta_i b_i\right)\right) F(\mathsf{S}^*).$$

Proof. We first prove that

$$F(\mathsf{S}_i) \geq \left(1 - \prod_{i'=1}^{i} \left(1 - \theta_{i'} \frac{b_{i'}}{k^*}\right)\right) F(\mathsf{S}^*).$$

holds for $i = 1, \dots, m$ by induction. If $i = 1$, the inequality holds due to (A.2). Assume that we have

$$F(\mathsf{S}_{i-1}) \geq \left(1 - \prod_{i'=1}^{i-1} \left(1 - \theta_{i'} \frac{b_{i'}}{k^*}\right)\right) F(\mathsf{S}^*). \tag{A.3}$$

Then we obtain

$$\begin{aligned}
F(\mathsf{S}_i) &\geq \theta_i \frac{b_i}{k^*} (F(\mathsf{S}^*) - F(\mathsf{S}_{i-1})) + F(\mathsf{S}_{i-1}) && \because \text{(A.2)} \\
&\geq \theta_i \frac{b_i}{k^*} F(\mathsf{S}^*) + \left(1 - \theta_i \frac{b_i}{k^*}\right) \left(1 - \prod_{i'=1}^{i-1} \left(1 - \theta_{i'} \frac{b_{i'}}{k^*}\right)\right) F(\mathsf{S}^*) && \because \text{(A.3)} \\
&= \left(1 - \prod_{i'=1}^{i} \left(1 - \theta_{i'} \frac{b_{i'}}{k^*}\right)\right) F(\mathsf{S}^*).
\end{aligned}$$

Therefore, the above inequality holds for any $i \in [m]$ by induction. By setting $i = m$, we obtain the first inequality in Lemma A.2. We then prove the second inequality. Since $\theta_i \frac{b_i}{k^*} \in [0, 1]$ ($i \in [m]$), the arithmetic mean of $1 - \theta_1 \frac{b_1}{k^*}, \dots, 1 - \theta_m \frac{b_m}{k^*}$ is always lower bounded by their geometric mean thanks to AM–GM. Therefore, we have

$$\prod_{i=1}^{m} \left(1 - \theta_i \frac{b_i}{k^*}\right) \leq \left(1 - \frac{1}{m} \sum_{i=1}^{m} \theta_i \frac{b_i}{k^*}\right)^m \leq \exp\left(-\sum_{i=1}^{m} \theta_i \frac{b_i}{k^*}\right).$$

By plugging this into the first inequality, we obtain the second inequality. □

B.1.1 Multi-Greedy

Thanks to Lemma A.2, we can prove the guarantees of `Multi-Greedy` as follows:

Theorem 1. *Let $b_{\max}$ be an integer satisfying $1 \leq b_{\max} \leq k^*$. Set $b_1, \dots, b_m$ so as to satisfy $b_i \in [b_{\max}]$ for $i \in [m]$ and $\sum_{i\in[m]} b_i = k$. If S_m is the solution obtained with* `Multi-Greedy` *and F is $(\gamma_{\mathsf{S}_m,k^*}, \beta_{\mathsf{S}_m,b_{\max}})$-WM, we have*

$$F(\mathsf{S}_m) \geq \left(1 - \prod_{i=1}^{m} \left(1 - \gamma_{\mathsf{S}_{i-1},k^*}\beta_{\mathsf{S}_{i-1},b_i}\frac{b_i}{k^*}\right)\right) F(\mathsf{S}^*) \geq \left(1 - \exp\left(-\gamma_{\mathsf{S}_m,k^*}\beta_{\mathsf{S}_m,b_{\max}}\frac{k}{k^*}\right)\right) F(\mathsf{S}^*).$$

Proof. To prove the theorem, it suffices that B_i chosen by `Multi-Greedy` in each iteration satisfies (A.2) with $\theta_i = \gamma_{\mathsf{S}_{i-1},k^*}\beta_{\mathsf{S}_{i-1},b_i}$; then we can obtain the theorem by using Lemma A.2 and $\gamma_{\mathsf{S}_{i-1},k^*}\beta_{\mathsf{S}_{i-1},b_i} \geq \gamma_{\mathsf{S}_m,k^*}\beta_{\mathsf{S}_m,b_{\max}}$ ($i \in [m]$). Note that `Multi-Greedy` uses $G_{\mathsf{S}_{i-1}}(\mathsf{B}) = \sum_{j\in\mathsf{B}} F(j \mid \mathsf{S}_{i-1})$ as a surrogate function in each i-th iteration. From $b_i \leq k^* = |\mathsf{S}^*|$ and the greedy rule, we have

$$\frac{1}{b_i}\sum_{j\in\mathsf{B}_i} F(j \mid \mathsf{S}_{i-1}) \geq \frac{1}{k^*}\sum_{j\in\mathsf{S}^*\setminus\mathsf{S}_{i-1}} F(j \mid \mathsf{S}_{i-1}). \tag{A.4}$$

Therefore, we obtain

$$\begin{aligned}
&F(\mathsf{B}_i \mid \mathsf{S}_{i-1}) \\
&\geq \beta_{\mathsf{S}_{i-1},b_i}\sum_{j\in\mathsf{B}_i} F(j \mid \mathsf{S}_{i-1}) && \because \text{definition of } \beta_{\mathsf{S}_{i-1},b_i} \\
&\geq \beta_{\mathsf{S}_{i-1},b_i}\frac{b_i}{k^*}\sum_{j\in\mathsf{S}^*\setminus\mathsf{S}_{i-1}} F(j \mid \mathsf{S}_{i-1}) && \because \text{(A.4)} \\
&\geq \gamma_{\mathsf{S}_{i-1},|\mathsf{S}^*\setminus\mathsf{S}_{i-1}|}\beta_{\mathsf{S}_{i-1},b_i}\frac{b_i}{k^*}F(\mathsf{S}^*\setminus\mathsf{S}_{i-1} \mid \mathsf{S}_{i-1}) && \because \text{definition of } \gamma_{\mathsf{S}_{i-1},|\mathsf{S}^*\setminus\mathsf{S}_{i-1}|} \\
&\geq \gamma_{\mathsf{S}_{i-1},k^*}\beta_{\mathsf{S}_{i-1},b_i}\frac{b_i}{k^*}F(\mathsf{S}^* \mid \mathsf{S}_{i-1}) && \because \gamma_{\mathsf{S}_{i-1},|\mathsf{S}^*\setminus\mathsf{S}_{i-1}|} \geq \gamma_{\mathsf{S}_{i-1},k^*} \text{ and } F(\mathsf{S}^*\setminus\mathsf{S}_{i-1} \mid \mathsf{S}_{i-1}) = F(\mathsf{S}^* \mid \mathsf{S}_{i-1}) \\
&\geq \gamma_{\mathsf{S}_{i-1},k^*}\beta_{\mathsf{S}_{i-1},b_i}\frac{b_i}{k^*}(F(\mathsf{S}^*) - F(\mathsf{S}_{i-1})). && \because \text{monotonicity}
\end{aligned}$$

Thus the proof is completed. □

B.1.2 Multi-OMP

We then prove the following guarantee of `Multi-OMP`.

Theorem 2. *Suppose that F is defined as $F(\mathsf{S}) = l(0) - \min_{\mathrm{supp}(\boldsymbol{x}')\subseteq\mathsf{S}} l(\boldsymbol{x}')$ ($\forall \mathsf{S} \subseteq [d]$) and $b_1, \dots, b_m$ are set as in Theorem 1. Assume that l is μ_{k+k^*}-RSC and $\nu_{k,b_{\max}}$-RSM. If S_m is a solution obtained with* `Multi-OMP`*, then we have*

$$\begin{aligned}
F(\mathsf{S}_m) &\geq \left(1 - \prod_{i=1}^{m}\left(1 - \frac{\mu_{\mathsf{S}_{i-1},\mathsf{S}_{i-1}\cup\mathsf{S}^*}}{\tilde{\nu}_{\mathsf{S}_{i-1},\mathsf{S}_i}}\frac{b_i}{k^*}\right)\right) F(\mathsf{S}^*) \\
&\geq \left(1 - \exp\left(-\frac{1}{k^*}\sum_{i=1}^{m}\frac{\mu_{\mathsf{S}_{i-1},\mathsf{S}_{i-1}\cup\mathsf{S}^*}}{\tilde{\nu}_{\mathsf{S}_{i-1},\mathsf{S}_i}}b_i\right)\right) F(\mathsf{S}^*) \\
&\geq \left(1 - \exp\left(-\frac{\mu_{\mathsf{S}_m,\mathsf{S}_m\cup\mathsf{S}^*}}{\max_{i\in[m]}\tilde{\nu}_{\mathsf{S}_{i-1},\mathsf{S}_i}}\frac{k}{k^*}\right)\right) F(\mathsf{S}^*) \\
&\geq \left(1 - \exp\left(-\frac{\mu_{k+k^*}}{\nu_{k,b_{\max}}}\frac{k}{k^*}\right)\right) F(\mathsf{S}^*).
\end{aligned}$$

Consequently, solution $\boldsymbol{x} = \mathrm{argmin}_{\mathrm{supp}(\boldsymbol{x}')\subseteq\mathsf{S}_m} l(\boldsymbol{x}')$ satisfies

$$l(\boldsymbol{x}) \leq l(\boldsymbol{x}^*) + \prod_{i=1}^{m}\left(1 - \frac{\mu_{\mathsf{S}_{i-1},\mathsf{S}_{i-1}\cup\mathsf{S}^*}}{\tilde{\nu}_{\mathsf{S}_{i-1},\mathsf{S}_i}}\frac{b_i}{k^*}\right)(l(0) - l(\boldsymbol{x}^*))$$

$$\leq l(\boldsymbol{x}^*) + \exp\left(-\frac{\mu_{k+k^*}}{\nu_{k,b_{\max}}}\frac{k}{k^*}\right)(l(0) - l(\boldsymbol{x}^*))$$
$$\leq l(\boldsymbol{x}^*) + \exp\left(-\frac{1}{\kappa_{k+k^*}}\frac{k}{k^*}\right)(l(0) - l(\boldsymbol{x}^*)).$$

Proof. As in Section A.1, we define $f(\boldsymbol{x}) \coloneqq l(0) - l(\boldsymbol{x})$ and $\mathbf{b}^{(\mathsf{S})} \coloneqq \operatorname{argmax}_{\operatorname{supp}(\boldsymbol{x})\subseteq\mathsf{S}} f(\boldsymbol{x})$ for any given $\mathsf{S} \subseteq [d]$. Analogous with the proof of Theorem 1, we prove that (A.2) holds, where $\theta_i = \frac{\mu_{\mathsf{S}_{i-1},\mathsf{S}_{i-1}\cup\mathsf{S}^*}}{\tilde{\nu}_{\mathsf{S}_{i-1},\mathsf{S}_i}}$. Note that `Multi-OMP` uses $G_{\mathsf{S}_{i-1}}(\mathsf{B}) = \sum_{j\in\mathsf{B}} |\nabla l(\mathbf{b}^{(\mathsf{S}_{i-1})})_j|^2 = \sum_{j\in\mathsf{B}} |\nabla f(\mathbf{b}^{(\mathsf{S}_{i-1})})_j|^2 = \|\nabla f(\mathbf{b}^{(\mathsf{S}_{i-1})})_\mathsf{B}\|_2^2$ as a surrogate function. Therefore, by using $b_i \leq k^* = |\mathsf{S}^*|$, $\nabla f(\mathbf{b}^{(\mathsf{S}_{i-1})})_{\mathsf{S}_{i-1}} = 0$, and the greedy rule, we obtain

$$\frac{1}{b_i}\|\nabla f(\mathbf{b}^{(\mathsf{S}_{i-1})})_{\mathsf{B}_i}\|_2^2 \geq \frac{1}{k^*}\|\nabla f(\mathbf{b}^{(\mathsf{S}_{i-1})})_{\mathsf{S}^*\setminus\mathsf{S}_{i-1}}\|_2^2. \tag{A.5}$$

By using this inequality and Lemma A.1, we obtain

$$\begin{aligned}
F(\mathsf{B}_i \mid \mathsf{S}_{i-1}) &\geq \frac{1}{2\tilde{\nu}_{\mathsf{S}_{i-1},\mathsf{S}_i}}\|\nabla f(\mathbf{b}^{(\mathsf{S}_{i-1})})_{\mathsf{B}_i}\|_2^2 && \because \text{Lemma A.1}\\
&\geq \frac{1}{2\tilde{\nu}_{\mathsf{S}_{i-1},\mathsf{S}_i}}\cdot\frac{b_i}{k^*}\|\nabla f(\mathbf{b}^{(\mathsf{S}_{i-1})})_{\mathsf{S}^*\setminus\mathsf{S}_{i-1}}\|_2^2 && \because \text{(A.5)}\\
&\geq \frac{\mu_{\mathsf{S}_{i-1},\mathsf{S}_{i-1}\cup\mathsf{S}^*}}{\tilde{\nu}_{\mathsf{S}_{i-1},\mathsf{S}_i}}\cdot\frac{b_i}{k^*}F(\mathsf{S}^*\setminus\mathsf{S}_{i-1} \mid \mathsf{S}_{i-1}) && \because \text{Lemma A.1}\\
&\geq \frac{\mu_{\mathsf{S}_{i-1},\mathsf{S}_{i-1}\cup\mathsf{S}^*}}{\tilde{\nu}_{\mathsf{S}_{i-1},\mathsf{S}_i}}\cdot\frac{b_i}{k^*}(F(\mathsf{S}^*) - F(\mathsf{S}_{i-1})). && \because \text{monotonicity}
\end{aligned}$$

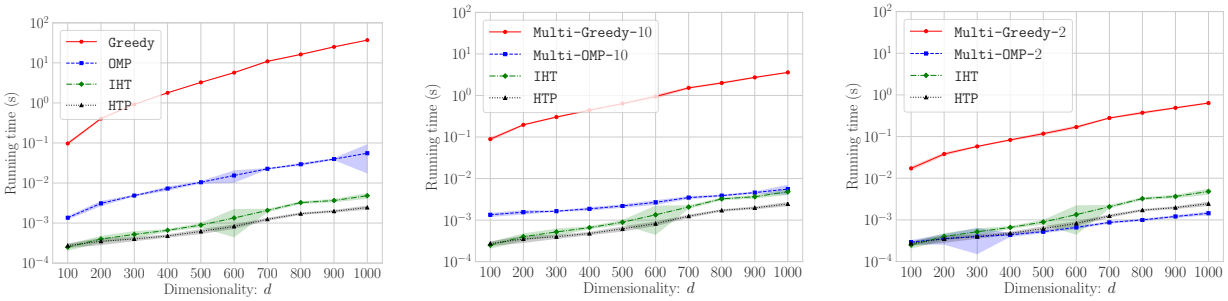
Thus the theorem holds thanks to Lemma A.2. □

B.2 Experiments with Synthetic ℓ_0-constrained Minimization Instances

We here evaluate the multi-stage algorithms with synthetic ℓ_0-constrained minimization instances.

Settings We consider well- and ill-conditioned sparse regression instances. Given design matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ and vector $\boldsymbol{y} \in \mathbb{R}^n$, we use the square loss function: $l(\boldsymbol{x}) \coloneqq \frac{1}{2n}\|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2$. We randomly generate well- and ill-conditioned instances as follows: We set the first k entries of the true sparse solution, $\boldsymbol{x}_{\text{true}}$, at 1 and the others at 0. In the well-conditioned case, we draw each entry of $\mathbf{A}$ from the standard normal distribution, denoted by $\mathcal{N}$. In the ill-conditioned case, we draw each row of $\mathbf{A}$ from a correlated d-dimensional normal distribution, whose mean and correlation coefficient are set at 0 and 0.3, respectively. Then, for both well- and ill-conditioned instances, we set $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}_{\text{true}} + 0.1\boldsymbol{u}$, where each entry of $\boldsymbol{u} \in \mathbb{R}^n$ is drawn from $\mathcal{N}$. We consider various dimensionalities: $d = 100, 200, \ldots, 1000$. We let $k = 0.1d$ and $n = \lfloor 10k\ln d \rfloor$. For each d value, we generate 100 random instances as above. We apply the multi-stage algorithms with $m = k, 10$, and 2 iterations to the instances, where $m = k$ corresponds to the standard `Greedy`/`OMP`. We use the projected-gradient-based methods (`IHT` and `IHT`) as baselines. We evaluate these methods in terms of running times and loss function values.

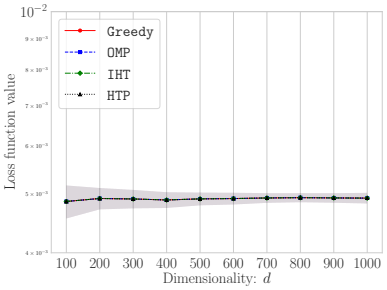
Results Figure 3 summarizes the results. We see that the multi-stage algorithms speed up as m decreases; they can become as fast as `IHT`/`HTP`. In the well-conditioned case, `Multi-OMP-2` is the fastest, and all the methods achieve the same loss function values, implying that the well-conditioned instances are so easy as to be solved almost optimally by all the methods. In the ill-conditioned case, as mentioned in Section 3.2.1, greedy-style methods achieve better loss function values than the projected-gradient-based methods. We see that the parameter, m, of multi-stage algorithms controls the trade-off between the running times and loss function values. To conclude, multi-stage algorithms with appropriate m values can outperform projected-gradient-based methods both in running time and solution quality, particularly when the instances are ill-conditioned.



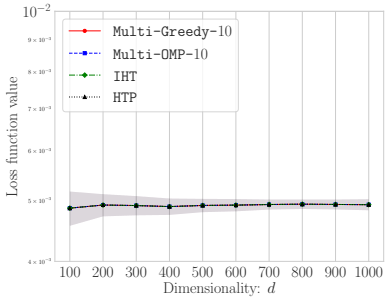
(a) $m = k$, Well-conditioned

(b) $m = 10$, Well-conditioned

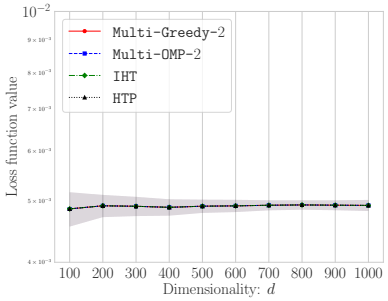
(c) $m = 2$, Well-conditioned



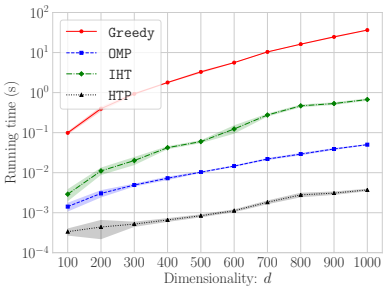
(d) $m = k$, Well-conditioned



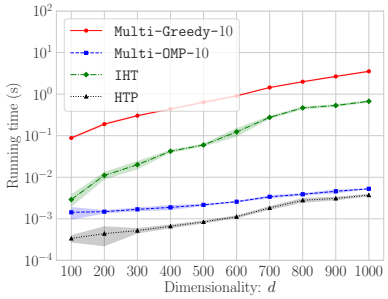
(e) $m = 10$, Well-conditioned



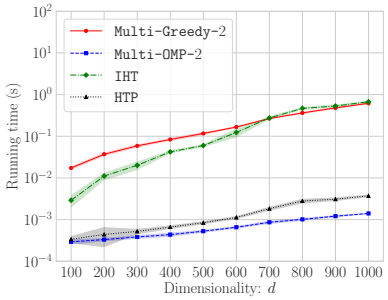
(f) $m = 2$, Well-conditioned



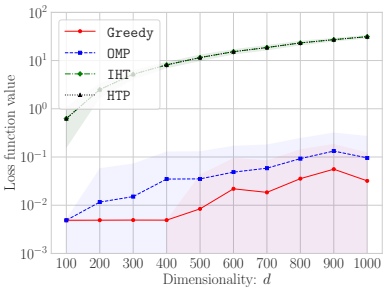
(g) $m = k$, Ill-conditioned



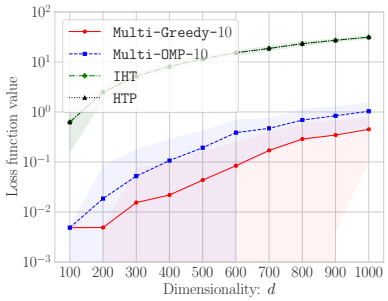
(h) $m = 10$, Ill-conditioned



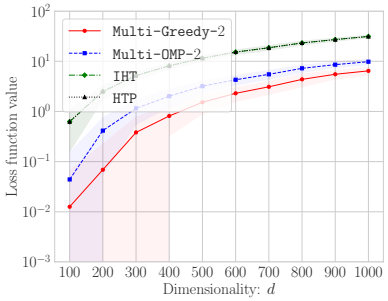
(i) $m = 2$, Ill-conditioned



(j) $m = k$, Ill-conditioned



(k) $m = 10$, Ill-conditioned



(l) $m = 2$, Ill-conditioned

Figure 3: Semi-log plots of running times and loss function values for well- and ill-conditioned instances. Figures (a)–(f) and (g)–(l) correspond to well- and ill-conditioned instances, respectively. The left, middle, and right figures show the results with $m = k, 10$, and 2, respectively. Each curve and error band indicate the mean and standard deviation, respectively, calculated over 100 random instances.

C FIXED-PARAMETER TRACTABILITY

In this section, we prove the guarantee of the randomized FPT algorithm for WMM. This result is an extension of (Skowron, 2017), which developed the randomized FPT algorithm for a subclass of monotone submodular maximization.

Algorithm 2 Randomized FPT algorithm

Execute `SingleRun()` T times and return the best solution.
function `SingleRun()`
    $\mathsf{S}_0 \leftarrow \emptyset$
    for $i = 1, \dots, k$ **do**
        Choose $j \in [d] \backslash \mathsf{S}_{i-1}$ randomly with probability proportional to $F(j \mid \mathsf{S}_{i-1})$.
        $\mathsf{S}_i \leftarrow \mathsf{S}_{i-1} \cup \{j\}$.
    return S_k

Let S^* be an optimal solution; i.e., $\mathsf{S}^* \in \operatorname{argmax}_{\mathsf{S}:|\mathsf{S}| \leq k} F(\mathsf{S})$. We first prove a key lemma, which provides a lower bound of the probability that $j \in \mathsf{S}^*$ is chosen in each iteration of `SingleRun()`.

Lemma A.3. *For $i \in [k]$, let S_{i-1} be the partial solution that is constructed in the loops of* `SingleRun()`*. Then the probability $p \in [0, 1]$ that newly chosen $j \in [d] \backslash \mathsf{S}_{i-1}$ is included in S^* is bounded from below as follows:*

$$p \geq \gamma_k \beta_{k,d} \cdot \frac{F(\mathsf{S}^* \mid \mathsf{S}_{i-1})}{F([d] \mid \mathsf{S}_{i-1})}.$$

Proof. The proof is obtained directly from the definitions of SBR and SPR as follows:

$$p = \frac{\sum_{j \in \mathsf{S}^* \backslash \mathsf{S}_{i-1}} F(j \mid \mathsf{S}_{i-1})}{\sum_{j \in [d] \backslash \mathsf{S}_{i-1}} F(j \mid \mathsf{S}_{i-1})} \geq \gamma_{\mathsf{S}_{i-1}, |\mathsf{S}^* \backslash \mathsf{S}_{i-1}|} \beta_{\mathsf{S}_{i-1}, |[d] \backslash \mathsf{S}_{i-1}|} \cdot \frac{F(\mathsf{S}^* \mid \mathsf{S}_{i-1})}{F([d] \mid \mathsf{S}_{i-1})} \geq \gamma_k \beta_{k,d} \cdot \frac{F(\mathsf{S}^* \mid \mathsf{S}_{i-1})}{F([d] \mid \mathsf{S}_{i-1})}.$$

□

Using this lemma we obtain the theorem as follows:

Theorem 3. *Assume F to be $(\gamma_k, \beta_{k,d})$-WM. Let S^* be an optimal solution for problem* (1) *and $\tilde{F} \coloneqq F([d]) - F(\mathsf{S}^*)$. For any $\epsilon > 0$, if $T \geq \left\lceil \left(\frac{1}{\gamma_k \beta_{k,d}} \cdot \frac{\tilde{F} + \epsilon}{\epsilon} \right)^k \ln \delta^{-1} \right\rceil$, then Algorithm 2 returns solution S satisfying $F(\mathsf{S}) \geq F(\mathsf{S}^*) - \epsilon$ with a probability of at least $1 - \delta$.*

Proof. We consider a single invocation of `SingleRun()`. In each i-th iteration ($i \in [k]$), one of the following two conditions occurs:

$$F(\mathsf{S}_{i-1}) \geq F(\mathsf{S}^*) - \epsilon, \tag{A.6}$$
$$F(\mathsf{S}_{i-1}) < F(\mathsf{S}^*) - \epsilon. \tag{A.7}$$

Once (A.6) occurs for some $i \in [k]$, then we have $F(\mathsf{S}_k) \geq F(\mathsf{S}_{i-1}) \geq F(\mathsf{S}^*) - \epsilon$ thanks to the monotonicity of F. If (A.7) occurs, we have

$$\frac{F(\mathsf{S}^* \mid \mathsf{S}_{i-1})}{F([d] \mid \mathsf{S}_{i-1})} \geq \frac{F(\mathsf{S}^*) - F(\mathsf{S}_{i-1})}{F([d]) - F(\mathsf{S}_{i-1})} = \frac{F(\mathsf{S}^*) - F(\mathsf{S}_{i-1})}{\tilde{F} + F(\mathsf{S}^*) - F(\mathsf{S}_{i-1})} > \frac{\epsilon}{\tilde{F} + \epsilon}.$$

Hence, newly chosen $j \in [d] \backslash \mathsf{S}_{i-1}$ is included in S^* with probability $p \geq \gamma_k \beta_{k,d} \cdot \frac{\epsilon}{\tilde{F} + \epsilon}$ thanks to Lemma A.3; if this occurs k times, we have $F(\mathsf{S}_k) = F(\mathsf{S}^*) \geq F(\mathsf{S}^*) - \epsilon$. Consequently, `SingleRun()` returns S_k that satisfies $F(\mathsf{S}_k) \geq F(\mathsf{S}^*) - \epsilon$ with a probability of at least $q \coloneqq \left(\gamma_k \beta_{k,d} \cdot \frac{\epsilon}{\tilde{F} + \epsilon} \right)^k$. Therefore, by setting $T \geq \left\lceil \frac{\ln \delta^{-1}}{q} \right\rceil = \left\lceil \left(\frac{1}{\gamma_k \beta_{k,d}} \cdot \frac{\tilde{F} + \epsilon}{\epsilon} \right)^k \ln \delta^{-1} \right\rceil$, Algorithm 2 finds a solution S such that $F(\mathsf{S}) \geq F(\mathsf{S}^*) - \epsilon$ with a probability of at least

$$1 - (1 - q)^T \geq 1 - (1 - q)^{-\frac{\ln \delta}{q}} \geq 1 - \mathrm{e}^{\ln \delta} = 1 - \delta.$$

Thus, the proof is completed. □

D HARDNESS OF IMPROVING APPROXIMATION RATIO

We first prove the hardness result (Theorem 4) in Appendix D.1. We then discuss the hardness for some easier subclasses of WMM in Appendix D.2 and a difficulty gap between two cases, $\beta_k = 1$ and $\beta_k < 1$, in Appendix D.3.

D.1 Proof of Theorem 4

As with the proof of (Nemhauser and Wolsey, 1978), we design objective function F appropriately and show that the problem of achieving an approximation guarantee that exceeds $1 - \mathrm{e}^{-\gamma_k}$ is at least as hard as another problem that cannot be solved in polynomial time: Roughly speaking, given an unknown subset M of size k, we consider seeking $\mathsf{S} \subseteq [d]$ such that $|\mathsf{S} \cap \mathsf{M}| \geq r+1$ and $|\mathsf{S}| \leq p_r^k \coloneqq 2k - r + 1$, where $r \leq k$ is any positive integer.

We explain how to design F. Fix the unknown subset $\mathsf{M} \subseteq [d]$ of size k. For any $\mathsf{S} \subseteq [d]$, we define the function value, $F(\mathsf{S})$, so that it depends only on $n_\mathsf{S} \coloneqq |\mathsf{S}|$, $m_\mathsf{S} \coloneqq |\mathsf{S} \cap \mathsf{M}|$, r, and k. We denote such a function by $G_r^k(m_\mathsf{S}, n_\mathsf{S})$, and we let $F(\mathsf{S}) \coloneqq G_r^k(|\mathsf{S} \cap \mathsf{M}|, |\mathsf{S}|) = G_r^k(m_\mathsf{S}, n_\mathsf{S})$. For any integers $m \in [0, k]$ and $n \in [0, d]$ such that $m \leq n$, we define the value of $G_r^k(m, n)$ so as to satisfy the following properties:

Property 1: $F(\cdot) = G_r^k(\cdot, \cdot)$ is monotone, and its SBR γ_k and SPR β_k satisfy $\gamma_k = 1$ and $\beta_k \geq \left(2 + \frac{r-1}{k-r+1}\right)^{-1} = \frac{1}{2} - \frac{1}{2} \cdot \frac{r-1}{2k-r+1}$, respectively.

Property 2: For any $m \in [0, r]$ and $n \in [0, d]$, the value of $G_r^k(m, n)$ is independent of m; i.e., $G_r^k(0, n) = G_r^k(1, n) = \cdots = G_r^k(r, n)$.

Property 3: $\max_{m,n:0 \leq m \leq n \leq k} G_r^k(m, n) = G_r^k(k, k) = k(k - r + 1)^{k-r}$.

Property 4: For any $n > p_r^k = 2k - r + 1$ and $m \in [0, k]$, we have $G_r^k(m, n) = k(k - r + 1)^{k-r}$.

Property 5: $\frac{G_r^k(0,k)}{G_r^k(k,k)} = \frac{G_r^k(1,k)}{G_r^k(k,k)} = \cdots = \frac{G_r^k(m,k)}{G_r^k(k,k)} = 1 - \left(\frac{k-r+1}{k}\right)\left(\frac{k-r}{k-r+1}\right)^{k-r+1} =: \alpha_k^{r-1}$.

As in (Nemhauser and Wolsey, 1978, Lemma 4.1), given monotone set function $F(\mathsf{S}) = G_r^k(m_\mathsf{S}, n_\mathsf{S})$ that satisfies Properties 2–5, to achieve an approximation guarantee that exceeds α_k^{r-1} is at least as hard as the following problem:

> For the unknown subset $\mathsf{M} \subseteq [d]$ of size k, find $\mathsf{S} \subseteq [d]$ that satisfies $|\mathsf{S} \cap \mathsf{M}| \geq r+1$ and $|\mathsf{S}| \leq p_r^k$ by using the following feedback: Once S is proposed, we are informed whether or not S satisfies $|\mathsf{S} \cap \mathsf{M}| \geq r+1$ and $|\mathsf{S}| \leq p_r^k$.

Intuitively, this can be proved as follows. From Properties 2, 3 and 5, if we are to achieve an approximation guarantee that exceeds α_k^{r-1}, we need at least to find S such that $m_\mathsf{S} \geq r+1$ and $n_\mathsf{S} \leq k$ ($\leq p_r^k$), while the information about G_r^k values is worthless as long as $m_\mathsf{S} \leq r$ and/or $n_\mathsf{S} > p_r^k$ due to Properties 2 and 4. These facts imply that to improve the α_k^{r-1}-approximation guarantee for the original maximization problem is at least as hard as to the above problem. Since M is unknown and no clue can be obtained by examining S if it violates $|\mathsf{S} \cap \mathsf{M}| \geq r+1$ and/or $|\mathsf{S}| \leq p_r^k$, the above problem cannot be solved via polynomially many queries. More precisely, the following proposition holds (see, the proof of (Nemhauser and Wolsey, 1978, Theorem 4.2)):

Proposition A.2. *Consider the maximization problem of form $\max_{\mathsf{S}:|\mathsf{S}| \leq k} F(\mathsf{S})$, where $F(\mathsf{S}) = G_r^k(m_\mathsf{S}, n_\mathsf{S})$ has monotonicity and Properties 2–5. For this problem, to achieve an approximation guarantee that exceeds α_k^{r-1} requires us to evaluate F at least $\Omega(d^{r+1}/k^{2r+2})$ times.*

By using the above properties and proposition, we obtain the main theorem.

Theorem 4. *Consider a class of problems of form $\max_{\mathsf{S}:|\mathsf{S}| \leq k} F(\mathsf{S})$ that satisfies the following conditions: F is monotone and has SBR $\gamma_k = 1$ and SPR $\beta_k \geq 1/2 - \Theta(1/k) \overset{k \to \infty}{\longrightarrow} 1/2$. For this class of problems, no algorithms that evaluate F only on polynomially many subsets can achieve an approximation guarantee that exceeds $1 - \mathrm{e}^{-1} = 1 - \mathrm{e}^{-\gamma_k}$ in general.*

Proof. The proof comprises two parts: (I) we prove the statement by assuming that there exists a function $F(\mathsf{S}) = G_r^k(m_\mathsf{S}, n_\mathsf{S})$ satisfying Properties 1–5, and (II) we show how to construct such a function.

Proof of the Statement Take k to be a monotone function of d that satisfies $\lim_{d\to\infty} k = \infty$ and $k = O(d^{\frac{1-c}{2}})$, where c is any constant such that $0 < c < 1$. Thanks to Property 1 and Proposition A.2, we have the following conditions:

- $\gamma_k = 1$ and $\beta_k \geq \frac{1}{2} - \frac{1}{2} \cdot \frac{r-1}{2k-r+1}$.
- To achieve an approximation guarantee that is better than α_k^{r-1} requires $\Omega(d^{c(r+1)})$ times function evaluation.

Since we can take r to be any fixed positive integer satisfying $r \leq k = O(d^{\frac{1-c}{2}})$, we see that $\Omega(d^{c(r+1)})$ is not polynomial in d. Furthermore, we have $\beta_k \overset{k\to\infty}{\longrightarrow} 1/2$ and $\alpha_k^{r-1} \overset{k\to\infty}{\longrightarrow} 1 - \mathrm{e}^{-1}$. Hence we obtain the statement by considering $d \to \infty$.

Construction of G_r^k Given any positive integer ℓ ($\leq k$), we define the following function $H^\ell(m, n)$ for integers $m \in [0, \ell]$ and $n \in [0, d]$ that satisfy $m \leq n$:

$$H^\ell(m,n) \coloneqq \begin{cases} \ell^\ell - \ell^{\ell-1}(\ell - m)\left(1 - \frac{1}{\ell}\right)^{n-m} & \text{if } n \leq k + \ell, \\ \ell^\ell & \text{otherwise.} \end{cases}$$

Note that the function is non-negative and that we have

$$H^\ell(0,0) = 0 \qquad \text{and} \qquad H^\ell(0,n) = H^\ell(1,n) = \ell^\ell \left(1 - \left(1 - \frac{1}{\ell}\right)^n\right).$$

Given any integers m_1, n_1, m_2, n_2 such that

$$0 \leq m_1 \leq n_1, \qquad 0 \leq m_2 \leq n_2, \qquad m_1 + m_2 \leq \ell, \qquad \text{and} \qquad n_1 + n_2 \leq d,$$

we define

$$\begin{aligned} H^\ell(m_2, n_2 \mid m_1, n_1) &\coloneqq H^\ell(m_1 + m_2, n_1 + n_2) - H^\ell(m_1, n_1) \\ &= \ell^{\ell-1}\left(1 - \frac{1}{\ell}\right)^{n_1 - m_1}\left(\ell - m_1 - (\ell - m_1 - m_2)\left(1 - \frac{1}{\ell}\right)^{n_2 - m_2}\right). \end{aligned}$$

When $(m_2, n_2) = (1, 1)$ and $(0, 1)$, for any m_1, n_1 satisfying the above conditions, we have

$$H^\ell(1,1 \mid m_1, n_1) = \ell^{\ell-1}\left(1 - \frac{1}{\ell}\right)^{n_1 - m_1} \quad \text{and} \quad H^\ell(0,1 \mid m_1, n_1) = \ell^{\ell-1}\left(1 - \frac{m_1}{\ell}\right)\left(1 - \frac{1}{\ell}\right)^{n_1 - m_1},$$

respectively. For later use, we prove the following lemma:

Lemma A.4. *For any integers m_1, n_1, m_2, n_2 that satisfy*

$$0 \leq m_1 \leq n_1 \leq \ell, \qquad 0 \leq m_2 \leq n_2 \leq k, \qquad m_1 + m_2 \leq \ell, \qquad \textit{and} \qquad n_1 + n_2 \leq d, \tag{A.8}$$

we have

$$\frac{H^\ell(m_2, n_2 \mid m_1, n_1)}{m_2 \times H^\ell(1,1 \mid m_1, n_1) + (n_2 - m_2) \times H^\ell(0,1 \mid m_1, n_1)} \geq \left(2 + \frac{k - \ell}{\ell}\right)^{-1},$$

where we regard $0/0 = 1$.

Proof. We rewrite the LHS of the target inequality as follows:

$$\begin{aligned} &\frac{H^\ell(m_2, n_2 \mid m_1, n_1)}{m_2 \times H^\ell(1,1 \mid m_1, n_1) + (n_2 - m_2) \times H^\ell(0,1 \mid m_1, n_1)} \\ &= \frac{\ell^{\ell-1}\left(1 - \frac{1}{\ell}\right)^{n_1 - m_1}\left(\ell - m_1 - (\ell - m_1 - m_2)\left(1 - \frac{1}{\ell}\right)^{n_2 - m_2}\right)}{m_2 \times \ell^{\ell-1}\left(1 - \frac{1}{\ell}\right)^{n_1 - m_1} + (n_2 - m_2) \times \ell^{\ell-1}\left(1 - \frac{1}{\ell}\right)^{n_1 - m_1}\left(1 - \frac{m_1}{\ell}\right)} \end{aligned}$$

$$= \frac{\ell - m_1 - (\ell - m_1 - m_2)\left(1 - \frac{1}{\ell}\right)^{n_2 - m_2}}{m_2 + (n_2 - m_2)\left(1 - \frac{m_1}{\ell}\right)}$$
$$= \frac{1 - \frac{m_1}{\ell} - \left(1 - \frac{m_1}{\ell} - \frac{m_2}{\ell}\right)\left(1 - \frac{1}{\ell}\right)^{\ell\left(\frac{n_2}{\ell} - \frac{m_2}{\ell}\right)}}{\frac{m_2}{\ell}\frac{m_1}{\ell} + \frac{n_2}{\ell}\left(1 - \frac{m_1}{\ell}\right)}.$$

By defining $x \coloneqq \frac{m_2}{\ell}$, $y \coloneqq \frac{n_2}{\ell}$, and $z \coloneqq 1 - \frac{m_1}{\ell}$, we obtain

$$\frac{H^\ell(m_2, n_2 \mid m_1, n_1)}{m_2 \times H^\ell(1,1 \mid m_1, n_1) + (n_2 - m_2) \times H^\ell(0,1 \mid m_1, n_1)} = \frac{z - (z - x)\left(1 - \frac{1}{\ell}\right)^{\ell(y-x)}}{x(1-z) + yz}, \tag{A.9}$$

where x, y, and z must satisfy the following inequalities from (A.8):

$$0 \le z \le 1, \qquad 0 \le x \le y \le \frac{k}{\ell} = 1 + \frac{k - \ell}{\ell}, \qquad x \le z, \qquad \text{and} \qquad y - z \le \frac{d}{\ell} - 1.$$

The RHS of (A.9) can be bounded from below by $\left(2 + \frac{k-\ell}{\ell}\right)^{-1}$ as follows:

$$\begin{aligned}
\frac{z - (z - x)\left(1 - \frac{1}{k}\right)^{k(y-x)}}{x(1-z) + yz} &\ge \frac{z - (z - x)\mathrm{e}^{-(y-x)}}{x(1-z) + yz} && \because (1 - 1/k)^k \le \mathrm{e}^{-1} \\
&\ge \frac{z - (z - x)\frac{1}{1+y-x}}{x(1-z) + yz} && \because \mathrm{e}^{-a} \le \frac{1}{1+a} \text{ for } a > -1 \\
&= \frac{1}{1 + y - x} \\
&\ge \left(2 + \frac{k - \ell}{\ell}\right)^{-1}. && \because x \ge 0 \text{ and } y \le 1 + \frac{k-\ell}{\ell}
\end{aligned}$$

Thus, the lemma holds. □

By using the above $H^\ell(m, n)$ with $\ell = k - r + 1$, we construct $G_r^k(m, n)$ for any integers $m \in [0, k]$ and $n \in [0, d]$ as follows:

$$G_r^k(m,n) \coloneqq \begin{cases} n \times H^{k-r+1}(0,1) & \text{if } 0 \le m \le n \le r, \\ (r-1) \times H^{k-r+1}(0,1) + H^{k-r+1}(0, n-r+1) & \text{if } 0 \le m \le r \text{ and } r \le n \le d, \\ (r-1) \times H^{k-r+1}(0,1) + H^{k-r+1}(m-r+1, n-r+1) & \text{if } r \le m \le k \text{ and } r \le n \le d. \end{cases}$$

By using G_r^k, we define $F(\mathsf{S}) = G_r^k(m_\mathsf{S}, n_\mathsf{S})$. We can confirm that G_r^k has Properties 2–5 as in the proof of (Nemhauser and Wolsey, 1978). Below we show that the function has Property 1. We let d satisfy $d \ge 2k$. The monotonicity can be confirmed easily by examining $F(j \mid \mathsf{S})$ for each case. Furthermore, by analogy with the proof in (Nemhauser and Wolsey, 1978), we can show that $F(\mathsf{S}) = G_r^k(m_\mathsf{S}, n_\mathsf{S})$ is submodular over all subsets of size at most $2k$: I.e., $F(j \mid \mathsf{S}) \ge F(j \mid \mathsf{T})$ for any $\mathsf{S} \subseteq \mathsf{T}$ satisfying $|\mathsf{T}| < 2k$ and $j \notin \mathsf{T}$. This suffices to prove that $\gamma_k = 1$ holds.

Below we prove $\beta_k \ge \left(2 + \frac{r-1}{k-r+1}\right)^{-1}$. Note that SPR can be written as

$$\beta_k = \min_{\substack{\mathsf{L}, \mathsf{S} \,:\, \mathsf{L} \cap \mathsf{S} = \emptyset, \\ |\mathsf{L}| \le k, |\mathsf{S}| \le k}} \frac{F(\mathsf{S} \mid \mathsf{L})}{\sum_{j \in \mathsf{S}} F(j \mid \mathsf{L})},$$

where we regard $0/0 = 1$. In what follows, for any disjoint $\mathsf{L}, \mathsf{S} \subseteq [d]$ of size at most k, we consider bounding $\frac{F(\mathsf{S}|\mathsf{L})}{\sum_{j \in \mathsf{S}} F(j|\mathsf{L})}$ from below. Depending on the values of $m_\mathsf{L} = |\mathsf{L} \cap \mathsf{M}|$, $m_\mathsf{S} = |\mathsf{S} \cap \mathsf{M}|$, $n_\mathsf{L} = |\mathsf{L}|$, and $n_\mathsf{S} = |\mathsf{S}|$, we have the following six cases. We first derive lower bounds of $\frac{F(\mathsf{S}|\mathsf{L})}{\sum_{j \in \mathsf{S}} F(j|\mathsf{L})}$ for all cases separately, and we then show that $\frac{F(\mathsf{S}|\mathsf{L})}{\sum_{j \in \mathsf{S}} F(j|\mathsf{L})} \ge \left(2 + \frac{r-1}{k-r+1}\right)^{-1}$ holds for all the cases.

Case 1: $n_{\mathsf{S}} + n_{\mathsf{L}} < r$**.** In this case, we have $F(\mathsf{S} \mid \mathsf{L}) = n_{\mathsf{S}} \times H^{k-r+1}(0,1)$; i.e., the function is modular. Therefore, we have $\frac{F(\mathsf{S}|\mathsf{L})}{\sum_{j\in\mathsf{S}} F(j|\mathsf{L})} = 1$.

Case 2: $n_{\mathsf{L}} < r$ **and** $m_{\mathsf{L}} + m_{\mathsf{S}} \leq r \leq n_{\mathsf{S}} + n_{\mathsf{L}}$**.** In this case, we have

$$\begin{aligned} F(\mathsf{S} \mid \mathsf{L}) &= H^{k-r+1}(0, n_{\mathsf{S}} + n_{\mathsf{L}} - r + 1) - (n_{\mathsf{L}} - r + 1)H^{k-r+1}(0,1), \\ F(j \mid \mathsf{L}) &= H^{k-r+1}(0,1). \end{aligned}$$

Note that we have $|\mathsf{L}| \leq k$ and $|\mathsf{S}| \leq k$. Due to the submodularity over all subsets of size at most $2k$, the more elements L includes, the smaller $F(\mathsf{S} \mid \mathsf{L})$ becomes, which means $F(\mathsf{S} \mid \mathsf{L})$ attains its minimum when $n_{\mathsf{L}} = r - 1$. Therefore, we have

$$\frac{F(\mathsf{S} \mid \mathsf{L})}{\sum_{j\in\mathsf{S}} F(j \mid \mathsf{L})} \geq \frac{H^{k-r+1}(0, n_{\mathsf{S}})}{n_{\mathsf{S}} \times H^{k-r+1}(0,1)} = \frac{H^{k-r+1}(0, n_{\mathsf{S}} \mid 0, 0)}{n_{\mathsf{S}} \times H^{k-r+1}(0, 1 \mid 0, 0)}.$$

Case 3: $n_{\mathsf{L}} < r$ **and** $r \leq m_{\mathsf{S}} + m_{\mathsf{L}}$**.** We have

$$\begin{aligned} F(\mathsf{S} \mid \mathsf{L}) &= H^{k-r+1}(m_{\mathsf{L}} + m_{\mathsf{S}} - r + 1, n_{\mathsf{S}} + n_{\mathsf{L}} - r + 1) - (n_{\mathsf{L}} - r + 1)H^{k-r+1}(0,1), \\ F(j \mid \mathsf{L}) &= H^{k-r+1}(0,1). \end{aligned}$$

By analogy with the above case, $F(\mathsf{S} \mid \mathsf{L})$ attains its minimum when $n_{\mathsf{L}} = r - 1$. Therefore,

$$\begin{aligned} \frac{F(\mathsf{S} \mid \mathsf{L})}{\sum_{j\in\mathsf{S}} F(j \mid \mathsf{L})} &\geq \frac{H^{k-r+1}(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}})}{n_{\mathsf{S}} \times H^{k-r+1}(0,1)} = \frac{H^{k-r+1}(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}} \mid 0, 0)}{n_{\mathsf{S}} \times H^{k-r+1}(0, 1 \mid 0, 0)} \\ &= \frac{H^{k-r+1}(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}} \mid 0, 0)}{(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1)H^{k-r+1}(1, 1 \mid 0, 0) + (n_{\mathsf{S}} - m_{\mathsf{S}} - m_{\mathsf{L}} + r - 1)H^{k-r+1}(0, 1 \mid 0, 0)}, \end{aligned}$$

where the last equality comes from $H^{k-r+1}(0, n) = H^{k-r+1}(1, n)$. Note that we have $m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1 \leq \ell = k - r + 1$ since $m_{\mathsf{S}} + m_{\mathsf{L}} = |\mathsf{S} \cap \mathsf{M}| + |\mathsf{L} \cap \mathsf{M}| \leq |\mathsf{M}| = k$, where the inequality comes from the fact that L and S are disjoint. Furthermore, we have $n_{\mathsf{S}} - m_{\mathsf{S}} - m_{\mathsf{L}} + r - 1 \geq 0$ since $n_{\mathsf{S}} \geq m_{\mathsf{S}}$ and $m_{\mathsf{L}} \leq n_{\mathsf{L}} \leq r - 1$.

Case 4: $m_{\mathsf{L}} < r \leq n_{\mathsf{L}}$ **and** $m_{\mathsf{S}} + m_{\mathsf{L}} \leq r$**.** We have

$$\begin{aligned} F(\mathsf{S} \mid \mathsf{L}) &= H^{k-r+1}(0, n_{\mathsf{S}} + n_{\mathsf{L}} - r + 1) - H^{k-r+1}(0, n_{\mathsf{L}} - r + 1) \\ &= H^{k-r+1}(0, n_{\mathsf{S}} \mid 0, n_{\mathsf{L}} - r + 1), \\ F(j \mid \mathsf{L}) &= H^{k-r+1}(0, 1 \mid 0, n_{\mathsf{L}} - r + 1), \end{aligned}$$

and thus we obtain

$$\frac{F(\mathsf{S} \mid \mathsf{L})}{\sum_{j\in\mathsf{S}} F(j \mid \mathsf{L})} = \frac{H^{k-r+1}(0, n_{\mathsf{S}} \mid 0, n_{\mathsf{L}} - r + 1)}{n_{\mathsf{S}} \times H^{k-r+1}(0, 1 \mid 0, n_{\mathsf{L}} - r + 1)}.$$

Case 5: $m_{\mathsf{L}} < r \leq n_{\mathsf{L}}$ **and** $m_{\mathsf{S}} + m_{\mathsf{L}} \geq r$**.** We have

$$\begin{aligned} F(\mathsf{S} \mid \mathsf{L}) &= H^{k-r+1}(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}} + n_{\mathsf{L}} - r + 1) - H^{k-r+1}(0, n_{\mathsf{L}} - r + 1) \\ &= H^{k-r+1}(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}} \mid 0, n_{\mathsf{L}} - r + 1), \\ F(j \mid \mathsf{L}) &= H^{k-r+1}(0, 1 \mid 0, n_{\mathsf{L}} - r + 1), \end{aligned}$$

and thus we obtain

$$\begin{aligned} &\frac{F(\mathsf{S} \mid \mathsf{L})}{\sum_{j\in\mathsf{S}} F(j \mid \mathsf{L})} \\ &= \frac{H^{k-r+1}(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}} \mid 0, n_{\mathsf{L}} - r + 1)}{n_{\mathsf{S}} \times H^{k-r+1}(0, 1 \mid 0, n_{\mathsf{L}} - r + 1)}. \\ &= \frac{H^{k-r+1}(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}} \mid 0, n_{\mathsf{L}} - r + 1)}{(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1)H^{k-r+1}(1, 1 \mid 0, n_{\mathsf{L}} - r + 1) + (n_{\mathsf{S}} - m_{\mathsf{S}} - m_{\mathsf{L}} + r - 1)H^{k-r+1}(0, 1 \mid 0, n_{\mathsf{L}} - r + 1)}, \end{aligned}$$

where we used $H^{k-r+1}(0, 1 \mid 0, n) = H^{k-r+1}(0, n+1) - H^{k-r+1}(0, n) = H^{k-r+1}(1, n+1) - H^{k-r+1}(0, n) = H^{k-r+1}(1, 1 \mid 0, n)$. Note that we can obtain $m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1 \leq \ell = k - r + 1$ and $n_{\mathsf{S}} - m_{\mathsf{S}} - m_{\mathsf{L}} + r - 1 \geq 0$ by analogy with Case 3.

Case 6: $m_{\mathsf{L}} \geq r$**.** We have

$$\begin{aligned} F(\mathsf{S} \mid \mathsf{L}) &= H^{k-r+1}(m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}} + n_{\mathsf{L}} - r + 1) - H^{k-r+1}(m_{\mathsf{L}} - r + 1, n_{\mathsf{L}} - r + 1) \\ &= H^{k-r+1}(m_{\mathsf{S}}, n_{\mathsf{S}} \mid m_{\mathsf{L}} - r + 1, n_{\mathsf{L}} - r + 1), \\ F(j \mid \mathsf{L}) &= \begin{cases} H^{k-r+1}(0, 1 \mid m_{\mathsf{L}} - r + 1, n_{\mathsf{L}} - r + 1) & \text{if } j \notin \mathsf{M}, \\ H^{k-r+1}(1, 1 \mid m_{\mathsf{L}} - r + 1, n_{\mathsf{L}} - r + 1) & \text{if } j \in \mathsf{M}. \end{cases} \end{aligned}$$

In this case, we obtain

$$\begin{aligned} &\frac{F(\mathsf{S} \mid \mathsf{L})}{\sum_{j \in \mathsf{S}} F(j \mid \mathsf{L})} \\ &= \frac{H^{k-r+1}(m_{\mathsf{S}}, n_{\mathsf{S}} \mid m_{\mathsf{L}} - r + 1, n_{\mathsf{L}} - r + 1)}{m_{\mathsf{S}} \times H^{k-r+1}(1, 1 \mid m_{\mathsf{L}} - r + 1, n_{\mathsf{L}} - r + 1) + (n_{\mathsf{S}} - m_{\mathsf{S}}) \times H^{k-r+1}(0, 1 \mid m_{\mathsf{L}} - r + 1, n_{\mathsf{L}} - r + 1)}. \end{aligned}$$

In Case 1, the value of $\frac{F(\mathsf{S}|\mathsf{L})}{\sum_{j \in \mathsf{S}} F(j|\mathsf{L})}$ is lower bounded by 1. In the other cases, the value of $\frac{F(\mathsf{S}|\mathsf{L})}{\sum_{j \in \mathsf{S}} F(j|\mathsf{L})}$ is lower bounded by $\left(2 + \frac{r-1}{k-r+1}\right)^{-1}$ thanks to Lemma A.4 with $\ell = k - r + 1$, where we let $(m_1, n_1, m_2, n_2) = (0, 0, 0, n_{\mathsf{S}}), (0, 0, m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}}), (0, n_{\mathsf{L}} - r + 1, 0, n_{\mathsf{S}}), (0, n_{\mathsf{L}} - r + 1, m_{\mathsf{S}} + m_{\mathsf{L}} - r + 1, n_{\mathsf{S}})$, and $(m_{\mathsf{L}} - r + 1, n_{\mathsf{L}} - r + 1, m_{\mathsf{S}}, n_{\mathsf{S}})$ in Cases 2, 3, 4, 5, and 6, respectively. Note that the conditions required in Lemma A.4 are satisfied in all cases. □

D.2 Discussion on Easier Subclasses of WMM

While Theorem 4 means that the existing $(1 - \mathrm{e}^{-\gamma_{\mathsf{S},k}})$-approximation guarantee (Das and Kempe, 2018) achieved by `Greedy` cannot be improved in polynomial time in general, there may exist easier subclasses of WMM that are not considered in Theorem 4, for which better guarantees may be possible. One such example is WMM such that SBR and SPR defined on the whole domain (i.e., γ_d and β_d) are lower bounded by constants. Therefore, as regards algorithms whose guarantees are proved by using the (weak) submodularity on the whole domain, their guarantees may be improved by using bounded β_d. One such algorithm is the *continuous greedy* algorithm (Calinescu et al., 2011), and so a better guarantee for WMM with bounded γ_d and β_d may be possible by using continuous-greedy-based methods. However, whether this approach works or not is non-trivial since we currently lack a guaranteed *rounding* scheme for WM functions. On the other hand, as in (Bian et al., 2017; Das and Kempe, 2018), the proofs of `Greedy` rely only on the weak submodularity defined on the restricted domain (or bounded γ_k). With regard to such algorithms, Theorem 4 suggests that, even if γ_d and β_d are bounded, it is hardly possible to obtain approximation ratios better than $1 - \mathrm{e}^{-\gamma_k}$ only via slight modification of the existing proofs. To conclude, there are the following two possibilities. It is hard to improve $1 - \mathrm{e}^{-\gamma_k}$ even for easier subclasses of WMM (e.g., γ_d and β_d are bounded), or there exist new techniques (e.g., a continuous-greedy-based one) and it is possible to obtain approximation guarantees that can exceed $1 - \mathrm{e}^{-\gamma_k}$ for some easier subclasses.

D.3 Discussion on the Gap between $\beta_k = 1$ and $\beta_k < 1$

If we have $\gamma_k = \beta_k = 1$, which guarantees that F is modular over all subsets of size at most k, then the greedy algorithm finds an optimal solution; therefore, one may expect that, if $\gamma_k = 1$, we can obtain a β_k-dependent approximation ratio that becomes close to 1 as β_k increases. Our hardness result disproves it, and so the result may seem to be counter-intuitive. We here give an example, which intuitively suggests that the existence of the difficulty gap between $\beta_k = 1$ and $\beta_k < 1$ ($\beta_k \approx 1/2$ in our hardness result) is not unusual.

We consider the set-cover instance that is often used to show the tightness of the greedy $\Theta(\ln d)$-approximation. Assume that we have $\sum_{i=1}^{m} 2^i$ elements, and let $d = m + 2$. Given m disjoint subsets $I_1, I_2, \ldots, I_m$ consisting of $2, 4, \ldots, 2^m$ elements, respectively, and additional two disjoint subsets I_{m+1} and I_{m+2}, each of which includes half of the elements from each I_i ($i \in [m]$), the approximation ratio of the greedy algorithm is asymptotically $\log_2 d/2$. Note that each element is covered twice, and so the coverage function satisfies $\beta_k \geq 1/2$ as discussed in Appendix A.3. On the other hand, if each subset I_j ($j \in [d]$) covers only one element, i.e., the coverage function is modular, the greedy algorithm finds an optimal solution. Therefore, although the hardness related to all polynomial-time algorithms are not considered unlike our hardness results, the above observation suggests that there can be a difficulty gap between the modular case ($\beta_k = 1$) and the non-modular case ($\beta_k < 1$) regarding set-cover problems.