Approximate Frank-Wolfe Algorithms over Graph-structured Support Sets

Baojian Zhou [1] **Yifan Sun** [2]

Abstract

In this paper, we consider approximate Frank-Wolfe (FW) algorithms to solve convex optimization problems over graph-structured support sets where the *linear minimization oracle* (LMO) cannot be efficiently obtained in general. We first demonstrate that two popular approximation assumptions (*additive and multiplicative gap errors)* are not applicable in that no cheap gap-approximate LMO oracle exists. Thus, *approximate dual maximization oracles* (DMO) are proposed, which approximate the inner product rather than the gap. We prove that the standard FW method using a δ-approximate DMO converges as $\mathcal{O}((1-\delta)\sqrt{s}/\delta)$ in the worst case, and as $\mathcal{O}(L/(\delta^2 t))$ over a δ-relaxation of the constraint set. Furthermore, when the solution is on the boundary, a variant of FW converges as $\mathcal{O}(1/t^2)$ under the quadratic growth assumption. Our empirical results suggest that even these improved bounds are pessimistic, showing fast convergence in recovering real-world images with graph-structured sparsity.

1. Introduction

This paper deals with the following graph-structured convex optimization (GSCO) problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x}), \text{ subject to } \boldsymbol{x} \in \mathcal{D}(C, \mathbb{M}), \qquad (1)$$

where $\mathcal{D}(C, \mathbb{M}) \triangleq \mathrm{conv}\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq C, \mathrm{supp}(\boldsymbol{x}) \in \mathbb{M}\}$ is a convex hull of the graph-structured support set described by $\mathbb{M}$, which contains a collection of allowed structures of the problem, and f is a convex differentiable function. The support of $\boldsymbol{x}$, i.e., $\mathrm{supp}(\boldsymbol{x}) \triangleq \{i : x_i \neq 0\}$ encodes the

---

[1]School of Data Science, Fudan University, Shanghai, China [2]Department of Computer Science, Stony Brook University, Stony Brook, New York, USA. Correspondence to: Baojian Zhou <bjzhou@fudan.edu.cn>.

sparsity pattern of $\boldsymbol{x}$, which can be defined by interesting graph structures such as a path, tree, or cluster over an underlying graph. Models $\mathbb{M}$ describe many interesting scenarios where graph structures serve as a powerful prior. Important applications of these include generalized s-sparsity (Argyriou et al., 2012; Lim & Wright, 2017), structured-sparsity (Bach et al., 2012a;b), clustered-sparsity (McDonald et al., 2016b), weighted graph models (WGM) (Hegde et al., 2015b), graph LASSO (Sharpnack et al., 2012; Hallac et al., 2015), marginal polytope (Krishnan et al., 2015), and many others (Baraniuk et al., 2010).

To solve the GSCO problem, a natural idea is to use projected gradient descent (PGD) where, a *projection oracle* finds a point in $\mathcal{D}$ at per-iteration. PGD-based methods for sparse and structure optimization have been well explored (Bahmani et al., 2013; Jain et al., 2014; Yuan et al., 2014; Nguyen et al., 2017; Hegde et al., 2015a;b; 2016). To obtain approximate convergence guarantees, existing works of this type assume projection oracles can be solved exactly or with very high approximations guarantees. However, projections satisfying these requirements are usually hard to find for problem (1). Furthermore, multiple projections may be needed at per-iteration (Hegde et al., 2015b).

Instead, the Frank-Wolfe (FW) algorithm (Frank et al., 1956) (a.k.a conditional gradient method) has been receiving increasing attention in recent years. Unlike PGD-based methods, FW-type methods, at each iteration, find a point using the *linear minimization oracle* (LMO), which for many constraints may enjoy a much cheaper per-iteration cost than the projection oracle (Combettes & Pokutta, 2021), and often obtain high-quality sparser solutions in early iterations. Hence, they are attractive for solving structured problems (Krishnan et al., 2015; Briol et al., 2015; Ping et al., 2016; Berthet & Perchet, 2017; Allen-Zhu et al., 2017; Abernethy & Wang, 2017). Yet, these attractive methods are less explored for graph-structured optimization problems.

To resolve problem (1) using FW-type methods, the main difficulty is that solving the LMO efficiently, even with $\mathcal{D}$ convex, is in general NP-hard for many structured models $\mathbb{M}$. A typical motivational example is a popular weighted graph model (WGM), where $\mathbb{M}$ contains all sets of g connected components of a specified weighted graph (Hegde et al., 2015b). Here, both the projection oracle and LMO are NP-

hard to compute. While convergence rates exist for FW with approximate LMOs, they tend to be limited to two kinds of approximations: *additive gap-approximate LMO* (Dunn & Harshbarger, 1978; Jaggi, 2013) and *multiplicative gap-approximate LMO* (Locatello et al., 2017; Pedregosa et al., 2020). As such, we ask the following crucial question

Do an additive and multiplication gap-approximate LMO exist for solving GSCO problems?

Our contributions. In particular, by answering this question negatively, we open and explore the space for inexact FW methods that are more appropriate for this class of problems. As demonstrated, for a WGM $\mathbb{M}$ of $\mathcal{D}$, one can always find adversarial examples to show that gap-additive and gap-multiplicative LMOs are as hard to resolve as exact LMOs. Therefore, the existing approximate-LMO FW convergence rates are inapplicable to GSCOs in general.[1]

Instead, we propose to use an approximate *dual maximization oracle* (DMO), which for several important GSCO problems can be easy to find in practice. This assumption is equivalent to multiplicatively approximating a key inner product, rather than the gap. We show that, for GSCOs, a simple heuristic method acts as an approximate DMO with a constant δ error guarantee, but no non-exact gap-additive or gap-multiplicative LMO exists.

The main theoretical contribution is then to give convergence rates of the approximate FW methods under a δ-approximate DMO assumption. We show that when f is L-smooth, a standard FW using a δ-approximate DMO converges as $\mathcal{O}((1-\delta)\sqrt{s}/\delta)$ where s is the maximum sparsity allowed in $\mathbb{M}$, and as $\mathcal{O}(L/(\delta^2 t))$ over $\mathcal{D}/\delta$. The convergence rate of the latter case is consistent with recent advances of generalized matching pursuit (MP) (Locatello et al., 2018).

Inspired by works concerning the *nearest extreme point oracle* (Garber & Wolf, 2021), we propose a new variant of FW, which is empirically faster. We also show that the convergence rate of this new variant is faster at $\mathcal{O}(1/t^2)$ when the solutions are on the boundary with $\delta = 1$. Empirically, we observe that these assumptions are not necessary to achieve this faster rate. Additionally, we show that an approximate version converges to a point in $\mathcal{D}/\delta$ at $\mathcal{O}(1/t)$.

The main contributions can be summarized as follows:

- We prove that the exact DMO for general GSCOs is NP-hard and no efficient additive and multiplicative gap-approximate LMO exists.

- We propose two FW-type methods and provide conver-

gence rate analysis when the problem is solved over $\mathcal{D}$ or a relaxed set $\mathcal{D}/\delta$.

- We empirically demonstrate that the proposed methods are more effective and efficient compared with PGD-based and MP-based methods on the graph-structured linear regression problem.

1.1. Related work

Existing works on sparse optimization considers models that limit the number of nonzeros in the solution. However, many real-world applications have more intricate graph structures as important priors; the tradeoff is that strictly enforcing them often loses convexity of the feasibility space.

FW method and its variants. The FW method (Frank et al., 1956) and its variants (Jaggi, 2013; Lacoste-Julien & Jaggi, 2015; Garber & Meshi, 2016; Bashiri & Zhang, 2017; Balasubramanian & Ghadimi, 2018; Kerdreux et al., 2018; Lei et al., 2019; Luise et al., 2019; Locatello et al., 2019; Thekumparampil et al., 2020; Garber, 2020; Combettes & Pokutta, 2020; Pedregosa et al., 2020; Sun & Bach, 2020) for convex constrained problems have recently received popularity mainly due to two advantages. First, it is projection free–the LMO is often much cheaper to compute than the projection oracle. Second, in applications with desired structured sparsity, early FW iterations tend to be naturally sparse. Inspired from these advantages, we seek to propose FW-type methods for GSCO problems.

Recent works put effort into accelerating FW with modifications. More specifically, Lacoste-Julien & Jaggi (2015) and Garber & Meshi (2016) propose away-step variants to reduce the computation overheads. An important related work (Garber & Hazan, 2015) shows that if $\mathcal{D}$ is a strongly convex set, then the FW rate can be improved to $\mathcal{O}(1/t^2)$. However, sparsity-inducing sets are generally not strongly convex, favoring low-dimensional facets of $\mathcal{D}$ as solutions; this is also true for our graph-sparsity application. We therefore achieve the $O(1/t^2)$ acceleration without necessitating strong convexity, drawing inspiration from the *nearest extreme point oracle* explored in Garber & Wolf (2021).

Connections with other methods. One of the reasons that FW is so heavily studied is its connections with other important greedy methods. For example, Bach (2015) shows that FW is closely related with mirror decent through duality; and Locatello et al. (2017) and Combettes & Pokutta (2019) explore the close connection between FW and MP.

Approximation of LMO. The study of inexact FW methods tend to center on two types of LMO errors: gap-additive (Dunn & Harshbarger, 1978; Jaggi, 2013) and gap-multiplicative (Locatello et al., 2017; Pedregosa et al., 2020).

---

[1]Although this paper focuses on a specific model, our proposed methods are applicable for any $\mathbb{M}$ whenever its DMO is available.

Under these two assumptions, the convergence rates are $\mathcal{O}(\delta/t)$ and $\mathcal{O}(1/(\delta^2 t))$ respectively. However, as we will demonstrate in Sec. 3, these two regimes do not adequately describe efficient methods for GSCO problems. Instead, we explore that is multiplicative with respect to an inner product; this is inspired by (Locatello et al., 2018) making a similar analysis for MP, and is related to the works of Hazan et al. (2018) and Garber (2017). Another related work is Kerdreux et al. (2018), which analyzes FW methods where the LMO is approximated by subsampling a subset of atoms at each iteration.

2. Preliminaries

We begin by introducing notations and setup. We then define the graph-structured support sets and the FW-type algorithm. *All proofs are postponed to the appendix.*

Notations and setup. We consider optimization over variables $\boldsymbol{x} \in \mathbb{R}^d$, which induce a ground set $[d] := \{1, 2, \ldots, d\}$. Uppercase letters (e.g., I, S) stand for subsets of $[d]$. $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ denotes a matrix and $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$ are column vectors. The masked vector $\boldsymbol{x}_S$ is defined as $\boldsymbol{x}_S(i) = x_i$ if $i \in S$, and 0 otherwise. We consider problems in Euclidean space equipped with an inner product $(\mathbb{R}^d, \langle \cdot, \cdot \rangle)$, where $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \sum_{i=1}^d x_i y_i$. The induced ℓ_2-norm is $\|\boldsymbol{x}\|_2 = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}$. An $1/\alpha$-scaling of set $\mathcal{D}$ is denoted as $\mathcal{D}/\alpha := \{\boldsymbol{x}/\alpha : \boldsymbol{x} \in \mathcal{D}\}$.

We define $\boldsymbol{x}$ on an underlying graph $\mathbb{G}(\mathbb{V}, \mathbb{E})$ where the node set $\mathbb{V} = [d]$ and edge set $\mathbb{E} \subseteq [d] \times [d]$. Each entry x_i is associated with each node v_i. Given $S' \subseteq \mathbb{V}, \mathbb{E}' \subseteq \mathbb{E}$, $\mathbb{G}(S', \mathbb{E}')$ represents a subgraph of $\mathbb{G}$, we say f is L-smooth if, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}$, there exists $L > 0$ such that

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \leq \frac{L}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2.$$

Graph-structured support sets. Before introducing the graph-structured support sets, we recall a more general form, structured support sets, as defined in the following.

Definition 2.1 (Structured support set). Given $\mathbb{M} := \{S_1, S_2, \ldots, S_m\}$, a collection of subsets of $[d]$, with $\cup_i S_i = [d]$, a structured support set $\mathcal{D}(C, \mathbb{M})$ is

$$\mathcal{D}(C, \mathbb{M}) := \text{conv}\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq C, \text{supp}(\boldsymbol{x}) \in \mathbb{M}\}.$$

The definition above is generally enough to model many interesting structures of $\boldsymbol{x}$. However, for graph-related problems, we specifically focus on where each S_i captures graph-structured information important for real-world applications. We define the graph-structured support sets as the following.

Definition 2.2 (Graph-structured support sets $\mathcal{D}$). Define a graph $\mathbb{G}(\mathbb{V} = [d], \mathbb{E})$ and consider a structured support set
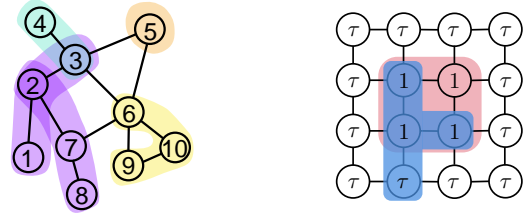

Figure 1. Left: An element of a g-subgraph model $\mathbb{M}(\mathbb{G}, s, g)$ defined on a 10-node graph where each colored region is a subgraph. $(g = 4, s = 5)$. *Right*: An example graph with $\mathbb{M} = \mathbb{M}(\mathbb{G}, s = 4, g = 1)$ where $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x}/2 - \boldsymbol{x}^\top \boldsymbol{b}$ with $\boldsymbol{b} = [1, 1, 1, 1, \tau, \ldots, \tau]^\top$ for some $0 < \tau < 1/2$ and . The red region is the optimal structure support of $\boldsymbol{x}^*$ while the blue is a best approximate structure support.

$\mathcal{D} := \mathcal{D}(C, \mathbb{M})$ where $C > 0$ and $\mathbb{M} = \{S_1, S_2, \ldots, S_m\}$. $\mathcal{D}$ is a *graph-structured support set* if each S_i is associated with a subgraph $\mathbb{G}(S_i, \mathbb{E}_i)$. We simply denote this set as $\mathcal{D}$.

Models of many real-world learning problems can be expressed as elements of the graph-structured support set. For example, learned structured polytopes are connected paths in a graph (Garber & Meshi, 2016), and models of multi-task learning are connected cliques (McDonald et al., 2016b). An important instance of Def. 2.2 is a simplified version of the weighted graph model $\mathbb{M}$ (Hegde et al., 2015b), which we call a g-subgraph model.

Definition 2.3 (g-subgraph model (Hegde et al., 2015b)). Given an underlying graph $\mathbb{G}(\mathbb{V}, \mathbb{E})$, the graph model $\mathbb{M} \triangleq \mathbb{M}(\mathbb{G}, s, g)$ is a union of g subgraphs

$$\mathbb{M}(\mathbb{G}, s, g) = \{S \triangleq S_1 \cup \cdots \cup S_g : |S| \leq s\}, \quad (2)$$

where subgraph $\mathbb{G}(S_i, \mathbb{E}_i)$ is connected.

We illustrate $\mathbb{M}(\mathbb{G}, s, g)$ in Fig. 1 (left).

FW-type algorithm. Given an initial $\boldsymbol{x}_0 \in \mathcal{D}$ and a learning rate $\eta_t \in [0, 1]$, for all $t \geq 0$, a FW-type algorithm for solving problem (1) uses the following updates

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t(\rho \boldsymbol{v}_t - \boldsymbol{x}_t), \quad (3)$$

where $\boldsymbol{v}_t$ is the minimizer of *graph-structured LMO*

$$\text{LMO} : \boldsymbol{v}_t \in \operatorname*{argmin}_{\boldsymbol{v} \in \mathcal{D}(C, \mathbb{M})} \langle a\boldsymbol{x}_t + b\nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle. \quad (4)$$

Setting $a = 0, b = 1, \eta_t = 2/(t + 2)$ with $\rho = 1$ recovers the standard FW. We will discuss more of its details in Sec. 4. In the rest, we focus on the graph-structured support set $\mathcal{D}(C, \mathbb{M})$ corresponding to g-subgraph model $\mathbb{M}(\mathbb{G}, s, g)$.

3. Graph-structured LMOs

This section deals with the computation aspects of our generalized LMO (4). We first present the hardness result in the

standard case and demonstrate how two popular gap-based LMOs are inappropriate in our setting.

3.1. Hardness of graph-structured LMOs

The computational barrier of (4) is mainly due to the combinatorial nature of $\mathbb{M}$. To illustrate this, we first limit our attention to the familiar LMO, where $a = 0, b = 1$. Note that in general, the LMO returns $\boldsymbol{v}_t$ an extreme point of $\mathcal{D}$. Hence, one can reformulate it as a subspace identification problem[2]

$$\min_{\boldsymbol{v} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle = \min_{S^* \in \mathbb{M}, \|\boldsymbol{v}\|_2 \leq C} \langle \nabla f(\boldsymbol{x}_t)_{S^*}, \boldsymbol{v} \rangle,$$

where S^* is an optimal support set minimizing the inner product in (4). Hence, the minimizer is

$$\boldsymbol{v}_t = -\frac{C \cdot \nabla f(\boldsymbol{x}_t)_{S^*}}{\|\nabla f(\boldsymbol{x}_t)_{S^*}\|_2}, \quad S^* \in \operatorname*{argmax}_{S \in \mathbb{M}} \|\nabla f(\boldsymbol{x}_t)_S\|_2^2. \quad (5)$$

From (5), we see that the computational complexity of graph-structured LMO is bottlenecked by finding S^*, whose complexity depends solely on $\mathbb{M}$. For example, in the simple instance when $\mathbb{M} = \{S \subseteq [d] : |S| \leq s\}$, the optimal s-sparse set can be discovered in $\mathcal{O}(d)$ time using a selection algorithm (Floyd & Rivest, 1975). However, when $\mathbb{M}$ is the g-subgraph model (2) or other types of models (Baldassarre et al., 2016; Lim & Wright, 2017), finding S^* is NP-hard in general as presented in Theorem 3.1.

Theorem 3.1 (Hardness of graph-structured LMO). *Given $C > 0, a = 0, b = 1$, convex differentiable function f, and the graph-structured model $\mathbb{M}$, the computation of an exact graph-structured LMO defined in (4) is NP-hard.*

Remark 3.2. The NP-hardness result is obtained via a reduction from the model selection problem (Baldassarre et al., 2016), which contains all instances of the maximum weighted coverage problem (Khuller et al., 1999).

3.2. Approximate LMOs and adversarial examples

To overcome the computational barrier, we resort to finding an approximate $\bar{\boldsymbol{v}}_t$ at each iteration. To obtain meaningful convergence rates for FW-type algorithms, one must first establish a metric of *approximation quality*, which will characterize the final bounds.

Additive approximate LMO. The gap-additive [3] approximate LMO (Dunn & Harshbarger, 1978) finds $\bar{\boldsymbol{v}}_t$ such that

$$\langle \nabla f(\boldsymbol{x}_t), \bar{\boldsymbol{v}}_t \rangle \leq \min_{\boldsymbol{v} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle + \mathcal{O}(\tfrac{\epsilon}{t}), \quad (6)$$

---

[2]Without loss of generality, we assume $a = 0, b = 1$ in (4). In the rest, we always assume $\mathcal{D} := \mathcal{D}(C, \mathbb{M})$.

[3]Given $\boldsymbol{x} \in \mathcal{D}$, the "gap" is to measure the difference between $\min_{\boldsymbol{v} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}), \boldsymbol{v} \rangle$ and $\langle \nabla f(\boldsymbol{x}), \bar{\boldsymbol{v}}_t \rangle$

where ϵ is the accuracy parameter and the approximate tolerance must decay over time in order to obtain an $\mathcal{O}(1/t)$ convergence rate.

Multiplicative approximate LMO. Another common approximation regime is the gap-multiplicative approximate LMO (Locatello et al., 2017), which returns $\bar{\boldsymbol{v}}_t$ such that

$$\underbrace{\langle \nabla f(\boldsymbol{x}_t), \bar{\boldsymbol{v}}_t - \boldsymbol{x}_t \rangle}_{\bar{g}_t(\boldsymbol{x}_t)} \leq \delta \cdot \underbrace{\min_{\boldsymbol{v} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} - \boldsymbol{x}_t \rangle}_{g_t(\boldsymbol{x}_t)}, \quad (7)$$

where $\delta \in (0, 1]$ is the approximation factor. Note that (7) is only possible if the approximation $\bar{g}_t(\boldsymbol{x}_t) < 0$ when $\boldsymbol{x}_t \notin \arg\min_{\mathcal{D}} f$. However, we show these two approximate LMOs (6) and (7) are generally impractical to obtain by constructing adversarial examples. That is, applied to the graph-structured support set $\mathcal{D}$; cases exist where any "approximation" is necessarily exact.

An adversarial problem setup. Consider a grid graph $\mathbb{G}$ (Fig. 1, right) and assume that $\mathbb{M} = \mathbb{M}(\mathbb{G}, s = 4, g = 1)$, i.e., a set of connected components of $\mathbb{G}$, each with at most four nodes. We map the four center nodes to the first 4 coordinates of $\boldsymbol{x}$. Suppose $\tau \in (0, 1/2)$ and

$$f(\boldsymbol{x}) = \tfrac{1}{2} \boldsymbol{x}^\top \boldsymbol{x} - \boldsymbol{x}^\top \boldsymbol{b}, \quad \boldsymbol{b} = [1, 1, 1, 1, \tau, \cdots, \tau]^\top.$$

Take $C = 1$ and then unique optimal solution of (1) has

$$\boldsymbol{x}^* = [\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, 0, \ldots, 0]^\top,$$
$$\nabla f(\boldsymbol{x}^*) = -[\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \tau, \cdots, \tau]^\top.$$

When $\boldsymbol{x}_t = \boldsymbol{x}^*$, $\boldsymbol{v}_t = \boldsymbol{x}^*$, and the duality gap $g_t(\boldsymbol{x}^*) = 0$.

Gap-additive bound cannot decay properly. In the ideal case, $\boldsymbol{x}_t \overset{t \to \infty}{\Longrightarrow} \boldsymbol{x}^*$. However, in the example above, pick any non-optimal $\bar{\boldsymbol{v}}_t$; for example, an instance, which only 1 element is suboptimal is

$$\bar{\boldsymbol{v}}_t = \frac{1}{\sqrt{3/4 + \tau^2}} [1, 1, 1, 0, \tau, 0, \cdots, 0]^\top$$

giving $g_t(\boldsymbol{x}^*) - \bar{g}_t(\boldsymbol{x}^*) = 1 - \sqrt{\frac{3}{4} + \tau^2} > 0$, which is strictly positive and constant in t, hence violating (6), even at $\boldsymbol{x}^*$. Moreover, this is a lower bound, since any more suboptimal $\bar{\boldsymbol{v}}_t$ can only increase this bound. The adversarial example at $\boldsymbol{x}^*$ shows that even if the algorithm starts arbitrarily close to the true solution $\boldsymbol{x}^*$, it still may fail with approximate LMOs. It is definitely not a unique problem point and not hard to extend this example not to $\boldsymbol{x}^*$ but any $\boldsymbol{x}_t$, for the appropriate problem. For example, one can find two adversarial examples at a non-optimal point $\boldsymbol{x}_t$ in Appendix B. Thus, any gap-additive assumption with a decaying tolerance will eventually require an exact LMO, which is NP-hard in general as stated in Thm. 3.1.

Gap-multiplicative estimate could be negative. Continuing with the above example, we notice that

$$\bar{g}_t(\boldsymbol{x}^*) = \nabla f(\boldsymbol{x}^*)^\top(\boldsymbol{x}^* - \bar{\boldsymbol{v}}_t) = \sqrt{3/4 + \tau^2} - 1 < 0,$$

which means no positive value of $\delta \in (0,1)$ can possibly satisfy (7); that is, the assumption is only satisfied if $\delta = 1$ and the LMO is exact, which again is NP-hard. Finally, we stress that the above adversarial examples are not contrived; they are generated using a very simple f and can be applied to very general types of graphs.

To further strengthen the argument that there exist interesting cases where no efficient approximate algorithms exist, note that one can map all instances of the maximum weighted coverage problem (Khuller et al., 1999) to our problem, but the best greedy algorithm for the maximum weighted coverage problem has approximation threshold of $1 - 1/e$ (Feige, 1998). This indicates a subset problems of (1) cannot have a better approximation bound.

4. Approximate FW-type methods

In practice, the above examples show that both (6) and (7) can be impossible to find unless we solve the LMO with exact support. Instead, rather than approximating the gap, we turn to approximating $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} \rangle$ using a *dual maximization oracle*; here, efficient inexact methods do exist and are easier to obtain. We first present the dual maximization oracle and then present our proposed approximate methods.

4.1. Approximate dual maximization oracle

Definition 4.1 (Inner Product Operator). *Given $\boldsymbol{z} \in \mathbb{R}^d$, $\mathcal{D} \subseteq \mathbb{R}^d$, and approximation factor $\delta \in (0,1]$, the approximated Inner Product Operator (IPO) returns $\boldsymbol{v}$ such that*

$$\text{Approximate IPO} \quad \langle \boldsymbol{z}, \boldsymbol{v} \rangle \leq \delta \cdot \min_{\boldsymbol{s} \in \mathcal{D}} \langle \boldsymbol{z}, \boldsymbol{s} \rangle. \quad (8)$$

We denote such set of $\boldsymbol{v}$ as $(\delta, \boldsymbol{z}, \mathcal{D})$-IPO.

Note that in general, if $\mathcal{D}$ is a symmetric nontrivial set then this quantity is negative. This approximate inner product operator is not new. Outside of FW-type methods, it is an important operator that can be used in approximate Matching Pursuit algorithms (Locatello et al., 2017; Mokhtari et al., 2018) and online linear optimization methods (Garber, 2017; 2021).

Definition 4.2 (Dual Maximization Oracle). *Given the structure support set $\mathbb{M}$, the dual approximation oracle finds an $S \in \mathbb{M}$ such that*

$$\|\boldsymbol{z}_S\|_* \geq \delta \cdot \max_{S' \in \mathbb{M}} \|\boldsymbol{z}_{S'}\|_*, \quad (9)$$

where $\delta \in (0,1]$ is the approximation factor and $\|\cdot\|_*$ corresponds to the dual norm ($\|\cdot\|_* = \|\cdot\|_2$ in our case). We denote such set S as the $(\delta, \boldsymbol{z}, \mathcal{D})$-DMO.

An important family of examples of this oracle are the approximate projections studied in Hegde et al. (2015b; 2016); Golbabaee & Davies (2018). A key observation from the above definitions is that the approximate IPO is equivalent to an approximate dual norm calculation. It is also in a recent work (Garber & Wolf, 2021): for certain norm balls, LMO is equivalent to the projection on extreme points.

One may notice that the exact LMO and DMO are identical. However, the approximation conditions are different: the usual assumptions for approximate LMO are not always realistic which we demonstrated this in previous section. In contrast, the DMO-inspired condition is weaker, but always satiable with existing methods (see the following Section 4.2). This weaker assumption complicates the convergence analysis (see Section 5), and we believe that the estimated error does not converge globally without making any assumptions.

Theorem 4.3. *Given the set $\mathcal{D}$ and suppose $S \in (\delta, \boldsymbol{z}, \mathcal{D})$-DMO. Define the approximate supporting vector $\tilde{\boldsymbol{v}}_t \triangleq -C \cdot \boldsymbol{z}_S/\|\boldsymbol{z}_S\|_2$. Then, $\tilde{\boldsymbol{v}}_t \in (\delta, \boldsymbol{z}, \mathcal{D})$-IPO.*

Thm. 4.3 provides an important way to obtain an approximate IPO via DMO, which is usually easier to obtain. That is, if one can find S such that $S \in (\delta, \boldsymbol{z}, \mathcal{D})$-DMO, then $\tilde{\boldsymbol{v}}_t$ is guaranteed in $(\delta, \boldsymbol{z}, \mathcal{D})$-IPO.

4.2. Efficient methods for DMO

Top-g + neighbor visiting. Let us first consider a simple method where we find an approximate $(\delta, z, \mathcal{D})$-DMO (9) for the g-subgraph model defined by (2) by first finding the indices of the g largest magnitude elements (denote as $I_g \subset S$) in the dual vector $\boldsymbol{z}$. Then add any feasible additional elements to S until $|S| = s$.[4] This can be done by just picking any adjacent nodes (without violating connectivity constraint) to the first g "seed" nodes. Then for any $S' \in \mathbb{M}$,

$$\left\lceil \frac{s}{g} \right\rceil \|\boldsymbol{z}_S\|_2^2 = \left\lceil \frac{s}{g} \right\rceil \left(\sum_{i \in I_g} |z_i|^2 + \sum_{j \in S \setminus I_g} |z_j|^2 \right) \geq \|\boldsymbol{z}_{S'}\|_2^2;$$

thus S approximates S^* with $\delta = \sqrt{1/\lceil s/g \rceil}$. This algorithm runs in $\mathcal{O}(|\mathbb{E}|)$.

In the adversarial example shown in Sec. 3.2, the 1-wrong examples of S' also achieve an approximate DMO-operator with $\delta = 3/4 + \tau^2$. In other words, unlike the gap-additive and gap-multiplicative approximations, *simple approximate solutions do not need to be exact in order to satisfy the IPO/DMO conditions*. Furthermore, one can find better methods from existing projection methods. For example, we can use the *head projection* (Hegde et al., 2015b), which provides a δ-DMO with $\delta = \sqrt{1/14}$ of a general weighted graph model, which runs in polynomial time $\mathcal{O}(|\mathbb{E}| \log^3 d)$.

---

[4]The exact algorithm description is in Appendix E.

4.3. Approximate FW-type methods via DMO

We present "standard" approximate FW-type methods in Alg. 1. There are 2 different design choices, which results in 4 variations. First, the approximate DMO can be performed on $z_t = -\nabla f(x_t)$, as in the usual FW (termed DMO-FW), or on $z_t = -(x_t - \frac{1}{L\eta_t}\nabla f(x_t))$ (termed DMO-ACCFW), which gives an accelerated variant based on the nearest extreme point oracle explored in Garber & Wolf (2021); Second, to obtain x_t, Option I (line 7) finds points within $\mathcal{D}$; while Option II (line 8) finds points over a relaxed constraint set $\mathcal{D}/\delta$. Specifically, DMO-ACCFW returns a convex combination of normalized supporting "gradient descent" S_t. One may treat it as a "hybrid" of FW and gradient descent. Hence, it has project-free property like FW but converges faster than the usual FW method.

Algorithm 1 FW-type methods for GSCOs

1: **Input:** step size $\{\eta_t\}$, δ, L, C, and $\mathbb{M}$
2: pick any point x_0 in $\mathcal{D}$
3: **for** $t = 0, 1, \ldots,$ **do**
4: $\quad z_t = \begin{cases} \text{DMO-FW} := -\nabla f(x_t) \\ \text{DMO-ACCFW} := -\left(x_t - \frac{\nabla f(x_t)}{L\eta_t}\right) \end{cases}$
5: $\quad S_t = (\delta, -z_t, \mathcal{D})\text{-DMO}$
6: $\quad \tilde{v}_t = \frac{C \cdot (z_t)_{S_t}}{\|(z_t)_{S_t}\|_2}$
7: $\quad$ Option I: $x_{t+1} = x_t + \eta_t(\tilde{v}_t - x_t)$
8: $\quad$ Option II: $x_{t+1} = x_t + \eta_t(\tilde{v}_t/\delta - x_t)$
9: **end for**
10: **Return** $(x_{\bar{t}}, f(x_{\bar{t}})), \bar{t} \in \underset{t}{\arg\min} f(x_t)$

---

5. Convergence analysis and sensing results

Denote the primal error $h(x_t) := f(x_t) - f(x^*)$ and assume the step size $\eta_t = 2/(t+2)$.[5] We first establish the convergence rates of the methods, and then showcase resulting on the graph-structured linear sensing problem. Throughout this section, we assume f to be L-smooth and denote $x^* \in \arg\min_{x \in \mathcal{D}} f(x)$ as a minimizer of (1).

5.1. Convergence rate of DMO-FW

By L-smoothness, for all $t \geq 0$, DMO-FW-I admits

$$h(x_{t+1}) \leq (1 - \delta\eta_t) h(x_t) + Q_t,$$

where

$$Q_t \triangleq (1-\delta)\eta_t \langle x_t, -\nabla f(x_t)\rangle + \frac{L\eta_t^2}{2}\|\tilde{v}_t - x_t\|_2^2.$$

Clearly, if $\delta = 1$, $Q_t = \mathcal{O}(1/t^2)$ and thus we can show recursively that $h(x_{t+1}) = \mathcal{O}(1/t)$. However, if $\delta < 1$, the

---

[5]Note that the solution $x_{\bar{t}}$ always has $h(x_{\bar{t}}) \leq h(x_t)$.

convergence of h is dominated by $\eta_t|\langle x_t, \nabla f(x_t)\rangle|$, which is not easy to bound in general. Thus, when $\delta < 1$, the standard proof technique fails. Instead, we first establish a δ-dependent convergence rate.

Theorem 5.1 (Universal rate of DMO-FW-I). *For all $t \geq 1$, the primal error of* DMO-FW-I *satisfies*

$$h(x_t) \leq \begin{cases} \min\left\{2C\sqrt{s}\|\nabla f(x_t)\|_\infty, P(\delta, 2\delta)\right\}, \delta \in (0, \frac{1}{2}] \\ \min\left\{P(\delta, 1), \frac{3(1-\delta)h(x_0)+A_t}{(2\delta-1)(t+2)}\right\}, \quad \delta \in (\frac{1}{2}, 1], \end{cases}$$

where s is maximal allowed sparsity, $P(\delta, \nu)$ is

$$P(\delta, \nu) \triangleq \frac{(1-\delta)9^\delta}{(t+2)^{2\delta}} \cdot h(x_0) + \frac{\ln(t+1)+1}{(t+2)^\nu} \cdot A_t$$

and $A_t \triangleq \max_{i \in [t]} 4(1-\delta)(i+2)|\langle x_i, \nabla f(x_i)\rangle| + 8LC^2$.

When $\delta = 1$, Thm. 5.1 recovers the standard convergence rate of the exact FW method. In the approximate case ($\delta < 1$), the bound is more involved. With an additional assumption on the decay of the magnitude of gradient, the following corollary shows the overall practical bound.

Corollary 5.2 (Practical rate of DMO-FW-I). *Suppose $\nu \in (0, 1]$ and $\|\nabla f(x)\|_\infty \leq B$, the primal error of* DMO-FW-I *admits the following practical bound*

$$h(x_t) \leq \begin{cases} \mathcal{O}\left(\frac{BC\sqrt{s}}{t^\nu}\right), & \|\nabla f(x_t)\|_\infty \leq \frac{B}{t^\nu} \\ \mathcal{O}\left(\frac{BC\sqrt{s}(1-\delta)}{\delta}\right), & Otherwise. \end{cases} \quad (10)$$

where $s \triangleq \max_{S \in \mathbb{M}} |S|$ is the maximal allowed sparsity.

Remark 5.3. Our practical bound (10) reveals the essential advantage of FW-type algorithms: sparse solutions provide $\sqrt{s}$ better than $\sqrt{d}$. Fig. 2 (left) also numerically substantiates our bound on the task of graph-structured linear regression. Fig. 2 (right) illustrates the decay of $\|\nabla f(x_t)\|_\infty$. Additionally, while the extra gradient decay assumption may seem unsatisfying, we observe it to be true in both our numerical simulations and real-world applications. In the other case, the gradient does not decay but assuming properly bounded; we have worst bound case $\mathcal{O}(BC\sqrt{s}(1-\delta)/\delta)$. Intuitively, the gradient decays when these NP-hard problems are "easy"; otherwise, they are truly "hard". In the latter case, it is better to use the convergence rate over the relaxed ball. This assumption is therefore not needed in our set expansion scenario (Option II) below.

Expanding $\mathcal{D}$ to $\mathcal{D}/\delta$. Even with the gradient decay assumption, the above analysis shows that even after infinite iterations, the approximation error could be large when δ is small. Alternatively, by allowing $x_t \in \mathcal{D}/\delta$, an expanded set of $\mathcal{D}$, we show that DMO-FW-II finds x_t in this *relaxed* problem at rate $\mathcal{O}(1/t)$. To see this, note that x_{t+1} is a
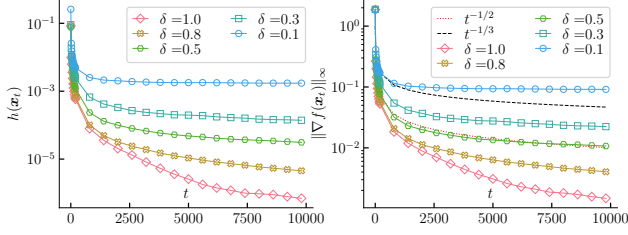
Figure 2. Left: The primal error $h(\boldsymbol{x}_t)$ as a function of t for DMO-FW-I with different δ on s-support norm ($s = 5$) under the task of graph-structured linear sensing. *Right*: The corresponding gradient norms as a function of t. More details of experimental setup can be found in Appendix C.

convex combination of approximate vectors $\tilde{\boldsymbol{v}}_i$. For each $\tilde{\boldsymbol{v}}_t$, we enlarge its length to $\|\tilde{\boldsymbol{v}}_t\|/\delta$ so that $\langle\nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t\rangle$ is a lower bound of $\min_{\boldsymbol{s}\in\mathcal{D}/\delta}\langle\nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{x}_t\rangle$. The rate of DMO-FW-II is stated in the following theorem.

Theorem 5.4 (Convergence of DMO-FW-II). *Under the same assumptions as in Thm. 5.1, for any $t \geq 1$, the primal error of* DMO-FW-II *satisfies*

$$h(\boldsymbol{x}_t) = f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{8LC^2}{\delta^2(t+2)}, \quad (11)$$

where $\boldsymbol{x}_t \in \mathcal{D}/\delta$ and $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$.

Remark 5.5. The rate in (11) is comparable with that of the two variants in Matching Pursuit algorithms (Locatello et al., 2018), which converge at $\mathcal{O}((\delta^2(t+2))^{-1})$. To compare, the aforementioned variants must estimate parameter L, while DMO-FW-II δ.

5.2. Convergence rate of DMO-AccFW

Note that DMO-FW achieves $\mathcal{O}(1/t)$ convergence rate even when $\delta = 1$. Inspired by Garber & Wolf (2021), we propose a variant FW-type method which uses an extreme point oracle, and can achieve a slightly better rate. The key idea is that rather than computing the approximate DMO over $\boldsymbol{z}_t = -\nabla f(\boldsymbol{x}_t)$, we instead compute the LMO over $\boldsymbol{z}_t = -\left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right)$, which, if $\|\boldsymbol{z}_t\|_2$ is constant, is equivalent to finding the Euclidean projection of $\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}$ on $\mathcal{D}$, that is, it is a projected gradient method with the per-iteration cost of a FW method. This new variant is named DMO-AccFW. With proper assumptions, we prove that this variant converges at $\mathcal{O}(1/t^2)$. Before we state our main result, recall the quadratic growth condition.

Definition 5.6 (μ-Quadratic Growth Condition). Let f be continuous differentiable. $\mathcal{X}^* \triangleq \arg\min_{\boldsymbol{x}\in\mathcal{D}} f(\boldsymbol{x})$. We say f satisfies the quadratic growth condition on $\mathcal{D}$ if there exists a constant $\mu > 0$ such that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\mu}{2}\|\boldsymbol{x} - [\boldsymbol{x}]_{\mathcal{X}^*}\|_2^2,$$

for all $\boldsymbol{x} \in \mathcal{D}$ where $[\boldsymbol{x}]_{\mathcal{X}^*} \triangleq \arg\min_{\boldsymbol{z}\in\mathcal{X}^*}\|\boldsymbol{z} - \boldsymbol{x}\|_2^2$.

The quadratic growth condition is weaker than the strongly convex condition. For the remainder of this subsection we assume the convex function f is L-smooth and satisfies a μ-quadratic growth condition over $\mathcal{D}$ for option I, and over $\mathcal{D}/\delta$ for option II. We state our main result as follows:

Theorem 5.7. *Assume further that $\boldsymbol{x}^*$ is on the boundary, i.e. $\boldsymbol{x}^* \in \mathcal{X}^*$ implies $\|\boldsymbol{x}^*\|_2 = C$. When $\delta = 1$, for all $t \geq 1$, the primal error of* DMO-AccFW *satisfies*

$$h(\boldsymbol{x}_t) \leq \frac{4e^{4L/\mu}h(\boldsymbol{x}_0)}{(t+2)^2}. \quad (12)$$

Standard FW method is well-known to have rate $\mathcal{O}(1/t^2)$ in general when the constraint set is a strongly convex set and f is strongly convex (Garber & Hazan, 2015). Our key extension is to cases where $\mathcal{D}$ may not be strongly convex, as is often the case for obtaining sparsity. Note that quadratic growth condition is a slightly weaker condition than strongly convex; in particular, nonconvex functions may still have quadratic growth. When $\boldsymbol{x}^*$ is not on the boundary, we have the following general convergence rate.

Corollary 5.8. *Denote $D_* = \min_{\boldsymbol{x}^*\in\mathcal{X}^*}\|\boldsymbol{x}^*\|_2$. When $\delta = 1$, for all $t \geq 1$,* DMO-AccFW-I *has the following convergence rate*

$$h(\boldsymbol{x}_t) \leq \min\left\{\frac{3Le^{2L/\mu}(C^2 - D_*^2)}{t+2} + \frac{4e^{4L/\mu}h(\boldsymbol{x}_0)}{(t+2)^2}, Z_t\right\},$$

where $Z_t = 2L(5C^2 - D_*^2)/(t+2)$.

Similar to the practical rate of DMO-FW-I, the practical rate and the rate on the expanded set $\mathcal{D}/\delta$ are stated in Thm. 5.9 and 5.10.

Theorem 5.9 (Practical rate of DMO-AccFW-I). *When $\delta \in (0,1)$ and $\|\nabla f(\boldsymbol{x})\|_\infty \leq B$, then* DMO-AccFW-I *admits*

$$h(\boldsymbol{x}_t) \leq \mathcal{O}\left(\sqrt{s}BC(1-\delta)/\delta\right).$$

Moreover, when $\|\nabla f(\boldsymbol{x})\|_\infty \leq B/t^\nu$ with $\nu \in (0,1]$, we have

$$h(\boldsymbol{x}_t) \leq \mathcal{O}\left(\sqrt{s}BC/t^\nu\right).$$

Theorem 5.10 (Convergence of DMO-AccFW-II). *When $\delta \in (0,1)$, then* DMO-AccFW-II *finds $\boldsymbol{x}_t \in \mathcal{D}/\delta$ and admits*

$$h(\boldsymbol{x}_t) \leq \frac{4e^{4L/\mu}}{(t+3)^2}h(\boldsymbol{x}_0) + \frac{28L^2(C^2/\delta^2 - D_*^2)}{5\mu(t+3)}. \quad (13)$$

The practical rate of DMO-AccFW shown in Thm 5.9 is the same as of DMO-FW as both two methods share the same DMO operator in worst case. For more comparisons between the above rates and the standard FW method, see Table 4 in the Appendix E.2.

5.3. Case study: Graph-structured linear sensing

The goal of graph-structured linear sensing is to recover a graph-sparse model $\tilde{\boldsymbol{x}}^*$ using fewer measurements than d. Specifically, measurements are generated as

$$\boldsymbol{y} = \langle \boldsymbol{A}, \tilde{\boldsymbol{x}}^* \rangle + \boldsymbol{e}, \qquad \boldsymbol{e} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d).$$

The sensing matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is Gaussian random ($a_{ij} \sim \mathcal{N}(0, 1/\sqrt{n})$ independently). Our goal is to recover $\tilde{\boldsymbol{x}}^*$ using regression $f(\boldsymbol{x}) := \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 / 2$ subject to graph-structured constraint $\mathcal{D}$. The following corollary provides the parameter estimation error bound for graph-structured linear sensing problem.

Corollary 5.11 (High probability parameter estimation). *Let $h(\boldsymbol{x}_t)$ be primal error for* DMO-FW *or* DMO-ACCFW. *The estimation error of the graph-structured linear sensing problem admits*

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t\|_2 \le \sqrt{\frac{2\sqrt{s}C\|\nabla f(\tilde{\boldsymbol{x}}^*)\|_\infty}{\mu}} + \sqrt{\frac{2h(\boldsymbol{x}_t)}{\mu}}, \quad (14)$$

Moreover, with large enough n, there exists an universal constant c such that with high probability:

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t\|_2 \le \left(\sqrt{\frac{4\sigma C\sqrt{s \log d/n}}{\frac{1}{2} - \frac{cs \log d}{n}}} + \sqrt{\frac{2h(\boldsymbol{x}_t)}{\frac{1}{2} - \frac{cs \log d}{n}}} \right).$$

6. Empirical results

We empirically evaluate our methods through the task of graph-structured linear regression problem over several graph-structured images. Our code and datasets will be made publically available upon publication, and is included in the submission. Our goal is to address two questions: **Q1.** Does DMO-ACCFW speed up DMO-FW? **Q2.** How does the approximation quality and efficiency compare with baseline methods? [6] Our code and datasets are accessible at https://github.com/baojian/dmo-fw.

DMO-AccFW converges faster than DMO-FW. To answer **Q1**, our empirical results as illustrated in Fig. 3 on task of graph-structured sparse recovery clearly demonstrate that DMO-ACCFW converges faster than DMO-FW and empirically matches the rate $\mathcal{O}(1/t^2)$ in (12) and estimation error matches $\mathcal{O}(1/t)$ in (14).

DMO-ACCFW *converges to a good local solution.* PGD-based methods and MP-based methods have been widely used on the task of sparse recovery. All methods shown linear convergence rate when training samples are sufficiently large. However, when the number of training samples are much less than p (extremely challenging to recover the original signal), our method demonstrates sensing superiority.

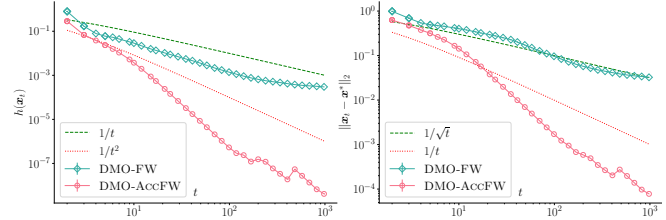[6]More experimental details are in Appendix D.

Figure 3. Left: Primal error $h(\boldsymbol{x}_t)$ as a function of t on task of the graph-structured sparse recovery of MNIST[7]. *Right*: The estimation error as a function of t.

Specifically, we pick $n = 2.5 \cdot |\operatorname{supp}(\boldsymbol{x}^*)|$ samples. We run each experiment for 20 trials, and compare our methods against the generalized MP (GEN-MP) discussed in Locatello et al. (2018) where each constant curvature is estimated by the maximal eigenvalue of $\boldsymbol{A}^\top \boldsymbol{A}$, COSAMP (Needell & Tropp, 2009), GraphCoSAMP (Hegde et al., 2015b), and GRAPH-IHT (Hegde et al., 2016) (the PGD-based method). The DMO we used is the head projection of Hegde et al. (2015b). We use Option I for DMO-ACCFW and simply set $L = 1$.
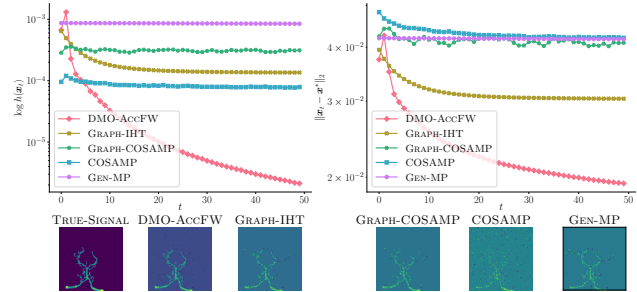

Figure 4. The performance of methods on the task of graph-structured linear regression task. *Top*: DMO-ACCFW vs other baseline methods on primal error $h(\boldsymbol{x}_t)$ *(left)* and variable suboptimality $\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2$ *(right)* as a function of t. *Bottom*: Recovered sparse images $\boldsymbol{x}_t$ vs truth image (bottom right) after 50 iterations.

The convergence comparison is shown in Fig. 4. Compared with the PGD-based method GRAPHIHT, DMO-ACCFW converges faster to a good local optimal while GRAPHIHT appears to be stuck in a local minimum. Three MP-based methods also failed to cleanly recover the original signal. Our DMO-ACCFW is the fastest one to converge to a good local minimum at very early stage. The bottom recovered image also demonstrate that DMO-ACCFW returns sparse solution at early stage. The efficiency of DMO-ACCFW has been demonstrated in Table 1 where it is several times faster than baselines because baselines either have multiple projections or need to have minimization step per-iteration.

Table 1. The comparison of run time of all methods.

Method	Run time (seconds)
GRAPHCOSAMP	662.10±252.21
COSAMP	531.57±45.47
GENMP	661.68±177.65
GRAPHIHT	445.89±22.63
DMO-ACCFW	**87.61±23.39**

7. Discussion and Conclusion

We have studied the approximate FW-type methods for GSCO problems. We first demonstrate that there exist adversarial examples such that two popular inexact LMOs are at least as hard to compute as the exact LMO. Instead, we consider an inexact DMO which is equivalent to an approximation on the inner product rather than the gap, and prove that the inexact DMO is equivalent to the inexact IPO. The standard FW admits $\mathcal{O}((1-\delta)\sqrt{s}/\delta)$ and our accelerated version has rate $\mathcal{O}(1/t^2)$ with a proper condition. We also prove that a relaxed version of FW admits $\mathcal{O}(LC^2/(\delta^2 t))$ for general convex functions where the iterates x_t may be infeasible, and converge to a relaxed set.

One weakness of the present work is that our established bound depends on the decay of the gradient with the fixed step size strategy, which is observed in practice but difficult to show theoretically. If x^* is on the boundary, then $\|\nabla f(x^*)\|_\infty$ may never decay to 0, and the proposed methods converge to an error neighborhood. We also do not consider line search, away steps, or fully-corrective versions, and limit our attention to the traditional fixed step size strategies. Finally, our initial experiments indicate that inexact FW-type methods are attractive for this type of problems and one is encouraged to find faster methods based on fully-corrective or other methods.

8. Acknowledgement

References

Abernethy, J. and Wang, J.-K. On Frank-Wolfe and equilibrium computation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6587–6596, 2017.

Aksoylar, C., Orecchia, L., and Saligrama, V. Connected subgraph detection with mirror descent on sdps. In *International Conference on Machine Learning*, pp. 51–59. PMLR, 2017.

Allen-Zhu, Z., Hazan, E., Hu, W., and Li, Y. Linear convergence of a Frank-Wolfe type algorithm over trace-norm balls. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6192–6201, 2017.

Argyriou, A., Foygel, R., and Srebro, N. Sparse prediction with the k-support norm. In *Advances in Neural Information Processing Systems*, pp. 1457–1465, 2012.

Arias-Castro, E., Candes, E. J., and Durand, A. Detection of an anomalous cluster in a network. *The Annals of Statistics*, pp. 278–304, 2011.

Bach, F. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1): 115–129, 2015.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012a.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012b.

Bahmani, S., Raj, B., and Boufounos, P. T. Greedy sparsity-constrained optimization. *Journal of Machine Learning Research*, 14(Mar):807–841, 2013.

Balasubramanian, K. and Ghadimi, S. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Advances in Neural Information Processing Systems*, pp. 3455–3464, 2018.

Baldassarre, L., Bhan, N., Cevher, V., Kyrillidis, A., and Satpathi, S. Group-sparse model selection: Hardness and relaxations. *IEEE Transactions on Information Theory*, 62(11):6508–6534, 2016.

Baraniuk, R. G., Cevher, V., Duarte, M. F., and Hegde, C. Model-based compressive sensing. *IEEE Transactions on information theory*, 56(4):1982–2001, 2010.

Bashiri, M. A. and Zhang, X. Decomposition-invariant conditional gradient for general polytopes with line search. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2687–2697, 2017.

Berthet, Q. and Perchet, V. Fast rates for bandit optimization with upper-confidence Frank-Wolfe. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 2222–2231, 2017.

Briol, F.-X., Oates, C. J., Girolami, M., and Osborne, M. A. Frank-Wolfe bayesian quadrature: probabilistic integration with theoretical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 1162–1170, 2015.

Candes, E. and Tao, T. The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351, 2007.

Combettes, C. and Pokutta, S. Boosting Frank-Wolfe by chasing gradients. In *International Conference on Machine Learning*, pp. 2111–2121. PMLR, 2020.

Combettes, C. W. and Pokutta, S. Blended matching pursuit. *arXiv preprint arXiv:1904.12335*, 2019.

Combettes, C. W. and Pokutta, S. Complexity of linear minimization and projection on some sets. *Operations Research Letters*, 2021.

Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A., and Sahinalp, S. C. Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27(13):i205–i213, 2011.

Dunn, J. C. and Harshbarger, S. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

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

Floyd, R. W. and Rivest, R. L. Algorithm 489: The algorithm select—for finding the ith smallest of n elements [m1]. *Commun. ACM*, 18(3):173, March 1975.

Frank, M., Wolfe, P., et al. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.

Garber, D. Efficient online linear optimization with approximation algorithms. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 627–635, 2017.

Garber, D. Revisiting Frank-Wolfe for polytopes: Strict complementarity and sparsity. *Advances in Neural Information Processing Systems*, 33, 2020.

Garber, D. Efficient online linear optimization with approximation algorithms. *Mathematics of Operations Research*, 46(1):204–220, 2021.

Garber, D. and Hazan, E. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *32nd International Conference on Machine Learning, ICML 2015*, 2015.

Garber, D. and Meshi, O. Linear-memory and decomposition-invariant linearly convergent conditional gradient algorithm for structured polytopes. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1009–1017, 2016.

Garber, D. and Wolf, N. Frank-Wolfe with a Nearest Extreme Point Oracle. In *Proceedings of Thirty Fourth Conference on Learning Theory (COLT)*, Proceedings of Machine Learning Research, pp. 2103–2132. PMLR, 2021.

Golbabaee, M. and Davies, M. E. Inexact gradient projection and fast data driven compressed sensing. *IEEE Transactions on Information Theory*, 64(10):6707–6721, 2018.

Hallac, D., Leskovec, J., and Boyd, S. Network lasso: Clustering and optimization in large graphs. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 387–396, 2015.

Hazan, E., Hu, W., Li, Y., and Li, Z. Online improper learning with an approximation oracle. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 5657–5665, 2018.

Hegde, C., Indyk, P., and Schmidt, L. A fast, adaptive variant of the goemans-williamson scheme for the prize-collecting steiner tree problem. In *Workshop of the 11th DIMACS Implementation Challenge*, volume 2, pp. 3. Workshop of the 11th DIMACS Implementation Challenge, 2014.

Hegde, C., Indyk, P., and Schmidt, L. Approximation algorithms for model-based compressive sensing. *IEEE Transactions on Information Theory*, 61(9):5129–5147, 2015a.

Hegde, C., Indyk, P., and Schmidt, L. A nearly-linear time framework for graph-structured sparsity. In *International Conference on Machine Learning*, pp. 928–937, 2015b.

Hegde, C., Indyk, P., and Schmidt, L. Fast recovery from a union of subspaces. In *Advances in Neural Information Processing Systems*, pp. 4394–4402, 2016.

Hochbaum, D. S. and Pathria, A. Node-optimal connected k-subgraphs. *University of California, Berkeley*, 1994.

Jacob, L., Obozinski, G., and Vert, J.-P. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pp. 433–440, 2009.

Jaggi, M. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th international conference on machine learning*, pp. 427–435, 2013.

Jain, P., Tewari, A., and Kar, P. On iterative hard thresholding methods for high-dimensional m-estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, pp. 685–693, 2014.

Karimi, H., Nutini, J., and Schmidt, M. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Kerdreux, T., Pedregosa, F., and d'Aspremont, A. Frank-Wolfe with subsampling oracle. In *International Conference on Machine Learning*, pp. 2591–2600. PMLR, 2018.

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

Krishnan, R. G., Lacoste-Julien, S., and Sontag, D. Barrier Frank-Wolfe for marginal inference. *Advances in Neural Information Processing Systems*, 2015:532–540, 2015.

Lacoste-Julien, S. and Jaggi, M. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in neural information processing systems*, pp. 496–504, 2015.

Lei, Q., Zhuo, J., Caramanis, C., Dhillon, I. S., and Dimakis, A. G. Primal-dual block generalized Frank-Wolfe. *Advances in Neural Information Processing Systems*, 32: 13866–13875, 2019.

Lim, C. H. and Wright, S. k-support and ordered weighted sparsity for overlapping groups: Hardness and algorithms. In *Advances in Neural Information Processing Systems*, pp. 284–292, 2017.

Locatello, F., Khanna, R., Tschannen, M., and Jaggi, M. A unified optimization view on generalized matching pursuit and Frank-Wolfe. In *Artificial Intelligence and Statistics*, pp. 860–868, 2017.

Locatello, F., Raj, A., Karimireddy, S. P., Rätsch, G., Schölkopf, B., Stich, S., and Jaggi, M. On matching pursuit and coordinate descent. In *International Conference on Machine Learning*, pp. 3198–3207. PMLR, 2018.

Locatello, F., Yurtsevert, A., Fercoq, O., and Cevhert, V. Stochastic Frank-Wolfe for composite convex minimization. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 32, 2019.

Luise, G., Salzo, S., Pontil, M., and Ciliberto, C. Sinkhorn barycenters with free support via Frank-Wolfe algorithm. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

McDonald, A., Pontil, M., and Stamos, D. Fitting spectral decay with the k-support norm. In *Artificial Intelligence and Statistics*, pp. 1061–1069. PMLR, 2016a.

McDonald, A. M., Pontil, M., and Stamos, D. New perspectives on k-support and cluster norms. *The Journal of Machine Learning Research*, 17(1):5376–5413, 2016b.

Mokhtari, A., Hassani, H., and Karbasi, A. Conditional gradient method for stochastic submodular maximization: Closing the gap. In *International Conference on Artificial Intelligence and Statistics*, pp. 1886–1895. PMLR, 2018.

Needell, D. and Tropp, J. A. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3):301–321, 2009.

Nguyen, N., Needell, D., and Woolf, T. Linear convergence of stochastic iterative greedy algorithms with sparse constraints. *IEEE Transactions on Information Theory*, 63 (11):6869–6895, 2017.

Pedregosa, F., Negiar, G., Askari, A., and Jaggi, M. Linearly convergent Frank-Wolfe with backtracking line-search. In *International Conference on Artificial Intelligence and Statistics*, pp. 1–10. PMLR, 2020.

Ping, W., Liu, Q., and Ihler, A. T. Learning infinite RBMs with Frank-Wolfe. *Advances in Neural Information Processing Systems*, 29:3063–3071, 2016.

Qian, J., Saligrama, V., and Chen, Y. Connected sub-graph detection. In *Artificial Intelligence and Statistics*, pp. 796–804. PMLR, 2014.

Riabov, A. and Liu, Z. Scalable planning for distributed stream processing systems. In *ICAPS*, pp. 31–41, 2006.

Sharpnack, J., Singh, A., and Rinaldo, A. Sparsistency of the edge lasso over graphs. In *Artificial Intelligence and Statistics*, pp. 1028–1036. PMLR, 2012.

Sun, Y. and Bach, F. Safe screening for the generalized conditional gradient method. *arXiv preprint arXiv:2002.09718*, 2020.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Projection efficient subgradient method and optimal nonsmooth Frank-Wolfe method. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12211–12224, 2020.

Yuan, X., Li, P., and Zhang, T. Gradient hard thresholding pursuit for sparsity-constrained optimization. In *International Conference on Machine Learning*, pp. 127–135. PMLR, 2014.

Section A provides all missing proofs. Section B presents the adversarial examples at non-optimal points $\boldsymbol{x}_t$. Section C gives the experimental details of Fig. 2. Section D provides more experimental results on graph-structured sparse recovery problem. Finally, in Section E, we present two DMOs for $\mathbb{M}(\mathbb{G}, s, g)$ and briefly discuss other DMOs. Our code, datasets, and results are also provided in the supplementary material and will be made available on publish.

A. Proofs

A.1. Proof of Theorem 3.1

Before proving the NP-hardness result, we first introduce *the group model selection (GMS) problem* (Baldassarre et al., 2016). Notations of this problem are: Let $\tau(\boldsymbol{y})$ be the indicator vector where

$$\tau(\boldsymbol{y})_i = \begin{cases} 1 & \text{if } y_i \neq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Given the ground set $[N]$, the group structure $\mathbb{B} \triangleq \{B_1, B_2, \ldots, B_T\}$ is a collection of index sets with

$$B_i \subseteq [N], \quad |B_i| \leq b_i, \quad \cup_{B_i \in \mathbb{B}} B_i = [N].$$

The group indicator vector $\boldsymbol{\omega}$ denotes activity of elements in $\mathbb{B}$, that is, $\omega_i = 1$ if B_i is active 0 otherwise. To encode the group structure $\mathbb{B}$, an indicator matrix $\boldsymbol{A} \in \mathbb{B}^{N \times T}$ is defined where each entry $a_{ij} = 1$ if $i \in B_j$ and 0 otherwise. Based on this definition, $\boldsymbol{A}\boldsymbol{\omega} \geq \tau(\boldsymbol{y})$ means that for each nonzero y_i, there at least one active group in $\mathbb{B}$ covers y_i. Given any $\boldsymbol{y} \in \mathbb{R}^n$, a best γ-group sparse approximation $\hat{\boldsymbol{y}}$ is given by

$$\hat{\boldsymbol{y}} \in \operatorname*{argmin}_{\boldsymbol{z} \in \mathbb{R}^N} \left\{ \|\boldsymbol{y} - \boldsymbol{z}\|_2^2 : \|\boldsymbol{z}\|_{\mathbb{B},0} \leq \gamma \right\}, \text{ where } \|\boldsymbol{z}\|_{\mathbb{B},0} \triangleq \min_{\boldsymbol{\omega} \in \mathbb{B}^T} \left\{ \sum_{j=1}^T \omega_j : \boldsymbol{A}\boldsymbol{\omega} \geq \tau(\boldsymbol{z}) \right\}, \tag{15}$$

where $\|\boldsymbol{z}\|_{\mathbb{B},0}$ expresses the minimal number of active groups that cover $\boldsymbol{z}$.

Problem A.1 (The group model selection (GMS) problem (Baldassarre et al., 2016)). *Let $[N]$ be the ground set. Assume the group structure $\mathbb{B} := \{B_1, B_2, \ldots, B_T\}$ where each $B_i \subseteq [N]$ and $\cup_{B_i \in \mathbb{B}} B_i = [N]$. Given any input $\boldsymbol{y} \in \mathbb{R}^N$, a γ-group cover for its γ-group sparse approximation is expressed as follows*

$$\mathcal{S}(\gamma, \hat{\boldsymbol{y}}) \in \operatorname*{argmax}_{\mathcal{S} \subseteq \mathbb{B}} \left\{ \|\boldsymbol{y}_I\|_2^2 : I = \cup_{S_i \in \mathcal{S}} S_i, |\mathcal{S}| \leq \gamma \right\}, \tag{16}$$

where the γ-group cover $\mathcal{S}(\gamma, \hat{\boldsymbol{y}})$ is a group cover for $\hat{\boldsymbol{y}}$ with at most γ groups, that is

$$\mathcal{S}(\gamma, \hat{\boldsymbol{y}}) = \left\{ B_i \in \mathbb{B} : \boldsymbol{\omega} \in \mathbb{B}^T, \omega_i = 1, \boldsymbol{A}\boldsymbol{\omega} \geq \tau(\hat{\boldsymbol{y}}), \sum_{j=1}^T \omega_j \leq \gamma \right\}. \tag{17}$$

The GMS problem is to find such cover $\mathcal{S}(\gamma, \hat{\boldsymbol{y}})$ for any given $\boldsymbol{y}$ and γ.

Lemma A.2 (The NP-hardness of the GMS problem (Baldassarre et al., 2016)). *Given the ground set $[N]$ and γ is a positive integer. Suppose $\boldsymbol{y} \in \mathbb{R}^N$ and $\mathbb{B} \in \mathcal{P}([N])$. The group model selection problem is NP-hard.*

To prove the NP-hardness of graph-structured LMO, we make a reduction from the GMS problem to it. The main observation is that, given any instance of GMS problem, one can construct $\mathbb{M}$ such that $\mathcal{S}(\gamma, \hat{\boldsymbol{y}})$ can be recovered from $\operatorname{supp}(\boldsymbol{v}_t)$.

Theorem 3.1 (Hardness of graph-structured LMO). Given $C > 0, a = 0, b = 1$, convex differentiable $f : \mathbb{R}^d \to \mathbb{R}$, and the graph-structured model $\mathbb{M}$, the computation of the graph-structured LMO defined in (4) is NP-hard.

Proof. For any given instance of the GMS problem defined in Table 2, one can find an instance of the graph structured LMO problem: Let $d = N$, fix $a = 0, b = 1$ and let the convex differentiable function $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{y}$, so that $\nabla f(\boldsymbol{x}_t) = \boldsymbol{y}$. The

Table 2. The comparison of parameter configurations between two problems.

	The group model selection problem	The computation of graph-structured LMO
Ground set	$[N] \triangleq \{1, 2, \ldots, N\}$	$[d] \triangleq \{1, 2, \ldots, d\}$
Input vector	$\boldsymbol{y}$	$a\boldsymbol{x}_t + b\nabla f(\boldsymbol{x}_t)$, $a, b \in \mathbb{R}^d$ and $\nabla f(\boldsymbol{x}_t) \in \mathbb{R}^d$
Structure model	$\mathbb{B} = \{B_1, B_2, \ldots, B_T\}$	$\mathbb{M} = \{S_1, S_2, \ldots, S_M\}$
Group size	γ	g
Solution	$\hat{\boldsymbol{y}}_I = \boldsymbol{y}_I$ where $I \subseteq \mathbb{B}$ with $|I| = \gamma$	$\boldsymbol{v}_t = \frac{-C \cdot \nabla f(\boldsymbol{x}_t)_{S^*}}{\|\nabla f(\boldsymbol{x}_t)_{S^*}\|_2}$ where $S^* \in \mathbb{M}$

rest is to construct $\mathbb{M}$ so that $\mathrm{supp}(\boldsymbol{v}_t)$ corresponds to I. Follow a similar argument presented in Sec. 3.1, we have

$$\min_{\boldsymbol{s} \in \mathcal{D}(C,\mathbb{M})} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} \rangle = \min_{\boldsymbol{s} \in \mathrm{conv}\{\cup_{S \in \mathbb{M}} \mathcal{B}(S,C)\}} \langle \boldsymbol{y}, \boldsymbol{s} \rangle$$

$$= \min_{\boldsymbol{s} \in \cup_{S \in \mathbb{M}} \mathcal{B}(S,C)} \langle \boldsymbol{y}, \boldsymbol{s} \rangle$$

$$= \min_{\boldsymbol{s} \in \mathcal{B}(S^*,C)} \langle \boldsymbol{y}_{S^*}, \boldsymbol{s} \rangle,$$

where S^* is an optimal support set and $\mathcal{B}(S, C) = \{\mathrm{supp}(\boldsymbol{x}) = S, \|\boldsymbol{x}\|_2 \le C\}$. Therefore, we have the following equivilent representation of $\boldsymbol{v}_t$

$$\boldsymbol{v}_t = -\frac{C \cdot \boldsymbol{y}_{S^*}}{\|\boldsymbol{y}_{S^*}\|_2}, S^* \in \underset{S \in \mathbb{M}}{\mathrm{argmax}} \|\boldsymbol{y}_S\|_2^2,$$

where we let $C = 1$. Now let $\mathbb{M} = \{S_i : S_i = \cup_{Q_j \in \mathcal{S}} Q_j, |\mathcal{S}| \le \gamma, \mathcal{S} \subseteq \mathbb{B}\}$. That is, $\mathbb{M}$ is problem space of (16). Notice further that, given the underlying graph $\mathbb{G}(\mathbb{V}, \mathbb{E})$, one can define $\mathbb{M}(s, g = \gamma) \triangleq \mathbb{M}$ with $s = \max_i |S_i|$. Therefore, with this specific configuration, the solution S^* of the graph-structured LMO is I. One can immediately recover $\mathcal{S}(\gamma, \hat{\boldsymbol{y}})$ from I by using the fact that $I = S^* = \cup_{Q_j \in \mathcal{S}} Q_j$ for a specific $\mathcal{S}$. $\square$

A.2. Proof of Theorem 4.3

Theorem 4.3. Given the set $\mathcal{D}$ and suppose $S \in (\delta, \boldsymbol{z}, \mathcal{D})$-DMO. Define the approximate supporting vector $\tilde{\boldsymbol{v}}_t \triangleq -C \cdot \boldsymbol{z}_S / \|\boldsymbol{z}_S\|_2$. Then, $\tilde{\boldsymbol{v}}_t \in (\delta, \boldsymbol{z}, \mathcal{D})$-IPO.

Proof. Since S is a support returned by $(\delta, \boldsymbol{z}, \mathcal{D})$-DMO and $\tilde{\boldsymbol{v}}_t = \frac{-C}{\|\boldsymbol{z}_S\|_2} \boldsymbol{z}_S$, we have the following

$$\langle \boldsymbol{z}, \tilde{\boldsymbol{v}}_t \rangle = \left\langle \boldsymbol{z}, \frac{-C}{\|\boldsymbol{z}_S\|_2} \boldsymbol{z}_S \right\rangle$$

$$= -C \cdot \|\boldsymbol{z}_S\|_2$$

$$\le -C \cdot \delta \max_{S' \in \mathbb{M}} \|\boldsymbol{z}_{S'}\|_2$$

$$= \delta \cdot \min_{S' \in \mathbb{M}} \|\boldsymbol{z}_{S'}\|_2 \cdot (-C), \tag{18}$$

where the last inequality due to the fact that S is in $(\delta, \boldsymbol{z}, \mathcal{D})$-DMO. As defined in (8), in the rest, we shall prove that $\langle \boldsymbol{z}, \tilde{\boldsymbol{v}}_t \rangle \le \delta \cdot \min_{\boldsymbol{s} \in \mathcal{D}} \langle \boldsymbol{z}, \boldsymbol{s} \rangle$. Recall $\mathcal{D} = \mathrm{conv}(\cup_{I \in \mathbb{M}} \{\boldsymbol{w} \in \mathbb{R}^d : \mathrm{supp}(\boldsymbol{w}) \subseteq I, \|\boldsymbol{w}\|_2 \le C\})$. Denote the ball induced by I as $\mathcal{B}(I) := \{\boldsymbol{x} \in \mathbb{R}^d : \mathrm{supp}(\boldsymbol{x}) \subseteq I, \|\boldsymbol{w}\|_2 \le C\}$. Notice that $\boldsymbol{s}$ be must in the set of extreme points. We have

$$\min_{\boldsymbol{s} \in \mathcal{D}} \langle \boldsymbol{z}, \boldsymbol{s} \rangle = \min_{\boldsymbol{s} \in \mathrm{conv}(\cup_{I \in \mathbb{M}} \mathcal{B}(I))} \langle \boldsymbol{z}, \boldsymbol{s} \rangle$$

$$= \min_{\boldsymbol{s} \in \cup_{I \in \mathbb{M}} \mathcal{B}(I)} \langle \boldsymbol{z}, \boldsymbol{s} \rangle$$

$$= \min_{\boldsymbol{s} \in \mathcal{B}(I^*)} \langle \boldsymbol{z}_{I^*}, \boldsymbol{s} \rangle,$$

where we denote I^* as the support of $\boldsymbol{s}$ in $\mathbb{M}$. By Cauchy-Schwarz inequality, we always have $-\|\boldsymbol{z}_{I^*}\|_2 \cdot \|\boldsymbol{s}\|_2 \le \langle \boldsymbol{z}_{I^*}, \boldsymbol{s} \rangle$. When $\boldsymbol{s} = -C \cdot \boldsymbol{z}_{I^*} / \|\boldsymbol{z}_{I^*}\|_2$, it attains the minimal value $-C \cdot \|\boldsymbol{z}_{I^*}\|_2$. Therefore, we have

$$\min_{\boldsymbol{s} \in \mathcal{D}} \langle \boldsymbol{z}, \boldsymbol{s} \rangle = \|\boldsymbol{z}_{I^*}\|_2 \cdot (-C) \text{ for some } I^* \in \mathbb{M}.$$

As $I^* \in \mathbb{M}$, we continue (18) to have

$$
\begin{aligned}
\langle \boldsymbol{z}, \tilde{\boldsymbol{v}}_t \rangle &\leq \delta \cdot \min_{S' \in \mathbb{M}} \|\boldsymbol{z}_{S'}\|_2 (-C) \\
&\leq \delta \cdot \|\boldsymbol{z}_{I^*}\|_2 \cdot (-C) \\
&= \delta \cdot \min_{\boldsymbol{s} \in \mathcal{D}} \langle \boldsymbol{z}, \boldsymbol{s} \rangle.
\end{aligned}
$$

The above inequality indicates that given S that satisfies DMO property, then $\tilde{\boldsymbol{v}}_t$ defined based on S is a solution of IPO operator. We prove the lemma. $\square$

Remark A.3. Theorem 4.3 provides an easier way to find a solution of IPO. It also indicates IPO operator and DMO operator are equivalent. That is, for an existing $\boldsymbol{v}_t \in (\delta, \boldsymbol{z}, \mathcal{D})$-IPO, one can find $\mathrm{supp}(\boldsymbol{v}_t) \in (\delta, \boldsymbol{z}, \mathcal{D})$. In case of k-support norm, the proof of exact equivalence (i.e., $\delta = 1$) appears in the Proposition 2 of McDonald et al. (2016a), Section 2.1 of Argyriou et al. (2012), and Lemma 2 of Jacob et al. (2009) where calculating the dual norm is equivalent to solving LMO. In particular, let I_k be the top-k largest magnitudes of $\boldsymbol{u}$, then any $\boldsymbol{s}$ such that $\langle \boldsymbol{u}, \boldsymbol{s} \rangle = \|\boldsymbol{u}_{I_k}\|$ will be a solution of LMO, with $\boldsymbol{s} = -\boldsymbol{u}_{I_k}/\|\boldsymbol{u}_{I_k}\|_2$.

A.3. Proof of Theorem 5.1

Denote the primal error $h(\boldsymbol{x}_t) := f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)$ and assume the step size $\eta_t = 2/(t+2)$. Note that the solution $\boldsymbol{x}_{\bar{t}}$ always has $h(\boldsymbol{x}_{\bar{t}}) \leq h(\boldsymbol{x}_t)$. We assume f is L-smooth and denote $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$ as a minimizer of (1). We begin to introduce a key lemma as the following

Lemma A.4. *Given $h : \mathbb{R}^d \to \mathbb{R}^+$ and the following first-order non-homogeneous recurrence relation*

$$
h(\boldsymbol{x}_{t+1}) \leq \left(1 - \frac{2\delta}{t+2}\right) h(\boldsymbol{x}_t) + \frac{A_t}{(t+2)^2}, \tag{19}
$$

where $\delta \in (0, 1]$, $t \geq 0$, and $A_t \in \mathbb{R}^+$. Then, $\forall t \geq 1$, we have

$$
h(\boldsymbol{x}_t) \leq
\begin{cases}
\dfrac{(1-\delta)9^\delta}{(t+2)^{2\delta}} \cdot h(\boldsymbol{x}_0) + \dfrac{\ln(t+1)+1}{(t+2)^{2\delta}} \cdot A & \text{if } \delta \in (0, \tfrac{1}{2}] \\[3mm]
\min\left\{ \dfrac{(1-\delta)9^\delta}{(t+2)^{2\delta}} \cdot h(\boldsymbol{x}_0) + \dfrac{\ln(t+1)+1}{t+2} \cdot A, \dfrac{3(1-\delta)h(\boldsymbol{x}_0)+A}{(2\delta-1)(t+2)} \right\} & \text{if } \delta \in (\tfrac{1}{2}, 1],
\end{cases} \tag{20}
$$

where A be such that $A_t \leq A$.

Proof. We directly expand the recurrence relation (19) as the following

$$
\begin{aligned}
h(\boldsymbol{x}_{t+1}) &\leq \left(1 - \frac{2\delta}{t+2}\right) h(\boldsymbol{x}_t) + \frac{A_t}{(t+2)^2} \\
&\leq \left(1 - \frac{2\delta}{t+2}\right)\left(1 - \frac{2\delta}{t+1}\right) h(\boldsymbol{x}_{t-1}) + \left(1 - \frac{2\delta}{t+2}\right)\left(\frac{A_{t-1}}{(t+1)^2}\right) + \frac{A_t}{(t+2)^2} \\
&\vdots \\
&\leq \prod_{i=0}^t \left(1 - \frac{2\delta}{i+2}\right) \cdot h(\boldsymbol{x}_0) + \sum_{j=0}^{t-1} \left\{ \frac{A_j}{(j+2)^2} \prod_{k=j+1}^t \left(1 - \frac{2\delta}{k+2}\right) \right\} + \frac{A_t}{(t+2)^2} \triangleq H(t, \delta), \tag{21}
\end{aligned}
$$

where we define the above tight bound (21) as $H(t, \delta)$. In the rest, we aim at getting an explainable upper bound of $H(t, \delta)$. First, note that

$$
\prod_{i=j}^t \left(1 - \frac{2\delta}{i+2}\right) \overset{(a)}{\leq} \left(e^{-2\delta}\right)^{\sum_{i=j}^t \frac{1}{i+2}} \overset{(b)}{\leq} \left(\frac{j+2}{t+3}\right)^{2\delta}. \tag{22}
$$

where (a) is from $1 - x \leq e^{-x}$ and (b) is from the integral bound:

$$
\sum_{i=j}^t \frac{1}{i+2} \geq \int_j^{t+1} \frac{1}{\tau+2} d\tau = \ln(t+3) - \ln(j+2).
$$

Hence, taking

$$\prod_{i=0}^{t}\left(1-\frac{2\delta}{i+2}\right)=(1-\delta)\prod_{i=1}^{t}\left(1-\frac{2\delta}{i+2}\right),$$

then $H(t,\delta)$ reduces to

$$
\begin{aligned}
H(t,\delta) &\leq (1-\delta)\left(\frac{3}{t+3}\right)^{2\delta}\cdot h(\boldsymbol{x}_0)+\sum_{j=0}^{t-1}\left\{\frac{A_j}{(j+2)^2}\left(\frac{j+3}{t+3}\right)^{2\delta}\right\}+\frac{A_t}{(t+2)^2} \\
&= (1-\delta)\left(\frac{3}{t+3}\right)^{2\delta}\cdot h(\boldsymbol{x}_0)+\underbrace{\left(\frac{1}{(t+3)^{2\delta}}\sum_{j=0}^{t}\frac{A_j(j+3)^{2\delta}}{(j+2)^2}\right)}_{(\star)}
\end{aligned}
\tag{23}
$$

Note that $\sum_{j=0}^{t}\frac{(j+3)^{2\delta}}{(j+2)^2}$ is increasing with respect to δ. Therefore, for $\delta\leq 1/2$,

$$
(\star)\overset{\delta\leq\frac{1}{2}}{\leq}\frac{1}{(t+3)^{2\delta}}\sum_{j=0}^{t}\underbrace{\left(\frac{1}{j+2}+\frac{1}{(j+2)^2}\right)}_{=\frac{j+3}{(j+2)^2}}\overset{(c)}{\leq}\frac{1}{(t+3)^{2\delta}}\left(\ln(t+2)+1-\frac{1}{t+2}\right)\leq\frac{\ln(t+2)+1}{(t+3)^{2\delta}},
\tag{24}
$$

where (c) is again because of the integral bound,

$$
\sum_{j=0}^{t}\left(\frac{1}{j+2}+\frac{1}{(j+2)^2}\right)\leq\int_{-1}^{t}\left(\frac{1}{j+2}+\frac{1}{(j+2)^2}\right)dj=\ln(t+2)-0-\frac{1}{t+2}+1.
\tag{25}
$$

On the other hand, $\sum_{j=0}^{t}\left(\frac{j+3}{t+2}\right)^{2\delta}$ is decreasing with respect to δ. Therefore, for $\delta\geq 1/2$, by similar logic,

$$
(\star)\overset{\delta\geq\frac{1}{2}}{\leq}\frac{1}{t+3}\sum_{j=0}^{t}\left(\frac{1}{j+2}+\frac{1}{(j+2)^2}\right)\leq\frac{\ln(t+2)+1}{t+3}.
\tag{26}
$$

Overall, this gives us

$$
h(\boldsymbol{x}_{t+1})\leq P(t+1,\delta)\triangleq
\begin{cases}
(1-\delta)\left(\frac{3}{t+3}\right)^{2\delta}\cdot h(\boldsymbol{x}_0)+\frac{\ln(t+2)+1}{(t+3)^{2\delta}}\cdot a & \text{if }\delta\in(0,1/2] \\
(1-\delta)\left(\frac{3}{t+3}\right)^{2\delta}\cdot h(\boldsymbol{x}_0)+\frac{\ln(t+2)+1}{t+3}\cdot a & \text{if }\delta\in(1/2,1).
\end{cases}
\tag{27}
$$

However, for the case of $\delta\geq 1/2$, this is not the best we can do. If we include the δ in the error term, we can achieve an $O(A_t/t)$ error rate. In particular, for $\delta\in(1/2,1]$ and $t\geq 1$, we have the following

$$
h(\boldsymbol{x}_t)\leq S(t,\delta)\triangleq\frac{3(1-\delta)h(\boldsymbol{x}_0)+A_t}{(2\delta-1)(t+2)}.
\tag{28}
$$

We prove (28) by the induction. For $t=0$, the initial recurrence relation can be expressed as

$$
\begin{aligned}
h(\boldsymbol{x}_1)\leq (1-\delta)h(\boldsymbol{x}_0)+\frac{A_0}{4}&\leq\frac{3}{3}\cdot\frac{(1-\delta)h(\boldsymbol{x}_0)}{(2\delta-1)}+\frac{A_0}{(2\delta-1)(3)} \\
&\leq\frac{3(1-\delta)h(\boldsymbol{x}_0)+A_0}{(2\delta-1)(t+2)}.
\end{aligned}
$$

Suppose $t = k$, (28) is true, and we consider $t = k + 1$. Then, defining $B_t = \frac{3(1-\delta)h(\boldsymbol{x}_0)+A}{2\delta-1}$,

$$
\begin{aligned}
h(\boldsymbol{x}_{k+1}) &\leq \left(1 - \frac{2\delta}{k+2}\right) h(\boldsymbol{x}_k) + \frac{A_k}{(k+2)^2} \\
&\leq \left(1 - \frac{2\delta}{k+2}\right) \frac{B_k}{k+2} + \frac{A_t}{(k+2)^2} \\
&= \frac{A}{k+3} + \left(\frac{A}{k+2} - \frac{A}{k+3}\right) - \frac{2\delta A}{(k+2)^2} + \frac{a}{(k+2)^2} \\
&= \frac{A}{k+3} + \underbrace{A\left(\frac{1}{(k+2)(k+3)} - \frac{2\delta}{(k+2)^2}\right) + \frac{a}{(k+2)^2}}_{B},
\end{aligned}
$$

where

$$
(2\delta - 1)(k+2)^2(k+3)B = 3(\delta - 1)(2\delta - 1)h(\boldsymbol{x}_0)k + 6(1 - 3\delta)(1 - \delta)h(\boldsymbol{x}_0) - a \leq 0
$$

taking $C = 3(1 - \delta)h(x_0)$ and thus $A = \frac{C}{2\delta-1} + \frac{a}{2\delta-1}$, then

$$
\begin{aligned}
(2\delta - 1)(k+2)^2(k+3)B &= (2\delta - 1)A((k+2) - 2\delta(k+3)) + (2\delta - 1)a(k+3) \\
&= C \underbrace{((1 - 2\delta)(k+3) - 1)}_{\leq 0 \text{ if } \delta \geq 1/2} \underbrace{-a((2\delta - 1)(k+3) + 1) + (2\delta - 1)a(k+3)}_{= -a} \\
&\leq 0
\end{aligned}
$$

provided $\delta \in (1/2, 1]$. Hence, we have the inequality (28).

Overall, this gives, for $t \geq 1$,

$$
h(\boldsymbol{x}_t) \leq \hat{P}(t, \delta) \triangleq \begin{cases} P(t, \delta) & \text{if } \delta \in (0, \frac{1}{2}] \\ \min\{P(t, \delta), S(t, \delta)\} & \text{if } \delta \in (\frac{1}{2}, 1] \end{cases} \tag{29}
$$

which concludes the proof. $\qquad\square$
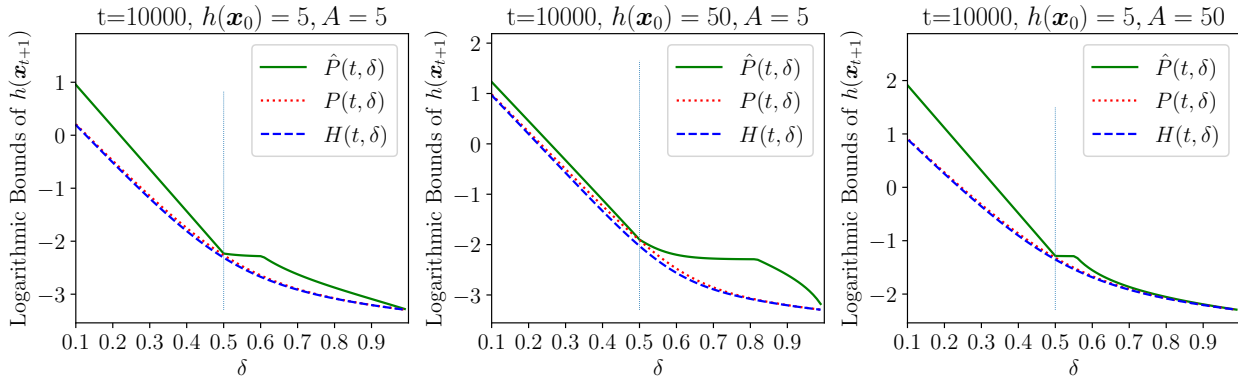
Figure 5. The comparison of three different upper bounds of $h(\boldsymbol{x}_{t+1})$ in Lemma A.4. $H(t, \delta)$ is defined in (21), $P(t, \delta)$ is defined in (27), and $\hat{P}(t, \delta)$ is defined in (29).

Figure 5 compares the rates $P(t, \delta)$ with $\hat{P}(t, \delta)$ in the regime of $\delta \geq 1/2$, as a way of bounding the error term $H(t, \delta)$. We are ready to prove our universal rate of DMO-FW-I.

Theorem 5.1 (Universal rate of DMO-FW-I). For all $t \geq 1$, the primal error of DMO-FW-I satisfies

$$
h(\boldsymbol{x}_t) \leq \begin{cases} \min\{2C\sqrt{s}\|\nabla f(\boldsymbol{x}_t)\|_\infty, P(\delta, 2\delta)\} & \delta \in (0, \frac{1}{2}] \\ \min\left\{P(\delta, 1), \frac{3(1-\delta)h(\boldsymbol{x}_0)+A_t}{(2\delta-1)(t+2)}\right\} & \delta \in (\frac{1}{2}, 1] \end{cases}
$$

where s is maximal allowed sparsity, $P(\delta, \nu)$ is

$$P(\delta, \nu) \triangleq \frac{(1-\delta)9^\delta}{(t+2)^{2\delta}} \cdot h(\boldsymbol{x}_0) + \frac{\ln(t+1)+1}{(t+2)^\nu} \cdot A_t$$

and $A_t \triangleq \max_{i \in [t]} 4(1-\delta)(i+2)|\langle -\boldsymbol{x}_i, \nabla f(\boldsymbol{x}_i)\rangle| + 8LC^2$.

Proof. When DMO provides an exact solution, i.e., $\delta = 1$, then for all $t \geq 0$, $h(\boldsymbol{x}_{t+1})$ of DMO-FW-I is recursively bounded as

$$h(\boldsymbol{x}_{t+1}) \leq \left(1 - \frac{2}{t+2}\right) h(\boldsymbol{x}_t) + \frac{2L(2C)^2}{(t+2)^2},$$

which eventually leads to $h(\boldsymbol{x}_t) \leq 8LC^2/(t+2)$ for all $t \geq 1$. In the rest of the proof, we assume $0 < \delta < 1$. Note that $\boldsymbol{x}_{t+1} = \frac{t\boldsymbol{x}_t}{t+2} + \frac{2\tilde{\boldsymbol{v}}_t}{t+2}$ and $\boldsymbol{x}_1 = \tilde{\boldsymbol{v}}_0$ by Line 7 of Algorithm 1. We rewrite $\boldsymbol{x}_{t+1}$ as a convex combination of $\tilde{\boldsymbol{v}}_t$ as follows

$$\boldsymbol{x}_{t+1} = \frac{t\boldsymbol{x}_t}{t+2} + \frac{2\tilde{\boldsymbol{v}}_t}{t+2} = \sum_{i=0}^{t} \frac{2(i+1)\tilde{\boldsymbol{v}}_i}{(t+1)(t+2)} = \sum_{i=0}^{t} \left\{ \frac{2(i+1)}{(t+1)(t+2)} \cdot \frac{-C\nabla f(\boldsymbol{x}_i)_{S_i}}{\|\nabla f(\boldsymbol{x}_i)_{S_i}\|_2} \right\},$$

where the last equality follows Line 6 of Algorithm 1. Since we assume $\boldsymbol{x}_0 \in \mathcal{D}$ and when $t \geq 1$, the inner product $\langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle$ can be expressed as

$$\langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle = \sum_{i=0}^{t-1} \frac{2(i+1)}{t(t+1)} \left\langle \frac{C\nabla f(\boldsymbol{x}_i)_{S_i}}{\|\nabla f(\boldsymbol{x}_i)_{S_i}\|_2}, \nabla f(\boldsymbol{x}_t) \right\rangle \tag{30}$$

First of all, note that when $\delta \in (0, 1/2]$, one can simply apply convexity, that is,

$$\begin{aligned}
h(\boldsymbol{x}_t) &= f(\boldsymbol{x}_t) - f(\boldsymbol{x}_*) \\
&\leq |\nabla f(\boldsymbol{x}_t)^\top (\boldsymbol{x}_t - \boldsymbol{x}_*)| \\
&\leq |\nabla f(\boldsymbol{x}_t)^\top \boldsymbol{x}_t| + |\nabla f(\boldsymbol{x}_t)^\top \boldsymbol{x}_*| \\
&\leq C\sqrt{s}\|\nabla f(\boldsymbol{x}_t)\|_\infty + C\sqrt{s}\|\nabla f(\boldsymbol{x}_t)\|_\infty \\
&= 2C\sqrt{s}\|\nabla f(\boldsymbol{x}_t)\|_\infty,
\end{aligned}$$

where the last inequality is due to (30) and follows the fact that $\|\boldsymbol{x}_*\|_0 \leq s$.

By L-smooth, we have

$$\begin{aligned}
f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) &\leq \eta_t \underbrace{\nabla f(\boldsymbol{x}_t)^\top \tilde{\boldsymbol{v}}_t}_{\leq \delta \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{v}_t} - \eta_t \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{x}_t + \frac{\eta_t^2 L \|\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t\|_2^2}{2} \\
&\leq -\eta_t \delta(f(\boldsymbol{x}_t) - f^*) - \eta_t(1-\delta)\nabla f(\boldsymbol{x}_t)^\top \boldsymbol{x}_t + \frac{\eta_t^2 L \|\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t\|_2^2}{2},
\end{aligned}$$

where the first inequality is due to $\nabla f(\boldsymbol{x}_t)^\top \tilde{\boldsymbol{v}}_t \leq \delta \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{v}_t$ and second inequality follows from convexity, i.e., $\nabla f(\boldsymbol{x}_t)^\top (\boldsymbol{v}_t - \boldsymbol{x}_t) \leq -(f(\boldsymbol{x}_t) - f^*)$. By setting $\eta_t = 2/(t+2)$, we reach

$$\begin{aligned}
h(\boldsymbol{x}_{t+1}) &\leq \left(1 - \frac{2\delta}{t+2}\right) h(\boldsymbol{x}_t) + \frac{2(1-\delta)}{t+2} \langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle + \frac{2L\|\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t\|^2}{(t+2)^2} \\
&\leq \left(1 - \frac{2\delta}{t+2}\right) h(\boldsymbol{x}_t) + \frac{2(1-\delta)}{t+2} \sum_{i=0}^{t-1} \frac{2(i+1)}{t(t+1)} \left\langle \frac{\nabla f(\boldsymbol{x}_i)_{S_i}}{\|\nabla f(\boldsymbol{x}_i)_{S_i}\|_2}, \nabla f(\boldsymbol{x}_t) \right\rangle + \frac{8LC^2}{(t+2)^2} \\
&= \left(1 - \frac{2\delta}{t+2}\right) h(\boldsymbol{x}_t) + \frac{1}{(t+2)^2} \left\{ \frac{4(1-\delta)(t+2)}{t(t+1)} \left\langle \sum_{i=0}^{t-1} \frac{(i+1)C\nabla f(\boldsymbol{x}_i)_{S_i}}{\|\nabla f(\boldsymbol{x}_i)_{S_i}\|_2}, \nabla f(\boldsymbol{x}_t) \right\rangle + 8LC^2 \right\} \tag{31} \\
&= \left(1 - \frac{2\delta}{t+2}\right) h(\boldsymbol{x}_t) + \frac{1}{(t+2)^2} \left\{ 4(1-\delta)(t+2) \langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle + 8LC^2 \right\}. \tag{32}
\end{aligned}$$

By applying Lemma A.4 with $A_t \triangleq \max_{i \in [t]} 4(1-\delta)(i+2)|\langle -\boldsymbol{x}_i, \nabla f(\boldsymbol{x}_i)\rangle| + 8LC^2$ and combine it with (31), we complete our proof. □

Corollary 5.2 (Practical rate of DMO-FW-I). Suppose $\nu \in (0, 1]$ and $\|\nabla f(\boldsymbol{x})\|_\infty \leq B$, the primal error of DMO-FW-I admits the following practical bound

$$h(\boldsymbol{x}_t) \leq \begin{cases} \mathcal{O}\left(\frac{BC\sqrt{s}}{t^\nu}\right) & \|\nabla f(\boldsymbol{x}_t)\|_\infty \leq \frac{B}{t^\nu} \\ \mathcal{O}\left(\frac{BC\sqrt{s}(1-\delta)}{\delta}\right) & \|\nabla f(\boldsymbol{x})\|_\infty \leq B. \end{cases} \tag{33}$$

where $s \triangleq \max_{S \in \mathbb{M}} |S|$ is the maximal allowed sparsity.

Proof. First of all, when $\|\nabla f(\boldsymbol{x}_t)\|_\infty \leq B/t^\nu$, one can simply apply Thm. 5.1 to show $h(\boldsymbol{x}_t) \leq \mathcal{O}(BC\sqrt{s}/t^\nu)$. In the rest, we show the second case when $\|\nabla f(\boldsymbol{x})\|_\infty \leq B$. Rewrite $\boldsymbol{x}_t$ where $\boldsymbol{x}_1 = \tilde{\boldsymbol{v}}_0$ and

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \frac{t\boldsymbol{x}_t}{t+2} + \frac{2\tilde{\boldsymbol{v}}_t}{t+2} \\ &= \sum_{i=0}^{t} \frac{2(i+1)\tilde{\boldsymbol{v}}_i}{(t+1)(t+2)} \\ &= \sum_{i=0}^{t} \left\{ \frac{2(i+1)}{(t+1)(t+2)} \cdot \frac{-C\nabla f(\boldsymbol{x}_i)_{S_i}}{\|\nabla f(\boldsymbol{x}_i)_{S_i}\|_2} \right\}, \end{aligned}$$

where the last equality follows Line 4 of Algorithm 1. When $t = 0$, we assume the initial point is $\boldsymbol{x}_0 = \boldsymbol{0}$. When $t \geq 1$ the inner product $\langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle$ can be bounded as the following

$$\begin{aligned} \langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle &= C \sum_{i=0}^{t-1} \frac{2(i+1)}{t(t+1)} \left\langle \frac{\nabla f(\boldsymbol{x}_i)_{S_i}}{\|\nabla f(\boldsymbol{x}_i)_{S_i}\|_2}, \nabla f(\boldsymbol{x}_t) \right\rangle \\ &\leq C \sum_{i=0}^{t-1} \frac{2(i+1)}{t(t+1)} \underbrace{\left\| \frac{\nabla f(\boldsymbol{x}_i)_{S_i}}{\|\nabla f(\boldsymbol{x}_i)_{S_i}\|_2} \right\|_2}_{=1} \cdot \underbrace{\|\nabla f(\boldsymbol{x}_t)_{S_i}\|_2}_{=1} \\ &\leq C\sqrt{s} \|\nabla f(\boldsymbol{x}_t)\|_\infty, \end{aligned}$$

where the first inequality follows by the Holder's inequality and the last inequality is the assumption of boundness of $\|\nabla f(\boldsymbol{x}_t)\|_\infty$. By L-smooth of f, we have

$$\begin{aligned} f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) &\leq \eta_t \underbrace{\nabla f(\boldsymbol{x}_t)^\top \tilde{\boldsymbol{v}}_t}_{\leq \delta \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{v}_t} - \eta_t \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{x}_t + \frac{L\eta_t^2}{2} \|\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t\|_2^2 \\ &\leq -\eta_t \delta(f(\boldsymbol{x}_t) - f^*) - (1-\delta)\eta_t \nabla f(\boldsymbol{x}_t)^\top \boldsymbol{x}_t + \frac{4\eta_t^2 LC^2}{2}. \end{aligned}$$

By setting the step size $\eta_t = \frac{2}{t+2}$. This leads to the following

$$\begin{aligned} h(\boldsymbol{x}_{t+1}) &\leq \left(1 - \frac{2\delta}{t+2}\right) h(\boldsymbol{x}_t) + \frac{2(1-\delta)}{(t+2)} \langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{8LC^2}{(t+2)^2} \\ &\leq \left(1 - \frac{2\delta}{t+2}\right) h(\boldsymbol{x}_t) + \frac{2(1-\delta)\sqrt{s}\|\nabla f(\boldsymbol{x}_t)\|_\infty C}{(t+2)} + \frac{8LC^2}{(t+2)^2}. \end{aligned} \tag{34}$$

Notice that the above recurrence (34) can be written as

$$\begin{aligned} h(\boldsymbol{x}_{t+1}) &\leq \prod_{i=0}^{t} \left(1 - \frac{2\delta}{i+2}\right) h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \left\{ \left(\frac{2(1-\delta)\sqrt{s}M_iC}{(i+2)} + \frac{8LC^2}{(i+2)^2}\right) \prod_{j=i+1}^{t} \left(1 - \frac{2\delta}{j+2}\right) \right\} \\ &\leq (1-\delta) \left(\frac{3}{t+3}\right)^{2\delta} \cdot h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \left\{ \left(\frac{2(1-\delta)\sqrt{s}M_iC}{(i+2)} + \frac{8LC^2}{(i+2)^2}\right) \prod_{j=i+1}^{t} \left(1 - \frac{2\delta}{j+2}\right) \right\} \\ &\leq (1-\delta) \left(\frac{3}{t+3}\right)^{2\delta} \cdot h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \left\{ \left(\frac{2(1-\delta)\sqrt{s}M_iC}{(i+2)} + \frac{8LC^2}{(i+2)^2}\right) \frac{(i+3)^{2\delta}}{(t+3)^{2\delta}} \right\}. \end{aligned}$$

Let $M_{\bar{t}} = \max_{i \in [t]} \|\nabla f(\boldsymbol{x}_t)\|_{\infty}$. Therefore, we reach

$$h(\boldsymbol{x}_{t+1}) \leq (1 - \delta) \left(\frac{3}{t+3} \right)^{2\delta} \cdot h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \left\{ \left(\frac{2(1-\delta)\sqrt{s}M_{\bar{t}}C}{(i+2)} + \frac{8LC^2}{(i+2)^2} \right) \frac{(i+3)^{2\delta}}{(t+3)^{2\delta}} \right\}$$

We consider three cases:

1) When $\delta = 1/2$, we have

$$\begin{aligned}
h(\boldsymbol{x}_{t+1}) &\leq \frac{3}{2(t+3)} \cdot h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \frac{\sqrt{s}M_{\bar{t}}C(i+3)}{(i+2)(t+3)} + \sum_{i=0}^{t} \frac{8LC^2(i+3)}{(i+2)^2(t+3)} \\
&\leq \frac{3}{2(t+3)} \cdot h(\boldsymbol{x}_0) + \frac{3\sqrt{s}M_{\bar{t}}C}{2(t+3)}(t+1) + \frac{12LC^2 \ln(t+2)}{(t+3)} \\
&\leq \mathcal{O}(\sqrt{s}M_{\bar{t}}C) \\
&= \mathcal{O}\left(\frac{\sqrt{s}(1-\delta)M_{\bar{t}}C}{\delta} \right).
\end{aligned}$$

2) when $\delta > 1/2$, we have

$$\begin{aligned}
h(\boldsymbol{x}_{t+1}) &\leq (1 - \delta) \left(\frac{3}{t+3} \right)^{2\delta} \cdot h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \left\{ \left(\frac{2(1-\delta)\sqrt{s}M_{\bar{t}}C}{(i+2)} + \frac{8LC^2}{(i+2)^2} \right) \frac{(\frac{3}{2})^{2\delta}(i+2)^{2\delta}}{(t+3)^{2\delta}} \right\} \\
&\leq (1 - \delta) \left(\frac{3}{t+3} \right)^{2\delta} \cdot h(\boldsymbol{x}_0) + \left(\frac{9}{4} \right)^{\delta} \int_{2}^{t+3} \frac{\sqrt{s}(1-\delta)M_{\bar{t}}C}{\delta} \frac{x^{2\delta}}{(t+3)^{2\delta}} + \frac{8LC^2}{(2\delta-1)} \frac{x^{2\delta-1}}{(t+3)^{2\delta}} dx \\
&\leq (1 - \delta) \left(\frac{3}{t+3} \right)^{2\delta} \cdot h(\boldsymbol{x}_0) + \left(\frac{9}{4} \right)^{\delta} \left(\frac{\sqrt{s}(1-\delta)M_{\bar{t}}C}{\delta} + \frac{8LC^2}{(2\delta-1)(t+3)} \right) \\
&\leq \mathcal{O}\left(\frac{\sqrt{s}(1-\delta)M_{\bar{t}}C}{\delta} \right).
\end{aligned}$$

3) when $\delta < 1/2$, we have

$$\begin{aligned}
&\leq (1 - \delta) \left(\frac{3}{t+3} \right)^{2\delta} \cdot h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \left\{ \left(\frac{3\sqrt{s}(1-\delta)M_{\bar{t}}C}{(i+3)} + \frac{18LC^2}{(i+3)^2} \right) \frac{(i+3)^{2\delta}}{(t+3)^{2\delta}} \right\} \\
&\leq (1 - \delta) \left(\frac{3}{t+3} \right)^{2\delta} \cdot h(\boldsymbol{x}_0) + \int_{2}^{t+3} \frac{3\sqrt{s}(1-\delta)M_{\bar{t}}Cx^{2\delta}}{2\delta(t+3)^{2\delta}} dx + \int_{2}^{t+3} \frac{18LC^2 x^{2\delta-1}}{(2\delta-1)(t+3)^{2\delta}} dx \\
&\leq (1 - \delta) \left(\frac{3}{t+3} \right)^{2\delta} \cdot h(\boldsymbol{x}_0) + \frac{3(1-\delta)\sqrt{s}M_{\bar{t}}C}{2\delta} + \frac{9LC^2 4^{\delta}}{(1-2\delta)(t+3)^{2\delta}} \\
&\leq \mathcal{O}\left(\frac{(1-\delta)\sqrt{s}M_{\bar{t}}C}{\delta} \right).
\end{aligned}$$

Combine the above three cases and use the fact that $M_{\bar{t}} \leq B$ by our assumption, we prove the corollary. $\qquad \square$

Remark A.5. The above bound is tight when $\delta = 1$, and it recovers the standard convergence of FW (Jaggi, 2013).

A.4. Proof of Theorem 5.4

Theorem 5.4 (Convergence of DMO-FW-II). Under the same assumptions as in Thm. 5.1, for any $t \geq 1$, the primal error of DMO-FW-II satisfies

$$h(\boldsymbol{x}_t) = f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{8LC^2}{\delta^2(t+2)}, \tag{35}$$

where $\boldsymbol{x}_t \in \mathcal{D}/\delta$ and $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$.

Proof. By L-smoothness of f, we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \eta_t \left\langle \nabla f(\boldsymbol{x}_t), \frac{\tilde{\boldsymbol{v}}_t}{\delta} - \boldsymbol{x}_t \right\rangle + \frac{L\eta_t^2 \|\frac{\tilde{\boldsymbol{v}}_t}{\delta} - \boldsymbol{x}_t\|^2}{2}.$$

Let $\eta_t = \frac{2}{(t+2)}$ and adding $-f(\boldsymbol{x}^*)$ (notice that $\boldsymbol{x}^* \in \arg\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$) on both sides, we have

$$h(\boldsymbol{x}_{t+1}) \leq h(\boldsymbol{x}_t) + \frac{2}{t+2} \left\langle \nabla f(\boldsymbol{x}_t), \frac{\tilde{\boldsymbol{v}}_t}{\delta} - \boldsymbol{x}_t \right\rangle + \frac{8LC^2}{\delta^2(t+2)^2},$$

where the last term follows from the scaling diameter of $\mathcal{D}/\delta$, i.e.

$$\boldsymbol{x}_t, \tilde{\boldsymbol{v}}_t/\delta \in \mathcal{D}/\delta \Rightarrow \|\boldsymbol{x}_t - \tilde{\boldsymbol{v}}_t/\delta\|_2^2 \leq (2C/\delta)^2.$$

Since $\tilde{\boldsymbol{v}}_t$ is a $(\delta, \nabla f(\boldsymbol{x}_t), \mathcal{D})$-DMO, it admits

$$\langle \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{v}}_t \rangle \leq \delta \min_{\boldsymbol{s} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} \rangle < 0.$$

Scaling by $1/\delta$ and then adding $\langle \nabla f(\boldsymbol{x}_t), -\boldsymbol{x}_t \rangle$ both sides, we reach

$$\langle \nabla f(\boldsymbol{x}_t), \tilde{\boldsymbol{v}}_t/\delta - \boldsymbol{x}_t \rangle \leq \min_{\boldsymbol{s} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{x}_t \rangle.$$

We continue to have the following

$$\begin{aligned}
h(\boldsymbol{x}_{t+1}) &\leq h(\boldsymbol{x}_t) + \frac{2}{t+2} \min_{\boldsymbol{s} \in \mathcal{D}} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{s} - \boldsymbol{x}_t \rangle + \frac{2LD^2}{\delta^2(t+2)^2} \\
&\leq h(\boldsymbol{x}_t) + \frac{2}{t+2} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}^* - \boldsymbol{x}_t \rangle + \frac{8LC^2}{\delta^2(t+2)^2} \\
&\leq \left(1 - \frac{2}{t+2}\right) h(\boldsymbol{x}_t) + \frac{8LC^2}{\delta^2(t+2)^2},
\end{aligned}$$

where the last inequality is due to the convexity of f, i.e. $\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}^* - \boldsymbol{x}_t \rangle \leq f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t)$ and $\boldsymbol{x}^* \in \mathcal{D}$. By a similar argument of induction shown in Lemma A.4, we can show the bound of

$$h(\boldsymbol{x}_t) \leq \frac{8LC^2}{\delta^2(t+2)}.$$

$\square$

Remark A.6. The above proof follows a similar proof strategy as in Jaggi (2013). Different from previous one, we show that when $\boldsymbol{x}_t$ extended to $\mathcal{D}/\delta$ with a δ-approximation DMO, we can still have a convergence rate inverse proportional to δ.

A.5. Proof of Theorem 5.10

Notations. Recall that our domain $\mathcal{D} = \text{conv}\{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq C, \text{supp}(\boldsymbol{x}) \in \mathbb{M}\}$. The corresponding set of extreme points is $\mathcal{V} := \{\boldsymbol{x} : \text{supp}(\boldsymbol{x}) \in \mathbb{M}, \|\boldsymbol{x}\|_2 = C\}$. Follow notations of Garber & Wolf (2021), we denote the set of optimal points $\mathcal{X}^* := \arg\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$. Recall the quadratic growth condition as the following.

Definition A.7 (Quadratic Growth Condition). Let f be continuous differentiable. $\mathcal{X}^* \triangleq \arg\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x})$. We say f satisfies quadratic growth condition on $\mathcal{D}$ if there exists a constant $\mu > 0$ such that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \geq \frac{\mu}{2} \|\boldsymbol{x} - [\boldsymbol{x}]_{\mathcal{X}^*}\|_2^2, \tag{36}$$

for all $\boldsymbol{x} \in \mathcal{D}$ where $[\boldsymbol{x}]_{\mathcal{X}^*} \triangleq \underset{\boldsymbol{z} \in \mathcal{X}^*}{\text{argmin}} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2$.

The quadratic growth condition is weaker than restricted strong convex and strongly convex. When f is convex differentiable, it has been observed, it is equivalent to others such as PL condition (Karimi et al., 2016). We start from the following key lemma.

Lemma A.8. *If $\delta = 1$, each iteration of* DMO-ACCFW *admits*

$$\eta_t \langle \boldsymbol{v}_t - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{L\eta_t^2}{2} \|\boldsymbol{v}_t - \boldsymbol{x}_t\|_2^2 \leq \min_{\boldsymbol{v} \in \mathcal{D}} \left\{ \eta_t \langle \boldsymbol{v} - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{L\eta_t^2}{2} \|\boldsymbol{v} - \boldsymbol{x}_t\|_2^2 + \frac{L\eta_t^2}{2} \left(C^2 - \|\boldsymbol{v}\|_2^2 \right) \right\}. \quad (37)$$

Let $\mathcal{X}^* = \underset{\boldsymbol{x} \in \mathcal{D}}{\operatorname{argmin}} f(\boldsymbol{x})$ and suppose optimal points are on the boundary, i.e., $\|\boldsymbol{x}^*\|_2 = C$ for all $\boldsymbol{x}^* \in \mathcal{X}^*$.

$$\eta_t \langle \boldsymbol{v}_t - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{L\eta_t^2}{2} \|\boldsymbol{v}_t - \boldsymbol{x}_t\|_2^2 \leq \eta_t \langle [\boldsymbol{x}_t]_{\mathcal{X}^*} - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{L\eta_t^2}{2} \|[\boldsymbol{x}_t]_{\mathcal{X}^*} - \boldsymbol{x}_t\|_2^2, \quad (38)$$

where $\boldsymbol{v}_t$ is a minimizer (recall that $\bar{\boldsymbol{v}}_t = \boldsymbol{v}_t$ when $\delta = 1$), e.g.,

$$\left\langle \boldsymbol{v}_t, - \left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right) \right\rangle \leq \min_{\boldsymbol{v} \in \mathcal{D}} \left\langle \boldsymbol{v}, - \left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right) \right\rangle$$

and where here $[\boldsymbol{x}]_{\mathcal{X}^*} \triangleq \underset{\boldsymbol{z} \in \mathcal{X}^*}{\operatorname{argmin}} \|\boldsymbol{z} - \boldsymbol{x}\|_2^2$.

Proof. Since $\boldsymbol{v}_t$ follows the above inequality, we have

$$L\eta_t^2 \left\langle \boldsymbol{v}_t, - \left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right) \right\rangle \leq \min_{\boldsymbol{v} \in \mathcal{D}} L\eta_t^2 \left\langle \boldsymbol{v}, - \left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right) \right\rangle$$

By adding both sides $L\eta_t^2(C^2 + \|\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\|_2^2)/2$, we reach

$$\frac{L\eta_t^2}{2} \left\| \boldsymbol{v}_t - \left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right) \right\|_2^2 \leq \min_{\boldsymbol{v} \in \mathcal{D}} \left\{ \frac{L\eta_t^2}{2} \left\| \boldsymbol{v} - \left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right) \right\|_2^2 + \frac{L\eta_t^2}{2} \left(C^2 - \|\boldsymbol{v}\|_2^2 \right) \right\},$$

where $\|\boldsymbol{v}_t\|_2^2 = C^2$. Since $[\boldsymbol{x}_t]_{\mathcal{X}^*} \in \mathcal{X}^* \subseteq \mathcal{D}$ and $\|[\boldsymbol{x}_t]_{\mathcal{X}^*}\|_2^2 = C^2$ by our assumption, we immediately get

$$\frac{L\eta_t^2}{2} \|\boldsymbol{v}_t - \boldsymbol{x}_t\|_2^2 + \eta_t \langle \boldsymbol{v}_t - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{\|\nabla f(\boldsymbol{x}_t)\|_2^2}{2L}$$

$$\leq \min_{\boldsymbol{v} \in \mathcal{D}} \left\{ \frac{L\eta_t^2}{2} \|\boldsymbol{v} - \boldsymbol{x}_t\|_2^2 + \eta_t \langle \boldsymbol{v} - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{\|\nabla f(\boldsymbol{x}_t)\|_2^2}{2L} + \frac{L\eta_t^2}{2} \left(C^2 - \|\boldsymbol{v}\|_2^2 \right) \right\}$$

$$\leq \frac{L\eta_t^2}{2} \|[\boldsymbol{x}_t]_{\mathcal{X}^*} - \boldsymbol{x}_t\|_2^2 + \eta_t \langle [\boldsymbol{x}_t]_{\mathcal{X}^*} - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{\|\nabla f(\boldsymbol{x}_t)\|_2^2}{2L}.$$

We immediately get

$$\frac{L\eta_t^2}{2} \|\boldsymbol{v}_t - \boldsymbol{x}_t\|_2^2 + \eta_t \langle \boldsymbol{v}_t - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{\|\nabla f(\boldsymbol{x}_t)\|_2^2}{2L}$$

$$\leq \min_{\boldsymbol{v} \in \mathcal{D}} \left\{ \frac{L\eta_t^2}{2} \|\boldsymbol{v} - \boldsymbol{x}_t\|_2^2 + \eta_t \langle \boldsymbol{v} - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{\|\nabla f(\boldsymbol{x}_t)\|_2^2}{2L} + \frac{L\eta_t^2}{2} \left(C^2 - \|\boldsymbol{v}\|_2^2 \right) \right\}.$$

Simplifying the above inequality, we have the general bound (37). When $\|\boldsymbol{x}^*\|_2^2 = C^2$, simplifying the above inequality, we have

$$\eta_t \langle \boldsymbol{v}_t - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{L\eta_t^2}{2} \|\boldsymbol{v}_t - \boldsymbol{x}_t\|_2^2 \leq \eta_t \langle [\boldsymbol{x}_t]_{\mathcal{X}^*} - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle + \frac{L\eta_t^2}{2} \|[\boldsymbol{x}_t]_{\mathcal{X}^*} - \boldsymbol{x}_t\|_2^2.$$

$\square$

Theorem A.9. *Let f be convex and satisfies quadratic growth condition. Assume that for all $\boldsymbol{x}^*$, $\|\boldsymbol{x}^*\|_2^2 = C$, then if $\delta = 1$ and $\eta_t = 2/(t+2)$,* DMO-ACCFW-I *has the following convergence rate*

$$h(\boldsymbol{x}_t) \leq \frac{4e^{4L/\mu}}{(t+2)^2} h(\boldsymbol{x}_0) \quad (39)$$

for all $t \geq 1$.

Proof. By the L-smoothness of f, we have

$$
\begin{aligned}
f(\boldsymbol{x}_{t+1}) &\leq f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^\top (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_2^2 \\
&= f(\boldsymbol{x}_t) + \eta_t \nabla f(\boldsymbol{x}_t)^\top (\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t) + \frac{L\eta_t^2}{2}\|\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t\|_2^2 \\
&\overset{(a)}{=} f(\boldsymbol{x}_t) + \eta_t \nabla f(\boldsymbol{x}_t)^\top (\boldsymbol{v}_t - \boldsymbol{x}_t) + \frac{L\eta_t^2}{2}\|\boldsymbol{v}_t - \boldsymbol{x}_t\|_2^2 \\
&\overset{(b)}{\leq} f(\boldsymbol{x}_t) + \eta_t \nabla f(\boldsymbol{x}_t)^\top ([\boldsymbol{x}_t]_{\mathcal{X}^*} - \boldsymbol{x}_t) + \frac{L\eta_t^2}{2}\|[\boldsymbol{x}_t]_{\mathcal{X}^*} - \boldsymbol{x}_t\|_2^2 \\
&\overset{(c)}{\leq} f(\boldsymbol{x}_t) - \eta_t \left(f(\boldsymbol{x}_t) - f([\boldsymbol{x}_t]_{\mathcal{X}^*})\right) + \frac{L\eta_t^2}{2}\|[\boldsymbol{x}_t]_{\mathcal{X}^*} - \boldsymbol{x}_t\|_2^2,
\end{aligned}
$$

where (a) is due to $\delta = 1$, (b) follows from Lemma A.8, and (c) uses the convexity of f. By setting the step size $\eta_t = 2/(t+2)$ and add $-f(\boldsymbol{x}^*)$ on both sides, we reach at

$$
h(\boldsymbol{x}_{t+1}) \leq \left(1 - \frac{2}{t+2}\right) h(\boldsymbol{x}_t) + \frac{2L\|\boldsymbol{x}_t - [\boldsymbol{x}_t]_{\mathcal{X}^*}\|_2^2}{(t+2)^2}.
$$

By the assumption of quadratic growth property property of f, we have

$$
f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \geq \frac{\mu}{2}\|\boldsymbol{x}_t - [\boldsymbol{x}]_{\mathcal{X}^*}\|_2^2.
$$

Specifically, for all $t \geq 0$

$$
\|\boldsymbol{x}_t - [\boldsymbol{x}]_{\mathcal{X}^*}\|_2^2 \leq \frac{2}{\mu}\left(f(\boldsymbol{x}_t) - f([\boldsymbol{x}]_{\mathcal{X}^*})\right) = \frac{2}{\mu}h(\boldsymbol{x}_t).
$$

We have

$$
\begin{aligned}
h(\boldsymbol{x}_{t+1}) &\leq \left(1 - \frac{2}{t+2}\right) h(\boldsymbol{x}_t) + \frac{2L\|\boldsymbol{x}_t - [\boldsymbol{x}]_{\mathcal{X}^*}\|_2^2}{(t+2)^2} \\
&\leq \left(1 - \frac{2}{t+2}\right) h(\boldsymbol{x}_t) + \frac{4Lh(\boldsymbol{x}_t)}{\mu(t+2)^2} \\
&= \left(1 - \frac{2}{t+2} + \frac{4L}{\mu(t+2)^2}\right) h(\boldsymbol{x}_t) \\
&= \prod_{i=0}^{t} \left(1 - \frac{2}{i+2} + \frac{4L}{\mu(i+2)^2}\right) h(\boldsymbol{x}_0) \\
&\leq \exp\left\{\sum_{i=0}^{t}\left(-\frac{2}{i+2} + \frac{4L}{\mu(i+2)^2}\right)\right\} \cdot h(\boldsymbol{x}_0) \\
&\leq \frac{4e^{4L/\mu}}{(t+3)^2} \cdot h(\boldsymbol{x}_0).
\end{aligned}
$$

$\square$

Theorem A.10. *Let f be convex and satisfies quadratic growth condition. Assume that $D_* = \min_{\boldsymbol{x}^* \in \mathcal{X}^*}\|\boldsymbol{x}^*\|_2$, then if $\delta = 1$ and $\eta_t = 2/(t+2)$, DMO-ACCFW-I has the following convergence rate*

$$
h(\boldsymbol{x}_t) \leq \min\left\{\frac{3Le^{2L/\mu}(C^2 - D_*^2)}{t+2} + \frac{4e^{4L/\mu}h(\boldsymbol{x}_0)}{(t+2)^2}, \frac{2L(5C^2 - D_*^2)}{(t+2)}\right\} \tag{40}
$$

for all $t \geq 1$.

Proof. By setting $\delta = 1$, by L-smooth, we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \eta_t \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t - \boldsymbol{x}_t \rangle + \frac{L\eta_t^2}{2} \|\boldsymbol{v}_t - \boldsymbol{x}_t\|_2^2$$

$$\leq f(\boldsymbol{x}_t) + \min_{\boldsymbol{v} \in \mathcal{D}} \left\{ \eta_t \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} - \boldsymbol{x}_t \rangle + \frac{L\eta_t^2}{2} \|\boldsymbol{v} - \boldsymbol{x}_t\|_2^2 + \frac{L\eta_t^2}{2} (C^2 - \|\boldsymbol{v}\|_2^2) \right\}$$

$$\leq f(\boldsymbol{x}_t) + \eta_t \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}^* - \boldsymbol{x}_t \rangle + \frac{L\eta_t^2}{2} \|\boldsymbol{x}^* - \boldsymbol{x}_t\|_2^2 + \frac{L\eta_t^2}{2} (C^2 - \|\boldsymbol{x}^*\|_2^2).$$

That is,

$$h(\boldsymbol{x}_{t+1}) \leq (1 - \eta_t) h(\boldsymbol{x}_t) + \frac{L\eta_t^2}{2} \left(\|\boldsymbol{x}^* - \boldsymbol{x}_t\|_2^2 + C^2 - \|\boldsymbol{x}^*\|_2^2 \right). \tag{41}$$

By the quadratic growth condition, we have

$$h(\boldsymbol{x}_{t+1}) \leq (1 - \eta_t) h(\boldsymbol{x}_t) + \frac{L\eta_t^2 h(\boldsymbol{x}_t)}{\mu} + \frac{L\eta_t^2}{2} \left(\|\boldsymbol{v}_t\|_2^2 - \|\boldsymbol{x}^*\|_2^2 \right).$$

Let $\eta_t = \frac{2}{(t+2)}$ and $D_* = \|\boldsymbol{x}^*\|_2$, we have

$$h(\boldsymbol{x}_{t+1}) \leq \left(1 - \frac{2}{t+2} + \frac{4L}{\mu(t+2)^2} \right) h(\boldsymbol{x}_t) + \frac{2L(C^2 - D_*^2)}{(t+2)^2}$$

$$\leq \prod_{i=0}^{t} \left(1 - \frac{2}{i+2} + \frac{4L}{\mu(i+2)^2} \right) h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \left\{ \frac{2L(C^2 - D_*^2)}{(i+2)^2} \cdot \prod_{j=i+1}^{t} \left(1 - \frac{2}{j+2} + \frac{4L}{\mu(j+2)^2} \right) \right\}$$

$$\leq \frac{4e^{4L/\mu}}{(t+3)^2} h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \left\{ \frac{2L(C^2 - D_*^2)}{(i+2)^2} \cdot \prod_{j=i+1}^{t} \left(1 - \frac{2}{j+2} + \frac{4L}{\mu(j+2)^2} \right) \right\}$$

$$\leq \frac{4e^{4L/\mu}}{(t+3)^2} h(\boldsymbol{x}_0) + \sum_{i=0}^{t} \left\{ \frac{2L(C^2 - D_*^2)}{(i+2)^2} \frac{(i+3)^2}{(t+3)^2} \cdot \exp\left\{ \frac{4L}{\mu} \left(\frac{1}{i+2} - \frac{1}{t+2} \right) \right\} \right\}$$

$$\leq \frac{4e^{4L/\mu}}{(t+3)^2} h(\boldsymbol{x}_0) + e^{2L/\mu} \sum_{i=0}^{t} \frac{2L(C^2 - D_*^2)(i+3)^2}{(i+2)^2(t+3)^2}$$

$$\leq \frac{4e^{4L/\mu}}{(t+3)^2} h(\boldsymbol{x}_0) + 2Le^{2L/\mu}(C^2 - D_*^2) \frac{t+1+2\ln(t+2)+1}{(t+3)^2}$$

$$\leq \frac{4e^{4L/\mu}}{(t+3)^2} h(\boldsymbol{x}_0) + \frac{3Le^{2L/\mu}(C^2 - D_*^2)}{t+3}.$$

On the other hand, from (41), we have

$$h(\boldsymbol{x}_{t+1}) \leq (1 - \frac{2}{t+2}) h(\boldsymbol{x}_t) + \frac{2L(5C^2 - D_*^2)}{(t+2)^2}. \tag{42}$$

Hence, for all $t \geq 1$, we have

$$h(\boldsymbol{x}_t) \leq \frac{2L(5C^2 - D_*^2)}{(t+2)}.$$

Combine these two bounds, we prove the theorem. $\qquad\square$

Theorem 5.9 (Practical rate of DMO-ACCFW-I). When $\delta \in (0, 1)$ and $\|\nabla f(\boldsymbol{x})\|_\infty \leq B$, then DMO-ACCFW-I admits

$$h(\boldsymbol{x}_t) \leq \mathcal{O}\left(\sqrt{s} BC(1 - \delta)/\delta \right).$$

Moreover, when $\|\nabla f(\boldsymbol{x})\|_\infty \leq B/t^\nu$ with $\nu \in (0, 1]$, we have

$$h(\boldsymbol{x}_t) \leq \mathcal{O}\left(\sqrt{s} BC/t^\nu \right).$$

Proof. Notice that, at t-th iteration, by the property of $(\delta, -(\boldsymbol{x}_t - \nabla f(\boldsymbol{x}_t)/(L\eta_t)), \mathcal{D})$-DMO operator, we have the following

$$\left\langle \tilde{\boldsymbol{v}}_t, -\left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right)\right\rangle \leq \delta \cdot \min_{\boldsymbol{v} \in \mathcal{D}} \left\langle \boldsymbol{v}, -\left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right)\right\rangle,$$

where the inequality is due to $(\delta, -\boldsymbol{x}_t + \nabla f(\boldsymbol{x}_t)/(L\eta_t), \mathcal{D})$-DMO oracle in Line 4 of Algorithm 1. We continue to have

$$L\eta_t^2 \left\langle \tilde{\boldsymbol{v}}_t, -\left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right)\right\rangle \leq \delta \cdot \min_{\boldsymbol{v} \in \mathcal{D}} L\eta_t^2 \left\langle \boldsymbol{v}, -\left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right)\right\rangle$$

$$\Leftrightarrow \quad \frac{C^2 L\eta_t^2}{2} + L\eta_t^2 \left\langle \tilde{\boldsymbol{v}}_t, -\left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right)\right\rangle + \frac{L\eta_t^2}{2}\left\|\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right\|_2^2$$

$$\leq \delta \cdot \min_{\boldsymbol{v} \in \mathcal{D}} \left\{\frac{L\eta_t^2 \|\boldsymbol{v}\|_2^2}{2} + L\eta_t^2 \left\langle \boldsymbol{v}, -\left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right)\right\rangle + \frac{L\eta_t^2}{2}\left\|\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right\|_2^2 - \frac{L\eta_t^2\|\boldsymbol{v}\|_2^2}{2}\right\}$$

$$+ \frac{C^2 L\eta_t^2}{2} + \frac{L\eta_t^2(1-\delta)}{2}\left\|\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right\|_2^2.$$

That is,

$$\frac{L\eta_t^2}{2}\|\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t\|_2^2 + \eta_t\langle \tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle + \frac{\|\nabla f(\boldsymbol{x}_t)\|_2^2}{2L}$$

$$\leq \delta \cdot \min_{\boldsymbol{v} \in \mathcal{D}} \left\{\frac{L\eta_t^2}{2}\|\boldsymbol{v} - \boldsymbol{x}_t\|_2^2 + \eta_t\langle \boldsymbol{v} - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle + \frac{\|\nabla f(\boldsymbol{x}_t)\|_2^2}{2L} - \frac{L\eta_t^2\|\boldsymbol{v}\|_2^2}{2}\right\}$$

$$+ \frac{C^2 L\eta_t^2}{2} + \frac{L\eta_t^2(1-\delta)}{2}\left\|\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right\|_2^2$$

$$\Leftrightarrow \quad \frac{L\eta_t^2}{2}\|\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t\|_2^2 + \eta_t\langle \tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle$$

$$\leq \delta \cdot \min_{\boldsymbol{v} \in \mathcal{D}} \left\{\frac{L\eta_t^2}{2}\|\boldsymbol{v} - \boldsymbol{x}_t\|_2^2 + \eta_t\langle \boldsymbol{v} - \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle - \frac{L\eta_t^2\|\boldsymbol{v}\|_2^2}{2}\right\}$$

$$+ \frac{C^2 L\eta_t^2}{2} + (1-\delta)\left(\frac{L\eta_t^2}{2}\left\|\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right\|_2^2 - \frac{\|\nabla f(\boldsymbol{x}_t)\|_2^2}{2L}\right) \tag{43}$$

Define $Q_t = (1-\delta)\left(\frac{L\eta_t^2}{2}\left\|\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right\|_2^2 - \frac{\|\nabla f(\boldsymbol{x}_t)\|_2^2}{2L}\right)$. By the L-smooth of f, we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^\top (\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_2^2$$

$$= f(\boldsymbol{x}_t) + \eta_t \nabla f(\boldsymbol{x}_t)^\top (\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t) + \frac{L\eta_t^2}{2}\|\tilde{\boldsymbol{v}}_t - \boldsymbol{x}_t\|_2^2$$

$$\overset{(a)}{\leq} f(\boldsymbol{x}_t) + \delta \cdot \min_{\boldsymbol{v} \in \mathcal{D}} \left\{\eta_t \nabla f(\boldsymbol{x}_t)^\top (\boldsymbol{v} - \boldsymbol{x}_t) + \frac{L\eta_t^2}{2}\|\boldsymbol{v} - \boldsymbol{x}_t\|_2^2 - \frac{\delta L\eta_t^2\|\boldsymbol{v}\|_2^2}{2}\right\} + \frac{C^2 L\eta_t^2}{2} + Q_t$$

$$\overset{(b)}{\leq} f(\boldsymbol{x}_t) + \delta \cdot \eta_t \nabla f(\boldsymbol{x}_t)^\top (\boldsymbol{x}^* - \boldsymbol{x}_t) + \delta \cdot \frac{L\eta_t^2}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_t\|_2^2 + \frac{C^2 L\eta_t^2}{2} - \frac{\delta L\eta_t^2\|\boldsymbol{x}^*\|_2^2}{2} + Q_t$$

$$\overset{(c)}{\leq} f(\boldsymbol{x}_t) - \delta \cdot \eta_t (f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) + \delta \cdot \frac{L\eta_t^2}{2}\|\boldsymbol{x}^* - \boldsymbol{x}_t\|_2^2 + \frac{C^2 L\eta_t^2}{2} - \frac{\delta L\eta_t^2\|\boldsymbol{x}^*\|_2^2}{2} + Q_t$$

$$\Leftrightarrow h(\boldsymbol{x}_{t+1}) \leq \left(1 - \delta\eta_t + \frac{\delta L\eta_t^2}{\mu}\right) h(\boldsymbol{x}_t) + \frac{L\eta_t^2(C^2 - \delta D_*^2)}{2} + Q_t,$$

where (a) is due to (43), (b) follows by $\boldsymbol{x}^* \in \mathcal{D}$, and (c) uses the convexity property. The last inequality is due to the quadratic growth condition and letting step size $\eta_t = 2/(t+2)$. Hence, we reach the following recurrence relation

$$h(\boldsymbol{x}_{t+1}) \leq \left(1 - \frac{2\delta}{t+2} + \frac{4\delta L}{\mu(t+2)^2}\right) h(\boldsymbol{x}_t) + \frac{2L(C^2 - \delta D_*^2)}{(t+2)^2} + (1-\delta)\left(\frac{2L\|\boldsymbol{x}_t\|_2^2}{(t+2)^2} + \frac{2\langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle}{(t+2)}\right)$$

Notice that

$$\langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle = \sum_{i=0}^{t-1}\frac{2(i+1)}{t(t+1)}\left\langle \frac{\boldsymbol{z}_{t S_i}}{\|\boldsymbol{z}_{t S_i}\|_2}, \nabla f(\boldsymbol{x}_t)\right\rangle,$$

where $\boldsymbol{z}_t = -C\left(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}\right)$. We continue to have

$$\langle -\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\rangle = \sum_{i=0}^{t-1}\frac{2(i+1)}{t(t+1)}\left\langle \frac{\boldsymbol{z}_{t S_i}}{\|\boldsymbol{z}_{t S_i}\|_2}, \nabla f(\boldsymbol{x}_t)\right\rangle$$

$$\leq \sum_{i=0}^{t-1}\frac{2(i+1)}{t(t+1)}\|\nabla f(\boldsymbol{x}_t)_{S_i}\|_2 \leq \sqrt{s}C\|\nabla f(\boldsymbol{x}_t)\|_\infty,$$

where s is the maximum number of nonzeros allowed in the sparsity pattern defined by $\mathbb{M}$. We have

$$h(\boldsymbol{x}_{t+1}) \leq \left(1 - \frac{2\delta}{t+2} + \frac{4\delta L}{\mu(t+2)^2}\right)h(\boldsymbol{x}_t) + \frac{2L(C^2 - \delta D_*^2)}{(t+2)^2} + (1-\delta)\left(\frac{2LC^2}{(t+2)^2} + \frac{2\sqrt{s}CM_t}{(t+2)}\right).$$

Again, let $T = 2L(C^2 - \delta D_*^2) + 2LC^2(1-\delta)$ using the recurrence relation, we have

$$h(\boldsymbol{x}_{t+1}) \leq \prod_{i=0}^{t}\left(1 - \frac{2\delta}{i+2} + \frac{4\delta L}{\mu(i+2)^2}\right)h(\boldsymbol{x}_0)$$

$$+ \sum_{i=0}^{t}\left\{\left(\frac{T}{(i+2)^2} + \frac{2\sqrt{s}CM_i(1-\delta)}{(i+2)}\right)\cdot \prod_{j=i+1}^{t}\left(1 - \frac{2\delta}{j+2} + \frac{4\delta L}{\mu(j+2)^2}\right)\right\}$$

$$\leq \frac{4e^{4\delta L/\mu}}{(t+3)^{2\delta}}h(\boldsymbol{x}_0) + \sum_{i=0}^{t}\left\{\left(\frac{T}{(i+2)^2} + \frac{2\sqrt{s}CM_i(1-\delta)}{(i+2)}\right)\cdot \prod_{j=i+1}^{t}\left(1 - \frac{2\delta}{j+2} + \frac{4\delta L}{\mu(j+2)^2}\right)\right\}$$

$$\leq \frac{4e^{4\delta L/\mu}}{(t+3)^{2\delta}}h(\boldsymbol{x}_0) + \sum_{i=0}^{t}\left\{\left(\frac{T}{(i+2)^2} + \frac{2\sqrt{s}CM_i(1-\delta)}{(i+2)}\right)\left(\frac{(i+3)^{2\delta}}{(t+3)^{2\delta}}\cdot \exp\left\{\frac{4\delta L}{\mu}\left(\frac{1}{i+2} - \frac{1}{t+2}\right)\right\}\right)\right\}$$

$$\leq \frac{4e^{4\delta L/\mu}}{(t+3)^{2\delta}}h(\boldsymbol{x}_0) + e^{2\delta L/\mu}\sum_{i=0}^{t}\left\{\frac{T(i+3)^{2\delta}}{(i+2)^2(t+3)^{2\delta}} + \frac{2\sqrt{s}CM_i(1-\delta)(i+3)^{2\delta}}{(i+2)(t+3)^{2\delta}}\right\}$$

$$\leq \frac{4e^{4\delta L/\mu}}{(t+3)^{2\delta}}h(\boldsymbol{x}_0) + e^{2\delta L/\mu}\mathcal{O}\left(\frac{2\sqrt{s}CM_i(1-\delta)}{2\delta}\right)$$

$$\leq \mathcal{O}\left(\frac{\sqrt{s}CM_{\bar{i}}(1-\delta)}{\delta}\right) = \leq \mathcal{O}\left(\frac{\sqrt{s}BC(1-\delta)}{\delta}\right),$$

where $M_{\bar{i}} = \max_{i\in[t]}\|\nabla f(\boldsymbol{x}_t)\|_\infty$. The above analysis indicates that the worse case bound is $\mathcal{O}(\sqrt{s}M_{\max})$. However, when $M_t \leq B/t^\mu$ with $\mu \in (0,1]$, we have an optimistic bound

$$h(\boldsymbol{x}_t) \leq \mathcal{O}\left(\frac{\sqrt{s}BC}{t^\mu}\right). \tag{44}$$

Combine two above bounds, we finish the proof. $\square$

Theorem 5.10 (Convergence of DMO-ACCFW-II). When $\delta \in (0,1)$, then DMO-ACCFW-II finds $\boldsymbol{x}_t \in \mathcal{D}/\delta$ and admits

$$h(\boldsymbol{x}_t) \leq \frac{4e^{4L/\mu}}{(t+3)^2}h(\boldsymbol{x}_0) + \frac{28L^2(C^2/\delta^2 - D_*^2)}{5\mu(t+3)}. \tag{45}$$

Proof. By L-smooth, we have

$$f(\boldsymbol{x}_{t+1}) \leq f(\boldsymbol{x}_t) + \nabla f(\boldsymbol{x}_t)^\top(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t) + \frac{L}{2}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_2^2$$

$$\leq f(\boldsymbol{x}_t) + \eta_t\nabla f(\boldsymbol{x}_t)^\top\left(\frac{\tilde{\boldsymbol{v}}_t}{\delta} - \boldsymbol{x}_t\right) + \frac{L\eta_t^2}{2}\left\|\frac{\tilde{\boldsymbol{v}}_t}{\delta} - \boldsymbol{x}_t\right\|_2^2 \tag{46}$$

By the $(\delta, -(\boldsymbol{x}_t - \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t}))$-DMO, we have

$$\left\langle \tilde{\boldsymbol{v}}_t, -\boldsymbol{x}_t + \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right\rangle \le \delta \cdot \min_{\boldsymbol{v} \in \mathcal{D}} \left\langle \boldsymbol{v}, -\boldsymbol{x}_t + \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right\rangle$$

We have the following equivalent form

$$\frac{L\eta_t^2}{2} \left\| \frac{\tilde{\boldsymbol{v}}_t}{\delta} - \boldsymbol{x}_t + \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right\|_2^2 \le \min_{\boldsymbol{v} \in \mathcal{D}} \left\{ \frac{L\eta_t^2}{2} \left\| \boldsymbol{v} - \boldsymbol{x}_t + \frac{\nabla f(\boldsymbol{x}_t)}{L\eta_t} \right\|_2^2 + \frac{L\eta_t^2 C^2}{2\delta^2} - \frac{L\eta_t^2}{2} \|\boldsymbol{v}\|_2^2 \right\},$$

which could be simplified as the following

$$\eta_t \left\langle \nabla f(\boldsymbol{x}_t), \frac{\tilde{\boldsymbol{v}}_t}{\delta} - \boldsymbol{x}_t \right\rangle + \frac{L\eta_t^2}{2} \left\| \frac{\tilde{\boldsymbol{v}}_t}{\delta} - \boldsymbol{x}_t \right\|_2^2 \le \min_{\boldsymbol{v} \in \mathcal{D}} \left\{ \eta_t \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} - \boldsymbol{x}_t \rangle + \frac{L\eta_t^2}{2} \|\boldsymbol{v} - \boldsymbol{x}_t\|_2^2 + \frac{L\eta_t^2 C^2}{2\delta^2} - \frac{L\eta_t^2}{2} \|\boldsymbol{v}\|_2^2 \right\}.$$

Therefore, we continue to have

$$f(\boldsymbol{x}_{t+1}) \le f(\boldsymbol{x}_t) + \min_{\boldsymbol{v} \in \mathcal{D}} \left\{ \eta_t \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v} - \boldsymbol{x}_t \rangle + \frac{L\eta_t^2}{2} \|\boldsymbol{v} - \boldsymbol{x}_t\|_2^2 + \frac{L\eta_t^2 C^2}{2\delta^2} - \frac{L\eta_t^2}{2} \|\boldsymbol{v}\|_2^2 \right\}$$

$$\le f(\boldsymbol{x}_t) - \eta_t(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) + \frac{L\eta_t^2}{2} \|\boldsymbol{x}^* - \boldsymbol{x}_t\|_2^2 + \frac{L\eta_t^2 C^2}{2\delta^2} - \frac{L\eta_t^2}{2} \|\boldsymbol{x}^*\|_2^2.$$

Set $\eta_t = 2/(t+2)$ and use the μ-quadratic growth condition, we have

$$h(\boldsymbol{x}_{t+1}) \le \left(1 - \frac{2}{t+2} + \frac{2L}{\mu(t+2)^2} \right) h(\boldsymbol{x}_t) + \frac{2L}{(t+2)^2} \left(\frac{C^2}{\delta^2} - D_*^2 \right).$$

Letting $B = 2L(\frac{C^2}{\delta^2} - D_*^2)$ and applying the above recurrence relation, we continue to have

$$h(\boldsymbol{x}_{t+1}) \le \prod_{i=0}^t \left(1 - \frac{2}{i+2} + \frac{4L}{\mu(i+2)^2} \right) h(\boldsymbol{x}_0) + \sum_{i=0}^t \left\{ \frac{B}{(i+2)^2} \cdot \prod_{j=i+1}^t \left(1 - \frac{2}{j+2} + \frac{4L}{\mu(j+2)^2} \right) \right\}$$

$$\le \frac{4e^{4L/\mu}}{(t+3)^2} h(\boldsymbol{x}_0) + \sum_{i=0}^t \frac{2LB(i+3)^2}{\mu(i+2)^2(t+3)^2}$$

$$\le \frac{4e^{4L/\mu}}{(t+3)^2} h(\boldsymbol{x}_0) + \frac{2LB(t+2+\ln(t+2))}{\mu(t+3)^2}$$

$$\le \frac{4e^{4L/\mu}}{(t+3)^2} h(\boldsymbol{x}_0) + \frac{14LB}{5\mu(t+3)},$$

where we use the fact that $t + 2 + \ln(t+2) \le 7(t+3)/5$. Replacing B by $2L\left(\frac{C^2}{\delta^2} - D_*^2 \right)$, we prove the theorem. $\qquad\square$

A.6. High probability parameter estimation

Corollary A.11 (High probability parameter estimation). *Let $h(\boldsymbol{x}_t)$ be primal error for* DMO-FW *or* DMO-ACCFW. *The estimation error of the graph-structured linear sensing problem admits*

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t\|_2 \le \sqrt{\frac{2\sqrt{s}C\|\nabla f(\tilde{\boldsymbol{x}}^*)\|_\infty}{\mu}} + \sqrt{\frac{2h(\boldsymbol{x}_t)}{\mu}}, \tag{47}$$

Moreover, with large enough n, there exists an universal constant c such that with high probability:

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t\|_2 \le \left(\sqrt{\frac{4\sigma C\sqrt{s\log d/n}}{\frac{1}{2}} - \frac{cs\log d}{n}} + \sqrt{\frac{2h(\boldsymbol{x}_t)}{\frac{1}{2} - \frac{cs\log d}{n}}} \right).$$

Proof. Let $\boldsymbol{x}^*$ be the optimal solution of the graph-structured linear regression problem, that is,

$$\boldsymbol{x}^* \in \underset{\boldsymbol{x} \in \mathcal{D}(C, \mathbb{M})}{\operatorname{argmin}} f(\boldsymbol{x}) \triangleq \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2, \tag{48}$$

where the measurements are obtained by $\boldsymbol{y} = \boldsymbol{A}\tilde{\boldsymbol{x}}^* + \boldsymbol{e}$ where each $a_{ij} \sim \mathcal{N}(0, 1)/\sqrt{n}$ and $e_i \sim \mathcal{N}(0, \sigma^2)$ for all $i \in [d]$. Notice that, let ϵ_t be the established bound of $h(\boldsymbol{x}_t)$ for DMO-FW and DMO-ACCFW. By μ-quadratic growth condition, we have

$$f(\tilde{\boldsymbol{x}}^*) \le f(\boldsymbol{x}_t) - \langle \nabla f(\tilde{\boldsymbol{x}}^*), \boldsymbol{x}_t - \tilde{\boldsymbol{x}}^* \rangle - \frac{\mu}{2} \|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t\|_2^2.$$

That is, we have the parameter estimation error

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t\|_2 \le \sqrt{\frac{2(\langle \nabla f(\tilde{\boldsymbol{x}}^*), \tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t \rangle + f(\boldsymbol{x}_t) - f(\tilde{\boldsymbol{x}}^*))}{\mu}}$$

$$\le \sqrt{\frac{2(\langle \nabla f(\tilde{\boldsymbol{x}}^*), \tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t \rangle + f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*))}{\mu}}$$

$$\le \sqrt{\frac{2|\langle \nabla f(\tilde{\boldsymbol{x}}^*), \tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t \rangle|}{\mu}} + \sqrt{\frac{2h(\boldsymbol{x}_t)}{\mu}}.$$

Notice that $|\langle \nabla f(\tilde{\boldsymbol{x}}^*), \tilde{\boldsymbol{x}}^* \rangle| \le \|\nabla f(\tilde{\boldsymbol{x}}^*)\|_\infty \cdot \|\tilde{\boldsymbol{x}}^*\|_1 \le \sqrt{s}C\|\nabla f(\tilde{\boldsymbol{x}}^*)\|_\infty$. For the term $|\langle \nabla f(\tilde{\boldsymbol{x}}^*), \boldsymbol{x}_t|$, we have

$$|\langle \nabla f(\tilde{\boldsymbol{x}}^*), \boldsymbol{x}_t \rangle| = C \sum_{i=0}^{t-1} \frac{2(i+1)}{t(t+1)} \left| \left\langle \frac{((-\boldsymbol{x}_i + \nabla f(\boldsymbol{x}_i)/(L\eta_i))_{S_i}}{\|(-\boldsymbol{x}_i + \nabla f(\boldsymbol{x}_i))_{S_i}\|_2}, \nabla f(\tilde{\boldsymbol{x}}^*) \right\rangle \right|$$

$$\le C \sum_{i=0}^{t-1} \frac{2(i+1)}{t(t+1)} \left\| \frac{((-\boldsymbol{x}_i + \nabla f(\boldsymbol{x}_i)/(L\eta_i))_{S_i}}{\|(-\boldsymbol{x}_i + \nabla f(\boldsymbol{x}_i))_{S_i}\|_2} \right\|_2 \cdot \|\nabla f(\tilde{\boldsymbol{x}}^*)\|_2$$

$$\le C\sqrt{s}\|\nabla f(\tilde{\boldsymbol{x}}^*)\|_\infty \cdot \sum_{i=0}^{t-1} \frac{2(i+1)}{t(t+1)} 1$$

$$= C\sqrt{s}\|\nabla f(\tilde{\boldsymbol{x}}^*)\|_\infty,$$

where s is the maximal allowed sparsity in $\mathbb{M}$, that is, $s = \max_{S \in \mathbb{M}} |S|$. Hence, we have

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t\|_2 \le \sqrt{\frac{2\sqrt{s}C\|\nabla f(\tilde{\boldsymbol{x}}^*)\|_\infty}{\mu}} + \sqrt{\frac{2h(\boldsymbol{x}_t)}{\mu}}.$$

The remaining part is to show f with the associated sensing matrix $\boldsymbol{A}$ satisfies μ-quadratic growth condition with high probability. Notice that $a_{ij} \sim \mathcal{N}(0, 1/\sqrt{n})$ and $e_i \sim \mathcal{N}(0, \sigma^2)$ and we know that with a large probability $\nabla f(\tilde{\boldsymbol{x}}^*)_i = (\boldsymbol{A}^\top \boldsymbol{e})_i \le 2\sigma\sqrt{\frac{\log d}{n}}$ (See Section 4 of Jain et al. (2014)). More specifically, $P(\|\nabla f(\tilde{\boldsymbol{x}}^*)\|_\infty > 2\sigma\sqrt{\log d/n}) = P(\sup_j \{\boldsymbol{A}^\top \boldsymbol{e})_j\} > 2\sigma\sqrt{\log d/n}) \le (de^{-\sigma^2 \log d/n}\sqrt{n})/(\sigma\sqrt{\pi \log d})$ (See Equ. 3.1 of Candes & Tao (2007)).

Furthermore, f defined in graph-structured linear sensing problem satisfies the μ-quadratic growth condition, i.e. $\mu \ge \frac{1}{2} - \frac{c_1 s \log d}{n}$ with probability at least $1 - e^{c_0 n}$ where c_0 and c_1 are two universal constants. Therefore, we continue to have

$$\|\tilde{\boldsymbol{x}}^* - \boldsymbol{x}_t\|_2 \le \sqrt{\frac{2\sqrt{s}C\|\nabla f(\tilde{\boldsymbol{x}}^*)\|_\infty}{\mu}} + \sqrt{\frac{2h(\boldsymbol{x}_t)}{\mu}}$$

$$\le \left(\sqrt{\frac{4\sigma C\sqrt{s \log d/n}}{\frac{1}{2}} - \frac{cs \log d}{n}} + \sqrt{\frac{2h(\boldsymbol{x}_t)}{\frac{1}{2} - \frac{cs \log d}{n}}} \right),$$

where $c = c_1$ and the last inequality is valid with high probability. $\qquad \square$

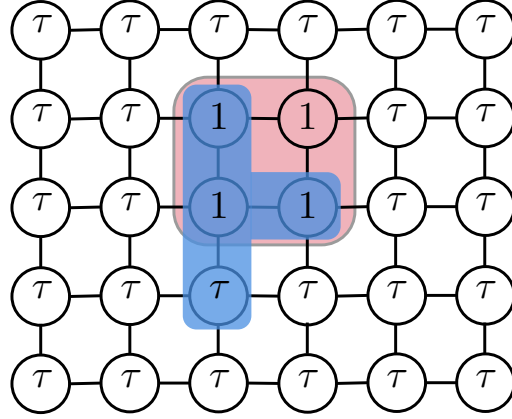# B. Adversarial examples at x_t

Figure 6. A toy example of S^* (4 nodes with 1 entry) with $s = 4$ and $d = 30$ where $\mathbb{G}[S^*]$ is the connected subgraph with up to 4 nodes (red region). Entries on nodes represent $\nabla f(\boldsymbol{x}_t)$. This example of $\nabla f(\boldsymbol{x}_t)$ provides an optimal solution $\langle \boldsymbol{v}_t, \nabla f(\boldsymbol{x}_t) \rangle = -\sqrt{4}$. The blue region is the best approximate solution with value $\langle \bar{\boldsymbol{v}}_t, \nabla f(\boldsymbol{x}_t) \rangle = -\sqrt{3 + \tau^2}$.

Adversarial example setup. Consider the graph $\mathbb{G}$ illustrated in Figure 6. Suppose $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x}/2 - \boldsymbol{x}^\top \boldsymbol{b}$. Clearly, $f(\boldsymbol{x})$ is 1-strongly convex and 1-smooth. Let $\boldsymbol{b}$ and $\boldsymbol{x}_t$ be such that $\nabla f(\boldsymbol{x}_t)^\top = [1, 1, 1, \ldots, 1, \tau, \tau, \ldots, \tau]^\top \in \mathbb{R}^{1 \times d}$ where the first $s = 4$ entries are 1 and rest $d - s = 30 - 4 = 26$ entries are τ with $0 < \tau < 1$ (This is always possible if $t = 0$ and one only needs to pick up such $\boldsymbol{x}_0$ given a predefined $\boldsymbol{b}$). Assume further that $\mathbb{M} = \mathbb{M}(\mathbb{G}, s = 4, g = 1)$ the set of connected components of $\mathbb{G}$ each with at most $s = 4$ nodes, and assume that nodes are numbered such that $S^* = \{1, 2, 3, 4\} \in \mathbb{M}$. Then, the optimal solution is

$$\boldsymbol{v}_t = \underset{\boldsymbol{v} \in \mathcal{D}(1, \mathbb{M})}{\operatorname{argmin}} \boldsymbol{v}^\top \nabla f(\boldsymbol{x}_t) = \frac{1}{\sqrt{4}}(-1, -1, -1, -1, 0, \ldots, 0),$$

where $\boldsymbol{v}_t^\top \nabla f(\boldsymbol{x}_t) = -\sqrt{4} = -2$. However, the best non-optimal-support LMO will select $\bar{\boldsymbol{v}}_t$ with at least one τ entry, and thus $\bar{\boldsymbol{v}}_t^\top \nabla f(\boldsymbol{x}_t) \leq \sqrt{3 + \tau^2}$.

Gap-additive adversarial example. The key problem with the gap-additive assumption is the requirement for the gap error to decay. Recall $\bar{g}_t(\boldsymbol{x}_t) = \langle \nabla f(\boldsymbol{x}_t), \bar{\boldsymbol{v}}_t - \boldsymbol{x}_t \rangle$, $g_t(\boldsymbol{x}_t) = \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{v}_t - \boldsymbol{x}_t \rangle$ In particular, $g_t(\boldsymbol{x}_t) - \bar{g}_t(\boldsymbol{x}_t) = \nabla f(\boldsymbol{x}_t)^\top (\boldsymbol{v}_t^* - \bar{\boldsymbol{v}}_t) = 2 - \sqrt{3 + \tau^2} > 0$, which is strictly positive and constant in t. Thus, any additive gap assumption with decaying tolerance will eventually require an exact LMO-support recovery.

Gap-multiplicative adversarial example. We further consider adversarial examples for satisfying (7), continuing the above example. In this scenario, (7) requires

$$\underbrace{\left(-\sqrt{3 + \tau^2} - \langle \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle\right)}_{A} \leq \delta \underbrace{\left(-\sqrt{4} - \langle \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle\right)}_{B}.$$

But for any $0 < \tau < 1$, suppose that

$$\boldsymbol{x}_t = \frac{1}{\sqrt{3 + \tau^2}} [-1, -1, -1, -\tau, 0, \ldots, 0]^\top \Rightarrow \sqrt{3 + \tau^2} < -\langle \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle < \sqrt{4}.$$

To see the above, the norm of $\boldsymbol{x}_t$ is unit, i.e. $\|\boldsymbol{x}_t\|_2 = 1$ and $-\langle \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle = \frac{3 + \tau}{\sqrt{3 + \tau^2}}$. Notice further that $\sqrt{3 + \tau^2} < -\langle \boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t) \rangle = \frac{3 + \tau}{\sqrt{3 + \tau^2}} < \sqrt{4}$. Then $A > 0$ but $B < 0$, and *no positive value of δ can possibly satisfy* (7); that is, the assumption is only satisfied if $\delta = 1$ and the LMO is exact, which is NP-hard.

Algorithm 2 Approximate $(\delta, \nabla f(\boldsymbol{x}_t), \mathcal{D})$-DMO for k-support-norm ball.

1: **Input**: approximation factor $\delta \in (0, 1]$, input vector $\nabla f(\boldsymbol{x}_t)$
2: $S^* = \arg\max_{S \in \mathbb{M}} \|\nabla f(\boldsymbol{x}_t)_S\|_2$
3: **if** $\delta = 1$ **then**
4:     **return** S^*
5: **end if**
6: $\bar{S} = \arg\min_{S \in \mathbb{M}} \|\nabla f(\boldsymbol{x}_t)_S\|_2$
7: **for** $j \in S^*$ **do**
8:     randomly remove an element from $\bar{S}$
9:     $\bar{S} = \bar{S} \cup j$
10:     **if** $\|\nabla f(\boldsymbol{x}_t)_{\bar{S}}\|_2 \geq \delta \|\nabla f(\boldsymbol{x}_t)_{S^*}\|_2$ **then**
11:         **return** $\bar{S}$
12:     **end if**
13: **end for**

C. Experimental details of Fig. 2

In Fig. 2, we consider the following optimization problem

$$\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x}) := \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2,$$

where $\mathcal{D} := \{\boldsymbol{x} : \operatorname{supp}(\boldsymbol{x}) \in \mathbb{M}, \|\boldsymbol{x}\|_2 \leq C\}$ with $C = 1$ and $\mathbb{M} = \{S \subseteq [d] : |S| \leq s = 5\}$. The LMO operator for this norm ball can be calculated as the following

$$\boldsymbol{v}_t = \arg\min_{\boldsymbol{x} \in \|\boldsymbol{x}\|_2 \leq 1} \langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x} \rangle = \frac{-\nabla f(\boldsymbol{x}_t)_{S^*}}{\|\nabla f(\boldsymbol{x}_t)_{S^*}\|_2}, \quad S^* \in \arg\max_{S \in \mathbb{M}} \|\nabla f(\boldsymbol{x}_t)_S\|_2. \tag{49}$$
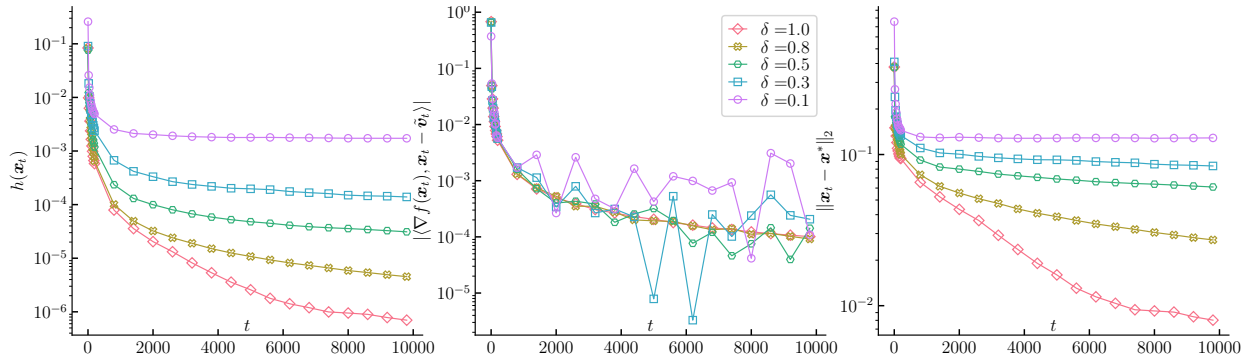

Figure 7. Left: The primal error $h(\boldsymbol{x}_t)$ as a function of t for DMO-FW-I with different δ. *Middle*: The absolute value of $|\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \tilde{\boldsymbol{v}}_t\rangle|$. *Right*: The estimation error $\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2$ as a function of t for DMO-FW-I with different δ. The number of measurements $n = 200$, i.e. $\boldsymbol{A} \in \mathbb{R}^{200 \times d}$.

To illustrate our approximation bound, instead of using the above exact operator (49), we obtain an approximate DMO operator for the above optimization problem and present the δ-approximate DMO in Alg. 2. It returns an S such that is at least δ-approximation DMO operator oracle. The key step to control the quality of S is Line 6 where S is returned whenever $\|\nabla f(\boldsymbol{x}_t)_S\|_2 \geq \delta \|\nabla f(\boldsymbol{x}_t)_{S^*}\|_2$ and S^* is the subset of maximal magnitudes of $\nabla f(\boldsymbol{x}_t)$.

Our experimental setting is as follows: We use a normalized Gaussian sensing matrix where each entry $A_{ij} \sim \mathcal{N}(0, 1/\sqrt{n})$. The number of samples $n = 200$ and the dimensionality $d = 500$. We fix sparsity $s = 50$ for $\boldsymbol{x}^*$, i.e. $\|\boldsymbol{x}^*\|_0 = 50$ where each nonzero entry is either 1 or -1 with same probability. We obtain $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^*$ with $\boldsymbol{x}^* = \boldsymbol{x}^*/\|\boldsymbol{x}^*\|_2$. We also plot the duality gap $|\langle \nabla f(\boldsymbol{x}_t), \boldsymbol{x}_t - \tilde{\boldsymbol{v}}_t\rangle|$ and estimation error of $\boldsymbol{x}^*$, i.e., $\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2$.

D. More experimental details

D.1. Experimental setup

Parameters of all methods. All methods except for CoSAMP use the same approximation operator, i.e., head projection proposed in (Hegde et al., 2015b) (See details in Sec. E). COSAMP uses the s-sparse thresholding operator. GraphCoSAMP share the same parameter setting as COSAMP but uses graph projection operator. Gen-MP has L-smooth parameter where we estimate L by finding the largest eigenvalue of $\boldsymbol{A}^\top \boldsymbol{A}$. The step size of Graph-IHT is then set to $\eta_t = 1/L$. The step size of both DMO-FW and DMO-AccFW are set to $\eta_t = 2/(t+2)$ for all $t \geq 0$.

Datasets. In our experiments, we use two datasets: 1) 10 MNIST images. We randomly select 10 MNIST images as our graph-structured signals $\boldsymbol{x}^* \in \mathbb{R}^{28 \times 28}$ and normalized them into a unit vector; 2) Angio image. We also choose a sparse angio image from (Hegde et al., 2015b) where $\boldsymbol{x}^* \in \mathbb{R}^{100 \times 100}$. These sparse images have 1 connected component. We run all methods on a sever with 246GB memory and 80 cores. All methods are implemented in Python-3.8. The graph projection operator is implemented in C++11.

Graph-structured sparse recovery. The goal of GS sparse recovery is to recovery a sparse image $\boldsymbol{x}^*$ with several small connected components as a prior. For example, in Angio image, we consider this $d = 100 \times 100$ sparse image where the true image $\boldsymbol{x}^*$ is shown in 4 (bottom left). The underlying sparsity pattern has $g = 1$ connected components. We then normalize $\boldsymbol{x}^*$ such that $\|\boldsymbol{x}^*\| = 1$. Measurements $\boldsymbol{y}$ are generated by $\boldsymbol{y} = \langle \boldsymbol{A}, \boldsymbol{x}^* \rangle + \boldsymbol{e}$ where $\boldsymbol{e} \sim \epsilon \cdot \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and ϵ controls the magnitude of noise $\boldsymbol{e}$. To summarize, the objective in our experiment is $\min_{\boldsymbol{x} \in \mathcal{D}} f(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_2^2$, where $\mathcal{D} = \text{conv} \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq 1, \text{supp}(\boldsymbol{x}) \in \mathbb{M}\}$, and , $\mathbb{M} = \{F = S_1 \cup S_2 \cup \cdots \cup S_{11} : S_i \text{ are CCs of } \mathbb{G}, |F| \leq s\}$. $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is a Gaussian sensing matrix where each entry $a_{ij} \sim \mathcal{N}(0, 1/\sqrt{n})$ independently. We run each experiment for 20 trials.

D.2. More results

In the MNIST image recovery task, we set $n = 5 \cdot |\text{supp}(\boldsymbol{x}^*)|$ where $\boldsymbol{x}^*$ is a specific normalized MNIST image. To compare DMO-FW with DMO-ACCFW on all ten sparse MNIST images. The prime error $h(\boldsymbol{x}_t)$ as a function of time t is illustrated in Fig. 8. These results indicate that DMO-ACCFW is DMO-FW on all of these sparse images. Similarly, Fig. 9 presents the estimation errors $\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2$ over time t.
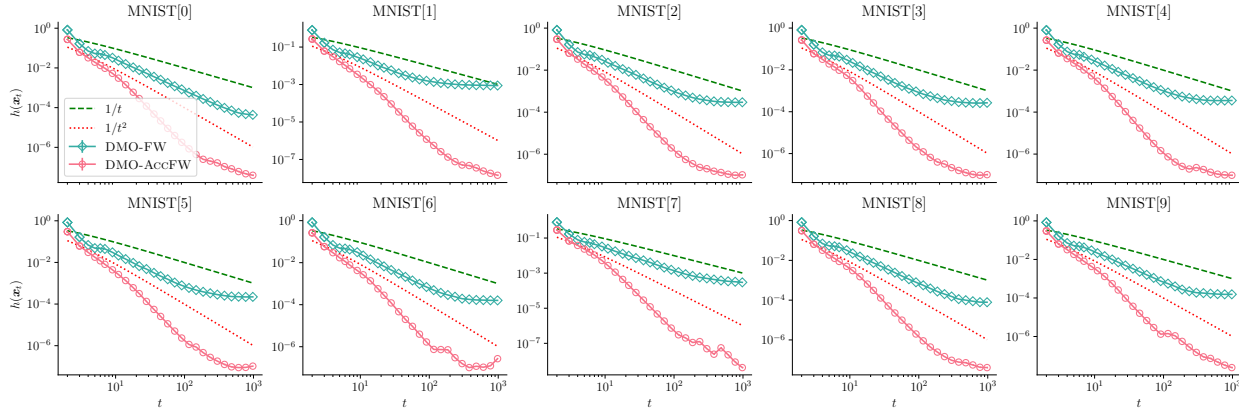

Figure 8. Primal error $h(\boldsymbol{x}_t)$ as a function of t on task of the graph-structured sparse recovery of ten MNIST images.
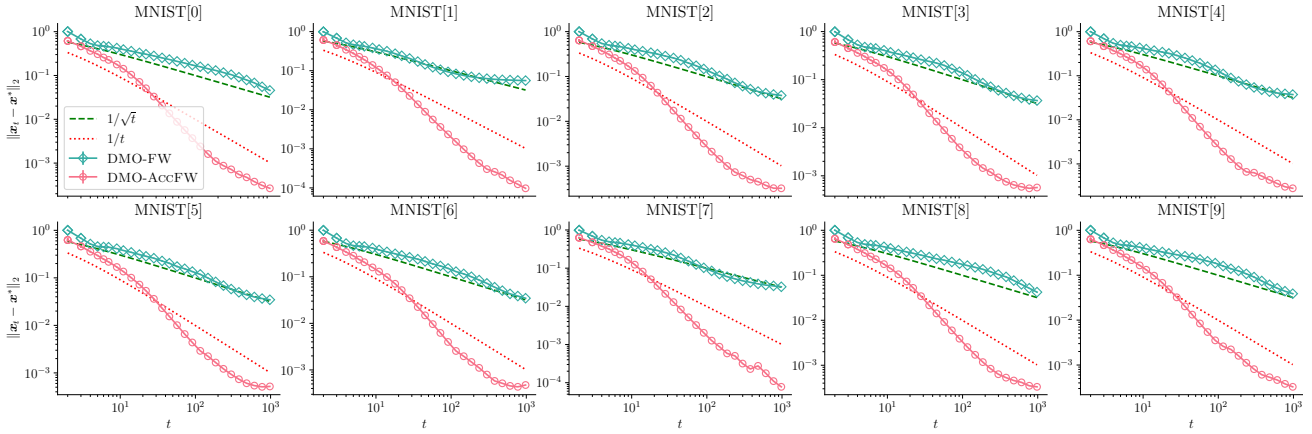
Figure 9. Estimation error $\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2$ as a function of t on task of the graph-structured sparse recovery of ten MNIST images.
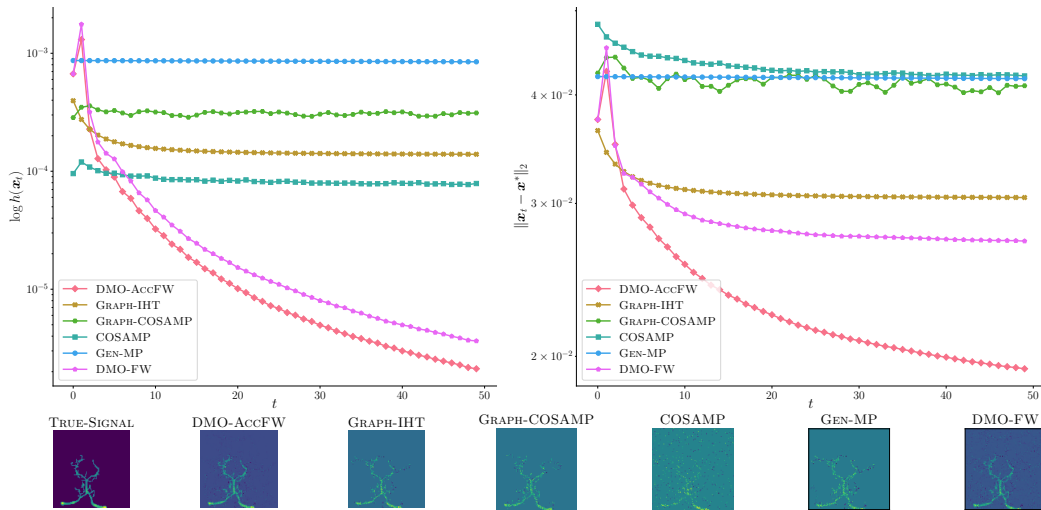

Figure 10. The performance of methods on the task of graph-structured linear regression task. *Top*: DMO-ACCFW vs other baseline methods on primal error $h(\boldsymbol{x}_t)$ (*left*) and variable suboptimality $\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_2$ (*right*) as a function of t. *Bottom*: Recovered sparse images $\boldsymbol{x}_t$ vs truth image (bottom right) after 50 iterations.

Performance of DMO-FW. We include the results of DMO-FW where its performance is between GRAPHIHT and DMO-ACCFW. One may notice that GRAPHIHT is stuck as a local minimum in Fig. 4 since it solves a nonconvex problem. It is a known issue with IHT methods, regardless of step size. The Fig. 10 on the right (gold line) demonstrates the same stalling behavior when $\eta_t = 1/(t+1)$ is used.
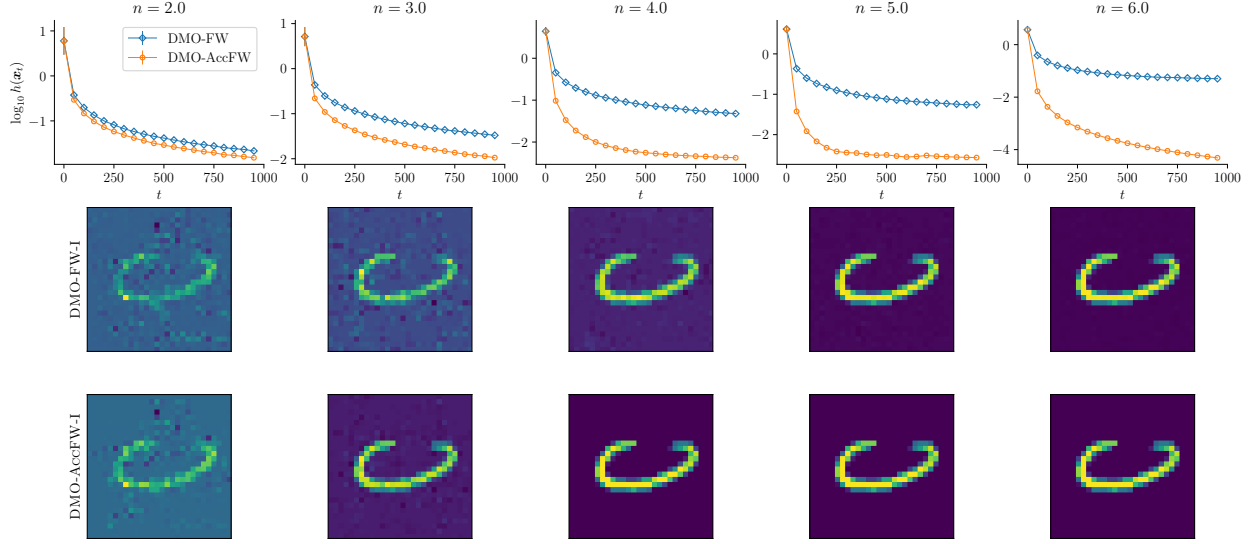
Figure 11. The comparison of DMO-FW-I with DMO-ACCFW-I for different sampling ratio.

Fig. 11 showcases the learning process as a function of training ratio from 2.0 to 6.0.

E. Dual maximization oracles

DMO via a heuristic method. A heuristic method with $\delta = \sqrt{1/\lceil s/g \rceil}$ for $\mathbb{M}(\mathbb{G}, s, g)$. In Section 4.1, we present a heuristic procedure that is for $\mathbb{M}(\mathbb{G}, s, g)$. Algorithm 3 presents this heuristic method, which has three main steps: **Step 1)** Let I_g be the indices of g largest magnitude $|z_i|$. Initialize a node set as $S = I_g$ (Line 2 and Line 3); **Step 2)** Next, iterate through the edges $(u, v) \in \mathbb{E}$, in any order. If $u \in S$, merge v into S; similarly, if $v \in S$, merge $u \in S$. If at any point $|S| = s$, terminate; **Step 3)** Repeat Step 2 until either no new edges are added, or $|S| = s$ (Line 10 to Line 24). This procedure finds a δ-DMO for $\mathbb{M}(\mathbb{G}, s, g)$ with $\delta = 1/\lceil s/g \rceil$, with runtime linear to the number of edges $\mathcal{O}(m)$.

We prove that S returned by Algorithm 3 satisfies $\|z_S\|_* \geq 1/\lceil s/g \rceil \max_{S' \in \mathbb{M}} \|z_{S'}\|_*$: First of all, S is in $\mathbb{M}$ and notice that $\|z_{S'}\|_*^2 = \sum_{i \in I_g} |z_i|^2 + \sum_{j \in S' \setminus I_g} |z_j|^2$, $\forall S' \in \mathbb{M}$. As S contains g largest magnitudes of z, we have $(\lceil s/g \rceil - 1) \sum_{i \in S} |z_i|^2 \geq \sum_{j \in S' \setminus I_g} |z_j|^2$. This inequality provides $\lceil s/g \rceil \|z_S\|_*^2 \geq \sum_{i \in I_g} |z_i|^2 + \sum_{j \in S' \setminus I_g} |z_j|^2 = \|z_{S'}\|_2^2$. Hence, we have $\lceil s/g \rceil \|z_S\|_*^2 \geq \max_{S' \in \mathbb{M}} \|z_{S'}\|_*^2$. Taking square root of both sides will provide a better approximation guarantee, i.e. $\delta = 1/\lceil s/g \rceil \geq \delta' = \sqrt{1/\lceil s/g \rceil}$. Clearly, the total run time is $\mathcal{O}(m)$ dominated by the for loop of Line 12.

Algorithm 3 A HEURISTIC DMO WITH $\delta = \sqrt{1/\lceil s/g \rceil}$ approximation guarantee

---

1: **Input**: underlying graph $\mathbb{G}$, sparsity k, number of CCs g, input vector $\mathbf{z}$
2: Sort entries of $\mathbf{z}$ by magnitudes such that $|z_{\tau_1}| \geq |z_{\tau_2}| \geq \ldots \geq |z_{\tau_g}| \geq |z_{\tau_{g+1}}|$ {Notice that this step can be done in $\mathcal{O}(d)$ time by using Floyd-Rivest selection algorithm (Floyd & Rivest, 1975).}
3: $I_g = [\tau_1, \tau_2, \ldots, \tau_g], S = I_g$
4: $\mathbf{c} = \mathbf{0}$ {Initially, all nodes have same connected component ID}
5: $i = 1$ // Tracking the ID of connected component
6: **for** $v \in S$ **do**
7:     $c_v = i$ // Node v has a component ID i
8:     $i = i + 1$
9: **end for**
10: $\mathbb{F} = \emptyset$ // Keep edges that are in g components
11: **if** $|S| = s$ **then**
12:     **Return** S // We assume $g \leq s$
13: **end if**
14: **for** $(u, v) \in \mathbb{E}$ **do**
15:     **if** $c_u == 0$ and $c_v \neq 0$ **then**
16:         $S = S \cup \{u\}$
17:         $\mathbb{F} = \mathbb{F} \cup (u, v)$
18:         $c_u = c_v$ // u is added to c_v-th component
19:     **end if**
20:     **if** $|S| = s$ **then**
21:         **Return** S
22:     **end if**
23:     **if** $c_u \neq 0$ and $c_v == 0$ **then**
24:         $S = S \cup \{v\}$
25:         $\mathbb{F} = \mathbb{F} \cup (u, v)$
26:         $c_v = c_u$ // v is added to c_u-th component
27:     **end if**
28:     **if** $|S| = s$ **then**
29:         **Return** S
30:     **end if**
31: **end for**

DMO via the head projection operator. Hegde et al. (2015b) presents an algorithm for $\mathbb{M}(\mathbb{G}, s, g)$ that has $\delta = \sqrt{1/14}$. We state a simplified version of it as the following: Consider $\mathbb{M}(\mathbb{G}, s, g)$-WGM and let $\mathbf{z} \in \mathbb{R}^d$. Then there is an algorithm that returns a support $S \subseteq [d]$ in $\mathbb{M}(\mathbb{G}, 2s + g, g)$-WGM satisfying that $\|\mathbf{z}_S\|_2 \geq \delta \cdot \max_{S' \in \mathbb{M}} \|\mathbf{b}_{S'}\|_2$, where $\delta = \sqrt{1/14}$ and it runs in $\mathcal{O}(m \log^3(d))$ where m is the number of edges in $\mathbb{G}$.

The operators have a budget $B = s - g$. The budget value is 1 for edge cost in our experiments. In this case, the cost budget will never be violated since total costs in a g forest are always not greater than $s - g$. The essential idea of this operator is a binary search over the Price-Collecting Steiner Forest problem (Hegde et al., 2014). It then prunes over the final forest so that the returning $\mathbb{G}[S]$ is "dense". A C++ implementation of PCSF-GW is publicly available at https://github.com/ludwigschmidt/cluster_approx. In our experiments, we implement a C-version, which is marginally faster.

Table 4. Comparison of convergence rates of FW-type methods with ours when $\mathcal{D}$ is a GS support set. In all cases, we assume the diameter of $\mathcal{D}$ is $D := \max_{\boldsymbol{x}, \boldsymbol{y} \in \mathcal{D}} \|\boldsymbol{x} - \boldsymbol{y}\|$ and f is convex differentiable. The column operator availability checks whether an approximate LMO/DMO-operator is efficiently obtainable for GSCOs. In our problem $D = 4C^2$.

Algorithm	Operator Availability	Solution	Condition	Convergence rate
Inexact gap-additive (Jaggi, 2013)	✗	$\boldsymbol{x}_t \in \mathcal{D}$	L-smooth	$\frac{2LD^2(1+\delta)}{(t+2)}$
Inexact gap-mult. (Pedregosa et al., 2020)	✗	$\boldsymbol{x}_t \in \mathcal{D}$	L-smooth	$\frac{2(LD^2+BD\delta)}{(\delta^2 t + 2\delta)}$
DMO-FW-I	✓	$\boldsymbol{x}_t \in \mathcal{D}$	L-smooth, $\|\nabla f(\boldsymbol{x})\|_\infty \leq B/t^\nu$	$\mathcal{O}\left(\frac{BC\sqrt{s}}{t^\nu}\right)$
DMO-FW-II	✓	$\boldsymbol{x}_t \in \frac{\mathcal{D}}{\delta}$	L-smooth	$\frac{2LD^2}{\delta^2(t+2)}$
DMO-AccFW-I	✗	$\boldsymbol{x}_t \in \mathcal{D}$	L-smooth, μ-quadratic, $\|\boldsymbol{x}^*\|_2 = C$	$\frac{4e^{4L/\mu}h(\boldsymbol{x}_0)}{(t+2)^2}$
DMO-AccFW-I	✓	$\boldsymbol{x}_t \in \mathcal{D}$	L-smooth, $\|\nabla f(\boldsymbol{x})\|_\infty \leq B/t^\nu$	$\mathcal{O}\left(\frac{BC\sqrt{s}}{t^\nu}\right)$
DMO-AccFW-II	✓	$\boldsymbol{x}_t \in \frac{\mathcal{D}}{\delta}$	L-smooth, μ-quadratic growth	$\mathcal{O}\left(\frac{L^2(C^2/\delta^2 - D_*^2)}{\mu(t+2)}\right)$

E.1. Other graph-structured models

Table 3. DMOs of different $\mathbb{M}$. DP is for Dynamic Programming.

$\mathbb{M}$	DMO	Complexity	δ-approx.
$\{H : H := \cup_{i=1}^s S_s\}$ where graph $\mathbb{G}$ is a tree and S_k is a subtree	Tree decomp. (Lim & Wright, 2017)	$\mathcal{O}(ms + d)$	$\delta = 1$
$\{S : \mathbb{T}_S$ is a subtree. $|S| \leq s\}$ where $\mathbb{T}$ is a tree and $\mathbb{G} = \mathbb{T}$	DP (Hochbaum & Pathria, 1994)	$\mathcal{O}(s^2 d)$	$\delta = 1$
$\mathbb{M}(\mathbb{G}, s, g)$	Algorithm 3	$\mathcal{O}(m)$	$\delta = \sqrt{1/\lceil s/g \rceil}$
$\mathbb{M}(\mathbb{G}, s, g)$	Head Proj. (Hegde et al., 2015b)	$\mathcal{O}(m \log^3 d)$	$\delta = \sqrt{1/14}$

Other operators and applications. We list GS models in Table 3 with time complexities and approximation guarantees. These operators consider connectivity constraints, a key property or requirement of subgraph detection. Connectivity and subgraph detection have been explored recently (Arias-Castro et al., 2011; Qian et al., 2014; Hegde et al., 2015a; Aksoylar et al., 2017). For example, if we assume $\mathbb{M} = \{S : |S| \leq s, \mathbb{G}[S]$ is connected.$\}$, DMO operator can be reformulated as s-maximum-weight subgraph problem, which has been considered in (Hochbaum & Pathria, 1994). This algorithm has been applied to identify subnetwork markers in protein-protein interaction (PPI) network (Dao et al., 2011) and automatic planning (Riabov & Liu, 2006).

E.2. Comparison of Convergence rate

This subsection summarizes and compares the convergences rate of different method as presented in Table 4.