# A Momentum-Guided Frank-Wolfe Algorithm

Bingcong Li, Mario Coutiño, Georgios B. Giannakis, and Geert Leus

***Abstract*—With the well-documented popularity of Frank Wolfe (FW) algorithms in machine learning tasks, the present paper establishes links between FW subproblems and the notion of momentum emerging in accelerated gradient methods (AGMs). On the one hand, these links reveal why momentum is unlikely to be effective for FW-type algorithms on general problems. On the other hand, it is established that momentum accelerates FW on a class of signal processing and machine learning applications. Specifically, it is proved that a momentum variant of FW, here termed accelerated Frank Wolfe (AFW), converges with a faster rate $\mathcal{O}(\frac{1}{k^2})$ on such a family of problems, despite the same $\mathcal{O}(\frac{1}{k})$ rate of FW on general cases. Distinct from existing fast convergent FW variants, the faster rates here rely on parameter-free step sizes. Numerical experiments on benchmarked machine learning tasks corroborate the theoretical findings.**

***Index Terms*—Frank Wolfe method, conditional gradient method, momentum, accelerated method, smooth convex optimization**

## I. Introduction

We consider efficient means of solving the following optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \tag{1}$$

where $f$ is a smooth convex function. The constraint set $\mathcal{X} \subset \mathbb{R}^d$ is assumed to be convex and compact, and $d$ is the dimension of the variable $\mathbf{x}$. We denote by $\mathbf{x}^* \in \mathcal{X}$ a minimizer of (1). Among problems across signal processing, machine learning, and other areas, the constraint set $\mathcal{X}$ can be structured but difficult or expensive to project onto. Examples include the nuclear norm ball constraint for matrix completion in recommender systems [1] and the total-variation norm ball adopted in image reconstruction tasks [2]. The computational inefficiency of the projection, especially for a large $d$, impairs the applicability of projected gradient descent (GD) [3] and projected Accelerated Gradient Method (AGM) [4], [5].

An alternative to GD for solving (1) is the Frank Wolfe (FW) method [6]–[8], also known as the conditional gradient approach. FW circumvents the projection in GD by first minimizing an affine function, which is the *supporting hyperplane* of $f(\mathbf{x})$ at $\mathbf{x}_k$, over $\mathcal{X}$ to obtain $\mathbf{v}_{k+1}$, and then updating $\mathbf{x}_{k+1}$ as a convex combination of $\mathbf{x}_k$ and $\mathbf{v}_{k+1}$. When dealing with structural constraints such as nuclear norm balls and total variation norm balls, an efficient implementation manner or even a closed-form solution for computing $\mathbf{v}_{k+1}$ is available [7], [9], resulting in reduced computational complexity compared with projection steps. In addition, when initializing well, FW directly promotes low rank (sparse) solutions when the constraint set is a nuclear norm ($\ell_1$ norm) ball [1]. Providing the easiness in implementation and enabling structural solutions, FW is of interest in various applications. Besides those mentioned earlier, other examples encompass structural SVM [10], video colocation [11], particle filtering [12], traffic assignment [13], and optimal transport [14], electronic vehicle charging [15], [16], and submodular optimization [17].

Although FW has well documented merits in several applications, it exhibits slower convergence when compared to AGM. Specifically, FW satisfies $f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}(\frac{1}{k})$. This convergence slowdown is confirmed by the lower bound, which indicates that the number of FW subproblems to solve in order to ensure $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$, is no less than $\mathcal{O}\big(\frac{1}{\epsilon}\big)$ [7], [18]. Thus, FW is a lower-bound-matching algorithm, in general. However, improved FW type algorithms are possible in speedup rates for certain subclasses of problems.

### A. Related works

There are three common approaches to select step sizes for FW and its variants: i) line search [7]; ii) minimizing a one-dimensional quadratic function over $[0, 1]$ for smooth step sizes [9], [19]; and iii) parameter-free step sizes; that is, $\mathcal{O}(\frac{1}{k})$ [7]. Most of the fast converging FW iterations rely on choices i) or ii), which require either the smoothness parameter or the function value of $f$. Step size i) is 'clumsy' when it is costly to access function values, e.g., in the big data regime. Concerns with choice ii) arise with how well the smoothness parameter is estimated. In addition, it is challenging to select the smoothness inducing norm, and each norm can result in a considerably different smoothness parameter [20]. The need thus arises for FW variants relying on parameter-free step sizes, especially those enabling faster convergence. To this end, we first briefly recap existing results on faster rates.

*Line search.* Jointly leveraging line search and 'away steps,' FW-type algorithms converge linearly for strongly convex problems when $\mathcal{X}$ is a polytope [8], [23]; see also [24], [25], and [21] where the memory efficiency of away steps is also improved.

*Smooth step sizes.* If $\mathcal{X}$ is strongly convex, and the optimal solution is at the boundary of $\mathcal{X}$, it is known that FW converges linearly [19]. For uniformly (and thus strongly) convex sets, faster rates are attained when the optimal solution is at the boundary of $\mathcal{X}$ [26]. When both $f$ and $\mathcal{X}$ are strongly convex, FW with the smooth step size converges at a rate of

This research is supported in part by NSF 1901134 and the ASPIRE project (project 14926 within the STW OTP programme), financed by the Netherlands Organization for Scientific Research (NWO). Mario Coutino is partially supported by CONACYT.

B. Li and G. B. Giannakis are with the Dept. of Electrical and Computer Engineering and the Digital Technology Center, University of Minnesota, Minneapolis, MN 55455 USA. Emails: {lixx5599, georgios}@umn.edu.

M. Coutino and G. Leus are with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft 2628 CD, The Netherlands. M. Coutino is now with Radar Technology, TNO, The Hague, The Netherlands. E-mails: {m.a.coutinominguez, g.j.t.leus}@tudelft.nl.

TABLE I
A COMPARISON OF FW VARIANTS WITH FASTER RATES

| Work | Assumptions on $f$ | Assumptions on $\mathcal{X}$ | Additional parameters in the step sizes | Convergence rate |
|---|---|---|---|---|
| [8], [21] | smooth and strongly convex | polytopes | function value | linear convergence |
| [19] | smooth and convex | active strongly convex sets, e.g., active $\ell_p$ norm balls with $p \in (1, 2]$ | smoothness constant | linear convergence |
| [9] | smooth and strongly convex | strongly convex sets | smoothness constant | $\mathcal{O}(\frac{1}{k^2})$ |
| [22] | smooth, convex, twice differentiable, and locally strongly convex around $\mathbf{x}^*$ | polytopes | – | $\mathcal{O}(\frac{1}{k^2})$ |
| This work | smooth and convex | active $\ell_p$ norm balls with $p \in [1, +\infty)$ | – | $\tilde{\mathcal{O}}(\frac{1}{k^2})$ |

$\mathcal{O}(\frac{1}{k^2})$, regardless of where the optimal solution resides [9]. A variant of smooth step size along with modifications on FW jointly enable faster rates on a strongly convex $f$ and Gauge set $\mathcal{X}$ [27], at the expense of requiring extra parameters besides the smoothness constant.

*Parameter-free step sizes.* Without any parameter involved here, there is no concern on the quality of parameter estimation, which saves time and effort because there is no need for tuning step sizes. Although implementation efficiency is ensured, theoretical guarantees are challenging to obtain. This is because $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ cannot be guaranteed without line search or smooth step sizes. Faster rates for parameter-free FW are rather limited in number. In a recent work [22], the behavior of FW when $k$ is large and $\mathcal{X}$ is a polytope is investigated under the strong assumptions on $f(\mathbf{x})$ being twice differentiable and locally strongly convex around $\mathbf{x}^*$. Hence, the analysis does not hold for e.g., the Huber loss, which is widely used in robust regression but is only once-differentiable. The faster rates, along with the assumptions on $f$ and $\mathcal{X}$, are summarized in Table I for comparison. To establish faster rates, our solution connects the FW subproblem with Nesterov's momentum, which is recapped next.

*Nesterov momentum.* After the $\mathcal{O}(\frac{1}{k^2})$ convergence rate was established in [3], [28], the efficiency of Nesterov momentum is proven almost universal; see e.g., the accelerated proximal gradient [5], [29], projected AGM [4], [5] for problems with constraints; accelerated mirror descent [4], [5], [30], and accelerated variance reduction for problems with finite-sum structures [31], [32]. Parallel to these works, AGM has been also investigated from an ordinary differential equation (ODE) perspective [30], [33]–[35]. However, the efficiency of Nesterov momentum on FW type algorithms is shaded given the lower bound on the number of subproblems [7], [18]. A means to bringing momentum into FW is to adopt conditional gradient sliding (CGS) [36], where the projection subproblem in the original AGM is substituted by gradient sliding which solves a sequence of FW subproblems. The faster rate $\mathcal{O}(\frac{1}{k^2})$ is obtained with the price of: i) the requirement of at most $\mathcal{O}(k)$ FW subproblems in the $k$th iteration; and ii) an inefficient implementation (e.g., the AGM subproblem has to be solved to certain accuracy, and it relies on other parameters that are not necessary in FW, such as the diameter of $\mathcal{X}$).

Although parameter-free FW is undoubtedly attractive in several applications, there are two main challenges in establishing faster rates for such step sizes: i) even AGM and most of its variants are not parameter-free since they involve a smoothness parameter; and ii) parameter-free FW in general cannot ensure per step descent, which is essential for faster rates. To overcome these challenges, we first unveil the links between the notion of momentum and the FW subproblem. Then, we leverage these connections to provide provable constraint-dependent faster rates.

### B. Our contributions

In succinct form, our contributions are as follows.

- We observe that the momentum update in AGM plays a similar role as the subproblem in FW, intuitively and analytically. Hence, the FW subproblem can be leveraged to play the role of Nesterov's momentum, thus enabling faster rates on a useful family of problems.
- We prove that a momentum-guided FW, termed accelerated Frank Wolfe (AFW), achieves a faster rate $\tilde{\mathcal{O}}(\frac{1}{k^2})$ on active $\ell_p$ norm ball constraints without knowledge of the smoothness parameter or the function value. We also establish that AFW converges no slower than FW on general problems.
- We corroborate the numerical efficiency of AFW on two benchmark tasks. We validate faster AFW rates on binary classification problems with different constraint sets. We further demonstrate that for matrix completion, AFW finds low-rank solutions with small optimality error more rapidly than FW.

**Notation**. Bold lowercase letters denote column vectors; $\|\mathbf{x}\|$ stands for the $\ell_2$ norm of a vector $\mathbf{x}$; and $\langle \mathbf{x}, \mathbf{y} \rangle$ denotes the inner product between vectors $\mathbf{x}$ and $\mathbf{y}$. All missing proofs can be found in the Appendix.

## II. PRELIMINARY

This section briefly reviews FW starting with the assumptions to clarify the class of problems we are focusing on.

**Assumption 1.** *(Lipschitz Continuous Gradient.) The function $f : \mathbb{R}^d \to \mathbb{R}$ has L-Lipchitz continuous gradients; that is, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|,\ \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

**Assumption 2.** *(Convex Objective Function.) The function $f : \mathbb{R}^d \to \mathbb{R}$ is convex; that is, $f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle,\ \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

**Assumption 3.** *(Constraint Set.) The constraint set $\mathcal{X}$ is convex and compact with diameter $D$, that is, $\|\mathbf{x} - \mathbf{y}\| \leq D,\ \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}$.*

**Algorithm 1** FW [6]

**Initialize:** $\mathbf{x}_0 \in \mathcal{X}$, $\delta_k = \frac{2}{k+2}, \forall k$.
**for** $k = 0, 1, \ldots, K-1$ **do**
  $\mathbf{v}_{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \nabla f(\mathbf{x}_k), \mathbf{x}\rangle$
  $\mathbf{x}_{k+1} = (1-\delta_k)\mathbf{x}_k + \delta_k \mathbf{v}_{k+1}$
**end for**
**Return:** $\mathbf{x}_K$

**Algorithm 2** AGM [3]

**Initialize:** $\mathbf{x}_0 = \mathbf{v}_0$, $\delta_k = \frac{2}{k+2}$, $\mu_0 = L$, $\mu_{k+1} = (1-\delta_k)\mu_k$.
**for** $k = 0, 1, \ldots, K-1$ **do**
  $\mathbf{y}_k = \delta_k \mathbf{v}_k + (1-\delta_k)\mathbf{x}_k$
  $\mathbf{x}_{k+1} = \mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)$
  $\mathbf{v}_{k+1} = \mathbf{v}_k - \frac{\delta_k}{\mu_{k+1}}\nabla f(\mathbf{y}_k)$
**end for**
**Return:** $\mathbf{x}_K$

Assumptions 1 – 3 are standard for FW type algorithms, and they are assumed to hold true throughout.

FW is summarized in Alg. 1. A subproblem with a linear loss needs to be solved to obtain $\mathbf{v}_{k+1}$ per iteration. This subproblem is also referred to as an *FW step*, and it admits a geometrical explanation. In particular, $\mathbf{v}_{k+1}$ can be rewritten as

$$\mathbf{v}_{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle. \tag{2}$$

Noticing that the RHS of (2) is a supporting hyperplane of $f(\mathbf{x})$ as $\mathbf{x}_k$, it is thus clear that $\mathbf{v}_{k+1}$ is a minimizer of this supporting hyperplane over $\mathcal{X}$. Note also that the supporting hyperplane in (2) is also a global lower bound of $f(\mathbf{x})$ due to the convexity of $f$, i.e., $f(\mathbf{x}) \geq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle$. Upon minimizing this lower bound in (2) to obtain $\mathbf{v}_{k+1}$, $\mathbf{x}_{k+1}$ is updated as a convex combination of $\mathbf{v}_{k+1}$ and $\mathbf{x}_k$ to eliminate the projection.

Next, we briefly recap the step sizes for FW to gain insights on why the parameter-free FW is challenging to analyze.

*Smooth step size.* At the $k$th iteration, the step size $\delta_k$ in Alg. 1 is obtained as

$$\delta_k = \arg\min_{\delta\in[0,1]} \delta\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\rangle + \frac{\delta^2 L}{2}\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2.$$

Clearly, it is imperative to estimate $L$ accurately because this estimate markedly influences the performance. It has also been argued that algorithms relying on a guess of $L$ are not robust [37]. Tuning to find the 'best' $L$ is employed in practice to optimize the performance empirically. On the other hand, smooth step sizes ensure descent per iteration, which is analytically attractive. Indeed, Assumption 1 implies that

$$\begin{aligned} & f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \qquad (3)\\ &\leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k\rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\overset{(a)}{=} \delta_k\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\rangle + \frac{\delta_k^2 L}{2}\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2 \overset{(b)}{\leq} 0 \end{aligned}$$

where (a) uses $\mathbf{x}_{k+1} = (1-\delta_k)\mathbf{x}_k + \delta_k\mathbf{v}_{k+1}$, and (b) holds because $\delta_k$ minimizes the RHS of (3) over $[0, 1]$.

*Line search.* An alternative to tune for the best $L$ is to employ line search for determining the local smoothness parameter. In particular, the step size is chosen as $\delta_k = \arg\min_{\delta\in[0,1]} f\big((1-\delta)\mathbf{x}_k + \delta\mathbf{v}_{k+1}\big)$. However, the price paid is the need to compute $f(\mathbf{x})$, which is inefficient when function evaluation is costly (e.g., in big-data regimes). Note that $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ is automatically ensured by line search.

*Parameter-free step size.* This type of step sizes does not rely on $L$ or other parameters, and hence it is extremely easy to implement. Two possible choices are $\delta_k = \frac{2}{k+2}$ or $\delta_k = \frac{1}{k+1}$. However, these step sizes do not guarantee descent per iteration, which becomes the bottleneck for establishing faster rates on specific constraint sets. Our insight to overcome this comes from the observation that the FW step is similar to the momentum in AGM for convex problems. Hence, the FW step itself can be used as an approximate momentum.

## III. Connecting momentum with FW

To bring intuition on how momentum can be helpful for FW type algorithms, we first recap AGM for unconstrained convex problems, i.e., $\mathcal{X} = \mathbb{R}^d$. Note that the reason for discussing the unconstrained problem here is only for the simplicity of exposition, and one can extend the arguments to constrained cases straightforwardly. AGM [3], [4], [28] is summarized in Alg. 2. We start this section by characterizing the behavior of $\{\mathbf{x}_k\}$, $\{\mathbf{y}_k\}$ and $\{\mathbf{v}_k\}$ in the next theorem.

**Theorem 1.** *Under Assumptions 1 and 2, with $\delta_k = \frac{2}{k+3}$, $\mu_0 = 2L$, and $\mu_{k+1} = (1-\delta_k)\mu_k$, AGM in Alg. 2 guarantees that*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}\Big(\frac{f(\mathbf{x}_0) - f(\mathbf{x}^*) + L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{k^2}\Big),\ \forall k.$$

$$\|\nabla f(\mathbf{y}_k)\|^2 \leq \mathcal{O}\bigg(\frac{L\big(f(\mathbf{x}_0) - f(\mathbf{x}^*) + L\|\mathbf{x}_0 - \mathbf{x}^*\|^2\big)}{(k+2)^2}\bigg),\ \forall k.$$

*In addition, it holds for any $k$ that $\|\mathbf{v}_k - \mathbf{x}^*\|^2 \leq \frac{1}{L}\big(f(\mathbf{x}_0) - f(\mathbf{x}^*) + L\|\mathbf{x}_0 - \mathbf{x}^*\|^2\big)$.*

Theorem 1 shows that $\|\nabla f(\mathbf{y}_k)\|^2 = \mathcal{O}(\frac{1}{k^2})$, which implies that $\mathbf{y}_k$ also converges to a minimizer as $k \to \infty$. Through the increasing step size $\frac{\delta_k}{\mu_{k+1}} = \mathcal{O}(\frac{k}{L})$, the update of $\mathbf{v}_k$ stays in the ball centered at $\mathbf{x}^*$ with radius depending on both $\mathbf{x}^*$ and $\mathbf{x}_0$.

One observation of AGM is that by substituting Line 5 in Alg. 2 with $\mathbf{v}_{k+1} = \mathbf{x}_{k+1}$, the modified algorithm boils down to GD. Hence, it is clear that the key behind AGMs acceleration is $\mathbf{v}_k$ and the way it is updated. We contend that the $\mathbf{v}_{k+1}$ is obtained by minimizing an approximated lower bound of $f(\mathbf{x})$ formed as the summation of a supporting hyperplane at $\mathbf{y}_k$ and a regularizer. To see this, one can rewrite Line 5 of AGM as

$$\mathbf{v}_{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^d} \underbrace{f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{x} - \mathbf{y}_k\rangle}_{\text{supporting hyperplane}} + \underbrace{\frac{\mu_{k+1}}{2\delta_k}\|\mathbf{x} - \mathbf{v}_k\|^2}_{\text{regularizer}} \tag{4}$$

where the linear part is the supporting hyperplane, and $\frac{\mu_{k+1}}{\delta_k} = \mathcal{O}(\frac{L}{k})$. As $k$ increases, the impact of the regularizer

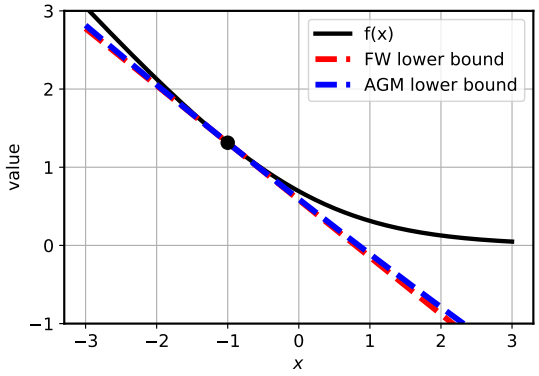



Fig. 1. Similarity between the RHS of (2) and (4).

$\frac{\mu_{k+1}}{2\delta_k}\|\mathbf{x}-\mathbf{v}_k\|^2$ in (4) will become limited. Thus the RHS can be viewed as an approximated lower bound of $f(\mathbf{x})$. Regarding the reasons to put a regularizer after the supporting hyperplane, it first guarantees the minimizer *exists* since directly minimize the supporting hyperplane over $\mathbb{R}^d$ yields no solution. In addition, $\mathbf{v}_{k+1}$ is ensured to be *unique* because the RHS of (4) is strongly convex thanks to the regularizer. Since $\mathbf{v}_{k+1}$ minimizes an approximated lower bound of $f(\mathbf{x})$, it can be used to estimate $f(\mathbf{x}^*)$. We explain in Theorem 4 in Appendix B that $f(\mathbf{y}_k)+\langle\nabla f(\mathbf{y}_k),\mathbf{v}_{k+1}-\mathbf{y}_k\rangle$ approximates $f(\mathbf{x}^*)$. Consequently, one can obtain an estimated suboptimality gap using $f(\mathbf{x}_{k+1})-f(\mathbf{y}_k)-\langle\nabla f(\mathbf{y}_k),\mathbf{v}_{k+1}-\mathbf{y}_k\rangle$.

**Momentum $\mathbf{v}_k$ update as an FW step.** It is observed that $\mathbf{v}_{k+1}$ in both FW and AGM (cf. (2) and (4)) are obtained by minimizing an (approximated) lower bound of $f(\mathbf{x})$, where the only difference lies on whether a regularizer with decreasing weights is utilized. The similarity between the RHS of (2) and (4) will be amplified when $k$ is large; see Fig. 1 for a graphical illustration on how (4) approaches to an affine function. In other words, the momentum update in (4) becomes similar to an FW step for a large $k$. In addition, there are also several other connections.

**Connection 1.** The $\mathbf{v}_{k+1}$ update via (4) is equivalent to

$$\mathbf{v}_{k+1}=\arg\min_{\mathbf{v}\in\mathcal{V}_k}\langle\nabla f(\mathbf{y}_k),\mathbf{v}-\mathbf{y}_k\rangle \tag{5}$$

for $\mathcal{V}_k:=\{\mathbf{v}|\|\mathbf{v}-\mathbf{v}_k\|^2\le r_k\}$ with $r_k$ denoting the time-varying radius of the norm ball. Clearly, $r_k$ depends on $\frac{\mu_{k+1}}{2\delta_k}$, and it is upper bounded by $\frac{2}{L}\big(f(\mathbf{x}_0)-f(\mathbf{x}^*)+L\|\mathbf{x}_0-\mathbf{x}^*\|^2\big)$ according to Theorem 1. By rewriting (4) in its constrained form (5), it can be readily recognized that for unconstrained problems *Nesterov momentum can be obtained via FW steps* with *time-varying* constraint sets.

**Connection 2.** Recall that in AGM, $\mathbf{v}_{k+1}$ obtained via (4) is used to construct an approximation of $f(\mathbf{x}^*)$, which is $f(\mathbf{y}_k)+\langle\nabla f(\mathbf{y}_k),\mathbf{v}_{k+1}-\mathbf{y}_k\rangle$. When a compact $\mathcal{X}$ is present, directly minimizing the supporting hyperplane $f(\mathbf{y}_k)+\langle\nabla f(\mathbf{y}_k),\mathbf{x}-\mathbf{y}_k\rangle$ over $\mathcal{X}$ also yields an estimate of $f(\mathbf{x}^*)$. Note that the latter is exactly an FW step. In addition, the FW step in Alg. 1 also results in a suboptimality gap (known as FW gap; see e.g., [7]), which is in line with the role of $\mathbf{v}_k$ in AGM. In a nutshell, both FW step and momentum update in AGM result in an estimated suboptimality gap.

**Algorithm 3** AFW

1: **Initialize:** $\mathbf{x}_0=\mathbf{v}_0\in\mathcal{X}$, $\boldsymbol{\theta}_0=\mathbf{0}$, $\delta_k=\frac{2}{k+3},\forall k$.
2: **for** $k=0,1,\ldots,K-1$ **do**
3: $\quad\mathbf{y}_k=(1-\delta_k)\mathbf{x}_k+\delta_k\mathbf{v}_k$
4: $\quad\boldsymbol{\theta}_{k+1}=(1-\delta_k)\boldsymbol{\theta}_k+\delta_k\nabla f(\mathbf{y}_k)$
5: $\quad\mathbf{v}_{k+1}=\arg\min_{\mathbf{x}\in\mathcal{X}}\langle\boldsymbol{\theta}_{k+1},\mathbf{x}\rangle$
6: $\quad\mathbf{x}_{k+1}=(1-\delta_k)\mathbf{x}_k+\delta_k\mathbf{v}_{k+1}$
7: **end for**
8: **Return:** $\mathbf{x}_K$

**Connection 3.** Connections between momentum and FW go beyond convexity. We discuss in Appendix C that AGM for strongly convex problems updates its momentum using exactly the same idea of FW, that is, both obtain a minimizer of a lower bound of $f(\mathbf{x})$, and then perform an update through a convex combination.

These links and similarities between momentum and FW naturally lead us to explore their connections, and see how momentum influences FW.

## IV. Momentum-guided FW

In this section we show that the momentum is beneficial for FW by proving that it is effective at least on certain constraint sets. Specifically, we will focus on the accelerated Frank Wolfe (AFW) summarized in Alg. 3, and analyze its convergence rate. Since we will see later that $\delta_k=\frac{2}{k+3}\in(0,1)$, $\forall k$, for which $\mathbf{y}_k$, $\mathbf{v}_k$ and $\mathbf{x}_k$ lie in $\mathcal{X}$ for all $k$, AFW is projection free. Albeit rarely, it is safe to choose $\mathbf{v}_{k+1}=\mathbf{v}_k$, and proceed when $\boldsymbol{\theta}_{k+1}=\mathbf{0}$. Note that the $\mathbf{x}_{k+1}$ update in AFW is slightly different with that of AGM. This is because AGM guarantees $f(\mathbf{x}_{k+1})\le f(\mathbf{y}_k)$, $\forall k$, taking advantage of the known $L$. However, the same guarantee is difficult to be replicated in a parameter-free algorithm.

The key to AFW is the $\mathbf{v}_{k+1}$ update, which plays the role of momentum. To see this, if one unrolls $\boldsymbol{\theta}_{k+1}$ (cf. (22) in Appendix) and plugs it into Line 5 of Alg. 3, $\mathbf{v}_{k+1}$ can be equivalently rewritten as

$$\mathbf{v}_{k+1}=\arg\min_{\mathbf{x}\in\mathcal{X}}\sum_{\tau=0}^{k}w_\tau\big[f(\mathbf{y}_\tau)+\langle\nabla f(\mathbf{y}_\tau),\mathbf{x}-\mathbf{y}_\tau\rangle\big] \tag{6}$$

where $w_\tau=\delta_\tau\prod_{j=\tau+1}^{k}(1-\delta_j)$ and $\sum_{\tau=0}^{k}w_\tau\approx1$ (the exact value of the sum depends on the choice of $\delta_\tau$). Note that $f(\mathbf{y}_\tau)+\langle\nabla f(\mathbf{y}_\tau),\mathbf{x}-\mathbf{y}_\tau\rangle$ is a supporting hyperplane of $f(\mathbf{x})$ at $\mathbf{y}_\tau$, hence the right-hand side (RHS) of (6) is a lower bound for $f(\mathbf{x})$ constructed through a weighted average of supporting hyperplanes at $\{\mathbf{y}_\tau\}$. In other words, $\mathbf{v}_{k+1}$ is a minimizer of a lower bound of $f(\mathbf{x})$, hence it is in line with the role of momentum. However, the momentum in AFW differs from AGM in two aspects. First, instead of relying on $\nabla f(\mathbf{y}_k)$, the update of $\mathbf{v}_{k+1}$ utilizes coefficient $\boldsymbol{\theta}_{k+1}$, which is (roughly) a weighted average of past gradients $\{\nabla f(\mathbf{y}_\tau)\}_{\tau=1}^{k}$ with more weight placed on recent ones. The second difference on the $\mathbf{v}_{k+1}$ update with AGM is whether a regularizer is used. As a consequence of the non-regularized lower bound (6), its minimizer is *not* guaranteed to be unique. A simple example is to consider the $i$th entry $[\boldsymbol{\theta}_{k+1}]_i=0$. The $i$th entry $[\mathbf{v}_{k+1}]_i$ can

then be chosen arbitrarily as long as $\mathbf{v}_{k+1} \in \mathcal{X}$. This subtle difference leads to a significant gap between the performance of AFW and AGM, that is, AFW cannot achieve acceleration on general problems, as will be illustrated shortly. However, we confirm that momentum is still helpful since it is effective on a class of problems.

### A. AFW convergence for general problems

The analysis of AFW relies on a tool known as estimate sequence (ES) introduced by [3]. ES is commonly adopted to analyze projection based algorithms; see e.g., [31], [32], [38], [39], but seldomly used for FW. Formally, ES is defined as follows.

**Definition 1.** *(ES.) A tuple* $\big(\{\Phi_k(\mathbf{x})\}_{k=0}^{\infty}, \{\lambda_k\}_{k=0}^{\infty}\big)$ *is called an estimate sequence of function* $f(\mathbf{x})$ *if* $\lim_{k\to\infty}\lambda_k = 0$*, and for any* $\mathbf{x} \in \mathbb{R}^d$ *we have*

$$\Phi_k(\mathbf{x}) \leq (1-\lambda_k)f(\mathbf{x}) + \lambda_k\Phi_0(\mathbf{x}).$$

ES is generally not unique and different constructions can be used to design different algorithms. To highlight our analysis technique, recall that quadratic surrogate functions $\{\Phi_k(\mathbf{x})\}$ are used for the derivation of AGM [3] (or see (12) in Appendix). Different from AGM, and taking advantage of the compact constraint set, here we consider *linear* surrogate functions for AFW

$$\Phi_0(\mathbf{x}) \equiv f(\mathbf{x}_0) \tag{7a}$$

$$\Phi_{k+1}(\mathbf{x}) = (1-\delta_k)\Phi_k(\mathbf{x}) + \delta_k\Big[f(\mathbf{y}_k) + \big\langle \nabla f(\mathbf{y}_k), \mathbf{x}-\mathbf{y}_k\big\rangle\Big],\ \forall\, k \geq 0. \tag{7b}$$

Evidenced by the terms in the bracket of (7b), i.e., it is a supporting hyperplane of $f(\mathbf{x})$, $\Phi_{k+1}(\mathbf{x})$ is an approximated lower bound of $f(\mathbf{x})$ constructed by weighting the supporting hyperplanes at $\{\mathbf{y}_\tau\}_{\tau=0}^{k}$. Next, we show that (7) together with proper $\{\lambda_k\}$ forms an ES for $f$. Through the ES based proof, it is also revealed that the link between the momentum in AGM and the FW step is also in the technical proof level.

**Lemma 1.** *With* $\lambda_0 = 1$ *and* $\lambda_k = \lambda_{k-1}(1-\delta_{k-1})$*, the tuple* $\big(\{\Phi_k(\mathbf{x})\}_{k=0}^{\infty}, \{\lambda_k\}_{k=0}^{\infty}\big)$ *in* (7) *is an ES of* $f(\mathbf{x})$*.*

Using properties of the functions in (7) (cf. Lemma 4 in Appendix E), the following lemma holds for AFW.

**Lemma 2.** *With* $\Phi_k^* := \min_{\mathbf{x}\in\mathcal{X}}\Phi_k(\mathbf{x})$*, AFW is guaranteed to satisfy* $f(\mathbf{x}_{k+1}) \leq \Phi_{k+1}^* + \xi_{k+1},\ \forall\, k$*, where* $\xi_{k+1} = (1-\delta_k)\xi_k + \frac{L\delta_k^2}{2}\|\mathbf{v}_{k+1}-\mathbf{v}_k\|^2$ *and* $\xi_0 = 0$*.*

Leveraging Lemma 2, the convergence rate of AFW for general problems can be established.

**Theorem 2.** *When Assumptions 1, 2 and 3 are satisfied, upon choosing* $\delta_k = \frac{2}{k+3}$ *and* $\boldsymbol{\theta}_0 = \mathbf{0}$*, AFW guarantees*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{2\big(f(\mathbf{x}_0)-f(\mathbf{x}^*)\big)}{(k+1)(k+2)} + \frac{2LD^2}{k+2},\ \forall\, k.$$

Theorem 2 asserts that the convergence rate of AFW is $\mathcal{O}(\frac{LD^2}{k})$, coinciding with that of FW [7]. Notwithstanding, AFW is tight in terms of the number of FW steps required. To see this, note that the convergence rate in Theorem 2 translates to requiring $\mathcal{O}(\frac{LD^2}{\epsilon})$ FW steps to guarantee $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$. This matches the lower bound [7], [40]. Similar to other FW variants, acceleration for AFW cannot be claimed for general problems. AFW however, is attractive numerically because it can alleviate the zig-zag behavior[1] of FW, as we will see in Section V.

**Why acceleration cannot be achieved in general?** Recall from Lemma 2, that critical to acceleration is ensuring a small $\xi_k$, which in turn requires $\mathbf{v}_{k+1}$ and $\mathbf{v}_k$ to stay sufficiently close. This is difficult in general because the non-uniqueness of $\mathbf{v}_k$ prevents one from ensuring a small upper bound of $\|\mathbf{v}_k - \mathbf{v}_{k+1}\|^2\ \forall\ \mathbf{v}_k,\ \forall\ \mathbf{v}_{k+1}$. The ineffectiveness of momentum in AFW in turn signifies the importance of the added regularizer in AGM momentum update (4).

### B. AFW acceleration for a class of problems

In this subsection, we provide constraint-dependent accelerated rates of AFW when $\mathcal{X}$ is a ball induced by some norm. Even for projection based algorithms, most accelerated rates are obtained with $L$-dependent step sizes [41]. Thus, faster rates for parameter-free algorithms are challenging to establish. An extra assumption is needed in this subsection.

**Assumption 4.** *The constraint is active; that is,* $\|\nabla f(\mathbf{x}^*)\|^2 \geq G > 0$*.*

To analyze convergence of FW iterations, it is reasonable to rely on the position of the optimal solution, which justifies why this assumption is also adopted in [19], [26], [42], [43]. For a number of signal processing and machine learning tasks, Assumption 4 is rather mild. Relying on Lagrangian duality, it can be seen that problem (1) with a norm ball constraint is equivalent to the regularized formulation $\min_{\mathbf{x}} f(\mathbf{x}) + \gamma g(\mathbf{x})$, where $\gamma \geq 0$ is the Lagrange multiplier, and $g(\mathbf{x})$ denotes some norm. In view of this, Assumption 4 simply requires $\gamma > 0$ in the equivalent regularized formulation, that is, the norm ball constraint plays the role of a regularizer. Given the prevalence of regularized formulations, it is worth investigating their equivalent constrained form (1) under Assumption 4. Next, we will use the $\ell_2$ norm ball constraints to illustrate the intuition behind the acceleration.

**$\ell_2$ norm ball constraint.** Consider $\mathcal{X} := \{\mathbf{x}|\|\mathbf{x}\|_2 \leq \frac{D}{2}\}$. In this case, $\mathbf{v}_{k+1}$ admits a closed-form solution

$$\mathbf{v}_{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}}\langle\boldsymbol{\theta}_{k+1}, \mathbf{x}\rangle = -\frac{D}{2\|\boldsymbol{\theta}_{k+1}\|_2}\boldsymbol{\theta}_{k+1}. \tag{8}$$

The uniqueness of $\mathbf{v}_{k+1}$ is ensured by its closed-form solution, wiping out the obstacle for a faster rate. In addition, through (8) it becomes possible to guarantee that $\mathbf{v}_{k+1}$ and $\mathbf{v}_k$ are close whenever $\boldsymbol{\theta}_k$ is close to $\boldsymbol{\theta}_{k+1}$.

**Theorem 3.** *If Assumptions 1, 2, 3 and 4 are satisfied, and* $\mathcal{X}$ *is an* $\ell_2$ *norm ball, choosing* $\delta_k = \frac{2}{k+3}$ *and* $\boldsymbol{\theta}_0 = \mathbf{0}$*, AFW guarantees acceleration with convergence rate*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}\bigg(\min\Big\{\frac{LD^2T + C\ln k}{k^2}, \frac{LD^2}{k}\Big\}\bigg)$$

[1]The change between $f(\mathbf{x}_{k+1})$ and $f(\mathbf{x}_k)$ is large with high frequency, so zig-zag emerges when plotting $f(\mathbf{x}_k) - f(\mathbf{x}^*)$ versus $k$.

TABLE II
A SUMMARY OF DATASETS USED IN NUMERICAL TESTS

| Dataset | $d$ | $n$ (train) | nonzeros |
|---|---|---|---|
| *a9a* | 123 | 32,561 | 11.28% |
| *covtype* | 54 | 406,709 | 22.12% |
| *mushroom* | 122 | 8,124 | 18.75% |
| *mnist* (digit 4) | 784 | 60,000 | 12.4% |

*where $C$ and $T$ are constants depending on $L$, $D$ and $G$.*

Theorem 3 demonstrates that momentum improves the convergence of FW by providing a faster rate. Roughly speaking, when the iteration number $k \geq T$, the rate of AFW dominates that of FW. We note that this matches our intuition, that is, the momentum in AGM (4) only behaves like an affine function when $k$ is large (so that the weight on the regularizer is small). In addition, the rate in Theorem 3 can be written compactly as $\tilde{\mathcal{O}}\big(\frac{TLD^2}{k^2}\big)$, $\forall k$, hence it achieves acceleration with a worse dependence on $D$ compared to vanilla FW. Note that the choice for $\delta_k$ and $\boldsymbol{\theta}_0$ remains the same as those used in general problems, leading to an identical implementation to non-accelerated cases. Compared with CGS, AFW sacrifices the $D$ dependence in the convergence rate to trade for i) the nonnecessity of the knowledge of $L$ and $D$, and ii) ensuring only one FW subproblem per iteration (whereas at most $\mathcal{O}(k)$ subproblems are needed in CGS).

**$\ell_1$ norm ball constraint.** For the sparsity-promoting constraint $\mathcal{X} := \{\mathbf{x} | \|\mathbf{x}\|_1 \leq R\}$, the FW steps can be solved in closed form. Taking $\mathbf{v}_{k+1}$ as an example, we have

$$\mathbf{v}_{k+1} = R \cdot [0, \ldots, 0, -\mathrm{sgn}[\boldsymbol{\theta}_{k+1}]_i, 0, \ldots, 0]^\top \quad \text{with} \quad i = \arg\max_j |[\boldsymbol{\theta}_{k+1}]_j|. \tag{9}$$

We show in the Appendix (Theorem 5) that when Assumption 4 holds and the set $\arg\max_j \big|[\nabla f(\mathbf{x}^*)]_j\big|$ has cardinality 1, a faster rate $\mathcal{O}(\frac{T_1 L D^2}{k^2})$ can be obtained. The additional assumption here is known as *strict complementarity*, and has been adopted also in, e.g., [44], [45] for analysis.

**$\ell_p$ norm ball constraint.** Consider an active $\ell_p$ norm ball constraint $\mathcal{X} := \{\mathbf{x} | \|\mathbf{x}\|_p \leq R\}$, where $p \in (1, +\infty)$ and $p \neq 2$. The $i$-th entry of $\mathbf{v}_{k+1}$ is found in closed form as

$$[\mathbf{v}_{k+1}]_i = -[\boldsymbol{\theta}_{k+1}]_i \frac{\big|[\boldsymbol{\theta}_{k+1}]_i\big|^{q-2}}{\|\boldsymbol{\theta}_{k+1}\|_q^{q-1}} \cdot R$$

where $1/p + 1/q = 1$. We discuss in Appendix K that faster rates are possible under mild conditions. Though not covering all cases, it still showcases that the momentum is partially helpful for parameter-free FW algorithms.

**Beyond $\ell_p$ norm balls.** In general, when a specific structure of $\mathbf{x}^*$ (e.g., sparsity) is promoted by $\mathcal{X}$ (so that $\mathbf{x}^*$ is likely to live on the boundary), and one can ensure the uniqueness of $\mathbf{v}_k$ through either a closed-form solution or a specific implementation, acceleration can be effected. A direct extension of the results in this subsection to matrix space is when the constraint is a Schatten $\ell_p$ norm ball. This is because $\|\mathbf{X}\|_p := \|\sigma_1(\mathbf{X}), \sigma_2(\mathbf{X}), \ldots, \sigma_r(\mathbf{X})\|_p$, where $\sigma_i(\mathbf{X})$ denotes the $i$th singular value of $\mathbf{X}$. Our numerical results confirm the acceleration in Section V-B.

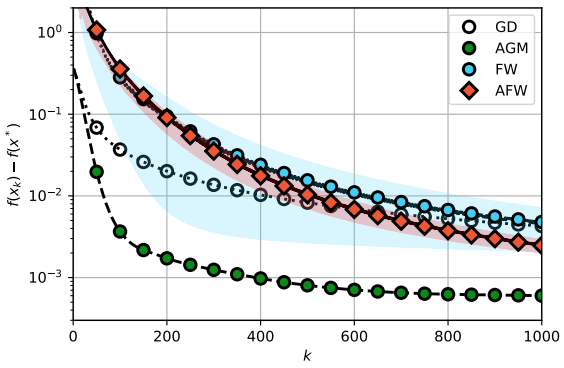



Fig. 2. Performance of AFW when the optimal solution is at interior.

## V. NUMERICAL TESTS

We validate our theoretical findings as well as the efficiency of AFW on two benchmarked machine learning problems, binary classification and matrix completion in this section. All numerical experiments are performed using Python 3.7 on a desktop equipped with Intel i7-4790 CPU @3.60 GHz (32 GB RAM). Additional numerical tests using other loss functions and constraints can be found in Appendix L.

### A. Binary classification

Logistic regression for binary classification is adopted to test AFW. The objective function is

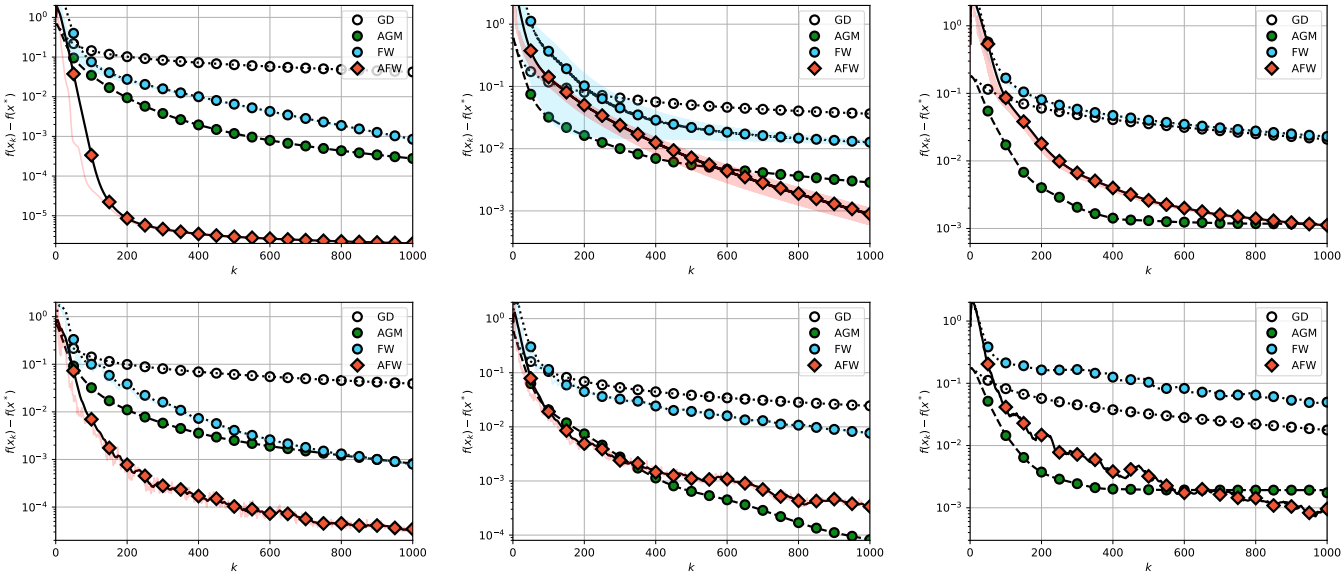
$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \ln\big(1 + \exp(-b_i \langle \mathbf{a}_i, \mathbf{x} \rangle)\big) \tag{10}$$

where $(\mathbf{a}_i, b_i)$ is the (feature, label) pair of datum $i$ and $n$ is the total number of data samples. Datasets from LIBSVM[2] are used in the numerical tests presented. Details regarding the datasets are summarized in Table II, where $d$ is the dimension of $\mathbf{x}$, $n$ is the number of data, and 'nonzeros' refers to the percentage of nonzero entries in $\{\mathbf{a}_i\}_{i=1}^n$ to reflect the sparsity of the dataset. The constraint sets considered include $\ell_1$ and $\ell_2$ norm balls. As benchmarks, the chosen algorithms are: projected GD with the standard step size $\frac{1}{L}$; parameter-free FW with step size $\frac{2}{k+2}$ [7]; and projected AGM with parameters according to [4]. The step size of AFW is $\delta_k = \frac{2}{k+3}$ according to Theorems 2 and 3. Note that both GD and AGM are not parameter-free.

We first let $\mathcal{X}$ be an $\ell_2$ norm ball with a large enough radius so that $\|\nabla f(\mathbf{x}^*)\| \approx 10^{-4}$. This case maps to our result in Theorem 2, where the convergence rate of AFW is $\mathcal{O}(\frac{1}{k})$. The performance of AFW is shown in Fig. 2. On dataset *a9a*, AFW slightly outperforms GD and FW, but is slower than AGM. Evidently, AFW is much more *stable* than FW, as one can see from the shaded areas that illustrate the zig-zag range.

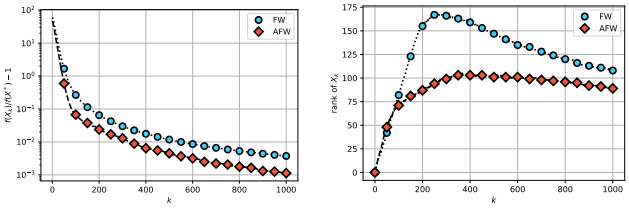
Next, we consider active $\ell_2$ norm ball constraints, where the diameter of $\mathcal{X}$ is chosen to maximize the generalization error on the validation dataset. In this case, our result in Theorem 3 applies and AFW achieves an $\tilde{\mathcal{O}}(\frac{1}{k^2})$ convergence rate. The performance of AFW is listed in the first row of Fig. 3. In all

[2]https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html.




(a) *mushroom* (b) *mnist* (c) *covtype*

Fig. 3. Performance of AFW on $\ell_2$ norm balls (first row) and $\ell_1$ norm balls (second row).




(a) optimality (b) rank

Fig. 4. Performance of AFW for matrix completion problems.

tested datasets, AFW significantly improves over FW, while on datasets other than *covtype*, AFW also outperforms AGM, especially on *mushroom*.

When the constraint set is an $\ell_1$ norm ball, the performance of AFW is depicted in the second row of Fig. 3. It can be seen that on datasets such as *covtype* and *mnist*, AFW exhibits performance similar to AGM, which is significantly faster than FW. While on dataset *mushroom*, AFW converges even faster than AGM. Note that comparing AFW with AGM is not fair since each FW step requires $d$ operations at most, while projection onto an $\ell_1$ norm ball in [46] takes $cd$ operations for some $c > 1$. This means that for the same running time, AFW will run more iterations than AGM. We stick to this unfair comparison to highlight how the optimality error of AFW and AGM evolves with $k$.

### B. Matrix completion

We then consider matrix completion problems that are ubiquitous in recommender systems. Consider a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with partially observed entries, that is, entries $A_{ij}$ for $(i,j) \in \mathcal{K}$ are known, where $\mathcal{K} \subset \{1,\ldots,m\} \times \{1,\ldots,n\}$. Note that the observed entries can also be contaminated by noise. The task is to predict the unobserved entries of $\mathbf{A}$. Although this problem can be approached in several ways, within the scope of recommender systems, a commonly adopted empirical observation is that $\mathbf{A}$ is low rank [47]–[49]. Hence the problem to be solved is

$$\min_{\mathbf{X}} \quad \frac{1}{2}\sum_{(i,j)\in\mathcal{K}} (X_{ij} - A_{ij})^2 \quad \text{s.t.} \quad \|\mathbf{X}\|_* \leq R \tag{11}$$

where $\|\mathbf{X}\|_*$ denotes the nuclear norm of $\mathbf{X}$, and it is leveraged to promote a low rank solution. Problem (11) is difficult to be solved via GD or AGM because projection onto a nuclear norm ball is expensive. On the contrary, FW and its variants are more suitable for (11) given that FW step can be solved easily and the update promotes low-rank solution directly [1].

We test AFW and FW on a widely used dataset, *MovieLens100K*[3], where 1682 movies are rated by 943 users with $6.30\%$ percent ratings observed. And the initialization and data processing are the same as those used in [1]. The numerical performance can be found in Fig. 4. In subfigures (a) and (b), we plot the optimality error and rank versus $k$ choosing $R = 3$. The choice of $R$ is based on the number of different movie categories. It is observed that AFW exhibits improvement in terms of both optimality error and rank of the solution. In particular, AFW roughly achieves $1.4$x performance improvement compared with FW in terms of optimality error, and finds solutions with much lower rank.

## VI. Conclusions

We built links between the momentum in AGM and the FW step by observing that they are both minimizing an (approximated) lower bound of the objective function. Exploring this link, we show how momentum benefits parameter-free FW. In particular, a momentum variant of FW, which we term AFW, was proved to achieve a faster rate on active $\ell_p$ norm ball constraints while maintaining the same convergence rate as

[3]Online available at https://grouplens.org/datasets/movielens/100k/

FW on general problems. AFW thus strictly outperforms FW providing the possibility for acceleration. Numerical experiments validate our theoretical findings, and suggest AFW is promising for binary classification and matrix completion.

## APPENDIX

### A. Proof of Theorem 1

The convergence on $\mathbf{x}_k$ is given in [41], and hence we do not repeat here. Next we show the behavior of $\mathbf{y}_k$ and $\mathbf{v}_k$.

We use the same surrogate functions with those in [41], i.e.,

$$\Phi_0(\mathbf{x}) = \Phi_0^* + \frac{\mu_0}{2}\|\mathbf{x}-\mathbf{x}_0\|^2 \tag{12a}$$

$$\Phi_{k+1}(\mathbf{x}) = (1-\delta_k)\Phi_k(\mathbf{x}) + \delta_k\Big[f(\mathbf{y}_k) + \big\langle \nabla f(\mathbf{y}_k), \mathbf{x}-\mathbf{y}_k\big\rangle\Big],\ \forall\, k\geq 0. \tag{12b}$$

In [41], it is shown that with $\lambda_0 = 1$ and $\lambda_k = \lambda_{k-1}(1-\delta_{k-1})$, the tuple $\big(\{\Phi_k(\mathbf{x})\}_{k=0}^{\infty}, \{\lambda_k\}_{k=0}^{\infty}\big)$ is an ES of $f(\mathbf{x})$. In addition, it is also shown that $\Phi_{k+1}(\mathbf{x})$ can be rewritten as $\Phi_k(\mathbf{x}) = \Phi_k^* + \frac{\mu_k}{2}\|\mathbf{x}-\mathbf{v}_k\|^2$, where $\mu_{k+1} = (1-\delta_k)\mu_k$, and $f(\mathbf{x}_k) \leq \Phi_k^* = \min_{\mathbf{x}} \Phi_k(\mathbf{x})$. We will use these conclusions directly. Rearranging the terms in $\Phi_k(\mathbf{x}) = \Phi_k^* + \frac{\mu_k}{2}\|\mathbf{x}-\mathbf{v}_k\|^2$, we arrive at

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{x}-\mathbf{v}_k\|^2 = \frac{1}{\mu_k}\Big(\Phi_k(\mathbf{x}) - \Phi_k^*\Big)\\
&= \frac{1}{\mu_k}\Big(\Phi_k(\mathbf{x}) - f(\mathbf{x}) + f(\mathbf{x}) - \Phi_k^*\Big)\\
&\overset{(a)}{\leq} \frac{\lambda_k}{\mu_k}\big[\Phi_0(\mathbf{x}) - f(\mathbf{x})\big] + \frac{1}{\mu_k}\big[f(\mathbf{x}) - f(\mathbf{x}_k)\big]\\
&= \frac{1}{2L}\big[\Phi_0(\mathbf{x}) - f(\mathbf{x})\big] + \frac{1}{\mu_k}\big[f(\mathbf{x}) - f(\mathbf{x}_k)\big]
\end{aligned}$$

where (a) is because $\Phi_k(\mathbf{x}) - f(\mathbf{x}) \leq \lambda_k\big(\Phi_0(\mathbf{x}) - f(\mathbf{x})\big)$ by Definition 1, and $f(\mathbf{x}_k) \leq \Phi_k^*$ shown in [3]. Choosing $\mathbf{x}$ as $\mathbf{x}^*$, we arrive at

$$\begin{aligned}
&\frac{1}{2}\|\mathbf{x}^*-\mathbf{v}_k\|^2\\
&\leq \frac{1}{2L}\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big] - \frac{1}{\mu_k}\big[f(\mathbf{x}_k) - f(\mathbf{x}^*)\big]\\
&\leq \frac{1}{2L}\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big],\ \forall k.
\end{aligned}$$

This further implies

$$\|\mathbf{x}^*-\mathbf{v}_k\|^2 \leq \frac{1}{L}\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big],\ \forall k. \tag{13}$$

Hence the behavior of $\mathbf{v}_k$ in Theorem 1 is proved.

To prove the convergence of $\mathbf{y}_k$, the following inequality is true as a result of (13)

$$\begin{aligned}
\|\mathbf{v}_{k+1}-\mathbf{v}_k\| &\leq \|\mathbf{v}_{k+1}-\mathbf{x}^*\| + \|\mathbf{x}^*-\mathbf{v}_k\|\\
&\leq 2\sqrt{\frac{1}{L}\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big]}.
\end{aligned}$$

Next, we link $\nabla f(\mathbf{y}_k)$ and $\mathbf{v}_{k+1}-\mathbf{v}_k$ through the update $\mathbf{v}_{k+1} = \mathbf{v}_k - \frac{\delta_k}{\mu_{k+1}}\nabla f(\mathbf{y}_k)$ to get

$$\begin{aligned}
\|\mathbf{v}_{k+1}-\mathbf{v}_k\|^2 &= \frac{(k+2)^2}{4L^2}\|\nabla f(\mathbf{y}_k)\|^2\\
&\leq \frac{4}{L}\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big],\ \forall k.
\end{aligned}$$

Rearranging the terms we can obtain the convergence of $\|\nabla f(\mathbf{y}_k)\|^2$, that is,

$$\|\nabla f(\mathbf{y}_k)\|^2 \leq \frac{16L}{(k+2)^2}\big[\Phi_0(\mathbf{x}^*) - f(\mathbf{x}^*)\big].$$

Plugging $\Phi_0(\mathbf{x}^*) = f(\mathbf{x}_0) + L\|\mathbf{x}_0-\mathbf{x}^*\|^2$ in completes the proof.

### B. $f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1}-\mathbf{y}_k\rangle$ approximates $f(\mathbf{x}^*)$

We show next that a weighted version of $f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1}-\mathbf{y}_k\rangle$ is no larger then $f(\mathbf{x}^*) + \mathcal{O}(\frac{1}{k^2})$ to elaborate that $f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1}-\mathbf{y}_k\rangle$ is (almost) an under-estimate of $f(\mathbf{x}^*)$.

**Theorem 4.** *If Assumptions 1 and 2 hold, and we choose $\frac{\mu_{k+1}}{\delta_k} = \frac{2L}{k+2}$; and per iteration $k$, we let $w_k^{(\tau)} = \frac{2(\tau+2)}{k(k+3)}$ for $\tau = 0, 1, \ldots, k-1$, then i) $\sum_{\tau=0}^{k-1} w_k^{(\tau)} = 1$; and, ii)*

$$\begin{aligned}
&\sum_{\tau=0}^{k-1} w_k^{(\tau)}\Big[f(\mathbf{y}_\tau) \quad + \langle \nabla f(\mathbf{y}_\tau), \mathbf{v}_{\tau+1}-\mathbf{y}_\tau\rangle\Big] - f(\mathbf{x}^*)\\
&\leq \frac{2L\|\mathbf{x}_0-\mathbf{x}^*\|^2}{k(k+3)}.
\end{aligned}$$

*Proof.* It is easy to verify that $\sum_{\tau=0}^{k-1} w_k^{(\tau)} = 1$. Next we have

$$\begin{aligned}
&f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1}-\mathbf{y}_k\rangle\\
&= f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1}-\mathbf{x}^*\rangle + \langle \nabla f(\mathbf{y}_k), \mathbf{x}^*-\mathbf{y}_k\rangle\\
&\overset{(a)}{\leq} f(\mathbf{x}^*) + \langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1}-\mathbf{x}^*\rangle\\
&= f(\mathbf{x}^*) + \frac{\mu_{k+1}}{\delta_k}\langle \mathbf{v}_k-\mathbf{v}_{k+1}, \mathbf{v}_{k+1}-\mathbf{x}^*\rangle\\
&\overset{(b)}{=} f(\mathbf{x}^*) + \frac{\mu_{k+1}}{2\delta_k}\Big[\|\mathbf{x}^*-\mathbf{v}_k\|^2\\
&\qquad - \|\mathbf{x}^*-\mathbf{v}_{k+1}\|^2 - \|\mathbf{v}_{k+1}-\mathbf{v}_k\|^2\Big]\\
&\overset{(c)}{=} f(\mathbf{x}^*) + \frac{L}{k+2}\Big[\|\mathbf{x}^*-\mathbf{v}_k\|^2\\
&\qquad - \|\mathbf{x}^*-\mathbf{v}_{k+1}\|^2 - \|\mathbf{v}_{k+1}-\mathbf{v}_k\|^2\Big]
\end{aligned} \tag{14}$$

where (a) follows from the convexity of $f$, that is, $\langle \nabla f(\mathbf{y}_k), \mathbf{x}^*-\mathbf{y}_k\rangle \leq f(\mathbf{x}^*) - f(\mathbf{y}_k)$; (b) uses $2\langle \mathbf{a}, \mathbf{b}\rangle = \|\mathbf{a}+\mathbf{b}\|^2 - \|\mathbf{a}\|^2 - \|\mathbf{b}\|^2$; and (c) is by plugging the value of $\frac{\mu_{k+1}}{\delta_k}$ in. Now, if we define $d_k := f(\mathbf{y}_k) + \langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1}-\mathbf{y}_k\rangle - f(\mathbf{x}^*)$, rearranging (14), we get

$$\begin{aligned}
&(k+2)d_k\\
&\leq L\Big[\|\mathbf{x}^*-\mathbf{v}_k\|^2 - \|\mathbf{x}^*-\mathbf{v}_{k+1}\|^2\Big] - L\|\mathbf{v}_{k+1}-\mathbf{v}_k\|^2\\
&\leq L\Big[\|\mathbf{x}^*-\mathbf{v}_k\|^2 - \|\mathbf{x}^*-\mathbf{v}_{k+1}\|^2\Big]
\end{aligned}$$

Summing over $k$ (and recalling $\mathbf{v}_0 = \mathbf{x}_0$), we arrive at

$$\sum_{\tau=0}^{k-1}(\tau+2)d_\tau \leq L\Big[\|\mathbf{x}^*-\mathbf{v}_0\|^2 - \|\mathbf{x}^*-\mathbf{v}_k\|^2\Big] \leq L\|\mathbf{x}^*-\mathbf{x}_0\|^2.$$

By the definition of $w_k^{(\tau)}$, which is $w_k^{(\tau)} = \frac{2(\tau+2)}{k(k+3)}$, we obtain

$$\sum_{\tau=0}^{k-1} w_k^{(\tau)} d_\tau \leq \frac{2L\|\mathbf{x}^*-\mathbf{x}_0\|^2}{k(k+3)} \tag{15}$$

which completes the proof. □

### C. AGM Links with FW in strongly convex case

We showcase the connection between the momentum update of AGM in strongly convex case and FW. We first formally define strong convexity, which is used in this subsection only.

**Assumption 5.** *(Strong convexity.) The function $f:\mathbb{R}^d\to\mathbb{R}$ is $\mu$-strongly convex; that is, $f(\mathbf{y})-f(\mathbf{x})\geq\langle\nabla f(\mathbf{x}),\mathbf{y}-\mathbf{x}\rangle+\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2,\ \forall\,\mathbf{x},\mathbf{y}\in\mathbb{R}^d$.*

Under Assumptions 1 and 5, the condition number of $f$ is $\kappa:=\frac{L}{\mu}$. To cope with strongly convex problems, Lines 4 – 6 in AGM (Alg. 2) should be modified to [3]

$$\mathbf{y}_k=\frac{1}{1+\delta}\mathbf{x}_k+\frac{\delta}{1+\delta}\mathbf{v}_k \tag{16a}$$

$$\mathbf{x}_{k+1}=\mathbf{y}_k-\frac{1}{L}\nabla f(\mathbf{y}_k) \tag{16b}$$

$$\mathbf{v}_{k+1}=(1-\delta)\mathbf{v}_k+\delta\mathbf{y}_k-\frac{\delta}{\mu}\nabla f(\mathbf{y}_k). \tag{16c}$$

where $\delta=\frac{1}{\sqrt{\kappa}}$. Here $\mathbf{v}_{k+1}$ in (16c) denotes the momentum and thus plays the critical role for acceleration. To see how $\mathbf{v}_{k+1}$ is linked with FW, we will rewrite $\mathbf{v}_{k+1}$ as

$$\begin{aligned}\mathbf{z}_{k+1}&=\arg\min_{\mathbf{x}} f(\mathbf{y}_k)+\langle\nabla f(\mathbf{y}_k),\mathbf{x}-\mathbf{y}_k\rangle+\frac{\mu}{2}\|\mathbf{x}-\mathbf{y}_k\|^2\\&=\mathbf{y}_k-\frac{1}{\mu}\mathbf{y}_k\end{aligned} \tag{17a}$$

$$\mathbf{v}_{k+1}=(1-\delta)\mathbf{v}_k+\delta\mathbf{z}_{k+1} \tag{17b}$$

Notice that $\mathbf{z}_{k+1}$ is the minimizer of a lower bound of $f(\mathbf{x})$ (due to strongly convexity). Therefore, the $\mathbf{v}_{k+1}$ update is similar to FW in the sense that it first minimizes a lower bound of $f(\mathbf{x})$, then update through convex combination (cf Alg. 1). This demonstrates that the momentum update in AGM shares the same idea of FW update.

A few basic lemmas for all the proofs in Section IV are provided below.

### D. Proof of Lemma 1.

*Proof.* We show this by induction. Because $\lambda_0=1$, it holds that $\Phi_0(\mathbf{x})=(1-\lambda_0)f(\mathbf{x})+\lambda_0\Phi_0(\mathbf{x})=\Phi_0(\mathbf{x})$. Suppose that $\Phi_k(\mathbf{x})\leq(1-\lambda_k)f(\mathbf{x})+\lambda_k\Phi_0(\mathbf{x})$ is true for some $k$. We have

$$\begin{aligned}\Phi_{k+1}(\mathbf{x})&=(1-\delta_k)\Phi_k(\mathbf{x})+\delta_k\Big[f(\mathbf{y}_k)+\big\langle\nabla f(\mathbf{y}_k),\mathbf{x}-\mathbf{y}_k\big\rangle\Big]\\&\overset{(a)}{\leq}(1-\delta_k)\Phi_k(\mathbf{x})+\delta_k f(\mathbf{x})\\&\leq(1-\delta_k)\Big[(1-\lambda_k)f(\mathbf{x})+\lambda_k\Phi_0(\mathbf{x})\Big]+\delta_k f(\mathbf{x})\\&=(1-\lambda_{k+1})f(\mathbf{x})+\lambda_{k+1}\Phi_0(\mathbf{x})\end{aligned}$$

where (a) is because the convexity of $f$; and the last equation is by definition of $\lambda_{k+1}$. Together with the fact that $\lim_{k\to\infty}\lambda_k=0$, the tuple $\big(\{\Phi_k(\mathbf{x})\}_{k=0}^{\infty},\{\lambda_k\}_{k=0}^{\infty}\big)$ satisfies the definition of an estimate sequence. □

### E. A few useful lemmas.

**Lemma 3.** *For $\{\Phi_k(\mathbf{x})\}$ in (7), if $f(\mathbf{x}_k)\leq\min_{\mathbf{x}\in\mathcal{X}}\Phi_k(\mathbf{x})+\xi_k$, it is true that*

$$f(\mathbf{x}_k)-f(\mathbf{x}^*)\leq\lambda_k\big(f(\mathbf{x}_0)-f(\mathbf{x}^*)\big)+\xi_k,\ \forall\,k.$$

*Proof.* If $f(\mathbf{x}_k)\leq\min_{\mathbf{x}\in\mathcal{X}}\Phi_k(\mathbf{x})+\xi_k$ holds, then we have

$$\begin{aligned}f(\mathbf{x}_k)&\leq\min_{\mathbf{x}\in\mathcal{X}}\Phi_k(\mathbf{x})+\xi_k\leq\Phi_k(\mathbf{x}^*)+\xi_k\\&\leq(1-\lambda_k)f(\mathbf{x}^*)+\lambda_k\Phi_0(\mathbf{x}^*)+\xi_k\end{aligned}$$

where the last inequality is because Definition 1. Subtracting $f(\mathbf{x}^*)$ on both sides, we arrive at

$$\begin{aligned}f(\mathbf{x}_k)-f(\mathbf{x}^*)&\leq\lambda_k\big(\Phi_0(\mathbf{x}^*)-f(\mathbf{x}^*)\big)+\xi_k\\&=\lambda_k\big(f(\mathbf{x}_0)-f(\mathbf{x}^*)\big)+\xi_k\end{aligned}$$

which completes the proof. □

**Lemma 4.** *Let $\mathbf{v}_0=\mathbf{x}_0$, $\boldsymbol{\theta}_0=\mathbf{0}$, $\Phi_0^*=f(\mathbf{x}_0)$, then $\Phi_{k+1}(\mathbf{x})$ in (7) can be rewritten as*

$$\Phi_{k+1}(\mathbf{x})=\Phi_{k+1}^*+\langle\mathbf{x}-\mathbf{v}_{k+1},\boldsymbol{\theta}_{k+1}\rangle \tag{18}$$

*with*

$$\boldsymbol{\theta}_{k+1}=\delta_k\nabla f(\mathbf{y}_k)+(1-\delta_k)\boldsymbol{\theta}_k \tag{19a}$$

$$\mathbf{v}_{k+1}:=\arg\min_{\mathbf{x}\in\mathcal{X}}\Phi_{k+1}(\mathbf{x})=\arg\min_{\mathbf{x}\in\mathcal{X}}\langle\mathbf{x},\boldsymbol{\theta}_{k+1}\rangle \tag{19b}$$

$$\begin{aligned}\Phi_{k+1}^*:=\min_{\mathbf{x}\in\mathcal{X}}\Phi_{k+1}(\mathbf{x})&=\Phi_{k+1}(\mathbf{v}_{k+1})\\&=(1-\delta_k)\Phi_k^*+\delta_k f(\mathbf{y}_k)+(1-\delta_k)\langle\boldsymbol{\theta}_k,\mathbf{v}_{k+1}-\mathbf{v}_k\rangle\\&\quad+\delta_k\langle\nabla f(\mathbf{y}_k),\mathbf{v}_{k+1}-\mathbf{y}_k\rangle.\end{aligned} \tag{19c}$$

*Proof.* We prove this lemma by induction. First $\Phi_0(\mathbf{x})=\Phi_0^*+\langle\mathbf{x}-\mathbf{v}_0,\boldsymbol{\theta}_0\rangle\equiv f(\mathbf{x}_0)$. From (7) it is obvious that $\Phi_k(\mathbf{x})$ is linear in $\mathbf{x}$, and hence suppose that $\Phi_k(\mathbf{x})=\Phi_k^*+\langle\mathbf{x}-\mathbf{v}_k,\boldsymbol{\theta}_k\rangle$ holds for some $k$. Then we will show that $\Phi_{k+1}(\mathbf{x})=\Phi_{k+1}^*+\langle\mathbf{x}-\mathbf{v}_{k+1},\boldsymbol{\theta}_{k+1}\rangle$ is true. Consider that

$$\begin{aligned}&\Phi_{k+1}(\mathbf{x})\\&=(1-\delta_k)\Phi_k(\mathbf{x})+\delta_k\Big[f(\mathbf{y}_k)+\big\langle\nabla f(\mathbf{y}_k),\mathbf{x}-\mathbf{y}_k\big\rangle\Big]\\&=(1-\delta_k)\Phi_k^*+(1-\delta_k)\langle\mathbf{x}-\mathbf{v}_k,\boldsymbol{\theta}_k\rangle+\delta_k f(\mathbf{y}_k)\\&\qquad+\delta_k\big\langle\nabla f(\mathbf{y}_k),\mathbf{x}-\mathbf{y}_k\big\rangle\\&=(1-\delta_k)\Phi_k^*+\delta_k f(\mathbf{y}_k)+\big\langle\mathbf{x},(1-\delta_k)\boldsymbol{\theta}_k+\delta_k\nabla f(\mathbf{y}_k)\big\rangle\\&\qquad-(1-\delta_k)\langle\mathbf{v}_k,\boldsymbol{\theta}_k\rangle-\delta_k\big\langle\nabla f(\mathbf{y}_k),\mathbf{y}_k\big\rangle.\end{aligned} \tag{20}$$

Clearly, since $\Phi_{k+1}(\mathbf{x})$ is linear in $\mathbf{x}$, the slope is $\boldsymbol{\theta}_{k+1}:=(1-\delta_k)\boldsymbol{\theta}_k+\delta_k\nabla f(\mathbf{y}_k)$. In addition, because $\mathbf{v}_{k+1}$ is defined as the minimizer of $\Phi_{k+1}(\mathbf{x})$ over $\mathcal{X}$, from (20) we have $\mathbf{v}_{k+1}=\arg\min_{\mathbf{x}\in\mathcal{X}}\langle\mathbf{x},\boldsymbol{\theta}_{k+1}\rangle$. Then, since $\Phi_{k+1}^*$ is defined as $\Phi_{k+1}^*:=\min_{\mathbf{x}\in\mathcal{X}}\Phi_{k+1}(\mathbf{x})$, by plugging $\mathbf{v}_{k+1}$ into $\Phi_{k+1}(\mathbf{x})$ in (20), we have

$$\begin{aligned}\Phi_{k+1}^*=\Phi_{k+1}(\mathbf{v}_{k+1})&=(1-\delta_k)\langle\mathbf{v}_{k+1}-\mathbf{v}_k,\boldsymbol{\theta}_k\rangle\\&\quad+(1-\delta_k)\Phi_k^*+\delta_k f(\mathbf{y}_k)+\delta_k\big\langle\nabla f(\mathbf{y}_k),\mathbf{v}_{k+1}-\mathbf{y}_k\big\rangle.\end{aligned}$$

The proof is thus completed. □

### F. Proof of Lemma 2.

*Proof.* We prove this lemma by induction. First by definition $f(\mathbf{x}_0)=\Phi_0^*+\xi_0$. Suppose now we have $f(\mathbf{x}_k)\leq\Phi_k^*+\xi_k$ for some $k$. Next, we will show that $f(\mathbf{x}_{k+1})\leq\Phi_{k+1}^*+\xi_{k+1}$.

Using (19c), we have

$$\begin{aligned}
&\Phi^*_{k+1} + (1-\delta_k)\xi_k \\
&= (1-\delta_k)\Phi^*_k + \delta_k f(\mathbf{y}_k) + (1-\delta_k)\langle \boldsymbol{\theta}_k, \mathbf{v}_{k+1} - \mathbf{v}_k\rangle \\
&\qquad + \delta_k\langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1} - \mathbf{y}_k\rangle + (1-\delta_k)\xi_k \\
&\overset{(a)}{\geq} (1-\delta_k) f(\mathbf{x}_k) + \delta_k f(\mathbf{y}_k) + (1-\delta_k)\langle \boldsymbol{\theta}_k, \mathbf{v}_{k+1} - \mathbf{v}_k\rangle \\
&\qquad + \delta_k\langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1} - \mathbf{y}_k\rangle \\
&\overset{(b)}{\geq} (1-\delta_k) f(\mathbf{x}_k) + \delta_k f(\mathbf{y}_k) + \delta_k\langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1} - \mathbf{y}_k\rangle \\
&= f(\mathbf{y}_k) + (1-\delta_k)\big[f(\mathbf{x}_k) - f(\mathbf{y}_k)\big] \\
&\qquad + \delta_k\langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1} - \mathbf{y}_k\rangle \\
&\overset{(c)}{\geq} f(\mathbf{y}_k) + (1-\delta_k)\big\langle \nabla f(\mathbf{y}_k), \mathbf{x}_k - \mathbf{y}_k\big\rangle \\
&\qquad + \delta_k\langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1} - \mathbf{y}_k\rangle \\
&\overset{(d)}{\geq} f(\mathbf{x}_{k+1}) - \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{y}_k\|^2 + \langle \nabla f(\mathbf{y}_k), \mathbf{y}_k - \mathbf{x}_{k+1}\rangle \\
&\quad + (1-\delta_k)\big\langle \nabla f(\mathbf{y}_k), \mathbf{x}_k - \mathbf{y}_k\big\rangle + \delta_k\langle \nabla f(\mathbf{y}_k), \mathbf{v}_{k+1} - \mathbf{y}_k\rangle \\
&\overset{(e)}{=} f(\mathbf{x}_{k+1}) - \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{y}_k\|^2
\end{aligned}$$

where (a) is because $\Phi^*_k \geq f(\mathbf{x}_k) - \xi_k$; (b) is by the fact $\mathbf{v}_k = \arg\min_{\mathbf{x}\in\mathcal{X}}\langle \boldsymbol{\theta}_k, \mathbf{x}\rangle$ so that $\langle \boldsymbol{\theta}_k, \mathbf{v}_{k+1} - \mathbf{v}_k\rangle \geq 0$; (c) is because of the convexity of $f$; (d) is by Assumption 1, that is $f(\mathbf{x}_{k+1}) - f(\mathbf{y}_k) \leq \langle \nabla f(\mathbf{y}_k), \mathbf{x}_{k+1} - \mathbf{y}_k\rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{y}_k\|^2$; (e) follows from the choice of $\mathbf{x}_{k+1} = (1-\delta_k)\mathbf{x}_k + \delta_k\mathbf{v}_{k+1}$. Finally by using $\mathbf{y}_k = (1-\delta_k)\mathbf{x}_k + \delta_k\mathbf{v}_k$, and plugging the definition of $\xi_{k+1}$, the proof is completed. □

### G. Proof of Theorem 2

*Proof.* Since Lemma 2 holds, one can directly apply Lemma 3 to have

$$\begin{aligned}
f(\mathbf{x}_k) - f(\mathbf{x}^*) &\leq \lambda_k\big(f(\mathbf{x}_0) - f(\mathbf{x}^*)\big) + \xi_k \qquad (21)\\
&= \frac{2\big(f(\mathbf{x}_0) - f(\mathbf{x}^*)\big)}{(k+1)(k+2)} + \xi_k
\end{aligned}$$

where $\xi_k$ is defined in Lemma 2. Clearly, $\xi_k \geq 0,\ \forall k$, and we can find an upper bound for it in the following manner.

$$\begin{aligned}
\xi_k &= (1-\delta_{k-1})\xi_{k-1} + \frac{L\delta^2_{k-1}}{2}\|\mathbf{v}_k - \mathbf{v}_{k-1}\|^2 \\
&\leq (1-\delta_{k-1})\xi_{k-1} + \frac{LD^2\delta^2_{k-1}}{2} \\
&= \frac{LD^2}{2}\sum_{\tau=0}^{k-1}\delta^2_\tau\Big[\prod_{j=\tau+1}^{k-1}(1-\delta_j)\Big] \\
&= \frac{LD^2}{2}\sum_{\tau=0}^{k-1}\frac{4}{(\tau+3)^2}\frac{(\tau+2)(\tau+3)}{(k+1)(k+2)} \leq \frac{2LD^2}{k+2}.
\end{aligned}$$

Plugging $\xi_k$ into (21) completes the proof. □

### H. Proof of Theorem 3

The basic idea is to show that under Assumptions 1, 2, 3 and 4, $\|\mathbf{v}_k - \mathbf{v}_{k+1}\|^2$ is small enough when $k$ is large. To this end, we will make use of the following lemmas.

**Lemma 5.** *[3, Theorem 2.1.5] If Assumptions 1 and 2 hold, then it is true that*

$$\frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle.$$

Next we show that the value of $\nabla f(\mathbf{x}^*)$ is unique.

**Lemma 6.** *If both $\mathbf{x}^*_1$ and $\mathbf{x}^*_2$ minimize $f(\mathbf{x})$ over $\mathcal{X}$, then we have $\nabla f(\mathbf{x}^*_1) = \nabla f(\mathbf{x}^*_2)$.*

*Proof.* From Lemma 5, we have

$$\begin{aligned}
&\frac{1}{2L}\|\nabla f(\mathbf{x}^*_2) - \nabla f(\mathbf{x}^*_1)\|_2^2 \\
&\leq f(\mathbf{x}^*_2) - f(\mathbf{x}^*_1) - \langle \nabla f(\mathbf{x}^*_1), \mathbf{x}^*_2 - \mathbf{x}^*_1\rangle \\
&\overset{(a)}{\leq} f(\mathbf{x}^*_2) - f(\mathbf{x}^*_1) = 0
\end{aligned}$$

where (a) is by the optimality condition, that is, $\langle \nabla f(\mathbf{x}^*_1), \mathbf{x} - \mathbf{x}^*_1\rangle \geq 0,\ \forall \mathbf{x} \in \mathcal{X}$. Hence we can only have $\nabla f(\mathbf{x}^*_2) = \nabla f(\mathbf{x}^*_1)$. This means that the value of $\nabla f(\mathbf{x}^*)$ is unique regardless of the uniqueness of $\mathbf{x}^*$. □

**Lemma 7.** *Choose $\delta_k = \frac{2}{k+3}$ and let $M := \max_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) - f(\mathbf{x}^*)$, then we have*

$$\|\nabla f(\mathbf{y}_k) - \nabla f(\mathbf{x}^*)\| \leq \frac{C_1}{\sqrt{k+3}}.$$

*where $C_1 = \sqrt{6LM + 4L^2D^2}$.*

*Proof.* By convexity

$$\begin{aligned}
&f(\mathbf{y}_k) - f(\mathbf{x}^*) \\
&\leq (1-\delta_k)\big[f(\mathbf{x}_k) - f(\mathbf{x}^*)\big] + \delta_k\big[f(\mathbf{v}_k) - f(\mathbf{x}^*)\big] \\
&\overset{(a)}{\leq} \frac{k+1}{k+3}\left[\frac{2\big(f(\mathbf{x}_0) - f(\mathbf{x}^*)\big)}{(k+1)(k+2)} + \frac{2LD^2}{k+2}\right] + \frac{2M}{k+3} \\
&\leq \frac{2M}{(k+2)(k+3)} + \frac{2LD^2}{k+3} + \frac{2M}{k+3} \\
&\leq \frac{3M + 2LD^2}{k+3}
\end{aligned}$$

where (a) is by Theorem 2. Next using Lemma 5, we have

$$\begin{aligned}
&\frac{1}{2L}\|\nabla f(\mathbf{y}_k) - \nabla f(\mathbf{x}^*)\|^2 \\
&\leq f(\mathbf{y}_k) - f(\mathbf{x}^*) - \langle \nabla f(\mathbf{x}^*), \mathbf{y}_k - \mathbf{x}^*\rangle \\
&\overset{(b)}{\leq} f(\mathbf{y}_k) - f(\mathbf{x}^*) \leq \frac{3M + 2LD^2}{k+3}
\end{aligned}$$

where (b) is by the optimality condition, that is, $\langle \nabla f(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^*\rangle \geq 0,\ \forall \mathbf{x} \in \mathcal{X}$. This further implies

$$\|\nabla f(\mathbf{y}_k) - \nabla f(\mathbf{x}^*)\| \leq \sqrt{\frac{2L(3M + 2LD^2)}{k+3}}.$$

The proof is thus completed. □

**Lemma 8.** *Choose $\delta_k = \frac{2}{k+3}$, it is guaranteed to have*

$$\|\boldsymbol{\theta}_{k+1} - \nabla f(\mathbf{x}^*)\| \leq \frac{4C_1}{3(\sqrt{k+3} - 1)} + \frac{2\sqrt{G}}{(k+2)(k+3)}.$$

*In addition, there exists a constant $C_2 \leq \frac{4}{3}C_1 + \frac{2}{3(\sqrt{3}+1)}\sqrt{G}$ such that*

$$\|\boldsymbol{\theta}_{k+1} - \nabla f(\mathbf{x}^*)\| \leq \frac{C_2}{\sqrt{k+3} - 1}.$$

*Proof.* First we have

$$\begin{aligned}\boldsymbol{\theta}_{k+1} &= (1-\delta_k)\boldsymbol{\theta}_k + \delta_k \nabla f(\mathbf{y}_k) \\ &= \sum_{\tau=0}^{k} \delta_\tau \nabla f(\mathbf{y}_\tau)\Big[\prod_{j=\tau+1}^{k}(1-\delta_j)\Big] \\ &= \sum_{\tau=0}^{k} \frac{2(\tau+2)}{(k+2)(k+3)}\nabla f(\mathbf{y}_\tau).\end{aligned} \tag{22}$$

Noticing that $2\sum_{\tau=0}^{k}(\tau+2) = (k+1)(k+4) = (k+2)(k+3)-2$, we have

$$\begin{aligned}&\|\boldsymbol{\theta}_{k+1} - \nabla f(\mathbf{x}^*)\| \\ &= \Big\| \sum_{\tau=0}^{k} \frac{2(\tau+2)}{(k+2)(k+3)}\big[\nabla f(\mathbf{y}_\tau) - \nabla f(\mathbf{x}^*)\big] \\ &\qquad\qquad - \frac{2}{(k+2)(k+3)}\nabla f(\mathbf{x}^*)\Big\| \\ &\le \sum_{\tau=0}^{k} \frac{2(\tau+2)}{(k+2)(k+3)}\big\|\nabla f(\mathbf{y}_\tau) - \nabla f(\mathbf{x}^*)\big\| \\ &\qquad\qquad + \frac{2}{(k+2)(k+3)}\big\|\nabla f(\mathbf{x}^*)\big\| \\ &\overset{(a)}{\le} \sum_{\tau=0}^{k} \frac{2(\tau+2)}{(k+2)(k+3)}\frac{C_1}{\sqrt{\tau+3}} + \frac{2\sqrt{G}}{(k+2)(k+3)} \\ &\le \frac{2C_1}{(k+2)(k+3)}\sum_{\tau=0}^{k}\sqrt{\tau+2} + \frac{2\sqrt{G}}{(k+2)(k+3)} \\ &\le \frac{4C_1}{3(k+2)(k+3)}(k+3)^{3/2} + \frac{2\sqrt{G}}{(k+2)(k+3)} \\ &= \frac{4C_1}{3(\sqrt{k+3}+1)(\sqrt{k+3}-1)}\sqrt{k+3} + \frac{2\sqrt{G}}{(k+2)(k+3)} \\ &\le \frac{4C_1}{3(\sqrt{k+3}-1)} + \frac{2\sqrt{G}}{(k+2)(k+3)}\end{aligned}$$

where (a) follows from Lemma 7 and Assumption 4.

Then to find $C_2$, we have

$$\begin{aligned}&\|\boldsymbol{\theta}_{k+1} - \nabla f(\mathbf{x}^*)\| \\ &\le \frac{4C_1}{3(\sqrt{k+3}-1)} + \frac{2\sqrt{G}}{(k+2)(k+3)} \\ &= \frac{4C_1}{3(\sqrt{k+3}-1)} + \frac{2\sqrt{G}}{(k+3)(\sqrt{k+3}+1)(\sqrt{k+3}-1)} \\ &\overset{(b)}{\le} \frac{4C_1}{3(\sqrt{k+3}-1)} + \frac{2\sqrt{G}}{3(\sqrt{3}+1)(\sqrt{k+3}-1)}\end{aligned}$$

where in (b) we use $k+3 \ge 3$ and $\sqrt{k+3}+1 \ge \sqrt{3}+1$. The proof is thus completed. □

**Lemma 9.** *There exists a constant $T \le \big(\frac{2C_2}{\sqrt{G}}+1\big)^2 - 3$, such that $\|\boldsymbol{\theta}_{k+1}\| \ge \frac{\sqrt{G}}{2}$, $\forall k \ge T$. In addition, it is guaranteed to have for any $k \ge T+1$*

$$\|\mathbf{v}_{k+1} - \mathbf{v}_k\| \le \frac{C_3}{\sqrt{k+2}-1}$$

*where $C_3 \le \frac{4R}{G}\big[4\sqrt{G}C_2 + \frac{2C_2^2}{\sqrt{T+4}-1}\big]$.*

*Proof.* Consider a specific $\tilde{k}$ with $\|\boldsymbol{\theta}_{\tilde{k}+1}\| < \frac{\sqrt{G}}{2}$ satisfied. In this case we have

$$\|\boldsymbol{\theta}_{\tilde{k}+1} - \nabla f(\mathbf{x}^*)\| \ge \|\nabla f(\mathbf{x}^*)\| - \|\boldsymbol{\theta}_{\tilde{k}+1}\| > \sqrt{G} - \frac{\sqrt{G}}{2} = \frac{\sqrt{G}}{2}.$$

From Lemma 8, we have

$$\frac{\sqrt{G}}{2} < \|\boldsymbol{\theta}_{\tilde{k}+1} - \nabla f(\mathbf{x}^*)\| \le \frac{C_2}{\sqrt{\tilde{k}+3}-1}.$$

From this inequality we can observe that $\|\boldsymbol{\theta}_{\tilde{k}+1}\|$ can be less than $\frac{\sqrt{G}}{2}$ only when $\tilde{k} < T = \big(\frac{2C_2}{\sqrt{G}}+1\big)^2 - 3$. Hence, the first part of this lemma is proved.

For the upper bound of $\|\mathbf{v}_{k+1} - \mathbf{v}_k\|$, we only consider the case where $\boldsymbol{\theta}_{k+1} \ne \mathbf{0}$ since otherwise $\mathbf{v}_{k+1} = \mathbf{v}_k$ and the lemma holds automatically. For any $k \ge T+1$, from (8), one can rewrite

$$\begin{aligned}&\|\mathbf{v}_{k+1} - \mathbf{v}_k\| \\ &= R\Big\|\frac{\boldsymbol{\theta}_{k+1}}{\|\boldsymbol{\theta}_{k+1}\|} - \frac{\boldsymbol{\theta}_k}{\|\boldsymbol{\theta}_k\|}\Big\| \\ &= \frac{R}{\|\boldsymbol{\theta}_{k+1}\|\|\boldsymbol{\theta}_k\|}\Big\|\|\boldsymbol{\theta}_k\|\boldsymbol{\theta}_{k+1} - \|\boldsymbol{\theta}_{k+1}\|\boldsymbol{\theta}_k\Big\| \\ &\overset{(a)}{\le} \frac{4R}{G}\Big\|\|\boldsymbol{\theta}_k\|\boldsymbol{\theta}_{k+1} - \|\boldsymbol{\theta}_{k+1}\|\boldsymbol{\theta}_k\Big\|\end{aligned} \tag{23}$$

where (a) is by $\boldsymbol{\theta}_k \ge \frac{\sqrt{G}}{2}$ for $k \ge T+1$. Next we rewrite $\boldsymbol{\theta}_k := \nabla f(\mathbf{x}^*) + \boldsymbol{\gamma}_k$. From Lemma 8 we have $\|\boldsymbol{\gamma}_k\| = \|\boldsymbol{\theta}_k - \nabla f(\mathbf{x}^*)\| \le \frac{C_2}{\sqrt{k+2}-1}$. Using this relation, the RHS of (23) becomes

$$\begin{aligned}&\Big\|\|\boldsymbol{\theta}_k\|\boldsymbol{\theta}_{k+1} - \|\boldsymbol{\theta}_{k+1}\|\boldsymbol{\theta}_k\Big\| \\ &= \Big\|\big\|\nabla f(\mathbf{x}^*) + \boldsymbol{\gamma}_k\big\|\big(\nabla f(\mathbf{x}^*) + \boldsymbol{\gamma}_{k+1}\big) \\ &\qquad - \big\|\nabla f(\mathbf{x}^*) + \boldsymbol{\gamma}_{k+1}\big\|\big(\nabla f(\mathbf{x}^*) + \boldsymbol{\gamma}_k\big)\Big\| \\ &\le \|\nabla f(\mathbf{x}^*)\|\Big\|\big\|\nabla f(\mathbf{x}^*) + \boldsymbol{\gamma}_k\big\| - \big\|\nabla f(\mathbf{x}^*) + \boldsymbol{\gamma}_{k+1}\big\|\Big\| \\ &\qquad + \Big\|\boldsymbol{\gamma}_{k+1}\big\|\nabla f(\mathbf{x}^*) + \boldsymbol{\gamma}_k\big\| - \boldsymbol{\gamma}_k\big\|\nabla f(\mathbf{x}^*) + \boldsymbol{\gamma}_{k+1}\big\|\Big\| \\ &\le \sqrt{G}\big(\|\boldsymbol{\gamma}_k\| + \|\boldsymbol{\gamma}_{k+1}\|\big) + \|\boldsymbol{\gamma}_{k+1}\|\big(\sqrt{G} + \|\boldsymbol{\gamma}_k\|\big) \\ &\qquad\qquad + \|\boldsymbol{\gamma}_k\|\big(\sqrt{G} + \|\boldsymbol{\gamma}_{k+1}\|\big) \\ &\le \frac{4\sqrt{G}C_2}{\sqrt{k+2}-1} + \frac{2C_2^2}{(\sqrt{k+2}-1)(\sqrt{k+3}-1)} \\ &\le \frac{4\sqrt{G}C_2}{\sqrt{k+2}-1} + \frac{2C_2^2}{(\sqrt{k+2}-1)(\sqrt{T+4}-1)}.\end{aligned}$$

Plugging back to (23), the proof can be completed. □

### I. Proof of Theorem 3.

*Proof.* We first consider the constraint set being an $\ell_2$ norm ball. From Lemma 2, we can write

$$\begin{aligned}
&\xi_{k+1} = (1-\delta_k)\xi_k + \frac{L\delta_k^2}{2}\|\mathbf{v}_{k+1}-\mathbf{v}_k\|^2 \\
&= \frac{L}{2}\sum_{\tau=0}^{k}\delta_\tau^2\|\mathbf{v}_{\tau+1}-\mathbf{v}_\tau\|^2\Big[\prod_{j=\tau+1}^{k}(1-\delta_\tau)\Big] \\
&\overset{(a)}{=} \frac{L}{2}\sum_{\tau=0}^{T}\delta_\tau^2\|\mathbf{v}_{\tau+1}-\mathbf{v}_\tau\|^2\Big[\prod_{j=\tau+1}^{k}(1-\delta_\tau)\Big] \\
&\qquad + \sum_{\tau=T+1}^{k}\delta_\tau^2\|\mathbf{v}_{\tau+1}-\mathbf{v}_\tau\|^2\Big[\prod_{j=\tau+1}^{k}(1-\delta_\tau)\Big] \\
&\overset{(b)}{\le} \frac{L}{2}\sum_{\tau=0}^{T}\delta_\tau^2 D^2\Big[\prod_{j=\tau+1}^{k}(1-\delta_\tau)\Big] \\
&\qquad + \sum_{\tau=T+1}^{k}\delta_\tau^2\frac{C_3^2}{(\sqrt{\tau+2}-1)^2}\Big[\prod_{j=\tau+1}^{k}(1-\delta_\tau)\Big] \\
&= \frac{L}{2}\sum_{\tau=0}^{T}\frac{4D^2}{(\tau+3)^2}\frac{(\tau+2)(\tau+3)}{(k+2)(k+3)} \\
&\qquad + \sum_{\tau=T+1}^{k}\frac{4}{(\tau+3)^2}\frac{C_3^2}{(\sqrt{\tau+2}-1)^2}\frac{(\tau+2)(\tau+3)}{(k+2)(k+3)} \\
&\le \frac{2LD^2(T+1)}{(k+2)(k+3)} + \frac{4C_3^2}{(k+2)(k+3)}\sum_{\tau=T+1}^{k}\frac{1}{(\sqrt{\tau+2}-1)^2} \\
&= \mathcal{O}\left(\frac{LD^2(T+1)+C_3^2\ln k}{(k+2)(k+3)}\right)
\end{aligned}$$

where in (a) $T$ is defined in Lemma 9; (b) is by Lemma 9 and Assumption 4; and in the last equation constants are hide in the big $\mathcal{O}$ notation.

Finally, applying Lemma 3, we have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \le \frac{2\big[f(\mathbf{x}_0)-f(\mathbf{x}^*)\big]}{(k+1)(k+2)} + \xi_k. \tag{24}$$

Plugging $\xi_k$ in the proof is completed.

When the constraint set is an $\ell_1$ norm ball, the basic proof idea is similar as the $\ell_2$ norm ball case, i.e., after $T$ iterations $\mathbf{v}_k$ and $\mathbf{v}_{k+1}$ are near to each other. The only difference is that a regularization condition should be satisfied to ensure the uniqueness of $\mathbf{v}_k$ (only for proof, not necessary for implementation). There are multiple kinds of regularization schemes, for example, $[\nabla f(\mathbf{x}^*)]_i - [\nabla f(\mathbf{x}^*)]_j = c > 0$, where $i, j$ are the largest and second largest entry of $\nabla f(\mathbf{x}^*)$, respectively. In this case, we only need to modify the $T$ in Lemma 9 as a $c$ dependent constant, and all the other proofs follow. □

### J. $\ell_1$ norm ball

In this subsection we focus on the convergence of AFW for $\ell_1$ norm ball constraint under the assumption that $\arg\max_j \big|[\nabla f(\mathbf{x}^*)]_j\big|$ has cardinality 1 (which naturally implies that the constraint is active). Note that in this case Lemma 6 still holds hence the value of $\nabla f(\mathbf{x}^*)$ is unique regardless the uniqueness of $\mathbf{x}^*$. This assumption directly leads to $\arg\max_j \big|[\nabla f(\mathbf{x}^*)]_j\big| - |[\nabla f(\mathbf{x}^*)]_i| \ge \lambda,\ \forall i$.

When $\mathcal{X} = \{\mathbf{x}|\|\mathbf{x}\|_1 \le R\}$, the FW steps for AFW can be solved in closed-form. We have $\mathbf{v}_{k+1} = [0,\dots,0,-\mathrm{sgn}[\boldsymbol{\theta}_{k+1}]_i R, 0,\dots,0]^\top$, i.e., only the $i$-th entry being nonzero with $i = \arg\max_j |[\boldsymbol{\theta}_{k+1}]_j|$.

**Lemma 10.** *There exist a constant $T$ (which is irreverent with $k$), whenever $k \ge T$, it is guaranteed to have*

$$\|\mathbf{v}_{k+1} - \mathbf{v}_{k+2}\| = 0$$

*Proof.* In the proof, we denote $i = \arg\max_j |[\nabla f(\mathbf{x}^*)]_j|$ for convenience. It can be seen that Lemma 8 still holds.

We show that there exist $T = (\frac{3C_2}{\lambda}+1)^2 - 3$, such that for all $k \ge T$, we have $\arg\max_j |[\boldsymbol{\theta}_{k+1}]_j| = i$, which further implies only the $i$-th entry of $\mathbf{v}_{k+1}$ is non-zero. Since Lemma 8 holds, one can see whenever $k \ge T$, it is guaranteed to have $\|\boldsymbol{\theta}_{k+1} - \nabla f(\mathbf{x}^*)\| \le \frac{\lambda}{3}$. Therefore, one must have $\big||[\boldsymbol{\theta}_{k+1}]_j| - |[\nabla f(\mathbf{x}^*)]_j|\big| \le \frac{\lambda}{3},\ \forall j$. Then it is easy to see that $|[\boldsymbol{\theta}_{k+1}]_i| - |[\boldsymbol{\theta}_{k+1}]_j| \ge \frac{\lambda}{3},\ \forall j$. Hence, we have $\arg\max_j |[\boldsymbol{\theta}_{k+1}]_j| = i$.

Then one can use the closed form solution of FW step to see that when $k \ge T$, we have $\mathbf{v}_{k+1} - \mathbf{v}_{k+2} = \mathbf{0}$. The proof is thus completed. □

**Lemma 11.** *Let $\xi_0 = 0$ and $T$ defined the same as in Lemma 10. Denote $\Phi_k^* := \Phi_k(\mathbf{v}_k)$ as the minimum value of $\Phi_k(\mathbf{x})$ over $\mathcal{X}$, then we have*

$$f(\mathbf{x}_k) \le \Phi_k(\mathbf{v}_k) = \Phi_k^* + \xi_k,\ \forall k \ge 0$$

*where for $k < T+1$, $\xi_{k+1} = (1-\delta_k)\xi_k + \frac{LD^2}{2}\delta_k^2$, and $\xi_{k+1} = (1-\delta_k)\xi_k$ for $k \ge T+1$.*

*Proof.* The proof for $k < T+1$ is similar as that in Lemma 2, hence it is omitted here. For $k \ge T+1$, using similar argument as in Lemma 2, we have

$$\begin{aligned}
\Phi_{k+1}^* &\ge f(\mathbf{x}_{k+1}) + \frac{L\delta_k^2}{2}\|\mathbf{v}_{k+1}-\mathbf{v}_k\|^2 - (1-\delta_k)\xi_k \\
&= f(\mathbf{x}_{k+1}) - (1-\delta_k)\xi_k
\end{aligned}$$

where the last equation is because of Lemma 10. □

**Theorem 5.** *Consider $\mathcal{X}$ is an $\ell_1$ norm ball. If $\arg\max_j \big|[\nabla f(\mathbf{x}^*)]_j\big|$ has cardinality 1, and Assumptions 1 - 3 are satisfied, AFW guarantees that*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}\Big(\frac{1}{k^2}\Big).$$

*Proof.* Let $T$ be defined the same as in Lemma 10. For convenience denote $\xi_{k+1} = (1-\delta_k)\xi_k + \zeta_k$. When $k < T+1$, we have $\zeta_k = \frac{LD^2}{2}\delta_k^2$; when $k \ge T+1$, we have $\zeta_k = 0$. Then we can write

$$\begin{aligned}
\xi_{k+1} &= (1-\delta_k)\xi_k + \theta_k \\
&= \sum_{\tau=0}^{k}\theta_\tau\prod_{j=\tau+1}^{k}(1-\delta_j) = \sum_{\tau=0}^{k}\theta_\tau\frac{(\tau+2)(\tau+3)}{(k+2)(k+3)} \\
&= \sum_{\tau=0}^{T}\frac{LD^2}{2}\delta_\tau^2\frac{(\tau+2)(\tau+3)}{(k+2)(k+3)} = \frac{2LD^2(T+1)}{(k+2)(k+3)}.
\end{aligned}$$

Finally, applying Lemma 3, we have

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{2\big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big]}{(k+1)(k+2)} + \xi_k.$$

Plugging $\xi_k$ in completes the proof. □

### K. $\ell_p$ norm ball

In this subsection we focus on AFW with an active $\ell_p$ norm ball constraint $\mathcal{X} := \{\mathbf{x} | \|\mathbf{x}\|_p \leq R\}$, where $p \in (1, +\infty)$ and $p \neq 2$. We show that if the magnitude of every entry in $\nabla f(\mathbf{x}^*)$ is bounded away from 0, i.e., $|[\nabla f(\mathbf{x}^*)]_i| = \lambda > 0,\ \forall i$, then AFW converges at $\mathcal{O}(\frac{1}{k^2})$.

In such cases, the FW step in AFW can be solved in closed-form, that is, the $i$-th entry of $\mathbf{v}_{k+1}$ can be obtained via

$$\begin{aligned}[\mathbf{v}_{k+1}]_i &= -\mathrm{sgn}\big([\boldsymbol{\theta}_{k+1}]_i\big) \frac{\big|[\boldsymbol{\theta}_{k+1}]_i\big|^{q-1}}{\|\boldsymbol{\theta}_{k+1}\|_q^{q-1}} \cdot R \qquad (25)\\ &= -[\boldsymbol{\theta}_{k+1}]_i \frac{\big|[\boldsymbol{\theta}_{k+1}]_i\big|^{q-2}}{\|\boldsymbol{\theta}_{k+1}\|_q^{q-1}} \cdot R\end{aligned}$$

where $1/p + 1/q = 1$. For simplicity we will emphasis on the $k$ dependence only and use $\mathcal{O}$ notation in this subsection. We will also use $\theta_k^i$ to replace $[\boldsymbol{\theta}_k]_i$ for notational simplicity. In other words, $\theta_k^i$ denotes the $i$-th entry of $\boldsymbol{\theta}_k$.

First according to Lemma 8, and use the equivalence of norms, we have $\|\boldsymbol{\theta}_k - \nabla f(\mathbf{x}^*)\|_q = \mathcal{O}(\frac{1}{\sqrt{k}})$. Hence, there must exist $T_1$, such that $\|\boldsymbol{\theta}_k\|_q \leq 2G,\ \forall k \geq T_1$. Next using similar arguments as the first part of Lemma 9, there must exist $T_2$, such that $\|\boldsymbol{\theta}_k\|_q \geq G/2,\ \forall k \geq T_2$. In addition, using again similar arguments as the first part of Lemma 9, we can find that there exist $T_3$, such that $|\theta_k^i| > \frac{\lambda}{2},\ \forall k \geq T_3$.

Let $T := \max\{T_1, T_2, T_3\}$. Next we will show that $\|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2 = \mathcal{O}(\frac{1}{k}),\ \forall k \geq T$. To start, using (25), one can have

$$\begin{aligned}&v_{k+1}^i - v_k^i\\ &= \frac{R}{\|\boldsymbol{\theta}_{k+1}\|_q^{q-1}\|\boldsymbol{\theta}_k\|_q^{q-1}}\Big[-\theta_{k+1}^i|\theta_{k+1}^i|^{q-2}\|\boldsymbol{\theta}_k\|_q^{q-1}\\ &\qquad + \theta_k^i|\theta_k^i|^{q-2}\|\boldsymbol{\theta}_{k+1}\|_q^{q-1}\Big]\\ &= \frac{R}{\|\boldsymbol{\theta}_{k+1}\|_q^{q-1}\|\boldsymbol{\theta}_k\|_q^{q-1}}\Big[\theta_{k+1}^i|\theta_{k+1}^i|^{q-2}\Big(\|\boldsymbol{\theta}_{k+1}\|_q^{q-1} - \|\boldsymbol{\theta}_k\|_q^{q-1}\Big)\\ &\qquad + \|\boldsymbol{\theta}_{k+1}\|_q^{q-1}\Big(\theta_k^i|\theta_k^i|^{q-2} - \theta_{k+1}^i|\theta_{k+1}^i|^{q-2}\Big)\Big].\end{aligned}$$

Next using $G/2 \leq \|\boldsymbol{\theta}_{k+1}\|_q \leq 2G,\ \forall k \geq T$, and $|\theta_{k+1}^i| \leq \|\boldsymbol{\theta}_{k+1}\|_q$, we have

$$\begin{aligned}&|v_{k+1}^i - v_k^i| \qquad (26)\\ &= \mathcal{O}\left(\Big|\|\boldsymbol{\theta}_{k+1}\|_q^{q-1} - \|\boldsymbol{\theta}_k\|_q^{q-1}\Big| + \Big|\theta_k^i|\theta_k^i|^{q-2} - \theta_{k+1}^i|\theta_{k+1}^i|^{q-2}\Big|\right).\end{aligned}$$

We first bound the first term in RHS of (26). Let $h(x) = (x)^{q-1}$. Then by mean value theorem we have $h(y) = h(x) + \nabla h(x)(y-x) + \nabla^2 h(z)\|x-y\|^2$, where $z = (1-\alpha)x + \alpha y$ for some $\alpha \in [0,1]$. Taking $x = \|\boldsymbol{\theta}_k\|_q$ and $y = \|\boldsymbol{\theta}_{k+1}\|_q$, and using the fact $G/2 \leq \|\boldsymbol{\theta}_k\|_q \leq 2G$ for $k \geq T$, we have

$$\begin{aligned}&\|\boldsymbol{\theta}_{k+1}\|_q^{q-1} \qquad (27)\\ &= \|\boldsymbol{\theta}_k\|_q^{q-1} + \mathcal{O}\big(\big|\|\boldsymbol{\theta}_k\|_q - \|\boldsymbol{\theta}_{k+1}\|_q\big| + \big|\|\boldsymbol{\theta}_k\|_q - \|\boldsymbol{\theta}_{k+1}\|_q\big|^2\big)\\ &= \|\boldsymbol{\theta}_k\|_q^{q-1} + \mathcal{O}\big(\frac{1}{\sqrt{k}}\big)\end{aligned}$$

Hence, one can find that the first term on the RHS of (26) is bounded by $\mathcal{O}\big(\frac{1}{\sqrt{k}}\big)$.

Next we focus on the second term of (26) by considering whether $\theta_k^i$ and $\theta_{k+1}^i$ have different signs.

*Case 1: $\theta_k^i$ and $\theta_{k+1}^i$ have the same sign.* Then we have

$$\begin{aligned}&\Big|\theta_k^i|\theta_k^i|^{q-2} - \theta_{k+1}^i|\theta_{k+1}^i|^{q-2}\Big|\\ &= \Big||\theta_k^i|^{q-1} - |\theta_{k+1}^i|^{q-1}\Big| \leq \mathcal{O}\big(\frac{1}{\sqrt{k}}\big) \qquad (28)\end{aligned}$$

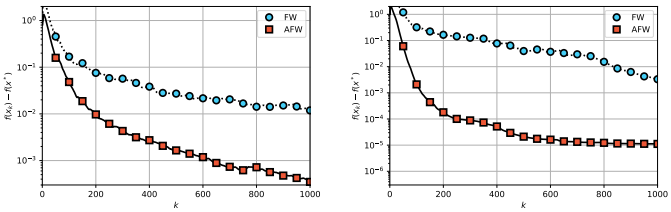
where the last inequality uses the same mean-value-theorem argument as (27) and the fact $|\theta_k^i| \geq \frac{\lambda}{2}$.

*Case 2: $\theta_k^i$ and $\theta_{k+1}^i$ have different signs.* We assume $\theta_{k+1}^i \geq 0$ w.l.o.g. In this case, by the update manner of $\boldsymbol{\theta}_{k+1}$, we have $|\theta_{k+1}^i| \leq |\delta_k[\nabla f(\mathbf{y}_k)]_i| = \mathcal{O}(\delta_k) = \mathcal{O}(\frac{1}{k})$. This is impossible given the fact $|\theta_{k+1}^i| > \frac{\lambda}{2}$ when $k \geq T$.

Therefore, we have the second term in (26) bounded by $\mathcal{O}(\frac{1}{\sqrt{k}})$. Hence, it is easy to see that

$$\|\mathbf{v}_{k+1} - \mathbf{v}_k\|^2 = \mathcal{O}\Big(\frac{1}{k}\Big).$$

Applying the same argument in the proof of Theorem 3, we have that when $k \geq T$, $\xi_{k+1} = \tilde{\mathcal{O}}(\frac{1}{k^2})$. This further implies $f(\mathbf{x}_k) - f(\mathbf{x}^*) = \tilde{\mathcal{O}}(\frac{1}{k^2})$ as well.



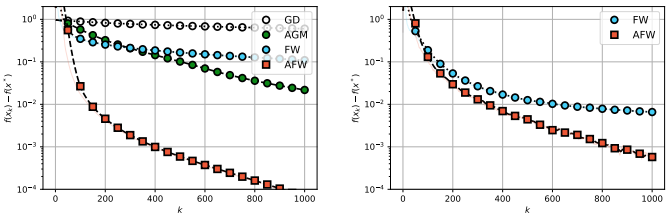

(a) *mnist* (b) *mushroom*

Fig. 5. Performance of AFW on $n$-support norm balls.

### L. Additional numerical tests

AFW is tested on other loss functions and constraints to demonstrate its efficiency.

**$n$-support norm ball constraint.** We first consider logistic regression over a $n$-support norm ball [50]. This is challenging due to the constraint $\mathcal{X} = \mathrm{conv}\{\mathbf{x} | \|\mathbf{x}\|_0 \leq n, \|\mathbf{x}\|_2 \leq R\}$, where $\mathrm{conv}\{\cdot\}$ denotes the convex hull. GD and AGM are expensive for such a constraint set since efficient projection is unclear, while the FW subproblem can be solved easily [51]. For this reason, we only compare FW with AFW, and the numerical results depicted in Fig. 5 demonstrate that AFW outperforms FW.




(a) $\ell_2$ norm ball (b) $n$-support norm ball

Fig. 6. Performance of AFW on log-sum-exp losses.

**Log-sum-exp loss.** We also test AFW using the log-sum-exp loss function, that is,

$$f(\mathbf{x}) = \ln\Big(\sum_{i=1}^{n} \exp\big(\langle \mathbf{a}_i, \mathbf{x}\rangle\big)\Big). \tag{29}$$

We set $n = 1,000$ and $d = 500$, and draw $\mathbf{a}_i$ from a standardized normal distribution. The $\ell_2$ norm ball and $n$-support norm balls are used as constraints. The results in Fig. 6 corroborate that AFW outperforms FW.

## REFERENCES


[1] R. M. Freund, P. Grigas, and R. Mazumder, "An extended frank–wolfe method with in-face directions, and its application to low-rank matrix completion," *SIAM Journal on Optimization*, vol. 27, no. 1, pp. 319–346, 2017.

[2] Z. Harchaoui, A. Juditsky, and A. Nemirovski, "Conditional gradient algorithms for norm-regularized smooth convex optimization," *Mathematical Programming*, vol. 152, no. 1-2, pp. 75–112, 2015.

[3] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2004, vol. 87.

[4] Z. Allen-Zhu and L. Orecchia, "Linear coupling: An ultimate unification of gradient and mirror descent," *arXiv preprint arXiv:1407.1537*, 2014.

[5] Y. Nesterov, "Universal gradient methods for convex optimization problems," *Mathematical Programming*, vol. 152, no. 1-2, pp. 381–404, 2015.

[6] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, no. 1-2, pp. 95–110, 1956.

[7] M. Jaggi, "Revisiting frank-wolfe: Projection-free sparse convex optimization." in *Proc. Intl. Conf. on Machine Learning*, 2013, pp. 427–435.

[8] S. Lacoste-Julien and M. Jaggi, "On the global linear convergence of frank-wolfe optimization variants," in *Proc. Advances in Neural Info. Process. Syst.*, 2015, pp. 496–504.

[9] D. Garber and E. Hazan, "Faster rates for the frank-wolfe method over strongly-convex sets," in *Proc. Intl. Conf. on Machine Learning*, 2015.

[10] S. Lacoste-Julien, M. Jaggi, M. W. Schmidt, and P. Pletscher, "Block-coordinate frank-wolfe optimization for structural svms," in *Proc. Intl. Conf. on Machine Learning*, no. CONF, 2013, pp. 53–61.

[11] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with frank-wolfe algorithm," in *Proc. European Conf. on Computer Vision*. Springer, 2014, pp. 253–268.

[12] S. Lacoste-Julien, F. Lindsten, and F. Bach, "Sequential kernel herding: Frank-wolfe optimization for particle filtering," in *Artificial Intelligence and Statistics*, 2015, pp. 544–552.

[13] M. Fukushima, "A modified frank-wolfe algorithm for solving the traffic assignment problem," *Transportation Research Part B: Methodological*, vol. 18, no. 2, pp. 169–177, 1984.

[14] G. Luise, S. Salzo, M. Pontil, and C. Ciliberto, "Sinkhorn barycenters with free support via frank-wolfe algorithm," in *Proc. Advances in Neural Info. Process. Syst.*, 2019, pp. 9318–9329.

[15] L. Zhang, V. Kekatos, and G. B. Giannakis, "Scalable electric vehicle charging protocols," *IEEE Trans. on Power Systems*, vol. 32, no. 2, pp. 1451–1462, 2016.

[16] L. Zhang, G. Wang, D. Romero, and G. B. Giannakis, "Randomized block frank–wolfe for convergent large-scale learning," *IEEE Trans. on Signal Processing*, vol. 65, no. 24, pp. 6448–6461, 2017.

[17] A. Mokhtari, H. Hassani, and A. Karbasi, "Stochastic conditional gradient methods: From convex minimization to submodular maximization," *arXiv preprint arXiv:1804.09554*, 2018.

[18] G. Lan, "The complexity of large-scale convex programming under a linear optimization oracle," *arXiv preprint arXiv:1309.5550*, 2013.

[19] E. S. Levitin and B. T. Polyak, "Constrained minimization methods," *USSR Computational mathematics and mathematical physics*, vol. 6, no. 5, pp. 1–50, 1966.

[20] B. Li, M. Coutiño, and G. B. Giannakis, "Revisit of estimate sequence for accelerated gradient methods," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3602–3606.

[21] D. Garber and O. Meshi, "Linear-memory and decomposition-invariant linearly convergent conditional gradient algorithm for structured polytopes," in *Proc. Advances in Neural Info. Process. Syst.*, 2016, pp. 1001–1009.

[22] F. Bach, "On the effectiveness of richardson extrapolation in machine learning," *arXiv preprint arXiv:2002.02835*, 2020.

[23] J. Guélat and P. Marcotte, "Some comments on wolfe's away step," *Mathematical Programming*, vol. 35, no. 1, pp. 110–119, 1986.

[24] F. Pedregosa, A. Askari, G. Negiar, and M. Jaggi, "Step-size adaptivity in projection-free optimization," *arXiv preprint arXiv:1806.05123*, 2018.

[25] G. Braun, S. Pokutta, D. Tu, and S. Wright, "Blended conditional gradients: the unconditioning of conditional gradients," *arXiv preprint arXiv:1805.07311*, 2018.

[26] T. Kerdreux, A. dAspremont, and S. Pokutta, "Projection-free optimization on uniformly convex sets," *arXiv preprint arXiv:2004.11053*, 2020.

[27] J. Abernethy, K. A. Lai, K. Y. Levy, and J.-K. Wang, "Faster rates for convex-concave games," in *Conference On Learning Theory*, 2018, pp. 1595–1625.

[28] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $1/k^2$," in *Soviet Math. Dokl*, vol. 27, 1983.

[29] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[30] W. Krichene, A. Bayen, and P. L. Bartlett, "Accelerated mirror descent in continuous and discrete time," in *Proc. Advances in Neural Info. Process. Syst.*, 2015, pp. 2845–2853.

[31] A. Nitanda, "Stochastic proximal gradient descent with acceleration techniques," in *Proc. Advances in Neural Info. Process. Syst.*, Montreal, Canada, 2014, pp. 1574–1582.

[32] H. Lin, J. Mairal, and Z. Harchaoui, "A universal catalyst for first-order optimization," in *Proc. Advances in Neural Info. Process. Syst.*, Montreal, Canada, 2015, pp. 3384–3392.

[33] W. Su, S. Boyd, and E. Candes, "A differential equation for modeling Nesterov accelerated gradient method: Theory and insights," in *Proc. Advances in Neural Info. Process. Syst.*, 2014, pp. 2510–2518.

[34] J. Zhang, A. Mokhtari, S. Sra, and A. Jadbabaie, "Direct runge-kutta discretization achieves acceleration," in *Proc. Advances in Neural Info. Process. Syst.*, 2018, pp. 3900–3909.

[35] B. Shi, S. S. Du, W. J. Su, and M. I. Jordan, "Acceleration via symplectic discretization of high-resolution differential equations," *arXiv preprint arXiv:1902.03694*, 2019.

[36] G. Lan and Y. Zhou, "Conditional gradient sliding for convex optimization," *SIAM Journal on Optimization*, vol. 26, no. 2, pp. 1379–1409, 2016.

[37] Y. Malitsky and K. Mishchenko, "Adaptive gradient descent without descent," in *Proc. Intl. Conf. on Machine Learning*, 2020.

[38] A. Kulunchakov and J. Mairal, "Estimate sequences for variance-reduced stochastic composite optimization," in *Proc. Intl. Conf. on Machine Learning*, 2019.

[39] B. Li, L. Wang, and G. B. Giannakis, "Almost tune-free variance reduction," in *Proc. Intl. Conf. on Machine Learning*, 2020.

[40] K. L. Clarkson, "Coresets, sparse greedy approximation, and the frank-wolfe algorithm," *ACM Transactions on Algorithms (TALG)*, vol. 6, no. 4, p. 63, 2010.

[41] A. Nemirovski, "Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems," *SIAM Journal on Optimization*, vol. 15, no. 1, pp. 229–251, 2004.

[42] J. C. Dunn, "Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals," *SIAM Journal on Control and Optimization*, vol. 17, no. 2, pp. 187–211, 1979.

[43] B. Li, L. Wang, G. B. Giannakis, and Z. Zhao, "Enhancing parameter-free frank wolfe with an extra subproblem," *arXiv preprint arXiv:2012.05284*, 2020.

[44] L. Ding, Y. Fei, Q. Xu, and C. Yang, "Spectral frank-wolfe algorithm: Strict complementarity and linear convergence," in *Proc. Intl. Conf. on Machine Learning*, 2020.

[45] D. Garber, "Revisiting frank-wolfe for polytopes: Strict complementary and sparsity," *arXiv preprint arXiv:2006.00558*, 2020.
[46] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l 1-ball for learning in high dimensions," in *Proc. Intl. Conf. on Machine Learning*. ACM, 2008, pp. 272–279.
[47] J. Bennett, S. Lanning *et al.*, "The netflix prize," in *Proc. KDD cup and workshop*, vol. 2007. New York, NY, USA., 2007, p. 35.
[48] R. M. Bell and Y. Koren, "Lessons from the netflix prize challenge." *SiGKDD Explorations*, vol. 9, no. 2, pp. 75–79, 2007.
[49] M. Fazel, "Matrix rank minimization with applications," 2002.
[50] A. Argyriou, R. Foygel, and N. Srebro, "Sparse prediction with the $k$-support norm," in *Proc. Advances in Neural Info. Process. Syst.*, 2012, pp. 1457–1465.
[51] B. Liu, X.-T. Yuan, S. Zhang, Q. Liu, and D. N. Metaxas, "Efficient k-support-norm regularized minimization via fully corrective frank-wolfe method." in *Proc. Intl. Joint Conf. on Artifical Intelligence*, 2016, pp. 1760–1766.

**Bingcong Li** received the B. Eng. degree (with highest honors) in Communication Science and Engineering from Fudan University, and the M.Sc. degree in Electrical and Computer Engineering (ECE) from the University of Minnesota (UMN), in 2017 and 2019, respectively. He is now pursuing his Ph.D. degree at UMN. His research interests lie in optimization and machine learning, with applications to cyber physical systems. He received the National Scholarship twice from China in 2014 and 2015, and UMN ECE Department Fellowship in 2017.

**Mario Coutino** (Student member, IEEE) received the M.Sc. and the Ph.D degree (cum laude) in electrical engineering in July 2016 and April 2021, respectively, from the Delft University of Technology, Delft, The Netherlands. Since October 2020, he has been working in TNO, The Netherlands, in the Radar Technology Department as a Signal Processing Researcher. He has held positions at Thales Nederlands, during 2015, and Bang & Olufsen, during 20152016. He received a Best Student Paper Award for his publication at the CAMSAP 2017 conference in Curacao and was a visiting researcher with RIKEN AIP and the Digital Technological Center, University of Minnesota, during 2018 and 2019, respectively. His research interests include array signal processing, signal processing on networks, submodular and convex optimization, and numerical linear algebra.

**Georgios B. Giannakis** (Fellow, IEEE) received his Diploma in Electrical Engr. from the Ntl. Tech. Univ. of Athens, Greece, 1981. From 1982 to 1986 he was with the Univ. of Southern California (USC), where he received his MSc. in Electrical Engineering, 1983, MSc. in Mathematics, 1986, and Ph.D. in Electrical Engr., 1986. He was a faculty member with the University of Virginia from 1987 to 1998, and as of 1999 he has been a professor with the Univ. of Minnesota, where he held an ADC Endowed Chair of Telecommunications, served as director of the Digital Technology Center from 2008 to 2021, and since 2016 he is a University of Minnesota McKnight Presidential Chair in ECE.

His general interests span the areas of statistical learning, signal processing, communications, and networking - subjects on which he has published more than 480 journal papers, 780 conference papers, 25 book chapters, two edited books and two research monographs. Current research focuses on Data Science, and Network Science with applications to the Internet of Things, and power networks with renewables. He is the (co-) inventor of 34 issued patents, and the (co-) recipient of 10 best journal paper awards from the IEEE Signal Processing (SP) and Communications Societies, including the G. Marconi Prize Paper Award in Wireless Communications. He also received the IEEE-SPS Norbert Wiener Society Award (2019); EURASIP's A. Papoulis Society Award (2020); Technical Achievement Awards from the IEEE-SPS (2000) and from EURASIP (2005); the IEEE ComSoc Education Award (2019); and the IEEE Fourier Technical Field Award (2015). He is a foreign member of the Academia Europaea, and Fellow of the National Academy of Inventors, the European Academy of Sciences, IEEE and EURASIP. He has served the IEEE in a number of posts, including that of a Distinguished Lecturer for the IEEE-SPS.

**Geert Leus** (Fellow, IEEE) received the M.Sc. and Ph.D. degree in Electrical Engineering from the KU Leuven, Belgium, in June 1996 and May 2000, respectively. Currently, Geert Leus is a Full Professor at the Faculty of Electrical Engineering, Mathematics and Computer Science of the Delft University of Technology, The Netherlands. Geert Leus received the 2021 EURASIP Individual Technical Achievement Award, a 2005 IEEE Signal Processing Society Best Paper Award, and a 2002 IEEE Signal Processing Society Young Author Best Paper Award. He is a Fellow of the IEEE and a Fellow of EURASIP. Geert Leus was a Member-at-Large of the Board of Governors of the IEEE Signal Processing Society, the Chair of the IEEE Signal Processing for Communications and Networking Technical Committee, and the Editor in Chief of the EURASIP Journal on Advances in Signal Processing. Currently, he is the Chair of the EURASIP Technical Area Committee on Signal Processing for Multisensor Systems and the Editor in Chief of EURASIP Signal Processing.