Heavy Ball Momentum for Conditional Gradient

Bingcong Li **Alireza Sadeghi** **Georgios B. Giannakis**

University of Minnesota - Twin Cities
Minneapolis, MN, USA
{lixx5599, sadeg012, georgios}@umn.edu

Abstract

Conditional gradient, aka Frank Wolfe (FW) algorithms, have well-documented merits in machine learning and signal processing applications. Unlike projection-based methods, momentum cannot improve the convergence rate of FW, in general. This limitation motivates the present work, which deals with heavy ball momentum, and its impact to FW. Specifically, it is established that heavy ball offers a unifying perspective on the primal-dual (PD) convergence, and enjoys a tighter *per iteration* PD error rate, for multiple choices of step sizes, where PD error can serve as the stopping criterion in practice. In addition, it is asserted that restart, a scheme typically employed jointly with Nesterov's momentum, can further tighten this PD error bound. Numerical results demonstrate the usefulness of heavy ball momentum in FW iterations.

1 Introduction

This work studies momentum in Frank Wolfe (FW) methods [9, 10, 16, 20] for solving

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

Here, f is a convex function with Lipschitz continuous gradients, and the constraint set $\mathcal{X} \subset \mathbb{R}^d$ is assumed convex and compact, where d is the dimension of variable $\mathbf{x}$. Throughout, we let $\mathbf{x}^* \in \mathcal{X}$ denote a minimizer of (1). FW and its variants are prevalent in various machine learning and signal processing applications, such as traffic assignment [12], non-negative matrix factorization [30], video colocation [17], image reconstruction [15], particle filtering [19], electronic vehicle charging [36], recommender systems [11], optimal transport [26], and neural network pruning [34]. The popularity of FW is partially due to the elimination of projection compared with projected gradient descent (GD) [29], leading to computational efficiency especially when d is large. In particular, FW solves a subproblem with a linear loss, i.e., $\mathbf{v}_{k+1} \in \arg\min_{\mathbf{v}\in\mathcal{X}} \langle \nabla f(\mathbf{x}_k), \mathbf{v}\rangle$ at kth iteration, and then updates $\mathbf{x}_{k+1}$ as a convex combination of $\mathbf{x}_k$ and $\mathbf{v}_{k+1}$. When dealing with a structured $\mathcal{X}$, a closed-form or efficient solution for $\mathbf{v}_{k+1}$ is available [13, 16], which is preferable over projection.

Unlike projection based algorithms [14, 32] though, momentum does not perform well with FW. Indeed, the lower bound in [16, 20] demonstrates that at least $\mathcal{O}(\frac{1}{\epsilon})$ linear subproblems are required to ensure $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$, which does not guarantee that momentum is beneficial for FW, because even vanilla FW achieves this lower bound. In this work, we contend that momentum is evidently useful for FW. Specifically, we prove that the *heavy ball momentum* leads to tightened and efficiently computed primal-dual error bound, as well as numerical improvement. To this end, we outline first the primal convergence.

Primal convergence. The primal error refers to $f(\mathbf{x}_k) - f(\mathbf{x}^*)$. It is guaranteed for FW that $f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}\big(1/k\big), \forall k \geq 1$ [16, 22]. This rate is tight in general since it matches to the lower bound [16, 20]. Other FW variants also ensure the same order of primal error; see e.g., [20, 21].

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

Table 1: A comparison of HFW with relevant works. The "computation" in the third column is short for "the number of required FW subproblems to calculate the PD error per iteration."

reference	computation	PD conv. type	PD conv. rate
[16]	1 subproblem	Type I	$\frac{27LD^2}{4(K+1)}$
[18]	2 subproblems	Type II	$\frac{2LD^2}{\sqrt{k+1}}, \forall k$
[28]	2 subproblems	Type II	$\frac{4LD^2}{k+1}, \forall k$
This work (Alg. 2)	1 subproblem	Type II	$\frac{2LD^2}{k+1}, \forall k$
This work (Alg. 3)	2 subproblems	Type II	$\frac{2LD^2}{k+1+c}, \forall k$ with $c \geq 0$

Primal-dual convergence. The primal-dual (PD) error quantifies the difference between both the primal and the 'dual' functions from the optimal objective, hence it is an upper bound on the primal error. When the PD error is shown to *converge*, it can be safely used as the stopping criterion: whenever the PD error is less than some prescribed $\epsilon > 0$, $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$ is ensured automatically. The PD error of FW is convenient to compute, hence FW is suitable for the requirement of "solving problems to some desirable accuracy;" see e.g., [33]. For pruning (two-layer) neural networks [34], the extra training loss incurred by removing neurons can be estimated via the PD error. However, due to technical difficulties, existing analyses on PD error are not satisfactory enough and lack of unification. It is established in [6, 10, 16] that the minimum PD error is sufficiently small, namely $\min_{k\in\{1,\ldots,K\}} \mathrm{PDError}_k = \mathcal{O}(\frac{1}{K})$, where K is the total number of iterations. We term such a bound for the minimum PD error as Type I guarantee. Another stronger guarantee, which directly implies Type I bound, emphasizes the per iteration convergence, e.g., $\mathrm{PDError}_k \leq \mathcal{O}(\frac{1}{k}), \forall k$. We term such guarantees as Type II bound. A Type II bound is reported in [18, Theorem 2], but with an unsatisfactory k dependence. This is improved by [7,28] with the price of extra computational burden since it involves solving *two* FW subproblems per iteration for computing this PD error. Several related works such as [10] provide a weaker PD error compared with [28]; see a summary in Table 1.

In this work, we show that a computationally affordable Type II bound can be obtained by simply relying on heavy ball momentum. Interestingly, FW based on heavy ball momentum (HFW) also maintains FW's neat geometric interpretation. Through unified analysis, the resultant type II PD error improves over existing bounds; see Table 1. This PD error of HFW is further tightened using *restart*. Although restart is more popular in projection based methods together with Nesterov's momentum [31], we show that restart for FW is natural to adopt jointly with heavy ball. In succinct form, our contributions can be summarized as follows.

• We show through unified analysis that HFW enables a tighter type II guarantee for PD error for multiple choices of the step size. When used as stopping criterion, no extra subproblem is needed.

• The Type II bound can be further tightened by restart triggered through a comparison between two PD-error-related quantities.

• Numerical tests on benchmark datasets support the effectiveness of heavy ball momentum. As a byproduct, a simple yet efficient means of computing local Lipschitz constants becomes available to improve the numerical efficiency of smooth step sizes [13, 22].

Notation. Bold lowercase (capital) letters denote column vectors (matrices); $\|\mathbf{x}\|$ stands for a norm of a vector $\mathbf{x}$, whose dual norm is denoted by $\|\mathbf{x}\|_*$; and $\langle \mathbf{x}, \mathbf{y} \rangle$ is the inner product of $\mathbf{x}$ and $\mathbf{y}$.

2 Preliminaries

This section outlines FW, starting with standard assumptions that will be taken to hold true throughout.

Assumption 1. *(Lipschitz continuous gradient.) The objective function $f : \mathcal{X} \to \mathbb{R}$ has L-Lipchitz continuous gradients; i.e., $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_* \leq L\|\mathbf{x} - \mathbf{y}\|, \forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$.*

Assumption 2. *(Convexity.) The objective function $f : \mathcal{X} \to \mathbb{R}$ is convex; that is, $f(\mathbf{y}) - f(\mathbf{x}) \geq \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$.*

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

FW for solving (1) under Assumptions 1 – 3 is listed in Alg. 1. The subproblem in Line 3 can be visualized geometrically as minimizing a supporting hyperplane of $f(\mathbf{x})$ at $\mathbf{x}_k$, i.e.,

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

Algorithm 1 FW [9]

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

For many constraint sets, efficient implementation or a closed-form solution is available for $\mathbf{v}_{k+1}$; see e.g., [16] for a comprehensive summary. Upon minimizing the supporting hyperplane in (2), $\mathbf{x}_{k+1}$ is updated as a convex combination of $\mathbf{v}_{k+1}$ and $\mathbf{x}_k$ in Line 4 so that no projection is required. The choices on the step size $\eta_k \in [0, 1]$ will be discussed shortly.

The PD error of Alg. 1 is captured by the so-termed *FW gap*, formally defined as

$$\bar{\mathcal{G}}_k := \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{v}_{k+1} \rangle = \underbrace{f(\mathbf{x}_k) - f(\mathbf{x}^*)}_{\text{primal error}} + \underbrace{f(\mathbf{x}^*) - \min_{\mathbf{v}\in\mathcal{X}} \Big[f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{v} - \mathbf{x}_k \rangle \Big]}_{\text{dual error}} \tag{3}$$

where the second equation is because of (2). It can be verified that both primal and dual errors marked in (3) are no less than 0 by appealing to the convexity of f. If $\bar{\mathcal{G}}_k$ converges, one can deduce that the primal error converges. For this reason, $\bar{\mathcal{G}}_k$ is typically used as a stopping criterion for Alg. 1. Next, we focus on the step sizes that ensure convergence.

Parameter-free step size. This type of step sizes does not rely on any problem dependent parameters such as L and D, and hence it is extremely simple to implement. The most commonly adopted step size is $\eta_k = \frac{2}{k+2}$, which ensures a converging primal error $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{2LD^2}{k+1}, \forall k \geq 1$, and a weaker claim on the PD error, $\min_{k\in\{1,\ldots,K\}} \bar{\mathcal{G}}_k = \frac{27LD^2}{4K}$ [16]. A variant of PD convergence has been established recently based on a modified FW gap [28]. Although Type II convergence is observed, the modified FW gap therein is inefficient to serve as stopping criterion because an additional FW subproblem has to be solved per iteration to compute its value.

Smooth step size. When the (estimate of) Lipschitz constant L is available, one can adopt the following step sizes in Alg. 1 [22]

$$\eta_k = \min \left\{ \frac{\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{v}_{k+1} \rangle}{L \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}, 1 \right\}. \tag{4}$$

Despite the estimated L is typically too pessimistic to capture the local Lipschitz continuity, such a step size ensures $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$; see derivations in Appendix A.1. The PD convergence is studied in [11], where the result is slightly weaker than that of [28].

3 FW with heavy ball momentum

After a brief recap of vanilla FW, we focus on the benefits of heavy ball momentum for FW under multiple step size choices, with special emphasis on PD errors.

3.1 Prelude

HFW is summarized in Alg. 2. Similar to GD with heavy ball momentum [14, 32], Alg. 2 updates decision variables using a weighted average of gradients $\mathbf{g}_{k+1}$. In addition, the update direction of Alg. 2 is no longer guaranteed to be a descent one. This is because in HFW, $\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{v}_{k+1} \rangle$ can be negative. Although a stochastic version of heavy ball momentum was adopted in [27] and its variants, e.g., [37], to reduce the mean square error of the gradient estimate, heavy ball is introduced here for a totally different purpose, that is, to improve the PD error. The most significant difference comes at technical perspectives, which is discussed in Sec. 3.4. Next, we gain some intuition on why heavy ball can be beneficial.

Consider $\mathcal{X}$ as an ℓ_2-norm ball, that is, $\mathcal{X} = \{\mathbf{x} | \|\mathbf{x}\|_2 \leq R\}$. In this case, we have $\mathbf{v}_{k+1} = -\frac{R}{\|\mathbf{g}_{k+1}\|_2} \mathbf{g}_{k+1}$ in Alg. 2. The momentum $\mathbf{g}_{k+1}$ can smooth out the changes of $\{\nabla f(\mathbf{x}_k)\}$, resulting

in a more concentrated sequence $\{\mathbf{v}_{k+1}\}$. Recall that the PD error is closely related to $\mathbf{v}_{k+1}$ [cf. equation (3)]. We hope the "concentration" of $\{\mathbf{v}_{k+1}\}$ to be helpful in reducing the changes of PD error among consecutive iterations so that a Type II PD error bound is attainable.

Algorithm 2 FW with heavy ball momentum

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

A few concepts are necessary to obtain a tightened PD error of HFW. First, we introduce the generalized FW gap associated with Alg. 2 that captures the PD error. Write $\mathbf{g}_{k+1}$ explicitly as $\mathbf{g}_{k+1} = \sum_{\tau=0}^{k} w_k^\tau \nabla f(\mathbf{x}_\tau)$, where $w_k^\tau = \delta_\tau \prod_{j=\tau+1}^{k}(1-\delta_j) > 0$, $\forall \tau \geq 1$, and $w_k^0 = \prod_{j=1}^{k}(1-\delta_j) > 0$. Then, define a sequence of linear functions $\{\Phi_k(\mathbf{x})\}$ as

$$\Phi_{k+1}(\mathbf{x}) := \sum_{\tau=0}^{k} w_k^\tau \big[f(\mathbf{x}_\tau) + \langle \nabla f(\mathbf{x}_\tau), \mathbf{x} - \mathbf{x}_\tau \rangle \big], \ \forall k \geq 0. \tag{5}$$

It is clear that $\Phi_{k+1}(\mathbf{x})$ is a weighted average of the supporting hyperplanes of $f(\mathbf{x})$ at $\{\mathbf{x}_\tau\}_{\tau=0}^k$. The properties of $\Phi_{k+1}(\mathbf{x})$, and how they relate to Alg. 2 are summarized in the next lemma.

Lemma 1. *For the linear function* $\Phi_{k+1}(\mathbf{x})$ *in* (5)*, it holds that: i)* $\mathbf{v}_{k+1}$ *minimizes* $\Phi_{k+1}(\mathbf{x})$ *over* $\mathcal{X}$*; and, ii)* $f(\mathbf{x}) \geq \Phi_{k+1}(\mathbf{x}), \forall k \geq 0, \forall \mathbf{x} \in \mathcal{X}$.

From the last lemma, one can see that $\mathbf{v}_k$ is obtained by minimizing $\Phi_k(\mathbf{x})$, which is an affine lower bound on $f(\mathbf{x})$. Hence, HFW admits a geometric interpretation similar to that of FW. In addition, based on $\Phi_k(\mathbf{x})$ we can define the generalized FW gap.

Definition 1. *(Generalized FW gap.) The generalized FW gap w.r.t.* $\Phi_k(\mathbf{x})$ *is*

$$\mathcal{G}_k := f(\mathbf{x}_k) - \min_{\mathbf{x}\in\mathcal{X}} \Phi_k(\mathbf{x}) = f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k). \tag{6}$$

In words, the generalized FW gap is defined as the difference between $f(\mathbf{x}_k)$ and the minimal value of $\Phi_k(\mathbf{x})$ over $\mathcal{X}$. The newly defined $\mathcal{G}_k$ also illustrates the PD error

$$\mathcal{G}_k = f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k) = \underbrace{f(\mathbf{x}_k) - f(\mathbf{x}^*)}_{\text{primal error}} + \underbrace{f(\mathbf{x}^*) - \Phi_k(\mathbf{v}_k)}_{\text{dual error}}. \tag{7}$$

For the dual error, we have $f(\mathbf{x}^*) - \Phi_k(\mathbf{v}_k) \geq \Phi_k(\mathbf{x}^*) - \Phi_k(\mathbf{v}_k) \geq 0$, where both inequalities follow from Lemma 1. Hence, $\mathcal{G}_k \geq 0$ automatically serves as an overestimate of both primal and dual errors. When establishing the convergence of $\mathcal{G}_k$, it can be adopted as the stopping criterion for Alg. 2. Related claims have been made for the generalized FW gap [20, 23, 28]. Lack of heavy ball momentum leads to inefficiency, because an additional FW subproblem is needed to compute this gap [28]. Works [20, 23] focus on Nesterov's momentum for FW, that incurs additional memory relative to HFW; see also Sec. 3.4 for additional elaboration. Having defined the generalized FW gap, we next pursue parameter choices that establish Type II convergence guarantees.

3.2 Parameter-free step size

We first consider a parameter-free choice for HFW to demonstrate the usefulness of heavy ball

$$\delta_k = \eta_k = \frac{2}{k+2}, \ \forall k \geq 0. \tag{8}$$

Such a choice on δ_k puts more weight on recent gradients when calculating $\mathbf{g}_{k+1}$, since $w_k^\tau = \mathcal{O}(\frac{\tau}{k^2})$. The following theorem specifies the convergence of $\mathcal{G}_k$.

Theorem 1. *If Assumptions 1-3 hold, then choosing* δ_k *and* η_k *as in* (8)*, Alg. 2 guarantees that*

$$\mathcal{G}_k = f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k) \leq \frac{2LD^2}{k+1}, \ \forall k \geq 1.$$

Theorem 1 provides a much stronger PD guarantee for all k than vanilla FW [16, Theorem 2]. In addition to a readily computable generalized FW gap, our rate is tighter than [28], where the provided bound is $\frac{4LD^2}{k+1}$. In fact, the constants in our PD bound even match to the best known primal error of vanilla FW. A direct consequence of Theorem 1 is the convergence of both primal and dual errors.

Corollary 1. *Choosing the parameters as in Theorem 1, then $\forall k \geq 1$, Alg.2 guarantees that*

$$\textit{primal conv.: } f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{2LD^2}{k+1}; \quad \textit{dual conv.: } f(\mathbf{x}^*) - \Phi_k(\mathbf{v}_k) \leq \frac{2LD^2}{k+1}.$$

Proof. Combine Theorem 1 with $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \mathcal{G}_k$ and $f(\mathbf{x}^*) - \Phi_k(\mathbf{v}_k) \leq \mathcal{G}_k$ [cf. (7)]. □

3.3 Smooth step size

Next, we focus on HFW with a variant of the smooth step size

$$\delta_k = \frac{2}{k+2} \quad \text{and} \quad \eta_k = \max\left\{0, \min\left\{\frac{\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{v}_{k+1}\rangle}{L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}, 1\right\}\right\}. \tag{9}$$

Comparing with the smooth step size for vanilla FW in (4), it can be deduced that the choice on η_k in (9) has to be trimmed to $[0, 1]$ manually. This is because $\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{v}_{k+1}\rangle$ is no longer guaranteed to be positive. The smooth step size enables an adaptive means of adjusting the weight for $\nabla f(\mathbf{x}_k)$. To see this, note that when $\eta_k = 0$, we have $\mathbf{x}_{k+1} = \mathbf{x}_k$. As a result, $\mathbf{g}_{k+2} = (1-\delta_{k+1})\mathbf{g}_{k+1} + \delta_{k+1}\nabla f(\mathbf{x}_{k+1}) = (1-\delta_{k+1})\mathbf{g}_{k+1} + \delta_{k+1}\nabla f(\mathbf{x}_k)$, that is, the weight on $\nabla f(\mathbf{x}_k)$ is adaptively increased to $\delta_k(1-\delta_{k+1}) + \delta_{k+1}$ if one further unpacks $\mathbf{g}_{k+1}$. Another analytical benefit of the step size in (9) is that it guarantees a non-increasing objective value; see Appendix A.2 for derivations. Convergence of the generalized FW gap is established next.

Theorem 2. *If Assumptions 1-3 hold, while η_k and δ_k are chosen as in* (9)*, Alg. 2 guarantees that*

$$\mathcal{G}_k = f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k) \leq \frac{2LD^2}{k+1}, \ \forall k \geq 1.$$

The proof of Theorem 2 follows from that of Theorem 1 after modifying just one inequality. This considerably simplifies the analysis on the (modified) FW gap compared to vanilla FW with smooth step size [11]. The PD convergence clearly implies the convergence of both primal and dual errors. A similar result to Corollary 1 can be obtained, but we omit it for brevity. We further extend Theorem 2 in Appendix B.4 by showing that if a slightly more difficult subproblem can be solved, it is possible to ensure *per step descent on the PD error; i.e., $\mathcal{G}_{k+1} \leq \mathcal{G}_k$.*

Line search. When choosing $\delta_k = \frac{2}{k+2}$ and η_k via line search, HFW can guarantee a Type II PD error of $\frac{2LD^2}{k+1}$; please refer to Appendix B.5 due to space limitation. For completeness, an iterative manner to update $\mathcal{G}_k$ for using as stopping criterion is also described in Appendix C.

3.4 Further considerations

There are more choices of δ_k and η_k leading to (primal) convergence. For example, one can choose $\delta_k \equiv \delta \in (0, 1)$ and $\eta_k = \mathcal{O}\big(\frac{1}{k}\big)$ as an extension of [27].[1] A proof is provided in Appendix B.7 for completeness. This analysis framework in [27], however, has two shortcomings: i) the convergence can be only established using ℓ_2-norm (recall that in Assumption 1, we do not pose any requirement on the norm); and, ii) the final primal error (hence PD error) can only be worse than vanilla FW because their analysis treats $\mathbf{g}_{k+1}$ as $\nabla f(\mathbf{x}_k)$ with errors but not momentum, therefore, it is difficult to obtain the same tight PD bound as in Theorem 1. Our analytical techniques avoid these limitations.

When choosing $\delta_k = \eta_k = \frac{1}{k+1}$, we can recover Algorithm 3 in [1]. Notice that such a choice on δ_k makes $\mathbf{g}_{k+1}$ a uniform average of all gradients. A slower convergence rate $f(\mathbf{x}_k) - f(\mathbf{x}^*) = \mathcal{O}\big(\frac{LD^2 \ln k}{k}\big)$ was established in [1] through a sophisticated derivation using no-regret online learning. Through our simpler analytical framework, we can attain the same rate while providing more options for the step size.

Theorem 3. *Let Assumptions 1-3 hold, and select $\delta_k = \frac{1}{k+1}$ with η_k using one of the following options: i) $\eta_k = \frac{1}{k+1}$; ii) as in* (9)*; or iii) line search as in* (26b)*. The generalized FW gap of Alg. 2 then converges with rate*

$$\mathcal{G}_k = f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k) \leq \frac{LD^2 \ln(k+1)}{2k}, \ \forall k \geq 1.$$

[1] We are unable to derive even a primal error bound using the same analysis framework in [27] for step sizes listed in Theorem 1.

The rate in Theorem 3 has worse dependence on k relative to Theorems 1 and 2, partially because too much weight is put on past gradients in $\mathbf{g}_{k+1}$, suggesting that large momentum may not be helpful.

Heavy ball versus Nesterov's momentum. A simple rule to compare these two momentums is whether gradient is calculated at the converging sequence $\{\mathbf{x}_k\}$. Heavy ball momentum follows this rule, while Nesterov's momentum computes the gradient at some extrapolation points that are not used in Alg. 2. It is unclear how the original Nesterov's momentum benefits the PD error, but the ∞-memory variant of Nesterov's momentum [20, 23, 24], which can be viewed as a combination of heavy ball and Nesterov's momentum, yields a Type II PD error. However, compared with HFW, additional memory should be allocated. In sum, these observations suggest that heavy ball momentum is essentially critical to improve the PD performance of FW. Nesterov's momentum, on the other hand, does not influence PD error when used alone; however, it gives rise to faster (local) primal rates under additional assumptions [20, 23].

3.5 A side result: Directional smooth step sizes

Common to both FW and HFW is that the globally estimated L might be too pessimistic for a local update. In this subsection, a local Lipschitz constant is investigated to further improve the numerical efficiency of smooth step sizes in (9). This easily computed local Lipschitz constant is another merit of (H)FW over projection based approaches.

Definition 2. *(Directional Lipschitz continuous.) For two points $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, the directional Lipschitz constant $L(\mathbf{x}, \mathbf{y})$ ensures $\|\nabla f(\hat{\mathbf{x}}) - \nabla f(\hat{\mathbf{y}})\|_* \leq L(\mathbf{x}, \mathbf{y})\|\hat{\mathbf{x}} - \hat{\mathbf{y}}\|$ for any $\hat{\mathbf{x}} = (1-\alpha)\mathbf{x} + \alpha\mathbf{y}, \hat{\mathbf{y}} = (1-\beta)\mathbf{x} + \beta\mathbf{y}$ with some $\alpha \in [0, 1]$ and $\beta \in [0, 1]$.*

In other words, the directional Lipschitz continuity depicts the local property on the segment between points $\mathbf{x}$ and $\mathbf{y}$. It is clear that $L(\mathbf{x}, \mathbf{y}) \leq L$. Using logistic loss for binary classification as an example, we have $L(\mathbf{x}, \mathbf{y}) \leq \frac{1}{4N}\sum_{i=1}^{N} \frac{\langle \mathbf{a}_i, \mathbf{x}-\mathbf{y}\rangle^2}{\|\mathbf{x}-\mathbf{y}\|_2^2}$, where N is the number of data, and $\mathbf{a}_i$ is the feature of the ith datum. As a comparison, the global Lipschitz constant is $L \leq \frac{1}{4N}\sum_{i=1}^{N}\|\mathbf{a}_i\|_2^2$. We show in Appendix E that at least for a class of functions, including widely used logistic loss and quadratic loss, $L(\mathbf{x}, \mathbf{y})$ has an analytical form.

Simply replacing L in (9) with $L(\mathbf{x}_k, \mathbf{v}_{k+1})$, i.e.,

$$\eta_k = \max\left\{0, \min\left\{\frac{\langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{v}_{k+1}\rangle}{L(\mathbf{x}_k, \mathbf{v}_{k+1})\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}, 1\right\}\right\} \tag{10}$$

we can obtain what we term *directionally smooth step size*. Upon exploring the collinearity of $\mathbf{x}_k$, $\mathbf{x}_{k+1}$ and $\mathbf{v}_{k+1}$, a simple modification of Theorem 2 ensures the PD convergence.

Corollary 2. *Choosing $\delta_k = \frac{2}{k+2}$, and η_k via* (10)*, Alg. 2 ensures*

$$\mathcal{G}_k = f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k) \leq \frac{2LD^2}{k+1},\ \forall k \geq 1.$$

The directional Lipschitz constant can also replace the global one in other FW variants, such as [13,22], with theories therein still holding. As we shall see in numerical tests, directional smooth step sizes outperform the vanilla one by an order of magnitude.

4 Restart further tightens the PD error

Up till now it is established that the heavy ball momentum enables a unified analysis for tighter Type II PD bounds. In this section, we show that if the computational resources are sufficient for solving two FW subproblems per iteration, the PD error can be further improved by restart when the standard FW gap is smaller than generalized FW gap. Restart is typically employed by Nesterov's momentum in projection based methods [31] to cope with the robustness to parameter estimates, and to capture the local geometry of problem (1). However, it is natural to integrate restart with heavy ball momentum in FW regime. In addition, restart provides an answer to the following question: *which is smaller, the generalized FW gap or the vanilla one?* Previous works using the generalized FW gap have not addressed this question [20, 23, 28].

Algorithm 3 FW with heavy ball momentum and restart

1: **Initialize:** $\mathbf{x}_0^0 \in \mathcal{X}, \mathbf{g}_0^0 = \nabla f(\mathbf{x}_0^0)$, $s \leftarrow 0$, $C^0 = 0$, $\mathcal{G}_0^0 = \bar{\mathcal{G}}_0^0$
2: **while** [not terminated] **do**
3: $k \leftarrow 0$, $\mathbf{g}_0^s = \nabla f(\mathbf{x}_0^s)$
4: **while** [$\mathcal{G}_k^s \leq \bar{\mathcal{G}}_k^s$ or $k = 0$] and [not terminated] **do** ▷ Check whether restart is needed
5: $\delta_k^s = \frac{2}{k+2+C^s}$
6: $\mathbf{g}_{k+1}^s = (1-\delta_k^s)\mathbf{g}_k^s + \delta_k^s \nabla f(\mathbf{x}_k^s)$
7: $\mathbf{v}_{k+1}^s = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \mathbf{g}_{k+1}^s, \mathbf{x}\rangle$
8: $\mathbf{x}_{k+1}^s = (1-\eta_k^s)\mathbf{x}_k^s + \eta_k^s \mathbf{v}_{k+1}^s$
9: $\bar{\mathbf{v}}_{k+1}^s = \arg\min_{\mathbf{x}\in\mathcal{X}} \langle \nabla f(\mathbf{x}_{k+1}^s), \mathbf{x}\rangle$
10: $\mathcal{G}_{k+1}^s = f(\mathbf{x}_{k+1}^s) - \Phi_{k+1}^s(\mathbf{v}_{k+1}^s)$ ▷ Generalized FW gap
11: $\bar{\mathcal{G}}_{k+1}^s = \langle \nabla f(\mathbf{x}_k^s), \mathbf{x}_k^s - \bar{\mathbf{v}}_{k+1}^s\rangle$ ▷ Vanilla FW gap
12: $k \leftarrow k+1$
13: **end while**
14: $K_s \leftarrow k$, $\mathbf{x}_0^{s+1} = \mathbf{x}_{K_s}^s$, $C^{s+1} = \frac{2LD^2}{\mathcal{G}_{K_s}^s}$, $s \leftarrow s+1$
15: **end while**

FW with heavy ball momentum and restart is summarized under Alg. 3. For exposition clarity, when updating the counters such as k and s, we use notation '$\leftarrow$'. Alg. 3 contains two loops. The inner loop is the same as Alg. 2 except for computing a standard FW gap (Line 11) in addition to the generalized one (Line 10). The variable K_s, depicting the iteration number of inner loop s, is of analysis purpose. Alg. 3 can be terminated immediately whenever $\min\{\mathcal{G}_k^s, \bar{\mathcal{G}}_k^s\} \leq \epsilon$ for a desirable $\epsilon > 0$. The restart happens when the standard FW gap is smaller than generalized FW gap. And after restart, $\mathbf{g}_{k+1}^s$ will be reset. For Alg. 3, the linear functions used for generalized FW gap are defined stage-wisely

$$\Phi_0^s(\mathbf{x}) = f(\mathbf{x}_0^s) + \left\langle \nabla f(\mathbf{x}_0^s), \mathbf{x} - \mathbf{x}_0^s \right\rangle \tag{11a}$$

$$\Phi_{k+1}^s(\mathbf{x}) = (1-\delta_k^s)\Phi_k^s(\mathbf{x}) + \delta_k^s\left[f(\mathbf{x}_k^s) + \left\langle \nabla f(\mathbf{x}_k^s), \mathbf{x} - \mathbf{x}_k^s \right\rangle\right], \forall\, k \geq 0. \tag{11b}$$

It can be verified that $\mathbf{v}_{k+1}^s$ minimizes $\Phi_{k+1}^s(\mathbf{x})$ over $\mathcal{X}$ for any $k \geq 0$. In addition, we also have $f(\mathbf{x}_0^s) - \Phi_0^s(\mathbf{v}_0^s) = \bar{\mathcal{G}}_{K_{s-1}}^{s-1}$ where $\mathbf{v}_0^s = \arg\min_{\mathbf{x}\in\mathcal{X}} \Phi_0^s(\mathbf{x})$.

There are two tunable parameters η_k^s and δ_k^s. The choice on δ_k^s has been provided directly in Line 5, where it is adaptively decided using a variable C^s relating to the generalized FW gap. Three choices are readily available for η_k^s: i) $\eta_k^s = \delta_k^s$; ii) smooth step size:

$$\eta_k^s = \max\left\{0, \min\left\{\frac{\langle \nabla f(\mathbf{x}_k^s), \mathbf{x}_k^s - \mathbf{v}_{k+1}^s\rangle}{L\|\mathbf{v}_{k+1}^s - \mathbf{x}_k^s\|^2}, 1\right\}\right\}; \tag{12}$$

and, iii) line search

$$\eta_k^s = \arg\min_{\eta\in[0,1]} f\big((1-\eta)\mathbf{x}_k^s + \eta\mathbf{v}_{k+1}^s\big). \tag{13}$$

Note that the directionally smooth step size, i.e., replacing L with $L(\mathbf{x}_k^s, \mathbf{v}_{k+1}^s)$ in (12) is also valid for convergence. We omit it to reduce repetition. Next we show how restart improves the PD error.

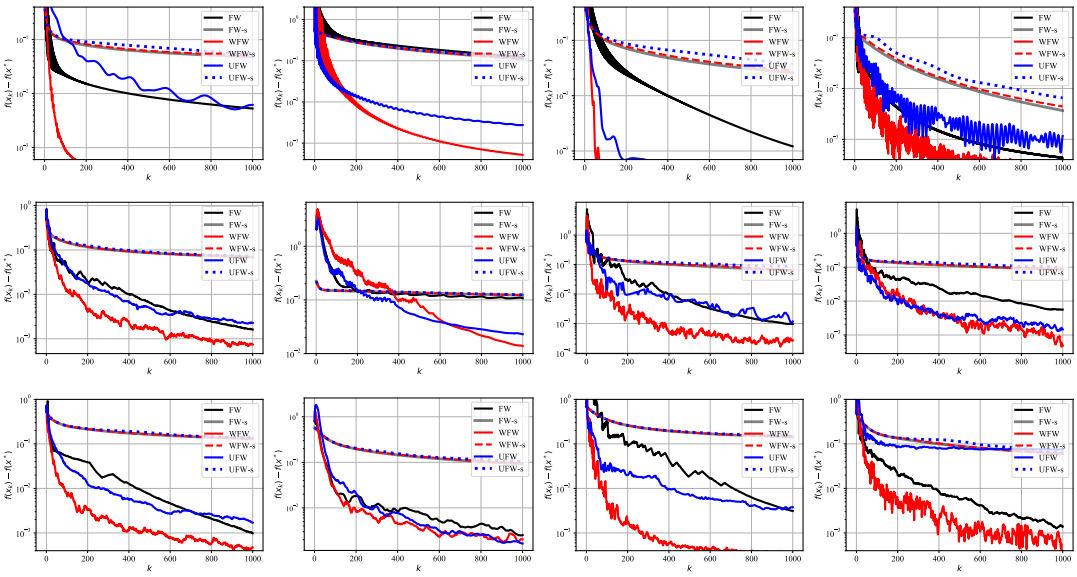
Theorem 4. *Choose η_k^s via one of the three manners: i) $\eta_k^s = \delta_k^s$; ii) as in* (12)*; or iii) as in* (13)*. If there is no restart (e.g., $s = 0$ when terminating), then Alg. 3 guarantees that*

$$\mathcal{G}_k^0 = f(\mathbf{x}_k^0) - \Phi_k(\mathbf{v}_k^0) \leq \frac{2LD^2}{k+1}, \forall k \geq 1. \tag{14a}$$

If restart happens, in additional to (14a)*, we have*

$$\mathcal{G}_k^s = f(\mathbf{x}_k^s) - \Phi_k(\mathbf{v}_k^s) < \frac{2LD^2}{k+C^s}, \forall k \geq 1, \forall s \geq 1, \;\; \textit{with} \;\; C^s \geq 1 + \sum_{j=0}^{s-1} K_j. \tag{14b}$$

Besides the convergence of both primal and dual errors of Alg. 3, Theorem 4 implies that when no restart happens, the generalized FW gap is smaller than the standard one, demonstrating that the



(a) *w7a* (b) *realsim* (c) *mushroom* (d) *ijcnn1*

Figure 1: Performance of FW variants for binary classification with the constraint being an ℓ_2-norm ball (first row), an ℓ_1-norm ball (second row), and an n-support norm ball (third row).

former is more suitable for the purpose of "stopping criterion". When restarted, Theorem 4 provides a strictly improved bound compared with Theorems 1, 2, and 6, since the denominator of the RHS in (14b) is no smaller than the total iteration number. An additional comparison with [28], where two subproblems are also required, once again confirms the power of heavy ball momentum to improve the constants in the PD error rate, especially with the aid of restart. The restart scheme (with slight modification) can also be employed in [23, 24, 28] to tighten their PD error.

5 Numerical tests

This section presents numerical tests to showcase the effectiveness of HFW on different machine learning problems. Since there are two parameters' choices for HFW in Theorems 1 and 3, we term them as weighted FW (WFW) and uniform FW (UFW), respectively, depending on the weight of $\{\nabla f(\mathbf{x}_k)\}$ in $\mathbf{g}_{k+1}$. When using smooth step size, the corresponding algorithms are marked as WFW-s and UFW-s. For comparison, the benchmark algorithms include: FW with $\eta_k = \frac{2}{k+2}$ (FW); and, FW with smooth step size (FW-s) in (4).

5.1 Binary classification

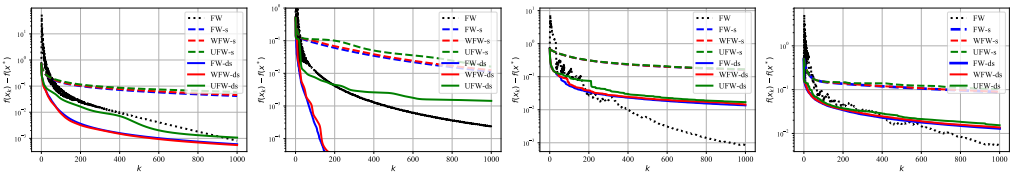
We first test the performance of Alg. 2 on binary classification using logistic regression

$$f(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \ln\left(1+\exp(-b_i\langle \mathbf{a}_i, \mathbf{x}\rangle)\right). \tag{15}$$

Here $(\mathbf{a}_i, b_i)$ is the (feature, label) pair of datum i, and N is the number of data. Datasets from LIBSVM[2] are used in the numerical tests, where details of the datasets are deferred to Appendix F due to space limitation.

ℓ_2-norm ball constraint. We start with $\mathcal{X} = \{\mathbf{x} | \|\mathbf{x}\|_2 \le R\}$. The primal errors are plotted in the first row of Fig. 1. We use primal error here for a fair comparison. It can be seen that the parameter-free step sizes achieve better performance compared with the smooth step sizes mainly because the quality of L estimate. Such a problem can be relived through directional smooth step sizes as we shall shortly. Among parameter-free step sizes, it can be seen that WFW consistently

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



(a) ℓ_2-norm ball (b) ℓ_2-norm ball (a) ℓ_1-norm ball (b) ℓ_1-norm ball

Figure 2: Performance of directionally smooth step sizes. (a) and (c) are tested on *mushroom*; and (b) and (d) use *ijcnn1*.

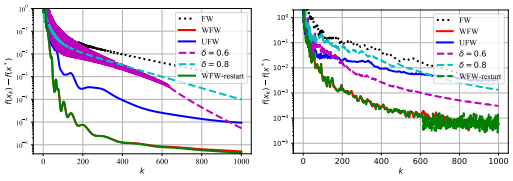
outperforms both UFW and FW on all tested datasets, while UFW converges faster than FW only on datasets *realsim* and *mushroom*. For smooth step sizes, the per-step-descent property is validated. The excellent performance of HFW can be partially explained by the similarity of its update, namely $\mathbf{x}_{k+1} = (1-\eta_k)\mathbf{x}_k + \eta_k R \frac{\mathbf{g}_{k+1}}{\|\mathbf{g}_{k+1}\|_2}$, with normalized gradient descent (NGD) one, that is given by $\mathbf{x}_{k+1} = \text{Proj}_{\mathcal{X}}\big(\mathbf{x}_k - \eta_k \frac{\mathbf{g}_{k+1}}{\|\mathbf{g}_{k+1}\|_2}\big)$. However, there is also a subtle difference between HFW and NGD updates. Indeed, when projection is in effect, $\mathbf{x}_{k+1}$ in NGD will lie on the boundary of the ℓ_2-norm ball. Due to the convex combination nature of the update in HFW, it is unlikely to have $\mathbf{x}_{k+1}$ on the boundary, though it can come arbitrarily close.

ℓ_1-norm ball constraint. Here $\mathcal{X} = \{\mathbf{x} | \|\mathbf{x}\|_1 \leq R\}$ denotes the constraint set that promotes sparse solutions. In the simulation, R is tuned for a solution with similar sparsity as the dataset itself. The results are showcased in the second row of Fig. 1. For smooth step sizes, FW-s, UFW-s, and WFW-s exhibit similar performances, and their curves are smooth. On the other hand, parameter-free step sizes eventually outperform smooth step sizes though the curves zig-zag. (The curves on *realsim* are smoothed to improve figure quality.) UFW has similar performance on *w7a* and *mushroom* with FW and faster convergence on other datasets. Once again, WFW consistently outperforms FW and UFW.

n-support norm ball constraint. The n-support norm ball is a tighter relaxation of a sparsity enforcing ℓ_0-norm ball combined with an ℓ_2-norm penalty compared with ElasticNet [38]. It gives rise to $\mathcal{X} = \text{conv}\{\mathbf{x} | \|\mathbf{x}\|_0 \leq n, \|\mathbf{x}\|_2 \leq R\}$, where $\text{conv}\{\cdot\}$ denotes the convex hull [3]. The closed-form solution of $\mathbf{v}_{k+1}$ is given in [25]. In the simulation, we choose $n = 2$ and tune R for a solution whose sparsity is similar to the adopted dataset. The results are showcased in the third row of Fig. 1. For smooth step sizes, FW-s and WFW-s exhibit similar performance, while UFW-s converges slightly slower on *ijcnn1*. Regarding parameter-free step sizes, UFW does not offer faster convergence compared with FW on the tested datasets, but WFW again has numerical merits.

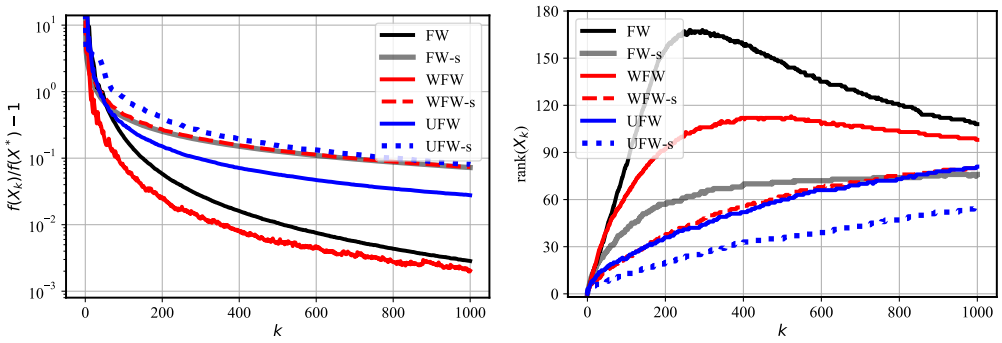
Directionally smooth step sizes. The results in Fig. 2 validate the effectiveness of directionally smooth (-ds) step sizes. For all datasets tested, the benefit of adopting $L(\mathbf{x}_k, \mathbf{v}_{k+1})$ is evident, as it improves the performance of smooth step sizes by an order of magnitude. In addition, it is also observed that UFW-ds performs worse than WFW-ds, which suggests that putting too much weight on past gradients could be less attractive in practice.

Additional comparisons. We also compare HFW with a generalized version of [27], where we set $\delta_k = \delta \in (0, 1), \forall k$ in Alg. 2. Two specific choices, i.e., $\delta = 0.6$, and $\delta = 0.8$, are plotted in Fig. 3, where the ℓ_2-norm ball and n-support norm ball are adopted as constraints. In both cases, WFW converges faster than the algorithm adapted from [27]. In addition, the choice of δ has major impact on convergence behavior, while WFW avoids this need for manual tuning of δ. The performance of WFW with restart, i.e., Alg. 3, is also shown in Fig. 3. Although it slightly outperforms WFW, restart also doubles the computational burden due to the need of solving two FW subproblems. From this point of view, WFW with restart is more of theoretical rather than practical interest. In addition, it is observed that Alg. 3 is not restarted after the first few iterations, which suggests that the generalized FW gap is smaller than the vanilla one, at least in the early stage of convergence. Thus, the generalized FW gap is attractive as a stopping criterion when a solution with moderate accuracy is desirable.



(a) ℓ_2 norm ball (b) n-supp norm ball

Figure 3: Comparison of HFW with other algorithms on *muchroom*.



(a) objective (b) rank

Figure 4: Performance of FW variants for matrix completion on *MovieLens100K*.

In a nutshell, the numerical experiments suggest that heavy ball momentum performs best with parameter-free step sizes with the momentum weight carefully adjusted. WFW is mainly recommended because it achieves improved empirical performance compared to UFW and FW, regardless of the constraint sets. The smooth step sizes on the other hand, eliminate the zig-zag behavior at the price of convergence slowdown due to the need of L, while directionally smooth step sizes can be helpful to alleviate this convergence slowdown.

5.2 Matrix completion

This subsection focuses on matrix completion problems for recommender systems. Consider a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ with partially observed entries, i.e., entries A_{ij} for $(i,j) \in \mathcal{K}$ are known, where $\mathcal{K} \subset \{1,\dots,m\} \times \{1,\dots,n\}$. Based on the observed entries that can be contaminated by noise, the goal is to predict the missing entries. Within the scope of recommender systems, a commonly adopted empirical observation is that $\mathbf{A}$ is low rank [4, 5, 8], leading to the following problem formulation.

$$\min_{\mathbf{X}} \ \frac{1}{2} \sum_{(i,j)\in\mathcal{K}} (X_{ij} - A_{ij})^2 \quad \text{s.t. } \|\mathbf{X}\|_{\text{nuc}} \le R. \tag{16}$$

Problem (16) is difficult to solve using GD because projection onto a nuclear norm ball requires a full SVD, which has complexity $\mathcal{O}\big(mn(m \wedge n)\big)$ with $(m \wedge n) := \min\{m, n\}$. In contrast, FW and its variants are more suitable for (16) since the FW subproblem has complexity less than $\mathcal{O}(mn)$ [2].

Heavy ball based FW are tested using dataset *MovieLens100K*[3]. Following the initialization of [11], the numerical results can be found in Fig. 4. Subfigures (a) and (b) depict the optimality error and rank versus k for $R = 3$. For parameter-free step sizes, WFW converges faster than FW while finding solutions with lower rank. The low rank solution of UFW is partially because it does not converge sufficiently. For smooth step sizes, UFW-s finds a solution with slightly larger objective value but much lower rank compared with WFW-s and FW-s. Overall, when a small optimality error is the priority, WFW is more attractive; while UFW-s is useful for finding low rank solutions.

6 Conclusions and future directions

This work demonstrated the merits of heavy ball momentum for FW. Multiple choices of the step size ensured a tighter Type II primal-dual error bound that can be efficiently computed when adopted as stopping criterion. An even tighter PD error bound can be achieved by relying jointly on heavy ball momentum and restart. A novel and general approach was developed to compute local Lipschitz constants in FW type algorithms. Numerical tests in the paradigms of logistic regression and matrix completion demonstrated the effectiveness of heavy ball momentum in FW.

Our future research agenda includes performance evaluation of heavy ball momentum for various learning tasks. For example, HFW holds great potential when fairness is to be accounted for [35].

[3] `https://grouplens.org/datasets/movielens/100k/`

Acknowledgement

This work is supported by NSF grants 1901134, 2126052, and 2128593. The authors would also like to thank the anonymous reviewers for their feedback.

References

[1] J. D. Abernethy and J.-K. Wang, "On Frank-Wolfe and equilibrium computation," in *Proc. Advances in Neural Info. Process. Syst.*, 2017, pp. 6584–6593.

[2] Z. Allen-Zhu, E. Hazan, W. Hu, and Y. Li, "Linear convergence of a Frank-Wolfe type algorithm over trace-norm balls," in *Proc. Advances in Neural Info. Process. Syst.*, 2017, pp. 6191–6200.

[3] 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.

[4] R. M. Bell and Y. Koren, "Lessons from the Netflix prize challenge." *SiGKDD Explorations*, vol. 9, no. 2, pp. 75–79, 2007.

[5] J. Bennett, S. Lanning *et al.*, "The Netflix prize," in *Proc. KDD cup and workshop*, vol. 2007. New York, NY, USA., 2007, p. 35.

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

[7] J. Diakonikolas and L. Orecchia, "The approximate duality gap technique: A unified theory of first-order methods," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 660–689, 2019.

[8] M. Fazel, "Matrix rank minimization with applications," 2002.

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

[10] R. M. Freund and P. Grigas, "New analysis and results for the Frank–Wolfe method," *Mathematical Programming*, vol. 155, no. 1-2, pp. 199–230, 2016.

[11] 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.

[12] 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.

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

[14] E. Ghadimi, H. R. Feyzmahdavian, and M. Johansson, "Global convergence of the heavy-ball method for convex optimization," in *Proc. of European control conference*, 2015, pp. 310–315.

[15] 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.

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

[17] 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.

[18] 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.

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

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

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

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

[23] B. Li, M. Coutino, G. B. Giannakis, and G. Leus, "A momentum-guided Frank-Wolfe algorithm," *IEEE Trans. on Signal Processing*, vol. 69, pp. 3597–3611, 2021.

[24] B. Li, L. Wang, G. B. Giannakis, and Z. Zhao, "Enhancing Frank Wolfe with an extra subproblem," in *Proc. of 35th AAAI Conf. on Artificial Intelligence*, 2021.

[25] 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.

[26] 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.

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

[28] Y. Nesterov, "Complexity bounds for primal-dual methods minimizing the model of objective function," *Mathematical Programming*, vol. 171, no. 1-2, pp. 311–330, 2018.

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

[30] T. Nguyen, X. Fu, and R. Wu, "Memory-efficient convex optimization for self-dictionary separable nonnegative matrix factorization: A frank-wolfe approach," *arXiv preprint arXiv:2109.11135*, 2021.

[31] B. O'donoghue and E. Candes, "Adaptive restart for accelerated gradient schemes," *Foundations of computational mathematics*, vol. 15, no. 3, pp. 715–732, 2015.

[32] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *Ussr computational mathematics and mathematical physics*, vol. 4, no. 5, pp. 1–17, 1964.

[33] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun, "Globally convergent parallel MAP LP relaxation solver using the Frank-Wolfe algorithm," in *Proc. Intl. Conf. on Machine Learning*, 2014, pp. 487–495.

[34] M. Ye, C. Gong, L. Nie, D. Zhou, A. Klivans, and Q. Liu, "Good subnetworks provably exist: Pruning via greedy forward selection," in *Proc. Intl. Conf. on Machine Learning*, 2020.

[35] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification," *The Journal of Machine Learning Research*, vol. 20, no. 1, pp. 2737–2778, 2019.

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

[37] M. Zhang, Z. Shen, A. Mokhtari, H. Hassani, and A. Karbasi, "One sample stochastic Frank-Wolfe," in *Proc. Intl. Conf. on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 4012–4023.

[38] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

Supplementary Document for "Heavy Ball Momentum for Conditional Gradient"

A Preludes

A.1 $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ for the smooth step sizes in Alg. 1

When using the step size (4) in Alg. 1, $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ is ensured automatically. To see this, we have from Assumption 1 that

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &\overset{(a)}{=} \eta_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \rangle + \frac{\eta_k^2 L}{2}\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2 \overset{(b)}{\leq} 0 \end{aligned} \tag{17}$$

where (a) uses $\mathbf{x}_{k+1} = (1-\eta_k)\mathbf{x}_k + \eta_k \mathbf{v}_{k+1}$; and (b) is because η_k minimizes the RHS of (17) over $[0, 1]$.

A.2 $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ for the smooth step sizes in Alg. 2

When using the step size (10) in Alg. 2, $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ is ensured.

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq \eta_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \rangle + \frac{\eta_k^2 L}{2}\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2 \leq 0$$

where the last ineqaulity is because η_k minimizes $\eta \langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \rangle + \frac{\eta^2 L}{2}\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2$ over $[0, 1]$.

B Missing proofs in Section 3.

B.1 Proof of Lemma 1

Proof. Using $\mathbf{g}_{k+1} = \sum_{\tau=0}^{k} w_k^\tau \nabla f(\mathbf{x}_\tau)$, we have

$$\arg\min_{\mathbf{x}\in\mathcal{X}} \Phi_{k+1}(\mathbf{x}) = \arg\min_{\mathbf{x}\in\mathcal{X}} \Big\langle \sum_{\tau=0}^{k} w_k^\tau \nabla f(\mathbf{x}_\tau), \mathbf{x} \Big\rangle = \arg\min_{\mathbf{x}\in\mathcal{X}} \big\langle \mathbf{g}_{k+1}, \mathbf{x} \big\rangle.$$

By comparing with Line 4 of Alg. 2, one can see that $\mathbf{v}_{k+1}$ is a minimizer of $\Phi_{k+1}(\mathbf{x})$ over $\mathcal{X}$. To prove that $\Phi_{k+1}(\mathbf{x})$ is a lower bound of $f(\mathbf{x})$, we appeal to convexity to write

$$\Phi_{k+1}(\mathbf{x}) = \sum_{\tau=0}^{k} w_k^\tau \big[f(\mathbf{x}_\tau) + \langle \nabla f(\mathbf{x}_\tau), \mathbf{x} - \mathbf{x}_\tau \rangle \big] \leq \sum_{\tau=0}^{k} w_k^\tau f(\mathbf{x}) = f(\mathbf{x})$$

where the last equation is because $\sum_{\tau=0}^{k} w_k^\tau = 1$ holds for any k. The proof is thus complete. □

B.2 Proof of Theorem 1

Proof. Using Assumption 1, we have

$$\begin{aligned} & f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \\ &\leq \big\langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \big\rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 \\ &= \eta_k \big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \big\rangle + \frac{\eta_k^2 L}{2}\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2. \end{aligned} \tag{18}$$

Inequality (18) is standard in the analysis of FW and its variants. Letting $\Phi_0(\mathbf{x}) \equiv 0$, and $\mathbf{v}_0$ be any point in $\mathcal{X}$, it can be verified that $\Phi_{k+1}(\mathbf{x}) = (1-\delta_k)\Phi_k(\mathbf{x}) + \delta_k \big[f(\mathbf{x}_k) + \big\langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \big\rangle \big]$,

from which we have

$$\begin{aligned}
&\Phi_{k+1}(\mathbf{v}_{k+1}) \qquad (19)\\
&= (1-\delta_k)\Phi_k(\mathbf{v}_{k+1}) + \delta_k\Big[f(\mathbf{x}_k) + \big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle\Big]\\
&\overset{(a)}{\geq} (1-\delta_k)\Phi_k(\mathbf{v}_k) + \delta_k\Big[f(\mathbf{x}_k) + \big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle\Big]
\end{aligned}$$

where (a) is because $1-\delta_k \geq 0$ and $\mathbf{v}_k$ minimizes $\Phi_k(\mathbf{x})$ over $\mathcal{X}$ (hence $\Phi_k(\mathbf{v}_k) \leq \Phi_k(\mathbf{v}_{k+1})$). Now subtracting $\Phi_{k+1}(\mathbf{v}_{k+1})$ on both sides of (18), we have

$$\begin{aligned}
&f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) \qquad (20)\\
&\overset{(b)}{\leq} (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2}\\
&\overset{(c)}{\leq} (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta_k^2 LD^2}{2}
\end{aligned}$$

where (b) uses $\eta_k = \delta_k$ and (19); and (c) relies on Assumption 3. For convenience, let $\Delta(i,j) := \prod_{\tau=i}^{j}(1-\delta_\tau)$, and unroll (20) to arrive at

$$\begin{aligned}
&f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1})\\
&\leq \Delta(0,k)\big[f(\mathbf{x}_0) - \Phi_0(\mathbf{v}_0)\big] + \sum_{\tau=0}^{k}\frac{LD^2\delta_\tau^2}{2}\Delta(\tau+1,k).
\end{aligned}$$

Plugging in the values of δ_k completes the proof. □

B.3 Proof of Theorem 2

Proof. The first a few steps are the same as the proof of Theorem 1; i.e., we have (18) and (19). Combining (18) and (19), we arrive at

$$\begin{aligned}
&f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) \qquad (21)\\
&\leq (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + (\eta_k - \delta_k)\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle + \frac{\eta_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2}.
\end{aligned}$$

It can be verified that the specific choice of η_k minimizes the RHS of (21) over $[0,1]$. Hence we have

$$\begin{aligned}
&f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) \qquad (22)\\
&\leq (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\eta_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2} + (\eta_k - \delta_k)\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle\\
&\overset{(a)}{\leq} (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\alpha_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2} + (\alpha_k - \delta_k)\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle\\
&\overset{(b)}{=} (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2}\\
&\leq \big[f(\mathbf{x}_0) - \Phi_0(\mathbf{v}_0)\big]\prod_{\tau=0}^{k}(1-\delta_\tau) + \sum_{\tau=0}^{k}\frac{LD^2\delta_\tau^2}{2}\prod_{j=\tau+1}^{k}(1-\delta_j)\\
&\leq \frac{2LD^2}{k+2}
\end{aligned}$$

where in (a) α_k can be chosen as any number in $[0,1]$; in (b) we set $\alpha_k = \delta_k$. This completes the proof. □

B.4 An extension of Theorem 2 for per step descent of $\mathcal{G}_k$

In this section, we show that it is possible to ensure per step descent on generalized FW gap when a more difficult subproblem can be solved. In particular, we will replace Line 4 of Alg. 2 and choose

parameters as

$$(\delta_k, \mathbf{v}_{k+1}) = \underset{\delta\in[0,1],\mathbf{v}\in\mathcal{X}}{\arg\min}\ (1-\delta)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta^2 L\|\mathbf{v}-\mathbf{x}_k\|^2}{2} \tag{23a}$$

$$\eta_k = \delta_k. \tag{23b}$$

It is clear that (23a) is harder to solve compared with a FW subproblem. The choice of δ_k enables an adaptive weights for $\nabla f(\mathbf{x}_k)$ in $\mathbf{g}_{k+1}$. Next we present the main result for such a parameter choice.

Theorem 5. *When Assumptions 1, 2 and 3 are satisfied, choosing $\mathbf{v}_{k+1}$, η_k and δ_k according to* (23)*, Alg. 2 guarantees that: i) $\mathcal{G}_{k+1} \leq \mathcal{G}_k$, and ii)*

$$\mathcal{G}_k = f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k) \leq \frac{2LD^2}{k+1},\ \forall k \geq 1.$$

Proof. It can be seen that (21) still holds, from which we have

$$\begin{aligned} & f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) \\ \leq\ & (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\eta_k^2 L\|\mathbf{v}_{k+1}-\mathbf{x}_k\|^2}{2} + (\eta_k-\delta_k)\big\langle\nabla f(\mathbf{x}_k), \mathbf{v}_{k+1}-\mathbf{x}_k\big\rangle \\ \overset{(a)}{=}\ & (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta_k^2 L\|\mathbf{v}_{k+1}-\mathbf{x}_k\|^2}{2} \end{aligned} \tag{24}$$

where (a) is because $\eta_k = \delta_k$. Then by the manner δ_k is chosen, we have

$$\begin{aligned} & f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) \\ =\ & (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta_k^2 L\|\mathbf{v}_{k+1}-\mathbf{x}_k\|^2}{2} \\ \overset{(b)}{\leq}\ & (1-\tilde{\delta}_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\tilde{\delta}_k^2 L\|\mathbf{v}_{k+1}-\mathbf{x}_k\|^2}{2} \end{aligned} \tag{25}$$

where in (b) $\tilde{\delta}_k \in [0,1]$. Choosing $\tilde{\delta}_k = 0$, we obtain $\mathcal{G}_{k+1} \leq \mathcal{G}_k$. Choosing $\tilde{\delta}_k = \frac{2}{k+2}$, we obtain the convergence rate. □

B.5 Line search for Alg. 2

We can also choose the step size η_k via line search, although this might be more computationally costly in practice because it requires computing the function value. The parameters are selected as

$$\delta_k = \frac{2}{k+2},\ \forall k \geq 0 \tag{26a}$$

$$\eta_k = \underset{\eta\in[0,1]}{\arg\min}\ f\big((1-\eta)\mathbf{x}_k + \eta\mathbf{v}_{k+1}\big). \tag{26b}$$

Such a parameter choice also ensures per step objective descent since

$$\begin{aligned} f(\mathbf{x}_{k+1}) &= \min_{\eta\in[0,1]} f\big((1-\eta)\mathbf{x}_k + \eta\mathbf{v}_{k+1}\big) \\ &\overset{(a)}{\leq} f\big((1-\theta)\mathbf{x}_k + \theta\mathbf{v}_{k+1}\big) \overset{(b)}{=} f(\mathbf{x}_k) \end{aligned}$$

where in (a) we have $\theta \in [0,1]$; and in (b) we set $\theta = 0$. Primal-dual convergence is established as follows.

Theorem 6. *If Assumptions 1-3 hold, while δ_k and η_k are chosen via* (26)*, Alg. 2 guarantees that*

$$\mathcal{G}_k = f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k) \leq \frac{2LD^2}{k+1},\ \forall k \geq 1.$$

Proof. Let $\tilde{\eta}_k = \frac{2}{k+2}, \forall k$. By the choice of η_k, we have

$$f(\mathbf{x}_{k+1}) = \min_{\eta\in[0,1]} f\big((1-\eta)\mathbf{x}_k + \eta\mathbf{v}_{k+1}\big) \leq f\big((1-\tilde{\eta}_k)\mathbf{x}_k + \tilde{\eta}_k\mathbf{v}_{k+1}\big). \tag{27}$$

Then using smoothness, we arrive at

$$\begin{aligned} & f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \\ \leq\ & f\big((1-\tilde{\eta}_k)\mathbf{x}_k + \tilde{\eta}_k\mathbf{v}_{k+1}\big) - f(\mathbf{x}_k) \\ \leq\ & \tilde{\eta}_k\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle + \frac{\tilde{\eta}_k^2 L}{2}\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2. \end{aligned} \tag{28}$$

Then combining (28) and (19), and following the same steps in (20), we can prove this theorem. □

Through Theorem 6 it is straightforward to derive the primal and dual convergence, respectively, following the same argument of Corollary 1. For this reason, it is omitted here.

B.6 Proof of Theorem 3

Proof. It can be seen that (21) still holds.

Parameter-free step size. Plugging in $\delta_k = \eta_k = \frac{1}{k+1}$ into (21), we arrive at

$$\begin{aligned} f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) &\leq (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2} \\ &\leq \Delta(0,k)\big[f(\mathbf{x}_0) - \Phi_0(\mathbf{v}_0)\big] + \sum_{\tau=0}^{k}\frac{LD^2\delta_\tau^2}{2}\Delta(\tau+1,k) \\ &= \mathcal{O}\Big(\frac{LD^2\ln(k+2)}{k+1}\Big) \end{aligned} \tag{29}$$

where $\Delta(i,j) := \prod_{\tau=i}^{j}(1-\delta_\tau)$, $\Phi_0(\mathbf{x}) \equiv 0$, and $\mathbf{v}_0$ is any point in $\mathcal{X}$.

Smooth step size. Notice that the choice of η_k minimizes the RHS of (21) when δ_k is fixed, then we have

$$\begin{aligned} & f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) \\ \leq\ & (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + (\eta_k - \delta_k)\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle + \frac{\eta_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2} \\ \overset{(a)}{\leq}\ & (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + (\tilde{\eta}_k - \delta_k)\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle + \frac{\tilde{\eta}_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2} \\ \overset{(b)}{\leq}\ & (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2} \\ =\ & \mathcal{O}\Big(\frac{LD^2\ln(k+2)}{k+1}\Big) \end{aligned} \tag{30}$$

where in (a) $\tilde{\eta}_k \in [0,1]$; and in (b) we set $\tilde{\eta}_k = \delta_k$.

Line search. When η_k is chosen via line search, we have for any $\tilde{\eta}_k \in [0,1]$

$$f(\mathbf{x}_{k+1}) = \min_{\eta\in[0,1]} f\big((1-\eta)\mathbf{x}_k + \eta\mathbf{v}_{k+1}\big) \leq f\big((1-\tilde{\eta}_k)\mathbf{x}_k + \tilde{\eta}_k\mathbf{v}_{k+1}\big). \tag{31}$$

Then by smoothness, we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &\leq f\big((1-\tilde{\eta}_k)\mathbf{x}_k + \tilde{\eta}_k\mathbf{v}_{k+1}\big) - f(\mathbf{x}_k) \\ &\leq \tilde{\eta}_k\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle + \frac{\tilde{\eta}_k^2 L}{2}\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2. \end{aligned} \tag{32}$$

Then using the same argument as the derivation of (21), we can obtain

$$\begin{aligned} & f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) \\ \leq\ & (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + (\tilde{\eta}_k - \delta_k)\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k\big\rangle + \frac{\tilde{\eta}_k^2 L\|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2}. \end{aligned} \tag{33}$$

Simply setting $\tilde{\eta}_k = \frac{1}{k+1}$, and using the same derivation as in (30), the proof can be completed. □

B.7 Proof for choosing $\delta_k = \delta$

When Assumptions 1 is satisfied w.r.t. ℓ_2-norm, we show the following parameter choice in Alg. 2 leads to convergence as well.

$$\delta_k = \delta,\ \eta_k = \frac{c}{k+k_0},\ \forall k \geq 0 \tag{34}$$

where $\delta \in (0,1)$, and c and k_0 are constants to be specified later. Due to the choice of $\delta_k = \delta$, $\mathbf{g}_{k+1}$ is an exponentially moving average of previous gradients. Note that the moving average was adopted in [27] for stochastic FW to reduce the mean square error of the noisy gradient. However, we use it in a totally different purpose.

Lemma 2. *Choose parameters as in* (34)*. Suppose there exist a constant* c_0 *that satisfies*

$$c_1^2 \leq \left[1-(1-\delta)\frac{(k_0+1)^2}{k_0^2}\right]\delta c_0^2 \tag{35}$$

then it is guaranteed that

$$\|\mathbf{g}_{k+1} - \nabla f(\mathbf{x}_k)\|_2^2 \leq \frac{c_0^2 L^2 D^2}{(k+k_0)^2}.$$

Proof.

$$\begin{aligned}
&\|\mathbf{g}_{k+1} - \nabla f(\mathbf{x}_k)\|_2^2 \qquad (36)\\
&= (1-\delta)^2\|\mathbf{g}_k - \nabla f(\mathbf{x}_k)\|_2^2\\
&= (1-\delta)^2\|\mathbf{g}_k - \nabla f(\mathbf{x}_{k-1}) + \nabla f(\mathbf{x}_{k-1}) - \nabla f(\mathbf{x}_k)\|_2^2\\
&\overset{(a)}{\leq} (1-\delta)^2(1+\theta)\|\mathbf{g}_k - \nabla f(\mathbf{x}_{k-1})\|_2^2 + (1-\delta)^2(1+\frac{1}{\theta})\|\nabla f(\mathbf{x}_{k-1}) - \nabla f(\mathbf{x}_k)\|_2^2\\
&\overset{(b)}{\leq} (1-\delta)^2(1+\theta)\|\mathbf{g}_k - \nabla f(\mathbf{x}_{k-1})\|_2^2 + (1-\delta)^2(1+\frac{1}{\theta})L^2\eta_{k-1}^2\|\mathbf{x}_{k-1} - \mathbf{v}_k\|_2^2\\
&\overset{(c)}{\leq} (1-\delta)^2(1+\theta)\|\mathbf{g}_k - \nabla f(\mathbf{x}_{k-1})\|_2^2 + (1-\delta)^2(1+\frac{1}{\theta})L^2D^2\eta_{k-1}^2\\
&\overset{(d)}{\leq} (1-\delta)\|\mathbf{g}_k - \nabla f(\mathbf{x}_{k-1})\|_2^2 + (1-\delta)^2(1+\frac{1}{\delta})L^2D^2\eta_{k-1}^2\\
&\overset{(e)}{\leq} (1-\delta)\|\mathbf{g}_k - \nabla f(\mathbf{x}_{k-1})\|_2^2 + L^2D^2\frac{\eta_{k-1}^2}{\delta}
\end{aligned}$$

where (a) is by Young's inequality with $\theta > 0$ to be specified later; (b) follows from Assumption 1; (c) is because Assumption 3; in (d) we choose $\theta = \delta < 1$ and use the fact that $(1-\delta)^2(1+\delta) \leq (1-\delta)$; and (e) uses $\delta \leq 1$ so that $(1-\delta)^2(1+\frac{1}{\delta}) = \frac{1}{\delta} - 1 + \delta^2 - 2\delta \leq \frac{1}{\delta}$.

We proof this lemma by induction. Given the choice of $\mathbf{g}_0 = \nabla f(\mathbf{x}_0)$, we must have $\mathbf{g}_1 = \nabla f(\mathbf{x}_0)$, which implies $\|\mathbf{g}_1 - \nabla f(\mathbf{x}_0)\|_2^2 = 0 \leq \frac{c_0^2 L^2 D^2}{k_0^2}$ directly. Next we assume that $\|\mathbf{g}_k - \nabla f(\mathbf{x}_{k-1})\|_2^2 \leq \frac{c_0^2 L^2 D^2}{(k-1+k_0)^2}$ holds for some $k \geq 1$. Using (36), we have

$$\begin{aligned}
\|\mathbf{g}_{k+1} - \nabla f(\mathbf{x}_k)\|_2^2 &\leq (1-\delta)\|\mathbf{g}_k - \nabla f(\mathbf{x}_{k-1})\|_2^2 + L^2D^2\frac{\eta_{k-1}^2}{\delta}\\
&\leq (1-\delta)\frac{c_0^2 L^2 D^2}{(k+k_0-1)^2} + L^2D^2\frac{\eta_{k-1}^2}{\delta}\\
&\leq (1-\delta)\frac{c_0^2 L^2 D^2}{(k+k_0-1)^2} + L^2D^2\frac{c_1^2}{\delta(k+k_0)^2}\\
&= (1-\delta)\frac{c_0^2 L^2 D^2}{(k+k_0)^2}\frac{(k+k_0)^2}{(k+k_0-1)^2} + L^2D^2\frac{c_1^2}{\delta(k+k_0)^2}\\
&\leq (1-\delta)\frac{c_0^2 L^2 D^2}{(k+k_0)^2}\frac{(k_0+1)^2}{k_0^2} + L^2D^2\frac{c_1^2}{\delta(k+k_0)^2}\\
&\leq \frac{c_0^2 L^2 D^2}{(k+k_0)^2} \qquad (37)
\end{aligned}$$

where the last inequality comes from the choice of c_1. The proof is thus completed. □

To avoid the complexity of choosing constants, we consider an instance where $k_0 = 2$, $\delta = 0.8$, $c_1 = 2$, and $c_0 \approx 3.05$. It can be verified that (35) is satisfied. Then applying Lemma 2, the convergence of Alg.2 can be obtained.

Theorem 7. *Let* $\mathbf{g}_0 = \nabla f(\mathbf{x}_0)$, $\eta_k = \frac{2}{k+3}$, *and* $\delta = 0.8$. *Then for* $\forall k \geq 1$, *the convergence rate of Alg. 2 with* (34) *is*

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

Proof. Using Assumption 1, we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) &\leq f(\mathbf{x}_k) - f(\mathbf{x}^*) + \big\langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \big\rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \qquad (38)\\ &= f(\mathbf{x}_k) - f(\mathbf{x}^*) + \eta_k \big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \big\rangle + \frac{\eta_k^2 L}{2}\|\mathbf{v}_{k+1} - \mathbf{x}_k\|_2^2 \\ &\leq f(\mathbf{x}_k) - f(\mathbf{x}^*) + \eta_k \big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \big\rangle + \frac{\eta_k^2 LD^2}{2}. \end{aligned}$$

Next we have

$$\begin{aligned} \big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \big\rangle &= \big\langle \nabla f(\mathbf{x}_k), \mathbf{x}^* - \mathbf{x}_k \big\rangle + \big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}^* \big\rangle \\ &\overset{(a)}{\leq} f(\mathbf{x}^*) - f(\mathbf{x}_k) + \big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}^* \big\rangle \\ &= f(\mathbf{x}^*) - f(\mathbf{x}_k) + \big\langle \mathbf{g}_{k+1}, \mathbf{v}_{k+1} - \mathbf{x}^* \big\rangle + \big\langle \nabla f(\mathbf{x}_k) - \mathbf{g}_{k+1}, \mathbf{v}_{k+1} - \mathbf{x}^* \big\rangle \\ &\overset{(b)}{\leq} f(\mathbf{x}^*) - f(\mathbf{x}_k) + \big\langle \nabla f(\mathbf{x}_k) - \mathbf{g}_{k+1}, \mathbf{v}_{k+1} - \mathbf{x}^* \big\rangle \\ &\leq f(\mathbf{x}^*) - f(\mathbf{x}_k) + D\|\nabla f(\mathbf{x}_k) - \mathbf{g}_{k+1}\|_2 \qquad (39) \end{aligned}$$

where (a) is by the convexity of $f(\mathbf{x})$; (b) is because $\mathbf{v}_{k+1}$ minimizes $\langle \mathbf{g}_{k+1}, \mathbf{x} \rangle$ over $\mathcal{X}$; and the last inequality relies on Cauchy-Schwarz inequality and Assumption 3. Plugging (39) into (38), we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq (1 - \eta_k)\big[f(\mathbf{x}_k) - f(\mathbf{x}^*)\big] + \eta_k D\|\nabla f(\mathbf{x}_k) - \mathbf{g}_{k+1}\|_2 + \frac{\eta_k^2 LD^2}{2}. \qquad (40)$$

Let $\xi_k = \frac{\eta_k c_0 LD^2}{k+k_0} + \frac{\eta_k^2 LD^2}{2}$, then we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) &\leq (1 - \eta_k)\big[f(\mathbf{x}_k) - f(\mathbf{x}^*)\big] + \eta_k D\|\nabla f(\mathbf{x}_k) - \mathbf{g}_{k+1}\|_2 + \frac{\eta_k^2 LD^2}{2} \\ &\leq (1 - \eta_k)\big[f(\mathbf{x}_k) - f(\mathbf{x}^*)\big] + \xi_k \\ &= \big[f(\mathbf{x}_0) - f(\mathbf{x}^*)\big] \prod_{\tau=0}^{k}(1 - \eta_\tau) + \sum_{\tau=0}^{k} \xi_\tau \prod_{j=\tau+1}^{k}(1 - \eta_j) \\ &= \mathcal{O}\Big(\frac{LD^2}{k}\Big). \qquad (41) \end{aligned}$$

The proof is thus completed. □

B.8 Additional discussions

Many of existing works, e.g., [14], study (projected) heavy ball momentum by introducing auxiliary variables $\mathbf{z}_k$ such that the update on variable $\mathbf{x}_k$ can be viewed as a "gradient update" on $\mathbf{z}_k$, i.e., $\mathbf{z}_{k+1} = \mathbf{z}_k - \eta \nabla f(\mathbf{x}_k)$. By constructing the $\{\mathbf{z}_k\}$ sequence, it is possible to view heavy ball momentum approximately as GD. Though this trick is smart and analytically convenient, it does not give too much insight for the heavy ball momentum itself.

By comparing the use of heavy ball momentum in FW and GD, it may suggest new perspectives. For example, one can view Alg.2 as the dual-averaging version of FW as well. This suggests that it is intriguing to study (projected) heavy ball momentum from dual-averaging point of view. This is slightly off the main theme of this work, and we leave it for future research.

C Stopping criterion

Recall that for a prescribed $\epsilon > 0$, having $f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k) \leq \epsilon$ directly implies $f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \epsilon$. Next, we show how to update $\Phi_k(\mathbf{v}_k)$ iteratively in order to obtain a stopping criterion. Let us note that

$$\begin{aligned}\Phi_{k+1}(\mathbf{x}) &= \sum_{\tau=0}^{k} w_k^\tau \big[f(\mathbf{x}_\tau) + \langle \nabla f(\mathbf{x}_\tau), \mathbf{x} - \mathbf{x}_\tau\rangle\big] \\ &= \sum_{\tau=0}^{k} w_k^\tau \big[f(\mathbf{x}_\tau) - \langle \nabla f(\mathbf{x}_\tau), \mathbf{x}_\tau\rangle\big] + \langle \mathbf{g}_{k+1}, \mathbf{x}\rangle \\ &:= C_{k+1} + \langle \mathbf{g}_{k+1}, \mathbf{x}\rangle,\ \forall k \geq 0.\end{aligned}$$

Hence, to compute $\Phi_{k+1}(\mathbf{v}_{k+1})$, we only need to update C_{k+1} iteratively. A simple derivation leads to

$$\begin{aligned}C_{k+1} = (1-\delta_k)C_k + \delta_k\Big[f(\mathbf{x}_k) - \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k\rangle\Big], \\ \text{with } C_1 = f(\mathbf{x}_0) - \langle \nabla f(\mathbf{x}_0), \mathbf{x}_0\rangle.\end{aligned} \tag{42}$$

In sum, one can efficiently obtain $\Phi_{k+1}(\mathbf{v}_{k+1})$ as

$$\Phi_{k+1}(\mathbf{v}_{k+1}) = C_{k+1} + \langle \mathbf{g}_{k+1}, \mathbf{v}_{k+1}\rangle \tag{43}$$

with C_{k+1} recursively updated via (42).

D Missing proofs in Section 4

D.1 Proof of Theorem 4

Proof. Consider the case where $\eta_k^s = \delta_k^s$. Using Assumption 1, we have

$$\begin{aligned}f(\mathbf{x}_{k+1}^s) - f(\mathbf{x}_k^s) &\leq \big\langle \nabla f(\mathbf{x}_k^s), \mathbf{x}_{k+1}^s - \mathbf{x}_k^s\big\rangle + \frac{L}{2}\|\mathbf{x}_{k+1}^s - \mathbf{x}_k^s\|^2 \\ &= \eta_k^s\big\langle \nabla f(\mathbf{x}_k^s), \mathbf{v}_{k+1}^s - \mathbf{x}_k^s\big\rangle + \frac{(\eta_k^s)^2 L}{2}\|\mathbf{v}_{k+1}^s - \mathbf{x}_k^s\|^2.\end{aligned} \tag{44}$$

Then we have

$$\begin{aligned}\Phi_{k+1}^s(\mathbf{v}_{k+1}^s) &= (1-\delta_k^s)\Phi_k^s(\mathbf{v}_{k+1}^s) + \delta_k^s\Big[f(\mathbf{x}_k^s) + \big\langle \nabla f(\mathbf{x}_k^s), \mathbf{v}_{k+1}^s - \mathbf{x}_k^s\big\rangle\Big] \\ &\geq (1-\delta_k^s)\Phi_k^s(\mathbf{v}_k^s) + \delta_k^s\Big[f(\mathbf{x}_k^s) + \big\langle \nabla f(\mathbf{x}_k^s), \mathbf{v}_{k+1}^s - \mathbf{x}_k^s\big\rangle\Big].\end{aligned} \tag{45}$$

Now subtracting $\Phi_{k+1}^s(\mathbf{v}_{k+1}^s)$ on both sides of (44), we have

$$\begin{aligned}&f(\mathbf{x}_{k+1}^s) - \Phi_{k+1}^s(\mathbf{v}_{k+1}^s) \\ &\leq f(\mathbf{x}_k^s) + \eta_k^s\big\langle \nabla f(\mathbf{x}_k^s), \mathbf{v}_{k+1}^s - \mathbf{x}_k^s\big\rangle + \frac{(\eta_k^s)^2 L\|\mathbf{v}_{k+1}^s - \mathbf{x}_k^s\|^2}{2} - \Phi_{k+1}^s(\mathbf{v}_{k+1}^s) \\ &\overset{(a)}{\leq} (1-\delta_k^s)\big[f(\mathbf{x}_k^s) - \Phi_k^s(\mathbf{v}_k^s)\big] + \frac{(\delta_k^s)^2 L\|\mathbf{v}_{k+1}^s - \mathbf{x}_k^s\|^2}{2} \\ &\overset{(b)}{\leq} (1-\delta_k^s)\big[f(\mathbf{x}_k^s) - \Phi_k^s(\mathbf{v}_k^s)\big] + \frac{(\delta_k^s)^2 LD^2}{2}\end{aligned} \tag{46}$$

where (a) uses $\eta_k^s = \delta_k^s$ and (45); and (b) relies on Assumption 3. For convenience, let us define $\Delta^s(i,j) := \prod_{\tau=i}^{j}(1-\delta_\tau^s)$. Then unrolling (46), we get

$$\begin{aligned}&f(\mathbf{x}_{k+1}^s) - \Phi_{k+1}^s(\mathbf{v}_{k+1}^s) \\ &\leq \Delta^s(0,k)\big[f(\mathbf{x}_0^s) - \Phi_0^s(\mathbf{v}_0^s)\big] + \sum_{\tau=0}^{k}\frac{LD^2(\delta_\tau^s)^2}{2}\Delta^s(\tau+1,k) \\ &\leq \frac{C^s(C^s+1)}{(k+1+C^s)(k+2+C^s)}\big[f(\mathbf{x}_0^s) - \Phi_0^s(\mathbf{v}_0^s)\big] + \frac{2(k+1)LD^2}{(k+1+C^s)(k+2+C^s)}.\end{aligned}$$

When $s = 0$, plugging $C^0 = 0$, we have

$$f(\mathbf{x}_{k+1}^0) - \Phi_{k+1}(\mathbf{v}_{k+1}^0) \leq \frac{2LD^2}{k+2}. \tag{47}$$

Hence (14a) in Theorem 4 is proved. Next consider $s \geq 1$. Using the observation that $f(\mathbf{x}_0^s) - \Phi_0^s(\mathbf{v}_0^s) = \bar{\mathcal{G}}_{K_{s-1}}^{s-1} < \mathcal{G}_{K_{s-1}}^{s-1}$, we then have

$$\begin{aligned} &\mathcal{G}_{k+1}^s = f(\mathbf{x}_{k+1}^s) - \Phi_{k+1}^s(\mathbf{v}_{k+1}^s) \qquad (48)\\ &< \frac{C^s(C^s+1)}{(k+1+C^s)(k+2+C^s)}\mathcal{G}_{K_{s-1}}^{s-1} + \frac{2(k+1)LD^2}{(k+1+C^s)(k+2+C^s)}\\ &\overset{(c)}{=} \frac{2LD^2(C^s+1)}{(k+1+C^s)(k+2+C^s)} + \frac{2(k+1)LD^2}{(k+1+C^s)(k+2+C^s)} = \frac{2LD^2}{k+1+C^s}. \end{aligned}$$

where (c) uses the definition of C^s. Hence (14b) in Theorem 4 is proved.

Finally, we only need to show that $C^s \geq 1 + \sum_{j=0}^{s-1} K_j$ by induction. First by definition of $C^1 = 2LD^2/(\mathcal{G}_{K_0}^0)$, with $\mathcal{G}_{K_0}^0 \leq \frac{2LD^2}{K_0+1}$, it is clear that $C^1 \geq 1 + K_0$. Then suppose $C^s \geq 1 + \sum_{j=0}^{s-1} K_j$ hold for some s, we will show that $C^{s+1} \geq 1 + \sum_{j=0}^{s} K_j$.

Using (48), we have $C^{s+1} = 2LD^2/(\mathcal{G}_{K_s}^s) \geq C^s + K_s \geq 1 + \sum_{j=0}^{s-1} K_j + K_s$. Hence (14b) is proved.

For the smooth step size (12) and line search (13), the same bound can be obtained by using the same arguments as in Theorems 2 and 6. Hence they are omitted here. □

E Directionally smooth step size

E.1 Proof of Corollary 2

Proof. Using Definition 2 and following the standard derivation of descent lemma [29, Lemma 1.2.3], we can show that

$$\begin{aligned} &f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \qquad (49)\\ &\leq \eta_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \rangle + \frac{\eta_k^2 L(\mathbf{x}_k, \mathbf{x}_{k+1})}{2} \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2\\ &\leq \eta_k \langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \rangle + \frac{\eta_k^2 L(\mathbf{x}_k, \mathbf{v}_{k+1})}{2} \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2. \end{aligned}$$

The reason for $L(\mathbf{x}_k, \mathbf{v}_{k+1}) \geq L(\mathbf{x}_k, \mathbf{x}_{k+1})$ is that $\mathbf{x}_{k+1}$ lives in between $\mathbf{x}_k$ and $\mathbf{v}_{k+1}$. Although $L(\mathbf{x}_k, \mathbf{x}_{k+1})$ can provide a tighter bound, it is not tractable.

Combining (49) and (19), we have

$$\begin{aligned} &f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) \qquad (50)\\ &\leq (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + (\eta_k - \delta_k)\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \big\rangle + \frac{\eta_k^2 L(\mathbf{x}_k, \mathbf{v}_{k+1}) \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2}. \end{aligned}$$

It can be verified that the specific choice of η_k in (10) minimizes the RHS of (50) over $[0, 1]$. Hence we have

$$\begin{aligned} &f(\mathbf{x}_{k+1}) - \Phi_{k+1}(\mathbf{v}_{k+1}) \qquad (51)\\ &\leq (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\eta_k^2 L(\mathbf{x}_k, \mathbf{v}_{k+1}) \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2} + (\eta_k - \delta_k)\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \big\rangle\\ &\overset{(a)}{\leq} (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\alpha_k^2 L(\mathbf{x}_k, \mathbf{v}_{k+1}) \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2} + (\alpha_k - \delta_k)\big\langle \nabla f(\mathbf{x}_k), \mathbf{v}_{k+1} - \mathbf{x}_k \big\rangle\\ &\overset{(b)}{=} (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta_k^2 L(\mathbf{x}_k, \mathbf{v}_{k+1}) \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2}\\ &\overset{(c)}{=} (1-\delta_k)\big[f(\mathbf{x}_k) - \Phi_k(\mathbf{v}_k)\big] + \frac{\delta_k^2 L \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2}{2}\\ &\leq \frac{2LD^2}{k+2} \end{aligned}$$

where in (a) α_k can be chosen as any number in $[0,1]$; in (b) we set $\alpha_k = \delta_k$; and (c) uses $L(\mathbf{x}_k, \mathbf{v}_{k+1}) \leq L$. This completes the proof. □

E.2 Computing directionally smooth constant

Define a one dimensional function $g(\eta) := f\big(\mathbf{x}_k + \eta(\mathbf{v}_{k+1} - \mathbf{x}_k)\big)$, where $\mathrm{dom}\ \eta = [0,1]$. Then it is clear that $\nabla g(\eta) = \langle \mathbf{v}_{k+1} - \mathbf{x}_k, \nabla f\big(\mathbf{x}_k + \eta(\mathbf{v}_{k+1} - \mathbf{x}_k)\big)\rangle$. Therefore, it is easy to see that $g(\eta)$ is smooth, i.e.,

$$\begin{aligned}\big|\nabla g(\eta_1) - \nabla g(\eta_2)\big| &= |\langle \mathbf{v}_{k+1} - \mathbf{x}_k, \nabla f\big(\mathbf{x}_k + \eta_1(\mathbf{v}_{k+1} - \mathbf{x}_k)\big) - \nabla f\big(\mathbf{x}_k + \eta_2(\mathbf{v}_{k+1} - \mathbf{x}_k)\big)\rangle| \\ &\leq \|\mathbf{v}_{k+1} - \mathbf{x}_k\| \big\|\nabla f\big(\mathbf{x}_k + \eta_1(\mathbf{v}_{k+1} - \mathbf{x}_k)\big) - \nabla f\big(\mathbf{x}_k + \eta_2(\mathbf{v}_{k+1} - \mathbf{x}_k)\big)\big\|_* \\ &\leq L(\mathbf{x}_k, \mathbf{v}_{k+1}) \|\mathbf{v}_{k+1} - \mathbf{x}_k\|^2 |\eta_1 - \eta_2| \end{aligned} \tag{52}$$

On the other hand, one can also analytically find L_g by definition; i.e., $\big|\nabla g(\eta_1) - \nabla g(\eta_2)\big| \leq L_g\big|\eta_1 - \eta_2\big|$. Comparing L_g with RHS of (52), we can obtain $L(\mathbf{x}_k, \mathbf{v}_{k+1})$. This method can be applied when f is e.g., quadratic loss and logistic loss.

F More on numerical tests

All numerical experiments are performed using Python 3.7 on an Intel i7-4790CPU @3.60 GHz (32 GB RAM) desktop.

F.1 Binary classification

Table 2: A summary of datasets used in numerical tests

Dataset	d	N (train)	nonzeros
w7a	300	24,692	3.89%
realsim	20,958	50,617	0.24%
mushromm	122	8,124	18.75%
ijcnn1	22	49,990	40.91%

Sparsity promoting property of FW variants for ℓ_1-norm ball constraint. FW in Alg. 1 directly promotes sparsity on the solution if it is initialized at $\mathbf{x}_0 = \mathbf{0}$. To see this, suppose that the i-th entry of $\nabla f(\mathbf{x}_k)$ has the largest absolute value, then we have $\mathbf{v}_{k+1} = [0, \ldots, -\mathrm{sgn}\big([\nabla f(\mathbf{x}_k)]_i\big)R, \ldots, 0]^\top$ with the i-th entry being non-zero. Hence, $\mathbf{x}_k$ has at most k non-zero entries given $k-1$ entries are non-zero in $\mathbf{x}_{k-1}$. This sparsity promoting property also holds for Alg. 2 for the same reason.

F.2 Matrix completion

The dataset used for the test is *MovieLens100K*, where 1682 movies are rated by 943 users with 6.30% ratings observed. The initialization and data processing are the same as those used in [11].

Besides the projection-free property, FW and its variants are more suitable for problem (16) compared to GD because they also guarantee $\mathrm{rank}(\mathbf{X}_k) \leq k+1$ [11, 15]. Take FW in Alg. 1 for example. First it is clear that $\nabla f(\mathbf{X}_k) = (\mathbf{X}_k - \mathbf{A})_{\mathcal{K}}$. Suppose that the SVD of $\nabla f(\mathbf{X}_k)$ is given by $\nabla f(\mathbf{X}_k) = \mathbf{P}_k \mathbf{\Sigma}_k \mathbf{Q}_k^\top$. Then the FW subproblem can be solved easily by

$$\boldsymbol{V}_{k+1} = -R\mathbf{p}_k\mathbf{q}_k^\top \tag{53}$$

where $\mathbf{p}_k$ and $\mathbf{q}_k$ denote the left and right singular vectors corresponding to the largest singular value of $\nabla f(\mathbf{X}_k)$, respectively. Clearly $\boldsymbol{V}_{k+1}$ in (53) has rank at most 1. Hence it is easy to see $\mathbf{X}_{k+1} = (1-\delta_k)\mathbf{X}_k + \delta_k \boldsymbol{V}_{k+1}$ has rank at most $k+2$ if $\mathbf{X}_k$ is a rank-$(k+1)$ matrix (i.e., $\mathbf{X}_0$ has rank 1). Using similar arguments, Alg. 2 also ensures $\mathrm{rank}(\mathbf{X}_k) \leq k+1$. Therefore, the low rank structure is directly promoted by FW variants.