Heavy Ball Momentum for Conditional Gradient

Bingcong Li  Alireza Sadeghi  Georgios B. Giannakis
University of Minnesota - Twin Cities
Minneapolis, MN, USA
{lxx5599, 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

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

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 $x$. Throughout, we let $x^* \in \mathcal{X}$ 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., $v_{k+1} \in \arg\min_{v \in \mathcal{X}} \langle \nabla f(x_k), v \rangle$ at $k$th iteration, and then updates $x_{k+1}$ as a convex combination of $x_k$ and $v_{k+1}$. When dealing with a structured $\mathcal{X}$, a closed-form or efficient solution for $v_{k+1}$ 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 $O(\frac{1}{\epsilon})$ linear subproblems are required to ensure $f(x_k) - f(x^*) \leq \epsilon$, 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 $f(x_k) - f(x^*)$. It is guaranteed for FW that $f(x_k) - f(x^*) = O(1/k)$, $\forall k \geq 1$ [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].

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.”

<table>
<thead>
<tr>
<th>reference</th>
<th>computation</th>
<th>PD conv. type</th>
<th>PD conv. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>[16]</td>
<td>1 subproblem</td>
<td>Type I</td>
<td>( \frac{2Ld^2}{k(R+1)} )</td>
</tr>
<tr>
<td>[18]</td>
<td>2 subproblems</td>
<td>Type II</td>
<td>( \frac{2Ld^2}{\sqrt{k+1}} )</td>
</tr>
<tr>
<td>[28]</td>
<td>2 subproblems</td>
<td>Type II</td>
<td>( \frac{LD^2}{k+1} )</td>
</tr>
<tr>
<td>This work (Alg. 2)</td>
<td>1 subproblem</td>
<td>Type II</td>
<td>( \frac{LD^2}{k+1} )</td>
</tr>
<tr>
<td>This work (Alg. 3)</td>
<td>2 subproblems</td>
<td>Type II</td>
<td>( \frac{2LD^2}{k+1+c} )</td>
</tr>
</tbody>
</table>

**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 \( \epsilon > 0 \), \( f(x_k) - f(x^*) \leq \epsilon \) 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 \( \min_{k \in \{1,...,K\}} \text{PDError}_k = O\left(\frac{1}{R}\right) \), 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., \( \text{PDError}_k \leq O\left(\frac{1}{k}\right) \), \( \forall k \). 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); \( \|x\| \) stands for a norm of a vector \( x \), whose dual norm is denoted by \( \|x\|_* \); and \( \langle x, y \rangle \) is the inner product of \( x \) and \( y \).

## 2 Preliminaries

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

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

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

**Assumption 3.** \( \text{(Convex and compact constraint set.)} \) The constraint set \( \mathcal{X} \subset \mathbb{R}^d \) is convex and compact with diameter \( D \), that is, \( \|x - y\| \leq D, \forall x, y \in \mathcal{X} \).
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 \( f(x) \) at \( x_k \), i.e.,
\[
\mathbf{v}_{k+1} \in \arg \min_{v \in \mathcal{X}} f(x_k) + \langle \nabla f(x_k), v - x_k \rangle. 
\]
(2)

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), \( x_{k+1} \) is updated as a convex combination of \( \mathbf{v}_{k+1} \) and \( x_k \) in Line 4 so that no projection is required. The choices on the step size \( \eta_k \) will be discussed shortly.

The PD error of Alg. 1 is captured by the so-called FW gap, formally defined as
\[
\mathcal{G}_k := \langle \nabla f(x_k), x_k - v_{k+1} \rangle = f(x_k) - f(x^*) + f(x^*) - \min_{v \in \mathcal{X}} \left[ f(x_k) + \langle \nabla f(x_k), v - x_k \rangle \right] 
\]
(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 \( \mathcal{G}_k \) converges, one can deduce that the primal error converges. For this reason, \( \mathcal{G}_k \) 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 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 \( f(x_k) - f(x^*) \leq \frac{2LDD^2}{k+1}, \forall k \geq 1 \), and a weaker claim on the PD error, \( \min_{k \in \{1, ..., K\}} \mathcal{G}_k = \frac{2LDD^2}{k} \) [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]
\[
\eta_k = \min \left\{ \frac{\langle \nabla f(x_k), x_k - v_{k+1} \rangle}{L\|v_{k+1} - x_k\|^2}, 1 \right\}. 
\]
(4)

Despite the estimated \( L \) is typically too pessimistic to capture the local Lipschitz continuity, such a step size ensures \( f(x_{k+1}) \leq f(x_k) \); 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, \( \langle \nabla f(x_k), x_k - v_{k+1} \rangle \) 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 \( \ell_2 \)-norm ball, that is, \( \mathcal{X} = \{ x | \| x \|_2 \leq R \} \). 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 \( \{ \nabla f(x_k) \} \), resulting
in a more concentrated sequence \(\{v_{k+1}\}\). Recall that the PD error is closely related to \(v_{k+1}\) [cf. equation (3)]. We hope the “concentration”, of \(\{v_{k+1}\}\) to be helpful in reducing the changes of PD error among consecutive iterations so that a Type II PD error bound is attainable.

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 \(g_{k+1}\) explicitly as \(g_{k+1} = \sum_{\tau=0}^{k} w_{\tau}^T \nabla f(x_{\tau})\), where \(w_{\tau}^T = \delta_{\tau} \prod_{\tau'=\tau+1}^{\tau}(1 - \delta_{j}) > 0\), \(\forall \tau \geq 1\), and \(w_{0}^T = \prod_{j=1}^{k}(1 - \delta_{j}) > 0\). Then, define a sequence of linear functions \(\{\Phi_{k}(x)\}\) as

\[
\Phi_{k+1}(x) := \sum_{\tau=0}^{k} w_{\tau}^T [f(x_{\tau}) + \langle \nabla f(x_{\tau}), x - x_{\tau} \rangle], \quad \forall k \geq 0.
\] (5)

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

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

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

**Definition 1.** (Generalized FW gap.) The generalized FW gap w.r.t. \(\Phi_{k}(x)\) is

\[
G_{k} := f(x_{k}) - \min_{x \in X} \Phi_{k}(x) = f(x_{k}) - \Phi_{k}(v_{k}).
\] (6)

In words, the generalized FW gap is defined as the difference between \(f(x_{k})\) and the minimal value of \(\Phi_{k}(x)\) over \(X\). The newly defined \(G_{k}\) also illustrates the PD error

\[
G_{k} = f(x_{k}) - \Phi_{k}(v_{k}) = f(x_{k}) - f(x^{*}) + f(x^{*}) - \Phi_{k}(v_{k}).
\] (7)

For the dual error, we have \(f(x^{*}) - \Phi_{k}(v_{k}) \geq \Phi_{k}(x^{*}) - \Phi_{k}(v_{k}) \geq 0\), where both inequalities follow from Lemma 1. Hence, \(G_{k} \geq 0\) automatically serves as an overestimate of both primal and dual errors. When establishing the convergence of \(G_{k}\), 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

\[
\delta_{k} = \eta_{k} = \frac{2}{k + 2}, \quad \forall k \geq 0.
\] (8)

Such a choice on \(\delta_{k}\) puts more weight on recent gradients when calculating \(g_{k+1}\), since \(w_{\tau}^T = \mathcal{O}(\frac{1}{\tau^{2}})\). The following theorem specifies the convergence of \(G_{k}\).

**Theorem 1.** If Assumptions 1-3 hold, then choosing \(\delta_{k}\) and \(\eta_{k}\) as in (8), Alg. 2 guarantees that

\[
G_{k} = f(x_{k}) - \Phi_{k}(v_{k}) \leq \frac{2LD^2}{k + 1}, \quad \forall k \geq 1.
\]

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.

---

**Algorithm 2 FW with heavy ball momentum**

1. **Initialize:** \(x_{0} \in \mathcal{X}, g_{0} = \nabla f(x_{0})\)
2. **for** \(k = 0, 1, \ldots, K - 1\) **do**
3. \(g_{k+1} = (1 - \delta_{k})g_{k} + \delta_{k}\nabla f(x_{k})\)
4. \(v_{k+1} = \arg \min_{v \in X} \langle g_{k+1}, v \rangle\)
5. \(x_{k+1} = (1 - \eta_{k})x_{k} + \eta_{k}v_{k+1}\)
6. **end for**
7. **Return:** \(x_{K}\)
Corollary 1. Choosing the parameters as in Theorem 1, then $\forall k \geq 1$, Alg. 2 guarantees that

$$
\text{primal conv.: } f(x_k) - f(x^*) \leq \frac{2LD^2}{k + 1}; \quad \text{dual conv.: } f(x^*) - \Phi_k(v_k) \leq \frac{2LD^2}{k + 1}.
$$

Proof. Combine Theorem 1 with $f(x_k) - f(x^*) \leq G_k$ and $f(x^*) - \Phi_k(v_k) \leq G_k$ [cf. (7)]. $\square$

3.3 Smooth step size

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

$$
\delta_k = \frac{2}{k + 2} \quad \text{and} \quad \eta_k = \max \left\{ 0, \min \left\{ \frac{\langle \nabla f(x_k), x_k - v_{k+1} \rangle}{L\|v_{k+1} - x_k\|^2}, 1 \right\} \right\}.
$$

(9)

Comparing with the smooth step size for vanilla FW in (4), it can be deduced that the choice on $\eta_k$ in (9) has to be trimmed to $[0, 1]$ manually. This is because $\langle \nabla f(x_k), x_k - v_{k+1} \rangle$ is no longer guaranteed to be positive. The smooth step size enables an adaptive means of adjusting the weight for $\nabla f(x_k)$. To see this, note that when $\eta_k = 0$, we have $x_{k+1} = x_k$. As a result, $g_{k+2} = (1 - \delta_{k+1})g_{k+1} + \delta_{k+1}\nabla f(x_{k+1}) = (1 - \delta_{k+1})g_{k+1} + \delta_{k+1}\nabla f(x_k)$, that is, the weight on $\nabla f(x_k)$ is adaptively increased to $\delta_{k}(1 - \delta_{k+1}) + \delta_{k+1}$ if one further unpacks $g_{k+1}$. 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 $\eta_k$ and $\delta_k$ are chosen as in (9), Alg. 2 guarantees that

$$
G_k = f(x_k) - \Phi_k(v_k) \leq \frac{2LD^2}{k + 1}, \quad \forall k \geq 1.
$$

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., $G_{k+1} \leq G_k$.

Line search. When choosing $\delta_k = \frac{2}{k + 2}$ and $\eta_k$ 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 $G_k$ for using as stopping criterion is also described in Appendix C.

3.4 Further considerations

There are more choices of $\delta_k$ and $\eta_k$ leading to (primal) convergence. For example, one can choose $\delta_k \equiv \delta \in (0, 1)$ and $\eta_k = O\left(\frac{1}{k}\right)$ 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 $\ell_2$-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 $g_{k+1}$ as $\nabla f(x_k)$ 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 $\delta_k$ makes $g_{k+1}$ a uniform average of all gradients. A slower convergence rate $f(x_k) - f(x^*) = O\left(\frac{LD^2 \ln k}{k}\right)$ 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 $\eta_k$ 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

$$
G_k = f(x_k) - \Phi_k(v_k) \leq \frac{LD^2 \ln(k + 1)}{2k}, \quad \forall k \geq 1.
$$

$^1$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 $g_{k+1}$, 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 $\{x_k\}$. 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 $\infty$-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 $x, y \in \mathcal{X}$, the directional Lipschitz constant $L(x, y)$ ensures $\|\nabla f(\hat{x}) - \nabla f(\hat{y})\|_\ast \leq L(x, y)\|x - y\|$ for any $\hat{x} = (1 - \alpha)x + \alpha y$, $\hat{y} = (1 - \beta)x + \beta y$ 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 $x$ and $y$. It is clear that $L(x, y) \leq L$. Using logistic loss for binary classification as an example, we have $L(x, y) \leq \frac{1}{N^2} \sum_{i=1}^{N} \frac{(a_i, x - y)^2}{\|x - y\|^2}$, where $N$ is the number of data, and $a_i$ is the feature of the $i$th datum. As a comparison, the global Lipschitz constant is $L \leq \frac{1}{N^2} \sum_{i=1}^{N} \|a_i\|^2 / 2$. We show in Appendix E that at least for a class of functions, including widely used logistic loss and quadratic loss, $L(x, y)$ has an analytical form.

Simply replacing $L$ in (9) with $L(x_k, v_{k+1})$, i.e.,

$$\eta_k = \max \left\{0, \min \left\{ \frac{\langle \nabla f(x_k), x_k - v_{k+1} \rangle}{L(x_k, v_{k+1})\|v_{k+1} - x_k\|^2}, 1 \right\} \right\}$$  (10)

we can obtain what we term *directionally smooth step size*. Upon exploring the collinearity of $x_k$, $x_{k+1}$ and $v_{k+1}$, a simple modification of Theorem 2 ensures the PD convergence.

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

$$g_k = f(x_k) - \Phi_k(v_k) \leq \frac{2LD^2}{k+1}, \forall k \geq 1.$$  

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].
where it is adaptively decided using a variable $C$

Algorithm 3

\[ \eta \]

Besides the convergence of both primal and dual errors of Alg. 3, Theorem 4 implies that when no \( s \) there is no restart (e.g., \( \Theta \)). Note that the directionally smooth step size, i.e., replacing \( L \) are readily available for \( f \).

It can be verified that \( \epsilon > 0 \)

and, iii) line search

\[ k \leftarrow k + 1 \]

end while

\[ K_s \leftarrow k, x_{0}^{s+1} = x_{K_s}^{s+1} = \frac{2LD^2}{\epsilon K_s} , s \leftarrow s + 1 \]

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 ‘\(~\)’. 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 \( \min \{ G_k^s, G_k^s \} \leq \epsilon \) for a desirable \( \epsilon > 0 \). The restart happens when the standard FW gap is smaller than generalized FW gap. And after restart, \( g_k^{s+1} \) will be reset. For Alg. 3, the linear functions used for generalized FW gap are defined stage-wisely

\[ \Phi_0^s(x) = f(x_0^s) + \langle \nabla f(x_0^s), x - x_0^s \rangle \]

\[ \Phi_k^{s+1}(x) = (1 - \delta_k^s) \Phi_k^s(x) + \delta_k^s \{ f(x_k^s) + \langle \nabla f(x_k^s), x - x_k^s \rangle \}, \forall k \geq 0. \]

It can be verified that \( v_k^{s+1} \) minimizes \( \Phi_k^{s+1}(x) \) over \( X \) for any \( k \geq 0 \). In addition, we also have \( f(x_k^s) - \Phi_0^s(v_0^s) = G_k^{s-1} \), where \( v_0^s = \arg \min_{x \in X} \Phi_0^s(x) \).

There are two tunable parameters \( \eta_k^s \) and \( \delta_k^s \). The choice on \( \delta_k^s \) 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 \( \eta_k^s \); i) \( \eta_k^s = \delta_k^s \); ii) smooth step size:

\[ \eta_k^s = \max \left\{ 0, \min \left\{ \frac{\langle \nabla f(x_k^s), x_k^s - v_{k+1}^s \rangle}{L\|v_{k+1}^s - x_k^s\|^2}, 1 \right\} \right\} ; \quad (12) \]

and, iii) line search

\[ \eta_k^s = \arg \min_{\eta \in [0,1]} f\left( (1 - \eta)x_k^s + \eta v_{k+1}^s \right). \quad (13) \]

Note that the directionally smooth step size, i.e., replacing \( L \) with \( L(x_k^s, v_{k+1}^s) \) 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 \( \eta_k^s \) 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

\[ G_k^0 = f(x_k^0) - \Phi_k(v_k^0) \leq \frac{2LD^2}{k + 1}, \forall k \geq 1. \quad (14a) \]

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

\[ G_k^s = f(x_k^s) - \Phi_k(v_k^s) < \frac{2LD^2}{k + C^s}, \forall k \geq 1, \forall s \geq 1, \text{ with } C^s \geq 1 + \sum_{j=0}^{s-1} K_j. \quad (14b) \]

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
We first test the performance of Alg. 2 on binary classification using logistic regression (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 (FW) and uniform FW (UFW), respectively, depending on the weight of \( \{ \nabla f(x_k) \} \) in \( g_{k+1} \). 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

\[
    f(x) = \frac{1}{N} \sum_{i=1}^{N} \ln \left( 1 + \exp(-b_i \langle a_i, x \rangle) \right)
\]

Here \((a_i, b_i)\) 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.

\(\ell_2\)-norm ball constraint. We start with \( \mathcal{X} = \{ x | \|x\|_2 \leq R \} \). 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

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 $x_{k+1} = (1 - \eta_k)x_k + \eta_k R \frac{g_k}{\|g_k\|^2}$, with normalized gradient descent (NGD) one, that is given by $x_{k+1} = \text{Proj}_{\mathcal{X}}(x_k - \eta_k \frac{g_k}{\|g_k\|^2})$. However, there is also a subtle difference between HFW and NGD updates. Indeed, when projection is in effect, $x_{k+1}$ in NGD will lie on the boundary of the $\ell_2$-norm ball. Due to the convex combination nature of the update in HFW, it is unlikely to have $x_{k+1}$ on the boundary, though it can come arbitrarily close.

**$\ell_1$-norm ball constraint.** Here $\mathcal{X} = \{x|\|x\|_1 \leq R\}$ 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 with 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 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.

**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 $\ell_2$-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 $\delta$ has major impact on convergence behavior, while WFW avoids this need for manual tuning of $\delta$. 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.
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 $\mathcal{K} \subset \{1,\ldots,m\} \times \{1,\ldots,n\}$. 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.

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

Problem (16) is difficult to solve using GD because projection onto a nuclear norm ball requires a full SVD, which has complexity $O(mn(m \wedge n))$ 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 $O(mn)$ [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


