Noname manuscript No.
(will be inserted by the editor)

Robustness of the Frank-Wolfe Method under Inexact Oracles and the Cost of Linear Minimization

Tao Hu

Abstract We investigate the robustness of the Frank-Wolfe method when gradients are computed inexactly and examine the relative computational cost of the linear minimization oracle (LMO) versus projection. For smooth nonconvex functions, we establish a convergence guarantee of order $\mathcal{O}(1/\sqrt{k}+\delta)$ for Frank-Wolfe with a δ–oracle. Our results strengthen previous analyses for convex objectives and show that the oracle errors do not accumulate asymptotically. We further prove that approximate projections cannot be computationally cheaper than accurate LMOs, thus extending to the case of inexact projections. These findings reinforce the robustness and efficiency of the Frank-Wolfe framework.

Keywords Frank-Wolfe method · Inexact oracle · Projection vs. LMO

1 Introduction

The Frank-Wolfe method is a projection-free method for constrained optimization. At each iteration, Frank-Wolfe solves a linear minimization subproblem instead of a projection, which can be advantageous on structured domains such as simplices or spectrahedra. Classical analyses guarantee an $\mathcal{O}(1/k)$ convergence rate for convex objectives [6,4] and $\mathcal{O}(1/\sqrt{k})$ for nonconvex ones [7].

Despite its popularity, the robustness of Frank-Wolfe under inexact gradient information remains incompletely understood. In deterministic optimization, the δ–oracle model [3] is used to bound the gradient error. We revisit this framework and derive new nonaccumulation bounds for both convex and nonconvex settings.

Tao Hu
School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, P.R. China
E-mail: hu_tao@stu.xjtu.edu.cn

Let $Q \subset \mathbb{R}^d$ be a compact convex set and $f : Q \to \mathbb{R}$ be the objective function. Denote $\|\cdot\|$ to be the l^2 norm. We consider the minimization problem:

$$\begin{aligned} &\min_x f(x) \\ &\text{s.t. } x \in Q \ . \end{aligned}$$

The Frank-Wolfe method, which has a linear minimization oracle as its basic module, is an effective way to address this problem. At the iteration point $x_k \in Q$, the Frank-Wolfe method solves the linear minimization subproblem

$$\tilde{x}_k \in \arg\min_{x \in Q} \left\{ f(x_k) + \nabla f(x_k)^\top (x - x_k) \right\}$$

and updates with $x_{k+1} = (1 - \overline{\alpha}_k)x_k + \overline{\alpha}_k \tilde{x}_k$ where $\overline{\alpha}_k \in [0, 1)$.

In the convergence analysis of the Frank–Wolfe method, the following auxiliary sequences are frequently used and will also appear in our proofs:

$$\beta_k = \frac{1}{\prod_{j=0}^{k-1}(1 - \overline{\alpha}_j)}, \qquad \alpha_k = \frac{\beta_k \overline{\alpha}_k}{1 - \overline{\alpha}_k}, \qquad k \geq 1. \tag{1}$$

Here, $\{\overline{\alpha}_k\}_{k=0}^{+\infty}$ denotes the sequence of step sizes used in our algorithm. We follow the conventions $\prod_{j=0}^{-1}(\cdot) := 1$ and $\sum_{i=0}^{-1}(\cdot) := 0$.

We denote the Frank-Wolfe gap at the point $x \in Q$ as

$$G(x) = \sup_{y \in Q} \nabla f(x)^T (x - y).$$

Besides the convergence guarantee, robustness and the efficiency of the Linear Minimization Oracle are also important aspects of the Frank-Wolfe method.

The performance of Frank-Wolfe under inexact gradient is a very interesting problem. With unbiased gradients and bounded variance (or sub-Gaussian tails), Stochastic Frank-Wolfe variants achieve a Frank-Wolfe gap of $\mathcal{O}(\varepsilon)$ with $\mathcal{O}(1/\varepsilon^4)$ gradient evaluations, and variance reduction accelerates finite-sum problems and can achieve the same Frank-Wolfe gap with $\mathcal{O}(1/\varepsilon^3)$ gradient evaluations [11,5,9,8,16,12]. For heavy-tailed noise, Stochastic Frank-Wolfe with clipping or robust estimation achieves high-probability guarantees [14, 13].

In the deterministic setting, there is situation where the gradient error is bounded by δ/D but can be arbitrarily chosen along the training trajectory. This case is often relaxed and referred to as obtaining a δ-oracle:

$$\left| (g_\delta(x) - \nabla f(x))^T (x - y) \right| \leq \delta, \, \forall \, y \in Q. \tag{2}$$

If we use an increasingly accurate gradient along the iterates, like if

$$\left| (g_\delta(x_k) - \nabla f(x_k))^T (x_k - y) \right| \leq \frac{1}{k+1} \delta L D^2, \, \forall \, y \in Q,$$

then

$$G(x_k) \leq \frac{27 L D^2}{4(k+2)} (1 + \delta).$$

For the more common scenario, where the error does not decrease, Freund and Grigas prove an $\mathcal{O}(1/k + \delta)$ convergence of the Frank-Wolfe gap[4], and we show an $\mathcal{O}(1/\sqrt{k} + \delta)$ convergence for nonconvex functions in this paper.

Sometimes we consider objective functions that are convex but non-smooth or the case when the gradients are computed at shifted points [2]. Those functions may not obtain a gradient, but they can be equipped with a (δ, L) oracle [3]:

$$0 \leq f(x) - (f_{\delta,L}(y) + g_{\delta,L}(y)^T(x-y)) \leq \frac{L}{2}\|x-y\|^2 + \delta, \, \forall \, x, y \in Q.$$

Since the first bound of the Frank-Wolfe gap under $(\delta, L)-$ oracle, which is $\mathcal{O}(1/k + k\delta)$, has been proposed[4], it has been an open problem for more than ten years whether the final guarantee of the Frank-Wolfe gap is optimal theoretically.

Besides robustness, the ease of computing the Linear Minimization Oracle is widely considered another major advantage of the Frank–Wolfe method, which makes it more prevalent than proximal gradient methods. However, this belief is currently supported mainly by intuition and set-specific comparisons [1,10]. Additionally, we address a recent question posed by Woodstock [15] on whether linear minimization oracles are inherently cheaper than projection operators. We show that even coarse, approximate projections cannot outperform accurate LMOs in computational complexity.

Our main contributions are as follows:

(i) **Nonconvex Frank-Wolfe with a δ-oracle.** We prove that for L-smooth nonconvex objectives, Frank-Wolfe with a δ-oracle achieves

$$\min_{0\leq k\leq K} G(x^k) \;\leq\; \sqrt{\frac{2C\,(f(x^0) - f^*)}{K+1}} \;+\; 2\delta.$$

(ii) **Projection vs. LMO.** We prove that a K-approximate projection at a scaled point $-\lambda x$ produces an ε-accurate LMO at x with $\varepsilon = \mathcal{O}((K + D^2)/\lambda)$, establishing that approximate projections can not be uniformly easier than LMOs.

2 Frank-Wolfe with an Inexact Oracle

Let $Q \subset \mathbb{R}^d$ be compact and convex, and $f : Q \to \mathbb{R}$ be convex with L-Lipschitz gradient. The Frank-Wolfe update in Algorithm 1 uses the δ-oracle g_δ defined in (2).

Lemma 1 *Under (2), for any $x_k \in Q$,*

$$f^\star \;\geq\; f(x_k) \;+\; \min_{x\in Q} g_\delta(x_k)^\top (x - x_k) \;-\; \delta.$$

Algorithm 1 Frank-Wolfe with a δ-oracle

1: Initialize $x_0 \in Q$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Query $g_\delta(x_k)$.
4: Solve $\tilde{x}_k = \arg\min_{x \in Q} g_\delta(x_k)^\top (x - x_k)$.
5: Set $x_{k+1} = x_k + \overline{\alpha}_k(\tilde{x}_k - x_k)$, $\overline{\alpha}_k \in [0, 1)$.
6: **end for**

Proof By convexity, $f(x) \geq f(x_k) + \nabla f(x_k)^\top (x - x_k)$ for any $x \in Q$. From (2), $\nabla f(x_k)^\top (x - x_k) \geq g_\delta(x_k)^\top (x - x_k) - \delta$. Therefore,

$$f(x) \geq f(x_k) + g_\delta(x_k)^T (x - x_k) - \delta.$$

Taking $\min_{x \in Q}$ on both sides yields the claim.

We also recall a subproblem-level accuracy transfer.

Proposition 2 ([4, Prop. 5.1]) *Fix $\overline{x} \in Q$ and $\delta \geq 0$. If $\tilde{x} \in \arg\min_{x \in Q} g_\delta(\overline{x})^\top x$, then*

$$\nabla f(\overline{x})^\top \tilde{x} \;\leq\; \min_{x \in Q} \nabla f(\overline{x})^\top x \;+\; 2\delta.$$

The convergence theorem of Frank-Wolfe with a δ- oracle on convex objectives is given by Freund and Grigas as follows (one can actually show that the result of Freund and Grigas actually applies to the widest range of step-sizes):

Theorem 3 (Nonaccumulation under δ-oracle, convex case[4]) *Let Q be compact convex with diameter D, and f be convex with L-Lipschitz gradient on Q. Let g_δ satisfy (2). For the Frank-Wolfe iterates of Algorithm 1 with stepsize satisfying $\sum_{k=0}^{+\infty} \overline{\alpha}_k = \infty$, $\sum_{k=0}^{+\infty} \overline{\alpha}_k^2 < \infty$ and $\overline{\alpha}_k \downarrow 0$, then*

$$f(x_{k+1}) - f^* \leq (1 - \overline{\alpha}_k)\big(f(x_k) - f^*\big) + 2\overline{\alpha}_k \delta + \tfrac{1}{2} L D^2 \overline{\alpha}_k^2, \tag{3}$$

and hence $\limsup_{k \to \infty} (f(x_k) - f^\star) \leq 2\delta$.

Example 4 (Tightness up to constants) *Let $Q = [-1, 1]$, $f(x) = \frac{1}{2}x^2$ (convex, $L = 1$, $D = 2$). Define a δ-oracle by $g_\delta(x) = \nabla f(x) - \frac{\delta}{D} \operatorname{sign}(x)$. Frank-Wolfe with $\overline{\alpha}_k = 2/(k + 2)$ converges to a neighborhood whose size is proportional to δ.*

3 Nonconvex Frank-Wolfe with an Inexact Oracle

We now consider *nonconvex* minimization over a compact convex set $S \subset \mathbb{R}^d$:

$$\min_{x \in S} f(x),$$

where f is differentiable and has L-Lipschitz gradient on S. Denote $D := \operatorname{Diam}(S)$ and set

$$C \;\triangleq\; \max\{\, LD^2,\; GD \,\} \quad \text{with} \quad G := \sup_{x \in S} \|\nabla f(x)\| < \infty.$$

The Frank-Wolfe gap at x is

$$G(x) \triangleq \max_{s\in S} \langle \nabla f(x),\, x - s\rangle.$$

We assume access to a δ-oracle for the gradient, i.e., for every $x \in S$ there exists $g_\delta(x)$ such that

$$\big|\langle \nabla f(x) - g_\delta(x),\, s - x\rangle\big| \;\le\; \delta \quad \forall\, s \in S. \tag{4}$$

Define the *approximate Frank-Wolfe gap*

$$\tilde{G}(x) \triangleq \max_{s\in S} \langle g_\delta(x),\, x - s\rangle.$$

From (4) it follows that

$$|\, G(x) - \tilde{G}(x)\,| \;\le\; \delta. \tag{5}$$

Algorithm 2 Nonconvex Frank-Wolfe with a δ-oracle

Input: $x^0 \in S$, curvature constant $C \ge \max\{LD^2,\, GD\}$, error level $\delta \ge 0$.
for $k = 0, 1, 2, \ldots$ **do**
  Query $g_\delta(x^k)$ that satisfies (4).
  Solve $s^k = \arg\min_{s\in S}\langle g_\delta(x^k),\, s - x^k\rangle$; $\tilde{g}_k := \tilde{G}(x^k) = \langle g_\delta(x^k),\, x^k - s^k\rangle$.
  Set $x^{k+1} = x^k + \overline{\alpha}_k(s^k - x^k)$, $\overline{\alpha}_k := \frac{(\tilde{g}_k - \delta)_+}{C}$, where $(u)_+ := \max\{u, 0\}$.
end for

Lemma 5 (One-step decrease) *The iterates of Algorithm 2 satisfy*

$$f(x^{k+1}) \;\le\; f(x^k) \;-\; \frac{(\tilde{g}_k - \delta)_+^2}{2C}. \tag{6}$$

Proof L-smoothness gives

$$\begin{aligned} f(x^{k+1}) &\le f(x^k) + \overline{\alpha}_k\langle \nabla f(x^k),\, s^k - x^k\rangle + \tfrac{L}{2}\overline{\alpha}_k^2\|s^k - x^k\|^2 \\ &\le f(x^k) + \overline{\alpha}_k\langle g_\delta(x^k), s^k - x^k\rangle + \overline{\alpha}_k\delta + \frac{C}{2}\overline{\alpha}_k^2 \\ &= f(x^k) - \overline{\alpha}_k\tilde{g}(x^k) + \overline{\alpha}_k\delta + \frac{C}{2}\overline{\alpha}_k^2 \\ &= f(x^k) - \overline{\alpha}_k(\tilde{g}_k - \delta) + \frac{C}{2}\overline{\alpha}_k^2 \end{aligned}$$

using (5) and $\|s^k - x^k\| \le D$.

With $\overline{\alpha}_k = (\tilde{g}_k - \delta)_+/C$,

$$f(x^{k+1}) \le f(x^k) - \frac{1}{2C}(\tilde{g}_k - \delta)_+^2,$$

since $\overline{\alpha}_k = 0$ if $\tilde{g}_k - \delta \le 0$.

Theorem 6 (Nonconvex Frank-Wolfe with a δ-oracle) *Let f be L-smooth on a compact convex set S of diameter D. Suppose the δ-oracle condition (4) holds. With curvature constant $C \geq \max\{LD^2, GD\}$ and step size $\overline{\alpha}_k = (\tilde{g}_k - \delta)_+/C$, the iterates of Algorithm 2 satisfy*

$$\min_{0\leq k\leq K} G(x^k) \;\leq\; \sqrt{\frac{2C\,(f(x^0)-f^*)}{K+1}} \;+\; 2\delta, \tag{7}$$

where $f^* := \inf_{x\in S} f(x)$. In particular, to reach a Frank-Wolfe gap at most $\varepsilon > 2\delta$, it suffices to take

$$K+1 \;\geq\; \frac{2C\,(f(x^0)-f^*)}{(\varepsilon-2\delta)^2}\,.$$

4 A stronger guarantee under a directionally relative δ-oracle

In this section we strengthen Section 3 by replacing the additive inexactness assumption with a *directionally relative* one:

$$\big|\langle \nabla f(x) - g_\delta(x),\, s-x\rangle\big| \;\leq\; \delta\,\|\nabla f(x)\| \qquad \forall\, s\in S,\ \forall\, x\in S. \tag{8}$$

Define the approximate Frank–Wolfe gap $\tilde{G}(x) := \max_{s\in S}\langle g_\delta(x),\, x-s\rangle$ and the true gap $G(x) := \max_{s\in S}\langle \nabla f(x),\, x-s\rangle$. From (8),

$$|\,G(x) - \tilde{G}(x)\,| \;\leq\; \delta\,\|\nabla f(x)\| \qquad \forall x\in S. \tag{9}$$

We analyze the same update as Algorithm 2 but with a stepsize that reflects the new assumption:

$$x^{k+1} = x^k + \overline{\alpha}_k(s^k - x^k), \qquad \overline{\alpha}_k := \frac{\big(\tilde{G}(x^k) - \delta\|\nabla f(x^k)\|\big)_+}{C}, \tag{10}$$

where $s^k \in \arg\min_{s\in S}\langle g_\delta(x^k),\, s-x^k\rangle$ and $C \geq \max\{LD^2,\ GD\}$ as in Section 3.

Lemma 7 (One-step decrease under (8)) *For the iterates (10),*

$$f(x^{k+1}) \;\leq\; f(x^k) \;-\; \frac{\big(\tilde{G}(x^k) - \delta\|\nabla f(x^k)\|\big)_+^2}{2C}\,.$$

Proof L-smoothness and $\|s^k - x^k\| \leq D$ give

$$f(x^{k+1}) \leq f(x^k) + \overline{\alpha}_k\langle \nabla f(x^k), s^k - x^k\rangle + \tfrac{L}{2}\overline{\alpha}_k^2\|s^k-x^k\|^2 \leq f(x^k) + \overline{\alpha}_k\langle \nabla f(x^k), s^k - x^k\rangle + \tfrac{C}{2}\overline{\alpha}_k^2.$$

By (8) and the choice of s^k, $\langle \nabla f(x^k), s^k - x^k\rangle \leq \langle g_\delta(x^k), s^k - x^k\rangle + \delta\|\nabla f(x^k)\| = -\tilde{G}(x^k) + \delta\|\nabla f(x^k)\|$. Plugging this in and minimizing the quadratic upper bound over $\overline{\alpha}_k \geq 0$ yields the claim with the choice (10).

Theorem 8 (Relative oracle: residual scales with $\|\nabla f\|$) *Let f be L-smooth on compact convex S of diameter D and suppose (8) holds. With the stepsize (10),*

$$\min_{0\le k\le K}\big(G(x^k)-2\delta\|\nabla f(x^k)\|\big)_+ \;\le\; \sqrt{\frac{2C\,(f(x^0)-f^*)}{K+1}}\,. \tag{11}$$

Proof Summing Lemma 7 for $k=0,\dots,K$ gives

$$\sum_{k=0}^{K}\big(\tilde G(x^k)-\delta\|\nabla f(x^k)\|\big)_+^2 \le 2C\big(f(x^0)-f(x^{K+1})\big)\le 2C(f(x^0)-f^*).$$

Hence

$$\min_{0\le k\le K}\big(\tilde G(x^k)-\delta\|\nabla f(x^k)\|\big)_+ \le \sqrt{2C(f(x^0)-f^*)/(K+1)}.$$

Using (9) yields

$$\big(G(x^k)-2\delta\|\nabla f(x^k)\|\big)_+ \le \big(\tilde G(x^k)-\delta\|\nabla f(x^k)\|\big)_+,$$

which implies (11).

The bound (11) is *strictly stronger* than Theorem 6 whenever the gradients encountered along the trajectory are small, since the residual term vanishes proportionally to $\|\nabla f(x^k)\|$. In particular, one obtains *asymptotically exact* Frank–Wolfe stationarity whenever $\|\nabla f(x^k)\|\to 0$.

We now show that a purely *gap-only* rate with no additive residual follows if the iterates remain at a positive distance from the boundary.

Corollary 9 (Residual-free rate under an interior margin) *Suppose there exists $r>0$ such that $\operatorname{dist}(x^k,\partial S)\ge r$ for all k. Then for all $x\in S$, $G(x)\ge r\|\nabla f(x)\|$, and if additionally $\delta<r/2$,*

$$\min_{0\le k\le K} G(x^k) \;\le\; \frac{1}{1-\frac{2\delta}{r}}\sqrt{\frac{2C\,(f(x^0)-f^*)}{K+1}}\,. \tag{12}$$

Consequently, $\min_{0\le k\le K} G(x^k)=\mathcal{O}(1/\sqrt{K})$ with no additive residual, and $\liminf_{k\to\infty} G(x^k)=0$.

Proof If $\operatorname{dist}(x,\partial S)\ge r$, then for the unit vector $u=\nabla f(x)/\|\nabla f(x)\|$ the point $x-ru\in S$, which yields $G(x)\ge\langle\nabla f(x),\,x-(x-ru)\rangle=r\|\nabla f(x)\|$. Using this in (11) gives $(1-2\delta/r)\min_k G(x^k)\le\sqrt{2C(f(x^0)-f^*)/(K+1)}$, and (12) follows because $\delta<r/2$.

Remark 10 (Implementability) *The step (10) uses $\|\nabla f(x^k)\|$ only in the analysis. A standard backtracking line-search based on the smoothness upper model $\varphi(\alpha)=f(x^k)+\alpha\langle g_\delta(x^k),s^k-x^k\rangle+\frac{L}{2}\alpha^2\|s^k-x^k\|^2$ (which uses only g_δ, L, and $\|s^k-x^k\|$) produces a stepsize $\hat\alpha_k$ satisfying the same decrease inequality as in Lemma 7 up to a harmless constant factor in C, so Theorem 8 and Corollary 9 continue to hold with possibly larger C.*

5 Projection vs. Linear Minimization Oracle

Let $(\cdot,\cdot)$ and $\|\cdot\|$ denote the Euclidean inner product and its norm. For a nonempty compact convex set $C \subset \mathbb{R}^d$, define projection $\mathrm{Proj}_C(x) = \arg\min_{c\in C} \frac{1}{2}\|c-x\|^2$ and linear minimization oracle $\mathrm{LMO}_C(z) = \arg\min_{c\in C}(c,z)$. $p' \in C$ is a *K-approximate projection* of x onto C if

$$\tfrac{1}{2}\|p'-x\|^2 \le \min_{c\in C} \tfrac{1}{2}\|c-x\|^2 + K.$$

Proposition 11 *Let $p' \in C$ be a K-approximate projection of x onto C, then for all $c \in C$,*

$$(c-p',\, x-p') \le K + \tfrac{1}{2}\|c-p'\|^2.$$

Proof From the definition of p', $\frac{1}{2}\|p'-x\|^2 \le \frac{1}{2}\|c-x\|^2 + K$ for all $c \in C$. Expanding the squares and simplifying gives $(c-p',\, x-p') \le K + \frac{1}{2}\|c-p'\|^2$.

The following theorem establishes an equivalence in computational effort between approximate projections and LMOs.

Theorem 12 (From K-projection to ε-LMO) *Let $C \subset \mathbb{R}^d$ be a nonempty compact convex set with diameter $\delta_C := \sup_{c_1,c_2\in C}\|c_1-c_2\|$ and radius $\mu_C := \sup_{c\in C}\|c\|$. $x \in \mathbb{R}^d$. Let $v \in \mathrm{LMO}_C(x)$ and $p' \in C$ be a K-approximate projection of $-\lambda x$ onto C for some $\lambda > 0$. Then*

$$0 \le (p',x) - (v,x) \le \frac{K + \frac{1}{2}\delta_C^2 + \mu_C\delta_C}{\lambda}.$$

In particular, choosing $\lambda \ge \big(K+\frac{1}{2}\delta_C^2+\mu_C\delta_C\big)/\varepsilon$ ensures $(p',x) \le \min_{c\in C}(c,x)+\varepsilon$, i.e., $p' \in \varepsilon\text{-LMO}_C(x)$.

Discussion. This extends the exact-projection implication of [15] to *inexact* projections: one K-projection at a scaled point yields an ε-accurate LMO. In particular, accurate linear minimization is *no slower* than coarse projection, uniformly over compact convex sets.

6 Conclusion

We have established tight robustness guarantees for the Frank-Wolfe method under inexact gradient oracles and extended recent insights on the relationship between projection and linear minimization. The results confirm that oracle errors do not accumulate and that approximate projections cannot be computationally superior to accurate LMOs.

7 Acknowledgment

I am grateful to Dr. Paul Grigas, who taught me a course on nonlinear optimization and led me to the topic of the Frank-Wolfe method.

8 Compliance with Ethical Standards

Conflict of interest. The authors declare that they have no competing interests.

Ethical approval. This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent. Not applicable.

9 Declarations

Funding. No funding was received for conducting this study.

Competing interests. The authors have no competing interests to declare.

Data availability. Not applicable.

Code availability. Not applicable.

Appendix A. Proof of Theorem 3

Let $D = \mathrm{Diam}(Q)$. Lipschitz smoothness of f and (2) yield, for the Frank-Wolfe step $x_{k+1} = x_k + \overline{\alpha}_k(\tilde{x}_k - x_k)$,

$$\begin{aligned}
f(x_{k+1}) &\leq f(x_k) + \nabla f(x_k)^\top (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
&= f(x_k) + \overline{\alpha}_k \nabla f(x_k)^\top (\tilde{x}_k - x_k) + \frac{L}{2}\overline{\alpha}_k^2 \|\tilde{x}_k - x_k\|^2 \\
&\leq f(x_k) + \overline{\alpha}_k g_\delta(x_k)^T (\tilde{x}_k - x_k) + \overline{\alpha}_k \delta + \frac{L}{2}\overline{\alpha}_k^2 \|\tilde{x}_k - x_k\|^2 \\
&\leq (1 - \overline{\alpha}_k) f(x_k) + \overline{\alpha}_k \big(f(x_k) + g_\delta(x_k)^\top (\tilde{x}_k - x_k) - \delta\big) + 2\overline{\alpha}_k \delta + \tfrac{L}{2} D^2 \overline{\alpha}_k^2 \\
&\leq (1 - \overline{\alpha}_k) f(x_k) + \overline{\alpha}_k f^\star + 2\overline{\alpha}_k \delta + \tfrac{L}{2} D^2 \overline{\alpha}_k^2,
\end{aligned}$$

where the third line uses $\|\tilde{x}_k - x_k\| \leq D$ and the last line uses Lemma 1. Subtracting both sides from $f^\star$ gives (3).

In order to continue, we multiply β_{k+1} by both sides of Equation (3); using Equations (1), we get that

$$\beta_{k+1}(f(x_{k+1}) - f^*) \leq \beta_k (f(x_k) - f^*) + 2\overline{\alpha}_k \beta_{k+1} \delta + \frac{L}{2} D^2 \overline{\alpha}_k^2 \beta_{k+1}.$$

By taking the summation, we get that

$$\begin{aligned}
\beta_{k+1}(f(x_{k+1}) - f^*) &\leq (f(x_0) - f^*) + 2\delta \sum_{j=0}^{k} \overline{\alpha}_j \beta_{j+1} + \frac{L}{2} D^2 \sum_{j=0}^{k} \overline{\alpha}_j^2 \beta_{j+1} \\
&\leq (f(x_0) - f^*) + 2\delta \sum_{j=0}^{k} (\beta_{j+1} - \beta_j) + \frac{L}{2} D^2 \sum_{j=0}^{k} \overline{\alpha}_j^2 \beta_{j+1},
\end{aligned}$$

since $\beta_{k+1} - \beta_k = \overline{\alpha}_k \beta_{k+1}$. The summation term telescopes as

$$\sum_{j=0}^{k} (\beta_{j+1} - \beta_j) = \beta_{k+1} - 1.$$

Substituting this back, we obtain

$$\beta_{k+1}(f(x_{k+1}) - f^*) \leq (f(x_0) - f^*) + 2\delta(\beta_{k+1} - 1) + \frac{L}{2}D^2 \sum_{j=0}^{k} \overline{\alpha}_j^2 \beta_{j+1}.$$

Dividing both sides by β_{k+1} yields

$$f(x_{k+1}) - f^* \leq \frac{f(x_0) - f^*}{\beta_{k+1}} + 2\delta\Big(1 - \frac{1}{\beta_{k+1}}\Big) + \frac{L}{2}D^2 \frac{\sum_{j=0}^{k} \overline{\alpha}_j^2 \beta_{j+1}}{\beta_{k+1}}.$$

Take any $1 < J < k$,

$$\begin{aligned}\frac{\sum_{j=0}^{k} \overline{\alpha}_j^2 \beta_{J+1}}{\beta_{k+1}} &= \sum_{j=0}^{J} \overline{\alpha}_j^2 \prod_{t=J+1}^{k} (1 - \overline{\alpha}_t) + \sum_{j=J+1}^{k} \overline{\alpha}_j^2 \prod_{t=j+1}^{k} (1 - \overline{\alpha}_t) \\ &\leq \sum_{j=0}^{J} \overline{\alpha}_j^2 \prod_{t=J+1}^{k} (1 - \overline{\alpha}_t) + \sum_{j=J+1}^{k} \overline{\alpha}_j^2.\end{aligned}$$

Therefore,

$$\limsup_{k \to +\infty} \frac{\sum_{j=0}^{k} \overline{\alpha}_j^2 \beta_{j+1}}{\beta_{k+1}} \leq \sum_{j=J+1}^{+\infty} \overline{\alpha}_j^2, \ \forall J > 1.$$

Hence,

$$\limsup_{k \to +\infty} \frac{\sum_{j=0}^{k} \overline{\alpha}_j^2 \beta_{j+1}}{\beta_{k+1}} = 0.$$

Hence

$$\limsup_{k \to \infty} (f(x_k) - f^*) \leq 2\delta.$$

This completes the proof.

Appendix B. Proof of Theorem 6

By Lemma 5 we have

$$f(x^{k+1}) \ \leq \ f(x^k) - \frac{(\tilde{g}_k - \delta)_+^2}{2C}.$$

Summing from $k = 0$ to K yields

$$\sum_{k=0}^{K} (\tilde{g}_k - \delta)_+^2 \ \leq \ 2C(f(x^0) - f(x^{K+1})) \ \leq \ 2C(f(x^0) - f^*).$$

Therefore

$$\min_{0 \leq k \leq K} (G(x^k) - 2\delta)_+ \leq \min_{0 \leq k \leq K} (\tilde{g}_k - \delta)_+ \leq \sqrt{2C(f(x^0) - f^*)/(K+1)}.$$

□

Appendix C. Proof of Theorem 12

Proof By proposition 11, we have that

$$(c - p', -\lambda x - p') \leq K + \frac{1}{2}\|c - p'\|^2, \forall\, c \in C.$$

Then, and take $c = v$,

$$\lambda(p', x) - \lambda(v, x) \leq K + \frac{1}{2}\|v - p'\|^2 + (p', v - p').$$

Next,

$$\begin{aligned}\lambda(p' - v, x) &\leq K + \frac{1}{2}\|v - p'\|^2 + (p', v - p') \\ &= K + \frac{1}{2}\|v - p'\|^2 + ((v, p') - (p', p')) \\ &\leq K + \frac{1}{2}\|v - p'\|^2 + \|p'\|(\|v\| - \|p'\|) \\ &\leq K + \frac{1}{2}\|v - p'\|^2 + \|p'\|\|v - p'\|.\end{aligned}$$

Therefore,

$$\lambda(p' - v, x) \leq K + \frac{1}{2}\delta_C^2 + \mu_C\delta_C.$$

Hence,

$$0 \leq (p', x) - (v, x) \leq \frac{K + \frac{1}{2}\delta_C^2 + \mu_C\delta_C}{\lambda},$$

where the first inequality is from the fact that $v \in \mathrm{LMO}_C(x)$. □

References

1. C. W. Combettes and S. Pokutta, *Complexity of linear minimization and projection on some sets*, arXiv preprint arXiv:2101.10040, (2021).
2. O. Devolder, F. Glineur, and Y. Nesterov, *First-Order Methods of Smooth Convex Optimization with Inexact Oracle*, Tech. Rep. 2013/19, Center for Operations Research and Econometrics (CORE), Louvain-la-Neuve, Belgium, 2013. CORE Discussion Paper.
3. O. Devolder, F. Glineur, and Y. Nesterov, *First-order methods of smooth convex optimization with inexact oracle*, Mathematical Programming, 146 (2014), pp. 37–75.
4. R. M. Freund and P. Grigas, *New analysis and results for the Frank–Wolfe method*, Mathematical Programming, 155 (2016), pp. 199–230. arXiv:1307.0873 (2013).
5. D. Goldfarb, G. Iyengar, and C. Zhou, *Linear convergence of stochastic frank–wolfe variants*, in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), 2017.
6. M. Jaggi, *Revisiting Frank–Wolfe: Projection-free sparse convex optimization*, in Proceedings of the 30th International Conference on Machine Learning (ICML), vol. 28, 2013, pp. 427–435.
7. S. Lacoste-Julien, *Convergence rate of frank-wolfe for non-convex objectives*, 2016.
8. F. Locatello et al., *Stochastic frank-wolfe for composite convex minimization*, in AISTATS, 2019.
9. H. Lu and R. M. Freund, *Generalized stochastic frank-wolfe with stochastic substitute gradient*, Optimization Online, (2018). 6748.
10. S. Pokutta, *The frank–wolfe algorithm: A short introduction*, Business & Information Systems Engineering, (2024).
11. S. J. Reddi, S. Sra, B. Poczos, and A. Smola, *Stochastic frank–wolfe methods for nonconvex optimization*, in Proceedings of the 33rd International Conference on Machine Learning (ICML), 2016.

12. M.-E. Sfyraki and J.-K. Wang, *Lions and muons: Optimization via stochastic frank-wolfe*, 2025.
13. M. E. Sfyraki and Y. Wang, *Lions and muons: Optimization via stochastic frank–wolfe*, arXiv preprint arXiv:2506.04192, (2025).
14. T. Tang, K. Balasubramanian, and T. C. M. Lee, *High-probability bounds for robust stochastic frank-wolfe algorithm*, in Proceedings of Machine Learning Research, vol. 180, 2022.
15. Z. Woodstock, *High-precision linear minimization is no slower than projection*, arXiv preprint arXiv:2501.18454, (2025).
16. M. Zhang et al., *One sample stochastic frank-wolfe*, in Proceedings of the 37th International Conference on Machine Learning (ICML), 2020.