Exploring the Robustness of the Frank-Wolfe method and the Effectiveness of Linear Minimization Oracle

Tao Hu

Abstract Keywords Frank-Wolfe · conditional gradients · inexact oracle · stochastic optimization · heavy-tailed noise · projection vs. LMO

1 Introduction

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. Our main purpose here is to consider the minimization problem here:

$$\begin{aligned} &\min_{\lambda} \ f(\lambda) \\ &\text{s.t. } \lambda \in Q \ . \end{aligned} \tag{1}$$

The Frank-Wolfe method is an effective way to address this problem which computes at $\lambda_k \in Q$ the point for linear minimization

$$\tilde{\lambda}_k \in \arg\min_{\lambda \in Q} \left\{ f(\lambda_k) + \nabla f(\lambda_k)^\top (\lambda - \lambda_k) \right\} \tag{2}$$

and updates with $\lambda_{k+1} = (1 - \bar{\alpha}_k)\lambda_k + \bar{\alpha}_k \tilde{\lambda}_k$ where $\bar{\alpha}_k \in [0, 1)$. Assuming that ∇f is L-Lipschitz on Q, and Q is of diameter D, then Frank-Wolfe achieves the classical $\mathcal{O}(LD^2/k)$ convergence rate for convex functions [6,4], and $\mathcal{O}(LD^2/\sqrt{k})$ convergence rate for nonconvex functions [7].

It is worth noticing that in the convergence analysis, those auxiliary sequences are frequently used and will also be used in our proof:

$$\beta_k = \frac{1}{\prod_{j=1}^{k-1} (1 - \bar{\alpha}_j)} \ , \qquad \alpha_k = \frac{\beta_k \bar{\alpha}_k}{1 - \bar{\alpha}_k} \ , \qquad k \geq 1 \ . \tag{3}$$

Tao Hu
Xi'an Jiaotong University
E-mail: tao_hu@berkeley.edu

Here $\{\overline{\alpha}_k\}_{k=1}^{+\infty}$ is sequence of stepsizes in our algorithm. We follow the conventions: $\prod_{j=1}^{0} \cdot = 1$ and $\sum_{i=1}^{0} \cdot = 0$.

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

To start with, the robustness of Frank-Wolfe, that is, how Frank-Wolfe performs 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, the situation that the noise is bounded by δ but can be arbitrarily chosen along the training trajectory is often referred to as obtaining a δ-oracle:

$$\left|(g_\delta(x) - \nabla f(x))^T (x - y)\right| \leq \delta, \forall\, y \in Q. \tag{4}$$

Freund and Grigas proves an $\mathcal{O}(1/k + \delta)$ convergence [4], and we shows an $\mathcal{O}(1/\sqrt{k} + \delta)$ convergence for nonconvex functions in this paper.

Another interesting problem occurs when considering objective functions that are convex but non-smooth. Those functions may not obtain a gradient, but they can be equipped with a (δ, L) oracle [3]:

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

Unlike inexact gradient, this (δ, L) oracle allows the error to interact with the local quadratic model and leads to error accumulation, which shows that the Frank-Wolfe method is only guaranteed to reach a Frank-Wolfe gap of $\mathcal{O}(\sqrt{\delta})$. However, it remains an open problem whether the final guarantee of Frank-Wolfe gap is optimal theoretically. In this paper, we show that

While Frank-Wolfe does not have the same guarantee on the (δ, L) oracle as proximal gradient descent. The ease of computing the Linear Minimization Oracle (LMO) is widely considered a major advantage of the Frank-Wolfe method. However, this belief is currently limited to intuition and set-specific comparisons [1,10]. Beyond such instances, Woodstock showed that exact projection is never easier than obtaining an ε-accurate LMO, uniformly over compact convex sets [15]. We extend this to *approximate* projections: a single K-projection at a scaled point yields an ε-accurate LMO.

Our contributions.

(i) **Frank-Wolfe with a δ-oracle (nonconvex).** We show that for L-smooth nonconvex f over a compact convex set, Frank-Wolfe with a directional δ-oracle achieves

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

where

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

(ii) **Frank-Wolfe with a** (δ, L)**-oracle.** We show that Frank-Wolfe method is theoretically guaranteed to reach a Frank-Wolfe gap of $\mathcal{O}(\sqrt{\delta})$. We also show that this final Frank-Wolfe gap can be reduced to $O(\delta)$ when f is convex.

(iii) **Projection vs. LMO.** We show that a K-approximate projection at $-\lambda x$ produces an ε-accurate LMO at x with $\varepsilon = \mathcal{O}((K + D_C^2)/\lambda)$, reinforcing that coarse projections are not cheaper than accurate LMOs.

2 Frank-Wolfe with a δ-oracle: main result and a tight example

We assume $Q \subset \mathbb{R}^d$ is compact and convex with diameter D, and $f : Q \to \mathbb{R}$ is convex with L-Lipschitz gradient on Q. We run Frank-Wolfe using the δ-oracle g_δ in Algorithm 1.

Algorithm 1 Frank-Wolfe with a gradient δ-oracle (maximization)

1: Initialize $\lambda_0 \in Q$.
2: **for** $k = 0, 1, 2, \dots$ **do**
3: Query $g_\delta(\lambda_k)$.
4: Compute $\tilde{\lambda}_k \in \arg\min_{\lambda \in Q} \{f(\lambda_k) + g_\delta(\lambda_k)^\top (\lambda - \lambda_k)\}$.
5: Update $\lambda_{k+1} = \lambda_k + \bar{\alpha}_k (\tilde{\lambda}_k - \lambda_k)$ with $\bar{\alpha}_k \in [0, 1)$.
6: **end for**

Lemma 1 *Under* (4)*, for any* $\lambda_k \in Q$*,*

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

Proof By convexity, $f(\lambda) \geq f(\lambda_k) + \nabla f(\lambda_k)^\top (\lambda - \lambda_k)$ for any $\lambda \in Q$. From (4), $\nabla f(\lambda_k)^\top (\lambda - \lambda_k) \geq g_\delta(\lambda_k)^\top (\lambda - \lambda_k) - \delta$. Therefore,

$$f(\lambda) \geq f(\lambda_k) + g_\delta(\lambda_k)^T (\lambda - \lambda_k) - \delta.$$

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

We also recall a subproblem-level accuracy transfer.

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

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

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

$$f(\lambda_{k+1}) - f^* \leq (1 - \bar{\alpha}_k)\big(f(\lambda_k) - f^*\big) + 2\bar{\alpha}_k\delta + \tfrac{1}{2}LD^2\bar{\alpha}_k^2, \tag{5}$$

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

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

3 Nonconvex objectives with a directional δ-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 $D := \mathrm{Diam}(S)$ and set

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

The Frank-Wolfe (FW) gap at x is

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

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

$$\big|\langle \nabla f(x) - g_\delta(x),\ s - x\rangle\big| \leq \delta \quad \forall s \in S. \tag{6}$$

Define the *approximate Frank-Wolfe gap*

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

From (6) it follows that

$$|\,g(x) - \tilde{g}(x)\,| \leq \delta, \tag{7}$$

where $s_\delta(x) \in \arg\max_{s\in S}\langle g_\delta(x),\ x - s\rangle$.

Algorithm 2 Nonconvex Frank-Wolfe with a directional δ-oracle

1: **Input:** $x^0 \in S$, curvature constant $C \geq \max\{LD^2,\ GD\}$, error level $\delta \geq 0$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Obtain $g_\delta(x^k)$ that satisfies (6); set $s^k \in \arg\max_{s\in S}\langle g_\delta(x^k),\ x^k - s\rangle$ and $\tilde{g}_k := \langle g_\delta(x^k),\ x^k - s^k\rangle = \tilde{g}(x_k)$.
4: Stepsize: $\overline{\alpha}_k := \min\left\{\frac{(\tilde{g}_k - \delta)_+}{C},\ 1\right\}$, where $(u)_+ := \max\{u, 0\}$.
5: Update: $x^{k+1} \leftarrow x^k + \overline{\alpha}_k(s^k - x^k)$.
6: **end for**

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

$$f(x^{k+1}) \;\le\; f(x^k) \;-\; \frac{(\tilde g_k - \delta)_+^2}{2C}\,. \tag{8}$$

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 (7) 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 directional δ-oracle) *Let f be L-smooth on a compact convex set S of diameter D and let $C \ge \max\{LD^2, GD\}$. Suppose the directional δ-oracle (6) is available. Then the iterates of Algorithm 2 satisfy, for all $K \ge 0$,*

$$\min_{0\le k\le K} g(x^k) \;\le\; \sqrt{\frac{2C\,(f(x^0) - f_{\inf})}{K+1}} \;+\; 2\delta, \tag{9}$$

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

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

Remark 7 (Discussion and special cases) *(i) When $\delta = 0$ the bound reduces to the classical nonconvex Frank-Wolfe rate. (ii) For $\delta > 0$, the method converges to an $O(\delta)$ neighborhood in the Frank-Wolfe gap; the error does not accumulate across iterations. (iii) The stepsize uses $\tilde g_k$ (computed "for free" while solving the LMO with g_δ), exactly mirroring the steepest-feasible steps in standard Frank-Wolfe. (iv) Any $C \ge LD^2$ works for (8); taking $C \ge \max\{LD^2, GD\}$ ensures $\alpha_k \le 1$ without extra capping.*

4 Frank-Wolfe with a (δ, L)-oracle: error does not accumulate for convex functions

Algorithm 3 Frank-Wolfe with a (δ, L)-oracle (maximization)

1: Initialize $\lambda_0 \in Q$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Query $(f_{\delta,L}(\lambda_k), g_{\delta,L}(\lambda_k))$.
4: Compute $\tilde{\lambda}_k \in \arg\min_{\lambda \in Q} \langle g_{\delta,L}(\lambda_k), \lambda - \lambda_k \rangle$. Denote $g(\lambda_k) = \nabla f(\lambda_k)^T(\lambda_k - \tilde{\lambda}_k)$.
5: Update $\lambda_{k+1} = \lambda_k + \bar{\alpha}_k(\tilde{\lambda}_k - \lambda_k)$ with $\bar{\alpha}_k \in [0, 1)$.
6: **end for**

We adopt the Devolder–Glineur–Nesterov (δ, L)-*oracle* for the function f[2]: for any $\bar{\lambda} \in Q$, the oracle returns $(f_{\delta,L}(\bar{\lambda}), g_{\delta,L}(\bar{\lambda}))$ such that for any $\lambda \in Q$

$$\begin{array}{rrcl} \text{(upper)} & f(\lambda) & \leq & f_{\delta,L}(\bar{\lambda}) + \langle g_{\delta,L}(\bar{\lambda}), \lambda - \bar{\lambda} \rangle + \frac{L}{2}\|\lambda - \bar{\lambda}\|^2 + \delta, \\ \text{(lower)} & f(\lambda) & \geq & f_{\delta,L}(\bar{\lambda}) + \langle g_{\delta,L}(\bar{\lambda}), \lambda - \bar{\lambda} \rangle. \end{array} \tag{10}$$

Theorem 8 ([4] Theorem 5.3) *If the step-size sequence $\{\bar{\alpha}_k\}$ is used in the iterate sequences of the Frank-Wolfe method with the (δ, L)-oracle (Algorithm 3),*

$$f(\lambda_{k+1}) - f^* \leq \frac{f(\lambda_0) - f^*}{\beta_{k+1}} + \frac{\frac{1}{2}C\sum_{i=1}^{k}\bar{\alpha}_i^2\beta_{i+1}}{\beta_{k+1}} + \frac{2\delta\sum_{i=1}^{k}\beta_{i+1}}{\beta_{k+1}}, \ \forall\, k \geq 0, \tag{11}$$

where $C = LD^2$.

Remark 9 *Consider the right hand side of Equation (11), we have that*

$$\begin{aligned} & \frac{f_0 - f^*}{\beta_{k+1}} + \frac{\frac{1}{2}C\sum_{i=1}^{k}\frac{\alpha_i^2}{\beta_{i+1}}}{\beta_{k+1}} + \frac{\sum_{i=1}^{k}\beta_{i+1}\delta}{\beta_{k+1}} \\ =& \frac{f_0 - f^*}{\beta_{k+1}} + \frac{\sum_{i=1}^{k}(\frac{1}{2}C_{h,Q}\frac{\alpha_i^2}{\beta_{i+1}} + \beta_{i+1}\delta)}{\beta_{k+1}} \\ \geq& \frac{f_0 - f^*}{\beta_{k+1}} + \frac{\sum_{i=1}^{k}\alpha_i\sqrt{2C_{h,Q}\delta}}{\beta_{k+1}} \\ =& \frac{f_0 - f^*}{\beta_{k+1}} + \sqrt{2C_{h,Q}\delta}\frac{\beta_{k+1} - 1}{\beta_{k+1}}. \end{aligned}$$

For k sufficiently large, we have that

$$0 < \frac{\beta_2 - 1}{\beta_2} \leq \frac{\beta_{k+1} - 1}{\beta_{k+1}} \leq 1.$$

Therefore, error accumulation means that, in order to achieve a first-order accurate solution, we need a second-order accurate gradient.

Example 10 *When we are using a constant step size $\bar{\alpha}_k = \bar{\alpha}$, and we take $\delta_i = \delta$, then we have that*

$$\beta_k = (1-\bar{\alpha})^{-k+1}, \alpha_k = \bar{\alpha}(1-\bar{\alpha})^{-k}.$$

We will show in the appendix that, if we take $\bar{\alpha} = 1-(k-1)^{-1/(k-2)}$, then

$$\bar{\alpha} = 1-(k-1)^{-1/(k-2)} = 1-e^{-\frac{\log(k-1)}{k-2}} \in \left[\frac{1}{e}\frac{\log(k-1)}{k-2}, \frac{\log(k-1)}{k-2}\right].$$

$$\begin{aligned}
&f(\lambda_{k+1}) - f^* \\
\le&\frac{1}{2}C\left((1-\bar{\alpha})^{k-1}+\bar{\alpha}\right)+\frac{\delta}{\bar{\alpha}} \\
\le&\frac{1}{2}C\left((1-\bar{\alpha})^{k-2}+\bar{\alpha}\right)+\frac{\delta}{\bar{\alpha}} \\
=&\frac{1}{2}C\left(\frac{1}{k-1}+1-(k-1)^{-\frac{1}{k-2}}\right)+\frac{\delta}{\bar{\alpha}} \\
=&\frac{1}{2}C\left(\frac{1}{k-1}+1-e^{-\frac{\log(k-1)}{k-2}}\right)+\frac{\delta}{\bar{\alpha}} \\
\le&\frac{1}{2}C\left(\frac{1}{k-1}+\frac{\log(k-1)}{k-2}\right)+\frac{\delta}{\bar{\alpha}}
\end{aligned}$$

Therefore, in order to achieve a convergence rate of $\widetilde{O}(\frac{1}{k})$, it suffices to let

$$\delta \sim O(\frac{\bar{\alpha}}{k}) = O(\frac{\log(k)}{k^2}).$$

However, if we assume that f is convex(not necessarily smooth), we can show that no error accumulates.

Theorem 11 (Nonaccumulation under convex functions[4]) *Let Q be compact convex with diameter D, and f be convex with (δ, L)-oracle on Q. Let $(f_{\delta,L}, g_{\delta,L})$ satisfy (10). For the Frank-Wolfe iterates of Algorithm 1 with stepsizes satisfying $\sum_{k=0}^{+\infty}\bar{\alpha}_k = \infty$ and $\bar{\alpha}_k \downarrow 0$, then*

$$f(\lambda_{k+1}) - f^* \le (1-\bar{\alpha}_k)\big(f(\lambda_k)-f^*\big) + 2\bar{\alpha}_k\delta + \tfrac{1}{2}LD^2\bar{\alpha}_k^2, \tag{12}$$

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

5 Projection vs. LMO: accurate linear minimization beats coarse projection

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

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

Proposition 12 *If $p' \in C$ is 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 $(c - p',\, x - p') \le K + \frac{1}{2}\|c - p'\|^2$.

Theorem 13 (From K-projection to accurate LMO) *Let $x \in \mathbb{R}^d$ and nonempty compact convex $C \subset \mathbb{R}^d$ with diameter $\delta_C := \sup_{c_1,c_2 \in C} \|c_1 - c_2\|$ and radius $\mu_C := \sup_{c\in C} \|c\|$. Let $v \in \mathrm{LMO}_C(x)$ and $p' \in C$ be a K-approximate projection of $-\lambda x$ for some $\lambda > 0$. Then*

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

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

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.

Appendix A. Proof of Theorem 3

Let $D = \mathrm{Diam}(Q)$. Lipschitz smoothness of f and (4) yield, for the Frank-Wolfe step $\lambda_{k+1} = \lambda_k + \bar\alpha_k(\tilde\lambda_k - \lambda_k)$,

$$\begin{aligned}
f(\lambda_{k+1}) &\le f(\lambda_k) + \nabla f(\lambda_k)^\top(\lambda_{k+1} - \lambda_k) + \frac{L}{2}\|\lambda_{k+1} - \lambda_k\|^2 \\
&= f(\lambda_k) + \bar\alpha_k\, \nabla f(\lambda_k)^\top(\tilde\lambda_k - \lambda_k) + \frac{L}{2}\bar\alpha_k^2\|\tilde\lambda_k - \lambda_k\|^2 \\
&\le f(\lambda_k) + \overline{\alpha}_k g_\delta(\lambda_k)^T(\widetilde{\lambda}_k - \lambda_k) + \overline{\alpha}_k\delta + \frac{L}{2}\bar\alpha_k^2\|\tilde\lambda_k - \lambda_k\|^2 \\
&\le (1 - \bar\alpha_k) f(\lambda_k) + \bar\alpha_k\big(f(\lambda_k) + g_\delta(\lambda_k)^\top(\tilde\lambda_k - \lambda_k) - \delta\big) + 2\bar\alpha_k\delta + \tfrac{L}{2}D^2\bar\alpha_k^2 \\
&\le (1 - \bar\alpha_k) f(\lambda_k) + \bar\alpha_k f^\star + 2\bar\alpha_k\delta + \tfrac{L}{2}D^2\bar\alpha_k^2,
\end{aligned}$$

where the third line uses $\|\tilde\lambda_k - \lambda_k\| \le D$ and the last line uses Lemma 1. Subtracting both sides from $f^\star$ gives (12).

In order to continue, we multiply β_k by both sides of Equation (12); we get that

$$\beta_{k+1}(f(\lambda_{k+1}) - f^*) \le \beta_k(f(\lambda_k) - f^*) + 2\bar\alpha_k\beta_{k+1}\delta + \frac{L}{2}D^2\bar\alpha_k^2\beta_{k+1}.$$

By taking summation, we get that

$$\begin{aligned}
\beta_{k+1}(f(\lambda_{k+1}) - f^*) &\le (f(\lambda_0) - f^*) + 2\delta\sum_{j=0}^{k}\bar\alpha_j\beta_{j+1} + \frac{L}{2}D^2\sum_{j=0}^{k}\bar\alpha_j^2\beta_{j+1} \\
&\le (f(\lambda_0) - f^*) + 2\delta\sum_{j=0}^{k}(\beta_{j+1} - \beta_j) + \frac{L}{2}D^2\sum_{j=0}^{k}\bar\alpha_j^2\beta_{j+1}
\end{aligned}$$

Since $\beta_{k+1} - \beta_k = \bar{\alpha}_k \beta_{k+1}$, the summation term telescopes as

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

Substituting this back, we obtain

$$\beta_{k+1}(f(\lambda_{k+1}) - f^*) \leq (f(\lambda_0) - f^*) + 2\delta(\beta_{k+1} - 1) + \frac{L}{2} D^2 \sum_{j=0}^{k} \bar{\alpha}_j^2 \beta_{j+1}.$$

Dividing both sides by β_{k+1} yields

$$f(\lambda_{k+1}) - f^* \leq \frac{f(\lambda_0) - f^*}{\beta_{k+1}} + 2\delta\Big(1 - \frac{1}{\beta_{k+1}}\Big) + \frac{L}{2} D^2 \frac{\sum_{j=0}^{k} \bar{\alpha}_j^2 \beta_{j+1}}{\beta_{k+1}}.$$

Take any $1 < J < k$,

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

Therefore,

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

Hence,

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

Hence

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

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

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

Therefore

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

$\square$

Appendix C. Proof of Theorem 8

$$\begin{aligned}
f(\lambda_{k+1}) \leq & f(\lambda_k) + g_{\delta,L}(\lambda_k)^T(\lambda_{k+1} - \lambda_k) + 2\delta + \frac{1}{2}L\|\lambda_{k+1} - \lambda_k\|^2 \\
= & f(\lambda_k) + \bar{\alpha}_k g_{\delta,L}(\lambda_k)^T(\widetilde{\lambda}_k - \lambda_k) + 2\delta + \frac{1}{2}\bar{\alpha}_k^2 L_{h,Q}\|\widetilde{\lambda}_k - \lambda_k\|^2 \\
\leq & (1 - \bar{\alpha}_k)f(\lambda_k) + \bar{\alpha}_k(f(\lambda_k) + g_{\delta,L}(\lambda_k)^T(\widetilde{\lambda}_k - \lambda_k) - \delta) + (2 + \bar{\alpha}_k)\delta + \frac{1}{2}C\bar{\alpha}_k^2 \\
\leq & (1 - \bar{\alpha}_k)h(\lambda_k) + \bar{\alpha}_k f^* + 3\delta + \frac{1}{2}C\bar{\alpha}_k^2 \\
\leq & (1 - \bar{\alpha}_k)h(\lambda_k) + \bar{\alpha}_k f^* + 3\delta + \frac{1}{2}C\bar{\alpha}_k^2
\end{aligned}$$

Hence,

$$f(\lambda_{k+1}) - f^* \leq (1 - \bar{\alpha}_k)(f(\lambda_k) - f^*) + 3\delta + \frac{1}{2}C\bar{\alpha}_k^2.$$

Therefore,

$$\beta_k(f(\lambda_{k+1}) - f^*) \leq \beta_{k-1}(f(\lambda_k) - f^*) + 3\delta\beta_k + \frac{1}{2}C\bar{\alpha}_k^2\beta_k.$$

By taking summation, we get that

$$\beta_k(f(\lambda_{k+1}) - f^*) \leq (f(\lambda_0) - f^*) + 3\delta\sum_{j=1}^{k}\beta_j + \frac{1}{2}C\sum_{j=1}^{k}\bar{\alpha}_j^2\beta_j.$$

Dividing both side by β_k yields the result. □

Appendix D. Proof of Theorem 11

Since $\lambda_{k+1} = (1 - \bar{\alpha}_k)\lambda_k + \bar{\alpha}_k\tilde{\lambda}_k$.

$$f(\tilde{\lambda}_k) \leq f(\lambda_k) + g_{\delta,L}(\lambda_k)^T(\tilde{\lambda}_k - \lambda_k) + 2\delta + \frac{1}{2}L\|\lambda_{k+1} - \lambda_k\|^2.$$

Since $f(\lambda_{k+1}) \leq (1 - \bar{\alpha}_k)f(\lambda_k) + \bar{\alpha}_k f(\tilde{\lambda}_k)$,

$$f(\lambda_{k+1}) \leq f(\lambda_k) + \bar{\alpha}_k g_{\delta,L}(\lambda_k)^T(\tilde{\lambda}_k - \lambda_k) + 2\bar{\alpha}_k\delta + \frac{1}{2}\bar{\alpha}_k^2 L\|\tilde{\lambda}_k - \lambda_k\|^2.$$

The rest parts are the same as the proof of theorem 3.

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.