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

Tao Hu

Abstract It is commonly believed that the Frank-Wolfe method is a cheap and robust method for constrained optimization. However, how the Frank-Wolfe method performs under an inexact gradient remains unsolved. Besides, Zev Woodstock recently questions whether the linear minimization oracle (LMO) is inherently cheaper than the proximal operator, irrespective of future advances in numerical algorithms. Our work solves the first open problem completely and provides the strongest result for the second problem.

Keywords Frank-Wolfe · 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:

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

The Frank-Wolfe method, which has a linear minimization oracle (LMO) 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

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

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

and updates with $x_{k+1} = (1-\overline{\alpha}_k)x_k + \overline{\alpha}_k \tilde{x}_k$ where $\overline{\alpha}_k \in [0,1)$. Assuming that ∇f is L-Lipschitz on Q and Q is of diameter D, 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 noting that, 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 (LMO) are also important aspects of the Frank-Wolfe method.

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, 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

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

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. In this paper, we improve the final guarantee of the final Frank-Wolfe gap to $\mathcal{O}(\delta)$, showing the non-accumulation of error.

Besides robustness, the ease of computing the Linear Minimization Oracle (LMO) 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]. Beyond such instances, Woodstock [15] showed that exact projection is never easier than obtaining an ε-accurate LMO uniformly over compact convex sets. We extend this result to *approximate* projections, showing that 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 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 show that a K-approximate projection at $-\lambda x$ produces an ε-accurate LMO at x with $\varepsilon = \mathcal{O}((K + D^2)/\lambda)$, 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.

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

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

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.

Algorithm 1 Frank-Wolfe with a gradient δ-oracle

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

We also recall a subproblem-level accuracy transfer.

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

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

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 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 (2). For the Frank-Wolfe iterates of Algorithm 1 with stepsizes 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*

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

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

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 $g_\delta(x) = \nabla f(x) - \frac{\delta}{D}\,\mathrm{sign}(x)$. Frank-Wolfe with $\overline{\alpha}_k = 2/(k + 2)$ converges to a neighborhood whose size is proportional to δ.*

3 Nonconvex objectives with a 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 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

1: **Input:** $x^0 \in S$, curvature constant $C \ge \max\{LD^2,\, GD\}$, error level $\delta \ge 0$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Obtain $g_\delta(x^k)$ that satisfies (4); set $s^k \in \arg\max_{s\in S}\langle g_\delta(x^k),\, x^k - s\rangle$ and $\tilde{g}_k := \tilde{G}(x^k) = \langle g_\delta(x^k),\, x^k - s^k\rangle$.
4: Stepsize: $\overline{\alpha}_k := \dfrac{(\tilde{g}_k - \delta)_+}{C}$, 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*

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

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 δ-oracle) *Let f be L-smooth on a compact convex set S of diameter D and let $C \geq \max\{LD^2, GD\}$. Suppose the δ-oracle (4) is available. Then the iterates of Algorithm 2 satisfy, for all $K \geq 0$,*

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

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 Projection vs. LMO: accurate linear minimization beats coarse projection

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

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

Proposition 7 *If $p' \in C$ is a K-approximate projection of x onto C, then for all $c \in C$,*

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

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

Theorem 8 (From K-projection to LMO with high precision) *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*

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

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

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 (2) yield, for the Frank-Wolfe step $x_{k+1} = x_k + \overline{\alpha}_k(\tilde{x}_k - x_k)$,

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

where the third line uses $\|\tilde{x}_k - x_k\| \le D$ 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

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

By taking summation, we get that

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

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

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

Dividing both sides by β_{k+1} yields

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

Take any $1 < J < k$,

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

Therefore,

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

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 8

Proof By proposition 7, 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.