Frank-Wolfe and friends:
a journey into projection-free optimization methods

Francesco Rinaldi

Università degli Studi di Padova
Dipartimento Matematica

10th AiroYoung Padova
February 9th, 2026

Frank–Wolfe and Friends

- **Frank–Wolfe method** (aka conditional gradient) is a simple **first-order iterative** optimization method.
- Introduced in **1956** by M. Frank & P. Wolfe for solving **quadratic problems**.
- Recently revived thanks to its **projection-free nature** and strong performance in **data science**.

Frank–Wolfe and Friends

- **Frank–Wolfe method** (aka conditional gradient) is a simple **first-order iterative** optimization method.
- Introduced in **1956** by M. Frank & P. Wolfe for solving **quadratic problems**.
- Recently revived thanks to its **projection-free nature** and strong performance in **data science**.

Goals

- Understand how the method works (and why it works).
- See why data scientists like it so much.

Our Problem

Setup

$$\min_{\mathsf{x}\in C} \; f(\mathsf{x}) \tag{1}$$

- $C \subset \mathbb{R}^n$ is **convex and compact**.
- f is **differentiable** with L-**Lipschitz continuous gradient**:

$$\|\nabla f(\mathsf{x}) - \nabla f(\mathsf{y})\| \leq L\|\mathsf{x} - \mathsf{y}\|, \qquad \forall \mathsf{x}, \mathsf{y} \in C.$$

Our Problem

Setup

$$\min_{x \in C} \; f(x) \tag{1}$$

- $C \subset \mathbb{R}^n$ is **convex and compact**.
- f is **differentiable** with L-**Lipschitz continuous gradient**:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \qquad \forall x, y \in C.$$

- This smoothness assumption is **key for first-order methods**.
- It implies (and, for convex f, is equivalent to) the **Descent Lemma**:

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2.$$

- x^*: global solution, $\quad f^* := f(x^*)$.

General First-Order Scheme

- We consider a broad class of **first-order methods**.
- At each iteration, a set of **feasible (descent) directions** $F(\mathsf{x}, \nabla f(\mathsf{x}))$, built using **local first-order information**.
- Direction $\mathsf{d}_k \in F(\mathsf{x}_k, \nabla f(\mathsf{x}_k))$ combined with a stepsize $\alpha_k \in (0, \alpha_k^{\max}]$.

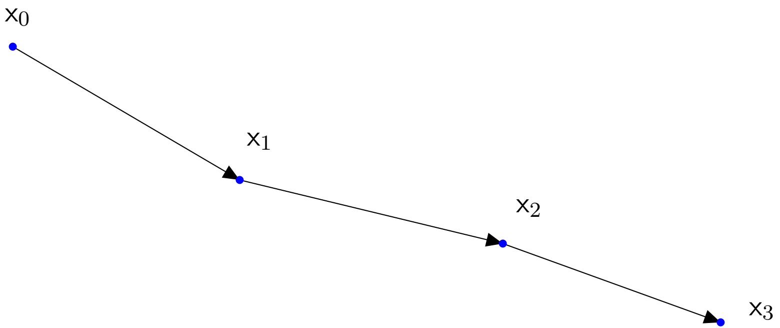
General First-Order Scheme

- We consider a broad class of **first-order methods**.
- At each iteration, a set of **feasible (descent) directions** $F(\mathsf{x}, \nabla f(\mathsf{x}))$, built using **local first-order information**.
- Direction $\mathsf{d}_k \in F(\mathsf{x}_k, \nabla f(\mathsf{x}_k))$ combined with a stepsize $\alpha_k \in (0, \alpha_k^{\max}]$.

Generic First-order method

1 Choose a point $x_0 \in C$
2 For $k = 0, \dots$
3 If x_k satisfies some specific condition, then STOP
4 Choose $\mathsf{d}_k \in F(\mathsf{x}_k, \nabla f(\mathsf{x}_k))$
5 Set $\mathsf{x}_{k+1} = \mathsf{x}_k + \alpha_k \mathsf{d}_k$, with $\alpha_k \in (0, \alpha_k^{\max}]$ a suitably chosen stepsize
6 End for

Iterative Algorithm



First-Order Descent Condition

Definition

A direction d is a first-order descent direction for f at x if

$$\nabla f(\mathrm{x})^\top \mathrm{d} < 0.$$

- Sufficient to guarantee decrease for small stepsizes,
- Central notion for constrained first-order methods (FW included).

Necessary Optimality Condition

Proposition (Necessary condition)

Let $\mathsf{x}^* \in C$ be a local minimum of

$$\min f(\mathsf{x}) \quad \text{s.t. } \mathsf{x} \in C,$$

with $C \subseteq \mathbb{R}^n$ convex and $f \in C^1(\mathbb{R}^n)$. Then

$$\nabla f(\mathsf{x}^*)^\top (\mathsf{x} - \mathsf{x}^*) \geq 0 \quad \forall \mathsf{x} \in C.$$

- No feasible direction at x^* can be a first-order descent direction.

Proposition (Convex case - N&S Condition)

Let $C \subseteq \mathbb{R}^n$ be convex and $f \in C^1(\mathbb{R}^n)$ convex. Then $\mathsf{x}^* \in C$ is a global minimum iff

$$\nabla f(\mathsf{x}^*)^\top (\mathsf{x} - \mathsf{x}^*) \geq 0 \quad \forall \mathsf{x} \in C.$$

Geometric Interpretation

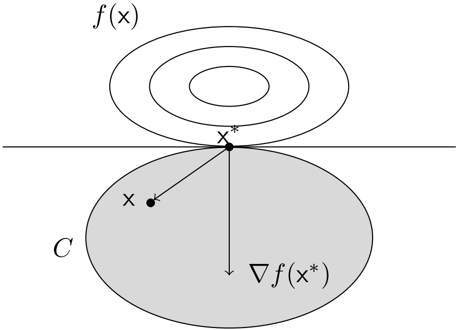


Figure: Geometric representation of first-order optimality conditions.

Classical Frank–Wolfe: Idea

- **Frank–Wolfe (FW)** minimizes a smooth f over a compact convex set C.
- At iteration k, it moves towards an **extreme point** of C.
- The direction is chosen by solving a problem with **linear objective** over C:

$$\mathsf{LMO}_C(\mathsf{g}) \in \operatorname{argmin}_{\mathsf{z}\in C}\langle \mathsf{g}, \mathsf{z}\rangle, \qquad \mathsf{g} = \nabla f(\mathsf{x}_k).$$

- This is known as the **Linear Minimization Oracle (LMO)**.

Classical Frank–Wolfe: Idea

- **Frank–Wolfe (FW)** minimizes a smooth f over a compact convex set C.
- At iteration k, it moves towards an **extreme point** of C.
- The direction is chosen by solving a problem with **linear objective** over C:

$$\mathsf{LMO}_C(\mathsf{g}) \in \operatorname{argmin}_{\mathsf{z}\in C}\langle \mathsf{g}, \mathsf{z}\rangle, \qquad \mathsf{g} = \nabla f(\mathsf{x}_k).$$

- This is known as the **Linear Minimization Oracle (LMO)**.

FW direction

$$\mathsf{d}_k^{FW} = \mathsf{s}_k - \mathsf{x}_k, \qquad \mathsf{s}_k \in \mathsf{LMO}_C(\nabla f(\mathsf{x}_k)).$$

Classical Frank–Wolfe: Update and Structure

- Update: $\mathsf{x}_{k+1} = \mathsf{x}_k + \alpha_k(\mathsf{s}_k - \mathsf{x}_k) = (1-\alpha_k)\mathsf{x}_k + \alpha_k \mathsf{s}_k, \quad \alpha_k \in (0,1].$
- x_{k+1} is a **convex combination** of elements in set $S_{k+1} := \{\mathsf{x}_0\} \cup \{\mathsf{s}_i\}_{0 \le i \le k}$.
- If $C = \text{conv}(A)$ with A set of atoms (points with some common feature) and $\mathsf{x}_0 \in A$, then x_k is a convex combination of at most $k+1$ atoms in A.
- By **Carathéodory's theorem**, only $\min\{k,n\}+1$ atoms are needed.

Classical Frank–Wolfe: Update and Structure

- Update: $\mathsf{x}_{k+1} = \mathsf{x}_k + \alpha_k(\mathsf{s}_k - \mathsf{x}_k) = (1-\alpha_k)\mathsf{x}_k + \alpha_k \mathsf{s}_k, \quad \alpha_k \in (0,1]$.
- x_{k+1} is a **convex combination** of elements in set $S_{k+1} := \{\mathsf{x}_0\} \cup \{\mathsf{s}_i\}_{0 \le i \le k}$.
- If $C = \text{conv}(A)$ with A set of atoms (points with some common feature) and $\mathsf{x}_0 \in A$, then x_k is a convex combination of at most $k+1$ atoms in A.
- By **Carathéodory's theorem**, only $\min\{k, n\} + 1$ atoms are needed.

Frank–Wolfe Algorithm

Choose a point $x_0 \in C$
For $k = 0, \ldots$
  Compute $\mathsf{s}_k \in \text{LMO}_C(\nabla f(x_k))$
  If x_k satisfies some specific condition, then STOP
  Set $\mathsf{d}_k^{FW} = \mathsf{s}_k - \mathsf{x}_k$
  Set $\mathsf{x}_{k+1} = \mathsf{x}_k + \alpha_k \mathsf{d}_k^{FW}$, with $\alpha_k \in (0,1]$ a suitably chosen stepsize
End for

Algorithmic Idea

- Start from an initial feasible point $x_0 \in C$.

Algorithmic Idea

- Start from an initial feasible point $\mathsf{x}_0 \in C$.
- At iteration k, build a search direction by solving the linearized subproblem

$$\min_{\mathsf{x} \in C} \ \nabla f(\mathsf{x}_k)^\top (\mathsf{x} - \mathsf{x}_k).$$

Algorithmic Idea

- Start from an initial feasible point $\mathsf{x}_0 \in C$.
- At iteration k, build a search direction by solving the linearized subproblem

$$\min_{\mathsf{x} \in C} \; \nabla f(\mathsf{x}_k)^\top (\mathsf{x} - \mathsf{x}_k).$$

- Equivalently, minimize the first-order Taylor approximation of f at x_k:

$$\min_{\mathsf{x} \in C} \; f(\mathsf{x}_k) + \nabla f(\mathsf{x}_k)^\top (\mathsf{x} - \mathsf{x}_k).$$

Algorithmic Idea

- Start from an initial feasible point $\mathsf{x}_0 \in C$.
- At iteration k, build a search direction by solving the linearized subproblem

$$\min_{\mathsf{x} \in C} \; \nabla f(\mathsf{x}_k)^\top (\mathsf{x} - \mathsf{x}_k).$$

- Equivalently, minimize the first-order Taylor approximation of f at x_k:

$$\min_{\mathsf{x} \in C} \; f(\mathsf{x}_k) + \nabla f(\mathsf{x}_k)^\top (\mathsf{x} - \mathsf{x}_k).$$

- Since C is compact, the linear problem admits a solution $\mathsf{s}_k \in C$.

Analysis of the Search Direction

- **Case 1:** $\nabla f(\mathsf{x}_k)^\top(\mathsf{s}_k - \mathsf{x}_k) = 0$.

  Then
  $$\nabla f(\mathsf{x}_k)^\top(\mathsf{x} - \mathsf{x}_k) \geq \nabla f(\mathsf{x}_k)^\top(\mathsf{s}_k - \mathsf{x}_k) = 0 \quad \forall \mathsf{x} \in C,$$
  and x_k satisfies the first-order optimality condition.

- **Case 1:** $\nabla f(\mathsf{x}_k)^\top(\mathsf{s}_k - \mathsf{x}_k) = 0.$

  Then
  $$\nabla f(\mathsf{x}_k)^\top(\mathsf{x} - \mathsf{x}_k) \geq \nabla f(\mathsf{x}_k)^\top(\mathsf{s}_k - \mathsf{x}_k) = 0 \quad \forall \mathsf{x} \in C,$$
  and x_k satisfies the first-order optimality condition.

- **Case 2:** $\nabla f(\mathsf{x}_k)^\top(\mathsf{s}_k - \mathsf{x}_k) < 0.$

  The vector
  $$\mathsf{d}_k := \mathsf{s}_k - \mathsf{x}_k$$
  is a feasible descent direction at x_k.

Analysis of the Search Direction

- **Case 1:** $\nabla f(\mathsf{x}_k)^\top(\mathsf{s}_k - \mathsf{x}_k) = 0$.

  Then
  $$\nabla f(\mathsf{x}_k)^\top(\mathsf{x} - \mathsf{x}_k) \geq \nabla f(\mathsf{x}_k)^\top(\mathsf{s}_k - \mathsf{x}_k) = 0 \quad \forall \mathsf{x} \in C,$$
  and x_k satisfies the first-order optimality condition.

- **Case 2:** $\nabla f(\mathsf{x}_k)^\top(\mathsf{s}_k - \mathsf{x}_k) < 0$.

  The vector
  $$\mathsf{d}_k := \mathsf{s}_k - \mathsf{x}_k$$
  is a feasible descent direction at x_k.

- A new iterate is obtained by
  $$\mathsf{x}_{k+1} = \mathsf{x}_k + \alpha_k \mathsf{d}_k,$$
  with stepsize $\alpha_k \in (0, 1]$ chosen by a suitable line-search rule.

Iteration of Frank-Wolfe Method

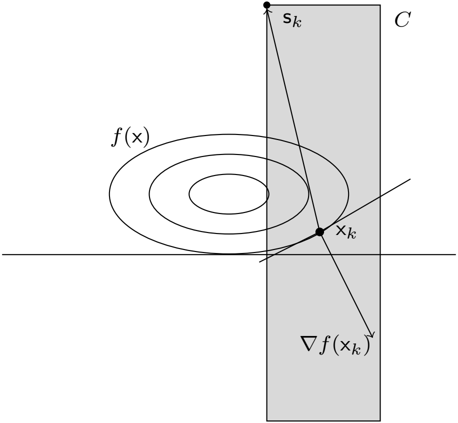


Figure: Iteration of the Frank-Wolfe method.

Examples Where Frank–Wolfe Excels

FW methods are particularly effective when:

- the feasible set is convex but projections are expensive;
- a Linear Minimization Oracle (LMO) is cheap.

Examples Where Frank–Wolfe Excels

FW methods are particularly effective when:

- the feasible set is convex but projections are expensive;
- a Linear Minimization Oracle (LMO) is cheap.

Examples of applications :

- Traffic assignment problem,
- Submodular optimization,
- LASSO problem,
- Matrix completion,
- Adversarial attacks,
- Minimum enclosing ball,
- SVM training,
- Maximal clique search in graphs,
- Sparse optimization.

Examples Where Frank–Wolfe Excels

FW methods are particularly effective when:

- the feasible set is convex but projections are expensive;
- a Linear Minimization Oracle (LMO) is cheap.

We focus on three representative examples:

- LASSO problem,
- Maximal clique search in graphs,
- Matrix completion.

LASSO: Sparse Linear Regression

Given training data (A, b), the LASSO problem reads

$$\min_{\mathsf{x}\in\mathbb{R}^n} \; f(\mathsf{x}) = \|\mathsf{A}\mathsf{x} - \mathsf{b}\|_2^2 \quad \text{s.t.} \quad \|\mathsf{x}\|_1 \le \tau.$$

GOAL: Find a sparse linear regressor.

The feasible set is the ℓ_1-ball:

$$C = \{\mathsf{x} : \|\mathsf{x}\|_1 \le \tau\} = \mathrm{conv}\{\pm\tau \mathsf{e}_i : i = 1, \ldots, n\}.$$

LASSO: Sparse Linear Regression

Given training data (A, b), the LASSO problem reads

$$\min_{\mathsf{x}\in\mathbb{R}^n} f(\mathsf{x}) = \|\mathsf{A}\mathsf{x} - \mathsf{b}\|_2^2 \quad \text{s.t.} \quad \|\mathsf{x}\|_1 \leq \tau.$$

GOAL: Find a sparse linear regressor.

The feasible set is the ℓ_1-ball:

$$C = \{\mathsf{x} : \|\mathsf{x}\|_1 \leq \tau\} = \text{conv}\{\pm\tau \mathsf{e}_i : i = 1, \ldots, n\}.$$

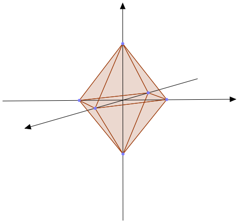


Figure: ℓ_1 ball.

LASSO: Sparse Linear Regression

Given training data (A, b), the LASSO problem reads

$$\min_{\mathsf{x}\in\mathbb{R}^n} f(\mathsf{x}) = \|\mathsf{A}\mathsf{x} - \mathsf{b}\|_2^2 \quad \text{s.t.} \quad \|\mathsf{x}\|_1 \leq \tau.$$

GOAL: Find a sparse linear regressor.

The feasible set is the ℓ_1-ball:

$$C = \{\mathsf{x} : \|\mathsf{x}\|_1 \leq \tau\} = \mathrm{conv}\{\pm\tau \mathsf{e}_i : i = 1, \ldots, n\}.$$

LMO

At iteration k:

$$\mathrm{LMO}_C(\nabla f(\mathsf{x}_k)) = \mathrm{sign}(-\nabla_{i_k} f(\mathsf{x}_k))\, \tau\, \mathsf{e}_{i_k}, \quad i_k \in \mathrm{argmax}_i |\nabla_i f(\mathsf{x}_k)|.$$

Key point: sparse LMO solutions and $\mathcal{O}(n)$ cost.

Maximal Clique in Graphs

Given a graph $G = (V, E)$ with adjacency matrix A_G, finding a maximal clique can be formulated as

$$\max_{\mathsf{x} \in \Delta_{n-1}} \; f(\mathsf{x}) = \mathsf{x}^\top \mathsf{A}_G \mathsf{x} + \frac{1}{2}\|\mathsf{x}\|^2.$$

GOAL: Identify a large (maximal) fully connected subgraph.

The feasible set is the simplex:

$$\Delta_{n-1} = \{\mathsf{x} \in \mathbb{R}^n_+ : \mathsf{e}^\top \mathsf{x} = 1\} = \mathrm{conv}\{\mathsf{e}_i : i = 1, \ldots, n\}.$$

Given a graph $G = (V, E)$ with adjacency matrix A_G, finding a maximal clique can be formulated as

$$\max_{\mathsf{x}\in\Delta_{n-1}} f(\mathsf{x}) = \mathsf{x}^\top \mathsf{A}_G \mathsf{x} + \tfrac{1}{2}\|\mathsf{x}\|^2.$$

GOAL: Identify a large (maximal) fully connected subgraph.

The feasible set is the simplex:

$$\Delta_{n-1} = \{\mathsf{x} \in \mathbb{R}^n_+ : \mathsf{e}^\top \mathsf{x} = 1\} = \mathrm{conv}\{\mathsf{e}_i : i = 1, \ldots, n\}.$$

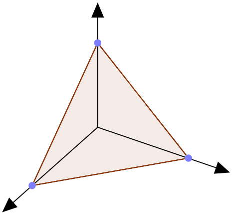


Figure: Unit simplex.

Maximal Clique in Graphs

Given a graph $G = (V, E)$ with adjacency matrix A_G, finding a maximal clique can be formulated as

$$\max_{\mathsf{x} \in \Delta_{n-1}} f(\mathsf{x}) = \mathsf{x}^\top \mathsf{A}_G \mathsf{x} + \frac{1}{2}\|\mathsf{x}\|^2.$$

GOAL: Identify a large (maximal) fully connected subgraph.

The feasible set is the simplex:

$$\Delta_{n-1} = \{\mathsf{x} \in \mathbb{R}^n_+ : \mathsf{e}^\top \mathsf{x} = 1\} = \mathrm{conv}\{\mathsf{e}_i : i = 1, \dots, n\}.$$

LMO

At iteration k:

$$\mathrm{LMO}_{\Delta_{n-1}}(\nabla f(\mathsf{x}_k)) = \mathsf{e}_{i_k}, \quad i_k \in \mathrm{argmax}_i \nabla_i f(\mathsf{x}_k).$$

Key point: sparse LMO solutions and $\mathcal{O}(n)$ cost.

Matrix Completion: Low-Rank Recovery

Given observed entries $\{U_{ij}\}_{(i,j)\in J}$, the problem reads

$$\min_{\mathsf{X}\in\mathbb{R}^{n_1\times n_2}} \sum_{(i,j)\in J} (X_{ij} - U_{ij})^2 \quad \text{s.t.} \quad \|\mathsf{X}\|_* \le \delta.$$

GOAL: Recover a low-rank matrix from partial observations.

The feasible set is the nuclear norm ball:

$$C = \{\mathsf{X} : \|\mathsf{X}\|_* \le \delta\} = \mathrm{conv}\{\delta\, \mathsf{u}\mathsf{v}^\top : \|\mathsf{u}\| = \|\mathsf{v}\| = 1\}.$$

Matrix Completion: Low-Rank Recovery

Given observed entries $\{U_{ij}\}_{(i,j)\in J}$, the problem reads

$$\min_{\mathsf{X}\in\mathbb{R}^{n_1\times n_2}} \sum_{(i,j)\in J} (X_{ij}-U_{ij})^2 \quad \text{s.t.} \quad \|\mathsf{X}\|_* \le \delta.$$

GOAL: Recover a low-rank matrix from partial observations.

The feasible set is the nuclear norm ball:

$$C = \{\mathsf{X} : \|\mathsf{X}\|_* \le \delta\} = \mathrm{conv}\{\delta\, \mathsf{u}\mathsf{v}^\top : \|\mathsf{u}\| = \|\mathsf{v}\| = 1\}.$$

LMO

At iteration k:

$$\mathrm{LMO}_C(\nabla f(\mathsf{X}_k)) = \delta\, \mathsf{u}_1\mathsf{v}_1^\top,$$

where $(\mathsf{u}_1, \mathsf{v}_1)$ are related to the top singular value of $-\nabla f(\mathsf{X}_k) = -2(\mathsf{X}_k - \mathsf{U})_J$.

Key point: each FW step adds at most one new rank-one component.

Stepsizes in First-Order Methods

At each iteration, FW-type methods update

$$\mathsf{x}_{k+1} = \mathsf{x}_k + \alpha_k \mathsf{d}_k,$$

where the choice of the stepsize α_k plays a crucial role.

Popular rules balance:

- theoretical guarantees,
- practical performance,
- computational cost.

Simple Stepsize Rules

- **Exact line search**

$$\alpha_k = \min \operatorname{argmin}_{\alpha \in (0, \alpha_k^{\max}]} f(\mathsf{x}_k + \alpha \mathsf{d}_k)$$

  Guarantees maximal decrease along d_k, but may be costly.

- **Armijo line search**
  - Choose $0 < \delta < 1, \ 0 < \gamma < \frac{1}{2}$.
  - Try $\alpha = \delta^m \alpha_k^{\max}$, $m = 0, 1, \dots$ and stop when

$$f(\mathsf{x}_k + \alpha \mathsf{d}_k) \leq f(\mathsf{x}_k) + \gamma \alpha \nabla f(\mathsf{x}_k)^\top \mathsf{d}_k$$

- **Diminishing stepsize**

$$\alpha_k = \frac{2}{k+2}$$

  Classic choice for Frank–Wolfe, widely used in theory.

- **Unit stepsize**

$$\alpha_k = 1$$

  Mainly used for concave objectives. Under suitable assumptions, finite convergence can be shown.

Lipschitz-Based Stepsize

If ∇f is L-Lipschitz continuous, a natural choice is

$$\alpha_k = \alpha_k(L) = \min\left\{-\frac{\nabla f(\mathsf{x}_k)^\top \mathsf{d}_k}{L\|\mathsf{d}_k\|^2},\ \alpha_k^{\max}\right\}.$$

- Requires knowledge (or estimate) of L,
- Closed-form and cheap,
- Central in convergence analysis.

Lipschitz-Based Stepsize

If ∇f is L-Lipschitz continuous, a natural choice is

$$\alpha_k = \alpha_k(L) = \min\left\{-\frac{\nabla f(\mathsf{x}_k)^\top \mathsf{d}_k}{L\|\mathsf{d}_k\|^2},\ \alpha_k^{\max}\right\}.$$

- Requires knowledge (or estimate) of L,
- Closed-form and cheap,
- Central in convergence analysis.

The Lipschitz-based stepsize minimizes the quadratic model

$$m_k(\alpha; L) = f(\mathsf{x}_k) + \alpha\, \nabla f(\mathsf{x}_k)^\top \mathsf{d}_k + \frac{L\alpha^2}{2}\|\mathsf{d}_k\|^2 \geq\ f(\mathsf{x}_k + \alpha \mathsf{d}_k),$$

where inequality follows by the Descent Lemma.

- **Interpretation:** step chosen by minimizing a local upper bound.
- In case L unknown, use backtracking rule.

Sufficient Decrease with Lipschitz Stepsize

Lemma

If α_k is chosen via the Lipschitz rule and $\alpha_k < \alpha_k^{\max}$, then

$$f(\mathsf{x}_{k+1}) \leq f(\mathsf{x}_k) - \frac{1}{2L}\left(\nabla f(\mathsf{x}_k)^\top \widehat{\mathsf{d}}_k\right)^2.$$

- Explicit decrease bound,
- Key ingredient for convergence rates,
- Links geometry of directions to progress.

Normalized vector

For a vector d we denote as $\widehat{\mathsf{d}} := \frac{1}{\|\mathsf{d}\|}\mathsf{d}$ its normalization, with the convention $\widehat{\mathsf{d}} = 0$ if $\mathsf{d} = 0$

Takeaway on Stepsizes

- Multiple choices available, no universal best rule,
- Lipschitz-based stepsize is analysis-friendly,
- Line searches trade accuracy for robustness,
- FW often benefits from simple, structured steps.

Stopping Condition via the Frank–Wolfe Gap

- A key quantity for measuring convergence of Frank–Wolfe methods is the Frank–Wolfe (FW) gap, defined as

$$G(\mathsf{x}) := \max_{\mathsf{s}\in C} \; -\nabla f(\mathsf{x})^\top (\mathsf{s}-\mathsf{x}) \; . \tag{2}$$

Stopping Condition via the Frank–Wolfe Gap

- A key quantity for measuring convergence of Frank–Wolfe methods is the Frank–Wolfe (FW) gap, defined as

$$G(\mathsf{x}) := \max_{\mathsf{s} \in C} \; -\nabla f(\mathsf{x})^\top (\mathsf{s} - \mathsf{x}) \; . \tag{2}$$

- The FW gap is always nonnegative and satisfies

$$G(\mathsf{x}) = 0 \quad \Longleftrightarrow \quad \mathsf{x} \text{ is a first-order stationary point.}$$

Stopping Condition via the Frank–Wolfe Gap

- A key quantity for measuring convergence of Frank–Wolfe methods is the Frank–Wolfe (FW) gap, defined as

$$G(\mathsf{x}) := \max_{\mathsf{s} \in C} \; -\nabla f(\mathsf{x})^\top (\mathsf{s} - \mathsf{x}) \; . \tag{2}$$

- The FW gap is always nonnegative and satisfies

$$G(\mathsf{x}) = 0 \quad \iff \quad \mathsf{x} \text{ is a first-order stationary point.}$$

- By construction, $G(\mathsf{x})$ is readily available during the algorithm, since it is obtained when solving the linear minimization subproblem.

Stopping Condition via the Frank–Wolfe Gap

- A key quantity for measuring convergence of Frank–Wolfe methods is the Frank–Wolfe (FW) gap, defined as

$$G(\mathsf{x}) := \max_{\mathsf{s}\in C} \; -\nabla f(\mathsf{x})^\top (\mathsf{s} - \mathsf{x}) \; . \tag{2}$$

- The FW gap is always nonnegative and satisfies

$$G(\mathsf{x}) = 0 \quad \Longleftrightarrow \quad \mathsf{x} \text{ is a first-order stationary point.}$$

- By construction, $G(\mathsf{x})$ is readily available during the algorithm, since it is obtained when solving the linear minimization subproblem.
- If f is convex, we have

$$G(\mathsf{x}) \geq -\nabla f(\mathsf{x})^\top (\mathsf{x}^* - \mathsf{x}) \geq f(\mathsf{x}) - f^*, \tag{3}$$

so the FW gap provides an upper bound on the primal optimality gap.

Convergence Rates of Frank–Wolfe

Let $G(\mathsf{x})$ denote the Frank–Wolfe gap.

- **Nonconvex f:**

$$\min_{i\in[0:k]} G(\mathsf{x}_i) = \mathcal{O}(k^{-1/2})$$

(stationarity guarantee).

- **Convex f:**

$$f(\mathsf{x}_k) - f^* = \mathcal{O}(k^{-1})$$

(true optimality gap).

Goal of this section

prove the $\mathcal{O}(1/k)$ rate for convex objectives using the Lipschitz-dependent stepsize.

$\mathcal{O}(1/k)$ Rate for Convex Objectives

Theorem

Assume f is convex with L-Lipschitz gradient and

$$\alpha_k = \min\left\{-\frac{\langle \nabla f(\mathsf{x}_k), \mathsf{d}_k\rangle}{L\|\mathsf{d}_k\|^2}, 1\right\}.$$

Then for all $k \geq 1$,

$$f(\mathsf{x}_k) - f^* \leq \frac{2LD^2}{k+2},$$

where $D = \max_{x,y\in C} \|x - y\|$ is the diameter of the feasible set.

Idea of the proof:

- separate analysis for full FW steps and short steps;
- combine descent estimates with convexity and diameter bounds.

Key Lemma: Full Frank–Wolfe Step

Lemma

If $\mathsf{d}_k = \mathsf{d}_k^{FW}$ and $\alpha_k = 1$, then

$$f(\mathsf{x}_{k+1}) - f^* \leq \tfrac{1}{2}\min\left\{L\|\mathsf{d}_k\|^2,\ f(\mathsf{x}_k) - f^*\right\}.$$

Proof sketch:

- From the stepsize rule and FW direction:

$$G(\mathsf{x}_k) = -\langle \nabla f(\mathsf{x}_k), \mathsf{d}_k\rangle \geq L\|\mathsf{d}_k\|^2.$$

- Descent Lemma gives:

$$f(\mathsf{x}_{k+1}) \leq f(\mathsf{x}_k) + \langle \nabla f(\mathsf{x}_k), \mathsf{d}_k\rangle + \tfrac{L}{2}\|\mathsf{d}_k\|^2.$$

- Convexity implies $f(\mathsf{x}_k) - f^* + \langle \nabla f(\mathsf{x}_k), \mathsf{d}_k\rangle \leq 0$ i.e. $G(\mathsf{x}_k) \geq f(\mathsf{x}_k) - f^*$.
- Combine inequalities to obtain the bound.

Descent for Short Steps

If $\alpha_k < 1$, the Lipschitz stepsize gives (Lemma earlier):

$$f(\mathsf{x}_{k+1}) \leq f(\mathsf{x}_k) - \frac{1}{2L} \left\langle \nabla f(\mathsf{x}_k), \widehat{\mathsf{d}}_k \right\rangle^2.$$

Using:

- $\|\mathsf{d}_k\| \leq D$,
- $G(\mathsf{x}_k) = \langle \nabla f(\mathsf{x}_k), \mathsf{d}_k \rangle \geq f(\mathsf{x}_k) - f^*$,

we obtain

$$f(\mathsf{x}_{k+1}) - f^* \leq (f(\mathsf{x}_k) - f^*)\left(1 - \frac{f(\mathsf{x}_k) - f^*}{2LD^2}\right).$$

Conclusion and Remarks

- By induction, the contraction yields

$$f(\mathsf{x}_k) - f^* \leq \frac{2LD^2}{k+2}.$$

- The rate holds in Banach spaces when C is convex and weakly compact.
- The bound is **tight**: zig-zagging near the boundary yields $\Omega(1/k)$ worst-case behavior.
- The FW gap also satisfies

$$\min_{i \leq k} G(\mathsf{x}_i) = \mathcal{O}(1/k).$$

- Some stepsizes give $\mathcal{O}(\log k/k)$ rates.

Frank–Wolfe Variants: Motivation

Classic FW enjoys a $\mathcal{O}(1/k)$ rate, but may:

- converge slowly near the boundary,
- fail to identify the optimal support in finite time.

Frank–Wolfe Variants: Motivation

Classic FW enjoys a $\mathcal{O}(1/k)$ rate, but may:

- converge slowly near the boundary,
- fail to identify the optimal support in finite time.

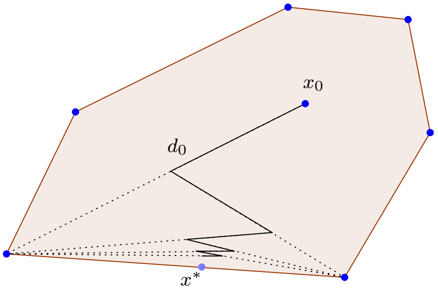


Figure: zig-zagging phenomenon.

Frank–Wolfe Variants: Motivation

Classic FW enjoys a $\mathcal{O}(1/k)$ rate, but may:

- converge slowly near the boundary,
- fail to identify the optimal support in finite time.

Idea: use **active sets** to build richer descent directions.

Active-set FW methods:

- maintain a set A_k such that $\mathsf{x}_k \in \mathrm{conv}(A_k)$,
- allow away or pairwise moves,
- can achieve faster rates and finite support identification.

Away-Step and Pairwise Frank–Wolfe

Assume $\mathsf{x}_k = \sum_{v \in A_k} \lambda_v v$, with $\sum_{v \in A_k} \lambda_v = 1,\ \lambda_v \geq 0$, and $A_k \subseteq C$.

Away vertex:

$$\mathsf{v}_k \in \operatorname{argmax}_{\mathsf{y} \in A_k} \langle \nabla f(\mathsf{x}_k), \mathsf{y} \rangle.$$

Away-step direction (AFW):

$$\mathsf{d}_k^{AS} = \mathsf{x}_k - \mathsf{v}_k, \qquad \mathsf{d}_k \in \operatorname{argmax}_{\mathsf{d} \in \{\mathsf{d}_k^{FW}, \mathsf{d}_k^{AS}\}} \langle -\nabla f(\mathsf{x}_k), \mathsf{d} \rangle.$$

Pairwise FW (PFW):

$$\mathsf{d}_k^{PFW} = \mathsf{s}_k - \mathsf{v}_k, \qquad \mathsf{s}_k \in \operatorname{argmin}_{\mathsf{x} \in C} \langle \nabla f(\mathsf{x}_k), \mathsf{x} \rangle.$$

Key point: mass is moved directly from a bad atom to a good one.

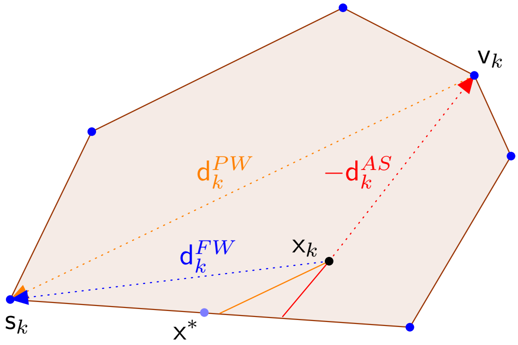


Figure: Behavior of the Frank-Wolfe variants.

Affine Invariance

The FW method and its variants are affine invariant.

Let P be a linear transformation, let $\hat{f}$ be such that

$$\hat{f}(\mathsf{P}\mathsf{x}) = f(\mathsf{x}), \qquad \hat{C} = \mathsf{P}(C).$$

Then, for every sequence $\{\mathsf{x}_k\}$ generated by the FW method applied to (f, C), the sequence

$$\{\mathsf{y}_k\} := \{\mathsf{P}\mathsf{x}_k\}$$

can be generated by the FW method applied to $(\hat{f}, \hat{C})$ using the same stepsizes.

Consequence

Algorithmic behavior is independent of the particular representation of the feasible set.

Reduction to the Simplex

Consider the special case where P is the matrix collecting the elements of a finite set A as columns.

By affine invariance:

- Results proved for $C = \Delta_{|A|-1}$ immediately extend to

$$\hat{C} := \mathrm{conv}(A).$$

- Convergence guarantees derived on the simplex apply to general polytopes.

Key idea

Affine invariance allows one to work on $\Delta_{|A|-1}$ without loss of generality.

Affine-Invariant Convergence Rates

An affine invariant convergence rate bound for convex objectives can be expressed using the curvature constant

$$\kappa_{f,C} := \sup\left\{2\frac{f(\alpha \mathsf{y} + (1-\alpha)\mathsf{x}) - f(\mathsf{x}) - \alpha \nabla f(\mathsf{x})^\top(\mathsf{y}-\mathsf{x})}{\alpha^2} : \{\mathsf{x},\mathsf{y}\} \subset C,\ \alpha \in (0,1]\right\}.$$

It holds that

$$\kappa_{f,C} \leq LD^2,$$

where D is the diameter of C.

Using the diminishing stepsize, we obtain for FW:

$$f(\mathsf{x}_k) - f^* \leq \frac{2\kappa_{f,C}}{k+2}.$$

Inexact Linear Oracle

In many applications, the linear minimization subproblem in FW methods can only be solved approximately. For this reason, convergence analyses often allow for inexact linear oracles.

Common Assumption

Oracle returns a point $\tilde{\mathsf{s}}_k \in C$ satisfying

$$\nabla f(\mathsf{x}_k)^\top (\tilde{\mathsf{s}}_k - \mathsf{x}_k) \leq \min_{\mathsf{s} \in C} \nabla f(\mathsf{x}_k)^\top (\mathsf{s} - \mathsf{x}_k) + \delta_k, \qquad (4)$$

where $\{\delta_k\}$ is a sequence of nonnegative approximation errors.

The exact FW oracle is recovered when $\delta_k = 0$ for all k.

If the approximation error is constant, $\delta_k \equiv \delta > 0$, then the FW method cannot converge beyond accuracy δ.

Using the stepsize $\alpha_k = \frac{2}{k+2}$, yields

$$f(\mathsf{x}_k) - f^* = \mathcal{O}\left(\frac{1}{k} + \delta\right).$$

$\mathcal{O}(1/k)$ rate

It can be recovered if the approximation errors decay sufficiently fast. In particular, assume $\delta_k = \frac{\delta \kappa_{f,C}}{k+2}$ for some constant $\delta > 0$.

Then, for the FW method with exact line search or $\alpha_k = \frac{2}{k+2}$, we have

$$f(\mathsf{x}_k) - f^* \le \frac{2\kappa_{f,C}}{k+2}(1+\delta).$$

Improved Rates

Stronger assumptions on the objective and/or the domain allow the FW method and its variants to achieve convergence rates faster than $\mathcal{O}(1/k)$.

Strongly convex function

objective function f is μ-strongly convex, i.e.,

$$f(\mathsf{y}) \geq f(\mathsf{x}) + \nabla f(\mathsf{x})^\top (\mathsf{y} - \mathsf{x}) + \frac{\mu}{2}\|\mathsf{x} - \mathsf{y}\|^2 \quad \forall \{\mathsf{x}, \mathsf{y}\} \subset C.$$

Strongly convex set

A closed convex set $C \subset \mathbb{R}^n$ is called β_C-**strongly convex** if there exists $\beta_C > 0$ such that for all $\mathsf{x}, \mathsf{y} \in C$ and all $\alpha \in [0, 1]$,

$$\alpha \mathsf{x} + (1 - \alpha)\mathsf{y} + \frac{\beta_C}{2}\alpha(1 - \alpha)\, \|\mathsf{x} - \mathsf{y}\|^2 \, \mathbb{B} \ \subseteq \ C,$$

where $\mathbb{B}$ denotes the unit Euclidean ball.

Linear Convergence for Strongly Convex f and $\mathsf{x}^* \in \mathrm{ri}(C)$

Linear Convergence

Assume:

- f strongly convex,
- $\mathsf{x}^* \in \mathrm{ri}(C)$.

Then FW with exact line search or the stepsize $\alpha_k(L)$ converges at a rate:

$$f(\mathsf{x}_k) - f^* \leq \left[1 - \frac{\mu}{L}\left(\frac{\mathsf{dist}(\mathsf{x}^*, \partial C)}{D}\right)^2\right]^k (f(\mathsf{x}_0) - f^*).$$

Strong Convexity of the Domain

A closed convex set $C \subset \mathbb{R}^n$ is called β_C-**strongly convex** if there exists $\beta_C > 0$ such that for all $\mathrm{x}, \mathrm{y} \in C$ and all $\alpha \in [0,1]$,

$$\alpha \mathrm{x} + (1-\alpha)\mathrm{y} + \frac{\beta_C}{2}\alpha(1-\alpha)\left\|\mathrm{x}-\mathrm{y}\right\|^2 \mathbb{B} \subseteq C, \tag{5}$$

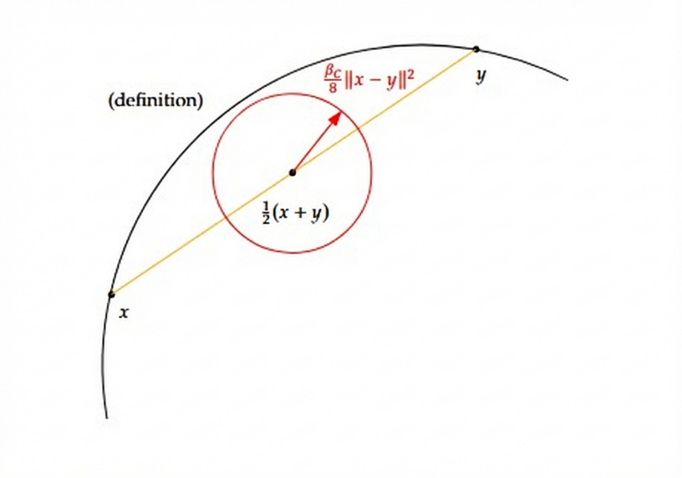
where $\mathbb{B}$ denotes the unit Euclidean ball.

Equivalently, C contains a ball of radius $\frac{\beta_C}{2}\alpha(1-\alpha)\left\|\mathrm{x}-\mathrm{y}\right\|^2$ centered at $\alpha \mathrm{x} + (1-\alpha)\mathrm{y}$.

Geometric interpretation

- ∂C has positive curvature everywhere
- C has no flat faces (unlike polytopes)
- Balls, ellipsoids, and ℓ_p balls with $p > 2$ are strongly convex

Curvature prevents the zig–zagging behavior typical of FW on polytopes.



Interpretation: the boundary of C bends inward with a quadratic curvature, ruling out flat faces and sharp edges.

Rates for Strongly Convex Domains

When C strongly convex, the classic FW method enjoys faster rates.

If C is β_C-strongly convex, then:

- $\mathcal{O}(1/k^2)$ convergence for strongly convex objectives,
- Linear convergence if $\|\nabla f(\mathrm{x})\| \geq c > 0$.

In the latter case,

$$h_{k+1} \leq \max\left\{\frac{1}{2},\ 1 - \frac{L}{2c\beta_C}\right\} h_k.$$

Intermediate rates can be obtained via:

- Hölderian error bounds (objective)
- Uniform convexity (domain)

Summary of Convergence Rates

Method	Objective	Domain	Assumptions	Rate
FW	NC	Generic	–	$\mathcal{O}(1/\sqrt{k})$
FW	C	Generic	–	$\mathcal{O}(1/k)$
FW	SC	Generic	$\mathrm{x}^* \in \mathrm{ri}(C)$	Linear
Variants	SC	Polytope	–	Linear
FW	SC	Strongly convex	–	$\mathcal{O}(1/k^2)$
FW	C	Strongly convex	$\min \|\nabla f(\mathrm{x})\| > 0$	Linear

Table: Known convergence rates for FW and its variants.

Overview of Generalized FW and Extensions

- **Generalized Frank–Wolfe (GFW):**
  - Composite optimization: $\min_{x \in \mathbb{R}^n} f(x) + g(x)$ with f smooth, g convex.
  - Generalized FW direction: linearization of f and proximal step on g.
  - FW gap extends naturally and coincides with the Fenchel duality gap.
  - Equivalence with mirror gradient descent on the Fenchel dual.
- **Extensions of FW methods:**
  - Block Coordinate FW (BCFW) for product domains, with stochastic, parallel and asynchronous variants.
  - Conditional Gradient Sliding (CGS): accelerated schemes achieving optimal gradient complexity.
  - FW methods for min-norm point problem and optimization over the trace norm ball.

Conclusions

- FW features: **projection-free updates**, **affine invariance**, and **sparse iterates**.
- For convex objectives, FW achieves a $\mathcal{O}(1/k)$ rate under mild assumptions.
- For strongly convex objectives or domains, better rates are attainable.
- Variants enable **support identification** and improve practical performance.
- Generalized and block-coordinate FW methods extend applicability to:
  - large-scale and distributed settings,
  - composite and non-smooth problems,
  - structured and low-rank optimization.
- FW represents a versatile and competitive first-order method.

Thanks for your attention!

For Further Details:

I.M. Bomze, F. Rinaldi, D. Zeffiro, *"Frank-Wolfe and friends: a journey into projection-free first-order optimization methods"*, Annals of OR, 2024

Complexity of Linear Minimization vs Projection

Set C	**LMO complexity**	**Projection complexity**
ℓ_p-ball, $p \in \{1, 2, \infty\}$	$\mathcal{O}(n)$	$\mathcal{O}(n)$
ℓ_p-ball, $p \in (1, 2) \cup (2, \infty)$	$\mathcal{O}(n)$	$\mathcal{O}\left(\frac{n\rho^2 \Vert \mathsf{y} - \bar{\mathsf{x}} \Vert_2^2}{\varepsilon^2}\right)$
Nuclear norm-ball	$\mathcal{O}\left(\nu \log(m+n) \sqrt{\sigma_1/\varepsilon}\right)$	$\mathcal{O}(mn \min\{m, n\})$
Flow polytope	$\mathcal{O}(m+n)$	$\widetilde{\mathcal{O}}(m^3 n + n^2)$
Birkhoff polytope	$\mathcal{O}(n^3)$	$\mathcal{O}\left(\frac{n^2 d_z^2}{\varepsilon^2}\right)$
Permutahedron	$\mathcal{O}(n \log n)$	$\mathcal{O}(n \log n + n)$

ν: number of nonzeros, σ_1: largest singular value, ε: target accuracy, ρ: curvature constant, y, $\bar{\mathsf{x}}$: original point and projection, d_z: Douglas–Rachford distance.
For further details [Combettes, Pokutta, 2021]

Assume $C = \Delta_{n-1}$ and let f be differentiable (not necessarily convex). Define the multiplier functions

$$\lambda_i(\mathsf{x}) = \nabla f(\mathsf{x})^\top(\mathsf{e}_i - \mathsf{x}), \qquad i \in [1:n].$$

If x^* is a stationary point, then $\{\lambda_i(\mathsf{x}^*)\}$ coincide with the Lagrange multipliers and satisfy the complementarity conditions

$$x_i^* \lambda_i(\mathsf{x}^*) = 0 \qquad \forall i \in [1:n].$$

Define

$$I(\mathsf{x}^*) := \{i \in [1:n] : \lambda_i(\mathsf{x}^*) = 0\}, \qquad \mathrm{supp}(\mathsf{x}^*) \subseteq I(\mathsf{x}^*).$$

Let

$$\delta_{\min} := \min_{i:\lambda_i(\mathsf{x}^*)>0} \lambda_i(\mathsf{x}^*), \qquad r_* := \frac{\delta_{\min}}{\delta_{\min} + 2L}.$$

If $\|\mathsf{x}_k - \mathsf{x}^*\|_1 < r_*$ and for every away step $\alpha_k \geq \alpha_k(L)$, then there exists

$$j \leq \min\{n - |I(\mathsf{x}^*)|,\ |\mathrm{supp}(\mathsf{x}_k)| - 1\}$$

such that

$$\mathrm{supp}(\mathsf{x}_{k+j}) \subseteq I(\mathsf{x}^*) \quad \text{and} \quad \|\mathsf{x}_{k+j} - \mathsf{x}^*\|_1 < r_*.$$

Support Identification for the AFW: General Polytopes

Let $C = \text{conv}(A)$ with $|A| < +\infty$ and assume f is μ-strongly convex:

$$f(\mathsf{y}) \geq f(\mathsf{x}) + \nabla f(\mathsf{x})^\top(\mathsf{y}-\mathsf{x}) + \frac{\mu}{2}\|\mathsf{x}-\mathsf{y}\|^2.$$

Define the exposed face at a stationary point x^* as

$$E_C(\mathsf{x}^*) := \operatorname{argmin}_{\mathsf{x}\in C} \nabla f(\mathsf{x}^*)^\top \mathsf{x}.$$

Let $\bar{\mathsf{A}}$ be the matrix collecting the atoms in A, $f_A(\mathsf{y}) := f(\bar{\mathsf{A}}\mathsf{y})$ on $\Delta_{|A|-1}$, and let L_A be the Lipschitz constant of ∇f_A. Define

$$\delta_{\min} := \min_{\mathsf{a}\in A\setminus E_C(\mathsf{x}^*)} \nabla f(\mathsf{x}^*)^\top(\mathsf{a}-\mathsf{x}^*), \qquad r_*(\mathsf{x}^*) := \frac{\delta_{\min}}{\delta_{\min}+2L_A}.$$

Let θ_A be the Hoffman constant and $\mu_A := \mu/(n\theta_A^2)$. If AFW converges linearly, $h_k \leq q^k h_0$, then AFW enters $E_C(\mathsf{x}^*)$ after at most

$$k \geq \max\left\{2\frac{\ln(h_0) - \ln(\mu_A r_*(\mathsf{x}^*)^2/2)}{\ln(1/q)},\ 0\right\}$$

iterations.