The Pivoting Framework: Frank-Wolfe Algorithms with Active Set Size Control

Sebastian Pokutta

Technische Universität Berlin
and
Zuse Institute Berlin

pokutta@zib.de
@spokutta

joint work with: Elias Wirth and Mathieu Besançon

AISTATS 2025

May 2025 · Phuket, Thailand

What is this talk about?

Introduction

A technique to control the active set size of Frank-Wolfe algorithms.

What is this talk about?

Introduction

A technique to control the active set size of Frank-Wolfe algorithms.

Why? Conditional gradients generate *sparse* iterates, now even *sparser*.

Important in many downstream applications, e.g.,

1. counterfactuals
2. local models
3. sparse regression

What is this talk about?

Introduction

A technique to control the active set size
of Frank-Wolfe algorithms.

Why? Conditional gradients generate *sparse* iterates, now even *sparser*.

Important in many downstream applications, e.g.,

1. counterfactuals
2. local models
3. sparse regression

Today: Simplied setup and general overview; see paper for details.

(Hyperlinked) References are not exhaustive; check references contained therein.

Conditional Gradients
a.k.a. the Frank-Wolfe algorithm

—The Basics—

Algorithm 1: Frank-Wolfe algorithm (FW) [with line-search] [Frank and Wolfe, 1956]

Input: Feasible Set $C \subseteq \mathbb{R}^n$, Initial point $\mathbf{x}^{(0)} \in \mathcal{V} = \text{vertex}(C)$.

for $t = 0, 1, 2, \ldots$ **do**

$\mathbf{v}^{(t)} \leftarrow \arg\min_{\mathbf{v} \in C} \langle \nabla f(\mathbf{x}^{(t)}), \mathbf{v} \rangle$

$\eta^{(t)} \leftarrow \arg\min_{\eta \in [0,1]} f(\mathbf{x}^{(t)} + \eta(\mathbf{v}^{(t)} - \mathbf{x}^{(t)}))$

$\mathbf{x}^{(t+1)} \leftarrow (1 - \eta^{(t)})\mathbf{x}^{(t)} + \eta^{(t)}\mathbf{v}^{(t)}$

end

- Convergence rate: $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) = \mathcal{O}(1/t)$ [Wolfe, 1970, Jaggi, 2013]
- Iterates are formed as convex combinations of extreme points of C.
- Simple to implement and very robust
- No parameters to tune

Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

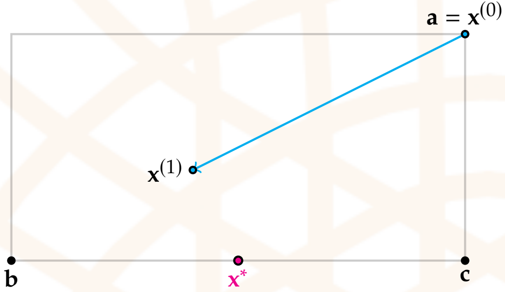


Figure: Frank-Wolfe with line-search.

Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

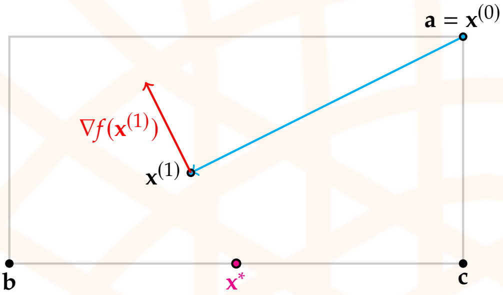


Figure: Frank-Wolfe with line-search.

At $\mathbf{x}^{(1)}$:

$$\nabla f(\mathbf{x}^{(1)}) = \mathbf{x}^{(1)} - \mathbf{x}^*$$

Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

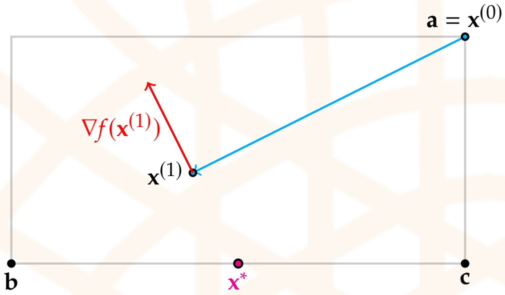


Figure: Frank-Wolfe with line-search.

At $\mathbf{x}^{(1)}$:

$\nabla f(\mathbf{x}^{(1)}) = \mathbf{x}^{(1)} - \mathbf{x}^*$

$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v}\in C}\langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v}\rangle$

Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

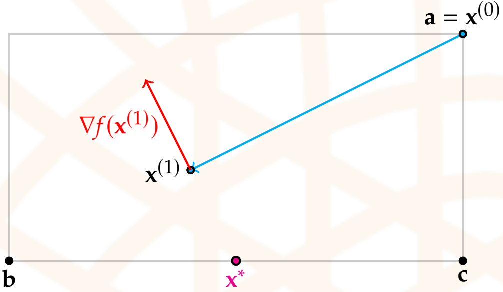


Figure: Frank-Wolfe with line-search.

At $\mathbf{x}^{(1)}$:

$\nabla f(\mathbf{x}^{(1)}) = \mathbf{x}^{(1)} - \mathbf{x}^*$
$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v} \in C} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v} \rangle$
$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$

Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

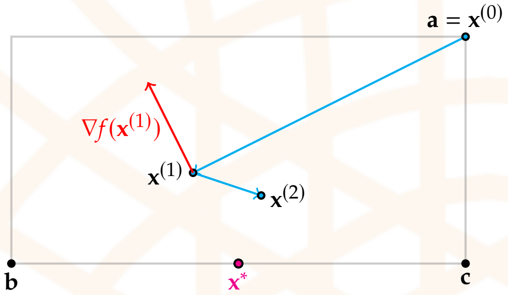


Figure: Frank-Wolfe with line-search.

At $\mathbf{x}^{(1)}$:

$\nabla f(\mathbf{x}^{(1)}) = \mathbf{x}^{(1)} - \mathbf{x}^*$

$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v} \in C} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v} \rangle$

$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$

Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

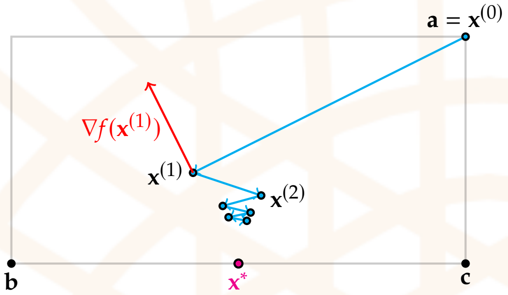


Figure: Frank-Wolfe with line-search.

At $\mathbf{x}^{(1)}$:

$\nabla f(\mathbf{x}^{(1)}) = \mathbf{x}^{(1)} - \mathbf{x}^*$
$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v} \in C} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v} \rangle$
$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$

Pairwise Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

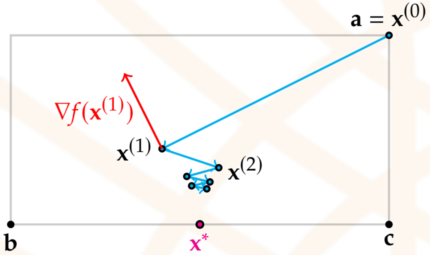


Figure: FW with line-search.

At $\mathbf{x}^{(1)}$:

$$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v}\in C} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v} \rangle$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$$

Pairwise Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

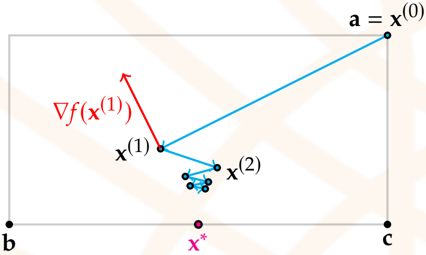


Figure: FW with line-search.

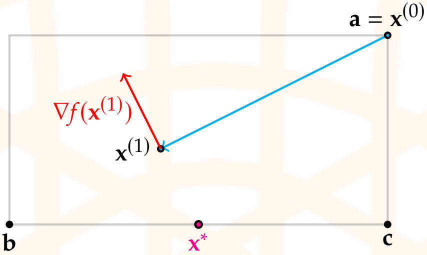


Figure: PFW with line-search.

At $\mathbf{x}^{(1)}$:

$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v} \in C} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v} \rangle$
$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$

Pairwise Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

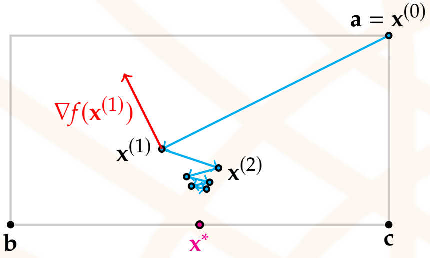


Figure: FW with line-search.

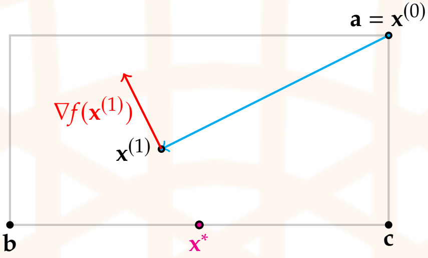


Figure: PFW with line-search.

At $\mathbf{x}^{(1)}$:

$$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v}\in C} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v}\rangle$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$$

At $\mathbf{x}^{(1)} = \alpha_{\mathbf{a}}^{(1)}\mathbf{a} + \alpha_{\mathbf{b}}^{(1)}\mathbf{b}$,
where $\alpha_{\mathbf{a}} + \alpha_{\mathbf{b}} = 1$ and $\alpha_{\mathbf{a}}, \alpha_{\mathbf{b}} \geq 0$:

Pairwise Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

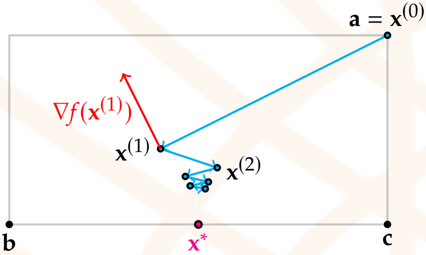


Figure: FW with line-search.

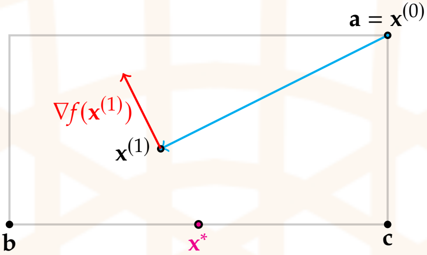


Figure: PFW with line-search.

At $\mathbf{x}^{(1)}$:

$$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v}\in C} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v}\rangle$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$$

At $\mathbf{x}^{(1)} = \alpha_{\mathbf{a}}^{(1)}\mathbf{a} + \alpha_{\mathbf{b}}^{(1)}\mathbf{b}$,
where $\alpha_{\mathbf{a}} + \alpha_{\mathbf{b}} = 1$ and $\alpha_{\mathbf{a}}, \alpha_{\mathbf{b}} \geq 0$:

$$\mathbf{w}^{(1)} = \arg\max_{\mathbf{w}\in\{\mathbf{a},\mathbf{b}\}} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{w}\rangle$$

Pairwise Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

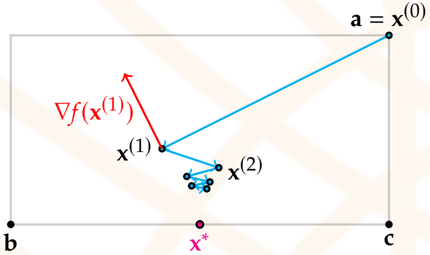


Figure: FW with line-search.

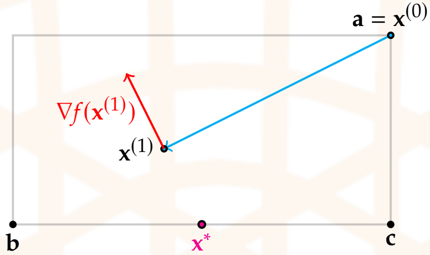


Figure: PFW with line-search.

At $\mathbf{x}^{(1)}$:

$$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v}\in C}\langle\nabla f(\mathbf{x}^{(1)}), \mathbf{v}\rangle$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$$

At $\mathbf{x}^{(1)} = \alpha_{\mathbf{a}}^{(1)}\mathbf{a} + \alpha_{\mathbf{b}}^{(1)}\mathbf{b}$,
where $\alpha_{\mathbf{a}} + \alpha_{\mathbf{b}} = 1$ and $\alpha_{\mathbf{a}}, \alpha_{\mathbf{b}} \geq 0$:

$$\mathbf{w}^{(1)} = \arg\max_{\mathbf{w}\in\{\mathbf{a},\mathbf{b}\}}\langle\nabla f(\mathbf{x}^{(1)}), \mathbf{w}\rangle$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{w}^{(1)})$$

Pairwise Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

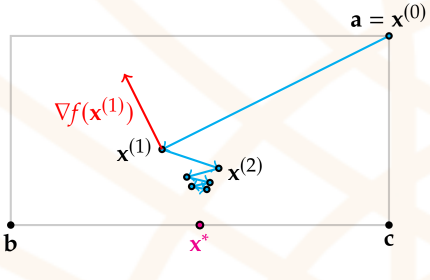


Figure: FW with line-search.

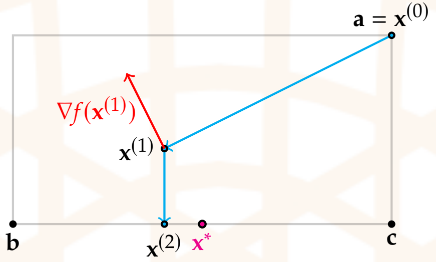


Figure: PFW with line-search.

At $\mathbf{x}^{(1)}$:

$$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v} \in C} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v} \rangle$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$$

At $\mathbf{x}^{(1)} = \alpha_{\mathbf{a}}^{(1)}\mathbf{a} + \alpha_{\mathbf{b}}^{(1)}\mathbf{b}$,
where $\alpha_{\mathbf{a}} + \alpha_{\mathbf{b}} = 1$ and $\alpha_{\mathbf{a}}, \alpha_{\mathbf{b}} \geq 0$:

$$\mathbf{w}^{(1)} = \arg\max_{\mathbf{w} \in \{\mathbf{a}, \mathbf{b}\}} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{w} \rangle$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{w}^{(1)})$$

Pairwise Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

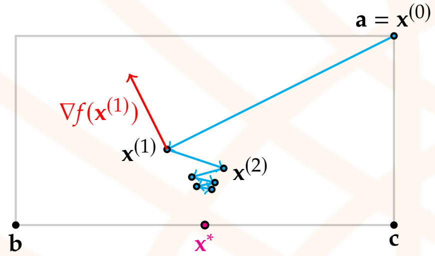


Figure: FW with line-search.

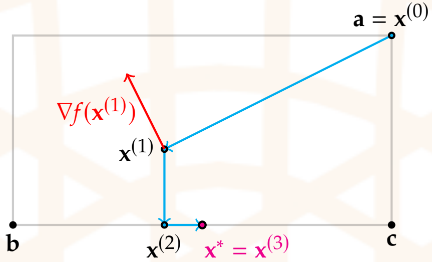


Figure: PFW with line-search.

At $\mathbf{x}^{(1)}$:

$$\mathbf{v}^{(1)} = \arg\min_{\mathbf{v} \in C} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{v} \rangle$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{x}^{(1)})$$

At $\mathbf{x}^{(1)} = \alpha_{\mathbf{a}}^{(1)}\mathbf{a} + \alpha_{\mathbf{b}}^{(1)}\mathbf{b}$,
where $\alpha_{\mathbf{a}} + \alpha_{\mathbf{b}} = 1$ and $\alpha_{\mathbf{a}}, \alpha_{\mathbf{b}} \geq 0$:

$$\mathbf{w}^{(1)} = \arg\max_{\mathbf{w} \in \{\mathbf{a},\mathbf{b}\}} \langle \nabla f(\mathbf{x}^{(1)}), \mathbf{w} \rangle$$
$$\mathbf{x}^{(2)} = \mathbf{x}^{(1)} + \eta^{(1)}(\mathbf{v}^{(1)} - \mathbf{w}^{(1)})$$

Pairwise Frank-Wolfe with line-search over a polytope

Example

Minimize $f(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{x}^*\|_2^2$ over the polytope.

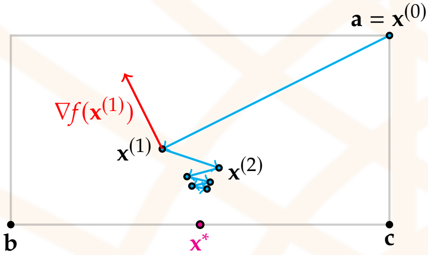


Figure: FW with line-search.

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) = \Theta(t^{-1})$$

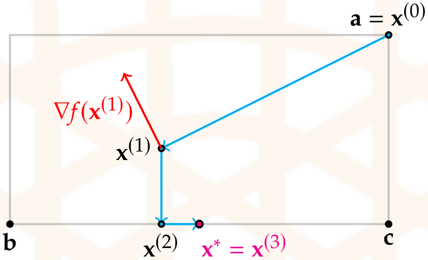


Figure: PFW with line-search.

$$f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) = \Theta(e^{-t})$$

The Pivoting Framework

Frank-Wolfe and Carathéodory's Theorem

Why pivoting?

Worst-case for FW and variants: one new vertex per iteration (bad sparsity).

Frank-Wolfe and Carathéodory's Theorem

Why pivoting?

Worst-case for FW and variants: one new vertex per iteration (bad sparsity).

⇒ Active set can be as large as $|\mathcal{S}^{(t)}| \leq \min\{|\mathcal{V}|, t+1\}$

Bad. Sparsity relevant in many contexts.

[Wolfe, 1970, Jaggi, 2013, Lacoste-Julien and Jaggi, 2015, Bomze et al., 2019, 2020, Combettes and Pokutta, 2020, Tsuji et al., 2022, Carderera et al., 2025, Designolle et al., 2024a, 2023, Filippozzi et al., 2023, Besançon et al., 2025, Sadiku et al., 2025, Designolle et al., 2024b, Macdonald et al., 2022, Carderera et al., 2021, 2024]

Frank-Wolfe and Carathéodory's Theorem

Why pivoting?

Worst-case for FW and variants: one new vertex per iteration (bad sparsity).

⇒ Active set can be as large as $|\mathcal{S}^{(t)}| \leq \min\{|\mathcal{V}|, t+1\}$

Bad. Sparsity relevant in many contexts.

[Wolfe, 1970, Jaggi, 2013, Lacoste-Julien and Jaggi, 2015, Bomze et al., 2019, 2020, Combettes and Pokutta, 2020, Tsuji et al., 2022, Carderera et al., 2025, Designolle et al., 2024a, 2023, Filippozzi et al., 2023, Besançon et al., 2025, Sadiku et al., 2025, Designolle et al., 2024b, Macdonald et al., 2022, Carderera et al., 2021, 2024]

[Carathéodory, 1907]

Theorem (Carathéodory's Theorem)

Let $C \subseteq \mathbb{R}^n$ be a compact convex set. Then, any $\mathbf{x} \in C$ can be represented as the convex combination of at most $n+1$ extreme points of C.

Frank-Wolfe and Carathéodory's Theorem

Why pivoting?

Worst-case for FW and variants: one new vertex per iteration (bad sparsity).

$\Rightarrow$ Active set can be as large as $|\mathcal{S}^{(t)}| \leq \min\{|\mathcal{V}|, t+1\}$

Bad. Sparsity relevant in many contexts.

[Wolfe, 1970, Jaggi, 2013, Lacoste-Julien and Jaggi, 2015, Bomze et al., 2019, 2020, Combettes and Pokutta, 2020, Tsuji et al., 2022, Carderera et al., 2025, Designolle et al., 2024a, 2023, Filippozzi et al., 2023, Besançon et al., 2025, Sadiku et al., 2025, Designolle et al., 2024b, Macdonald et al., 2022, Carderera et al., 2021, 2024]

[Carathéodory, 1907]

Theorem (Carathéodory's Theorem)

Let $C \subseteq \mathbb{R}^n$ be a compact convex set. Then, any $\mathbf{x} \in C$ can be represented as the convex combination of at most $n+1$ extreme points of C.

Goal. Improve $|\mathcal{S}^{(t)}| \leq \min\{|\mathcal{V}|, t+1\}$ to $|\mathcal{S}^{(t)}| \leq \min\{n+1\}$ (simplest variant)

Key observation
The pivoting framework

Suppose $\mathcal{S}^{(t)} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{n+2}\}$ is a set of $n+2$ vertices, i.e., too large.

$$\mathbf{x}^{(t)} = \sum_{i=1}^{n+2} \alpha_i^{(t)} \mathbf{v}_i$$

Goal. Find $\tilde{\mathcal{S}}^{(t)} \subsetneq \mathcal{S}^{(t)}$ such that $\mathbf{x}^{(t)} = \sum_{i=1}^{n+1} \beta_i^{(t)} \mathbf{v}_i$ with $\mathbf{v}_i \in \tilde{\mathcal{S}}^{(t)}$?

Key observation

The pivoting framework

Suppose $\mathcal{S}^{(t)} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{n+2}\}$ is a set of $n + 2$ vertices, i.e., too large.

$$\mathbf{x}^{(t)} = \sum_{i=1}^{n+2} \alpha_i^{(t)} \mathbf{v}_i$$

Goal. Find $\tilde{\mathcal{S}}^{(t)} \subsetneq \mathcal{S}^{(t)}$ such that $\mathbf{x}^{(t)} = \sum_{i=1}^{n+1} \beta_i^{(t)} \mathbf{v}_i$ with $\mathbf{v}_i \in \tilde{\mathcal{S}}^{(t)}$?

Linear program (objective is actually irrelevant):

$$\begin{aligned} \min_{\beta \in \mathbb{R}^{n+2}_{\geq 0}} \quad & \sum_{i=1}^{n+2} \beta_i \\ \text{subject to:} \quad & \sum_{i=1}^{n+2} \beta_i \mathbf{v}_i = \mathbf{x}^{(t)} \\ & \sum_{i=1}^{n+2} \beta_i = 1, \beta_i \geq 0 \quad \forall i \in [n+2] \end{aligned}$$

Key observation

The pivoting framework

Suppose $\mathcal{S}^{(t)} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{n+2}\}$ is a set of $n+2$ vertices, i.e., too large.

$$\mathbf{x}^{(t)} = \sum_{i=1}^{n+2} \alpha_i^{(t)} \mathbf{v}_i$$

Goal. Find $\tilde{\mathcal{S}}^{(t)} \subsetneq \mathcal{S}^{(t)}$ such that $\mathbf{x}^{(t)} = \sum_{i=1}^{n+1} \beta_i^{(t)} \mathbf{v}_i$ with $\mathbf{v}_i \in \tilde{\mathcal{S}}^{(t)}$?

Linear program (objective is actually irrelevant):

$$\begin{aligned} \min_{\beta \in \mathbb{R}^{n+2}_{\geq 0}} \quad & \sum_{i=1}^{n+2} \beta_i \\ \text{subject to:} \quad & \sum_{i=1}^{n+2} \beta_i \mathbf{v}_i = \mathbf{x}^{(t)} \\ & \sum_{i=1}^{n+2} \beta_i = 1, \beta_i \geq 0 \quad \forall i \in [n+2] \end{aligned}$$

Now. Any basic feasible solution = convex combination of at most $n+1$ vertices.

Key observation

The pivoting framework

Suppose $\mathcal{S}^{(t)} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{n+2}\}$ is a set of $n+2$ vertices, i.e., too large.

$$\mathbf{x}^{(t)} = \sum_{i=1}^{n+2} \alpha_i^{(t)} \mathbf{v}_i$$

Goal. Find $\tilde{\mathcal{S}}^{(t)} \subsetneq \mathcal{S}^{(t)}$ such that $\mathbf{x}^{(t)} = \sum_{i=1}^{n+1} \beta_i^{(t)} \mathbf{v}_i$ with $\mathbf{v}_i \in \tilde{\mathcal{S}}^{(t)}$?

Linear program (objective is actually irrelevant):

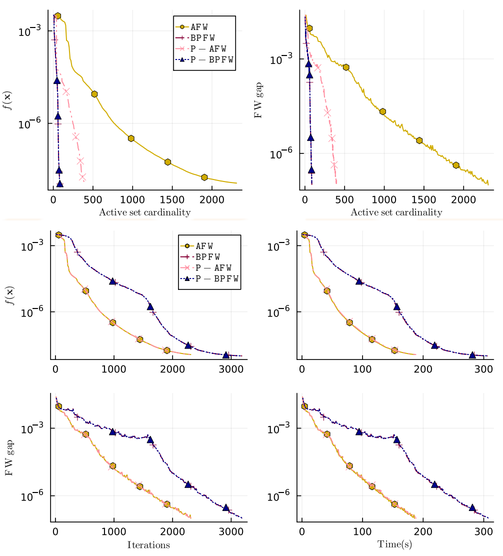
$$\begin{aligned} \min_{\beta \in \mathbb{R}^{n+2}_{\geq 0}} \quad & \sum_{i=1}^{n+2} \beta_i \\ \text{subject to:} \quad & \sum_{i=1}^{n+2} \beta_i \mathbf{v}_i = \mathbf{x}^{(t)} \\ & \sum_{i=1}^{n+2} \beta_i = 1, \beta_i \geq 0 \quad \forall i \in [n+2] \end{aligned}$$

Now. Any basic feasible solution = convex combination of at most $n+1$ vertices.

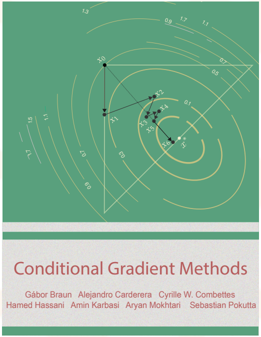
$\Rightarrow$ In any FW algorithm, simply pivot on the active set weights with simplex steps.

Sample Computational experiments

Logistic regression



If you want to learn more...



Thank you!

Conditional Gradient Methods

Gábor Braun, Alejandro Carderera, Cyrille W Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta

`https://conditional-gradients.org/`
`https://arxiv.org/abs/2211.14103`

to appear in MOS-SIAM Series on Optimization

The Pivoting Framework: Frank-Wolfe Algorithms with Active Set Size Control
`https://arxiv.org/abs/2407.11760`

References I

M. Besançon, S. Designolle, J. Halbey, D. Hendrych, D. Kuzinowicz, S. Pokutta, H. Troppens, D. Viladrich Herrmannsdoerfer, and E. Wirth. Improved algorithms and novel applications of the FrankWolfe.jl library. *preprint*, 1 2025.

I. M. Bomze, F. Rinaldi, and S. R. Bulo. First-order methods for the impatient: Support identification in finite time with convergent Frank–Wolfe variants. *SIAM Journal on Optimization*, 29(3):2211–2226, 2019.

I. M. Bomze, F. Rinaldi, and D. Zeffiro. Active set complexity of the away-step Frank–Wolfe algorithm. *SIAM Journal on Optimization*, 30(3):2470–2500, 2020.

G. Braun, A. Carderera, C. W. Combettes, H. Hassani, A. Karbasi, A. Mokhtari, and S. Pokutta. *Conditional Gradient Methods*. to appear in MOS-SIAM Series on Optimization, 2025.

C. Carathéodory. Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen. *Mathematische Annalen*, 64(1):95–115, 1907.

A. Carderera, M. Besançon, and S. Pokutta. Simple steps are all you need: Frank-Wolfe and generalized self-concordant functions. *Proceedings of NeurIPS*, 5 2021.

A. Carderera, M. Besançon, and S. Pokutta. Scalable Frank-Wolfe on Generalized Self-Concordant Functions via Simple Steps. *SIAM Journal on Optimization*, 34 (3), 4 2024.

A. Carderera, S. Pokutta, C. Schütte, and M. Weiser. An efficient first-order conditional gradient algorithm in data-driven sparse identification of nonlinear dynamics to solve sparse recovery problems under noise. *to appear in Journal of Computational and Applied Mathematics*, 3 2025.

C. W. Combettes and S. Pokutta. Boosting Frank-Wolfe by Chasing Gradients. *Proceedings of ICML*, 3 2020.

S. Designolle, G. Iommazzo, M. Besançon, S. Knebel, P. Gelß, and S. Pokutta. Improved local models and new Bell inequalities via Frank-Wolfe algorithms. *Physical Reviews Research*, 10 2023.

S. Designolle, T. Vértesi, and S. Pokutta. Symmetric multipartite Bell inequalities via Frank-Wolfe algorithms. *Physical Review A*, 109, 2 2024a.

S. Designolle, T. Vértesi, and S. Pokutta. Better bounds on Grothendieck constants of finite orders. *preprint*, 9 2024b.

R. Filippozzi, D. S. Gonçalves, and L.-R. Santos. First-order methods for the convex hull membership problem. *European Journal of Operational Research*, 306(1): 17–33, 2023.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, number CONF, pages 427–435. PMLR, 2013.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Proceedings of Advances in Neural Information Processing Systems*, pages 496–504, 2015.

J. Macdonald, M. Besançon, and S. Pokutta. Interpretable Neural Networks with Frank-Wolfe: Sparse Relevance Maps and Relevance Orderings. *Proceedings of ICML*, 5 2022.

S. Sadiku, M. Wagner, S. G. Nagarajan, and S. Pokutta. S-CFE: Simple Counterfactual Explanations. *to appear in Proceedings of AISTATS*, 1 2025.

K. Tsuji, K. Tanaka, and S. Pokutta. Pairwise Conditional Gradients without Swap Steps and Sparser Kernel Herding. *Proceedings of ICML*, 5 2022.

P. Wolfe. Convergence theory in nonlinear programming. *Integer and Nonlinear Programming*, pages 1–36, 1970.