Revisiting the Approximate Carathéodory Problem via the Frank-Wolfe Algorithm

Cyrille W. Combettes

Joint work with Sebastian Pokutta

Georgia Tech H. Milton Stewart School of Industrial and Systems Engineering

Conference on Data Science and Optimization
Fields Institute, Toronto, Canada

November 20th, 2019

Table of contents

Introduction

Theorem (Carathéodory [1907])

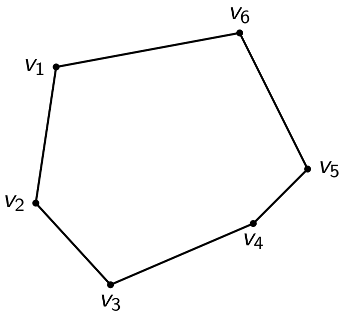
Every point in the convex hull of a set $\mathcal{V} \subset \mathbb{R}^n$ is the convex combination of at most $n+1$ points in $\mathcal{V}$.

Introduction

Theorem (Carathéodory [1907])

Every point in the convex hull of a set $\mathcal{V} \subset \mathbb{R}^n$ is the convex combination of at most $n+1$ points in $\mathcal{V}$.

- In $\mathbb{R}^2$, every point in $\text{conv}(\mathcal{V})$ is the convex combination of at most 3 points in $\mathcal{V}$

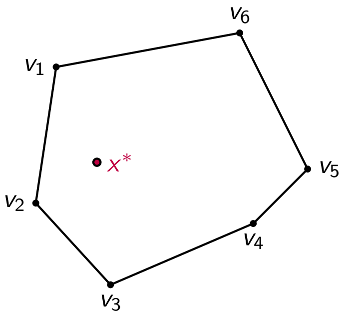


Introduction

Theorem (Carathéodory [1907])

Every point in the convex hull of a set $\mathcal{V} \subset \mathbb{R}^n$ is the convex combination of at most $n+1$ points in $\mathcal{V}$.

- In $\mathbb{R}^2$, every point in conv($\mathcal{V}$) is the convex combination of at most 3 points in $\mathcal{V}$

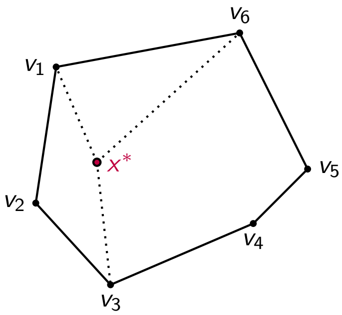


Introduction

Theorem (Carathéodory [1907])

Every point in the convex hull of a set $\mathcal{V} \subset \mathbb{R}^n$ is the convex combination of at most $n+1$ points in $\mathcal{V}$.

- In $\mathbb{R}^2$, every point in conv($\mathcal{V}$) is the convex combination of at most 3 points in $\mathcal{V}$

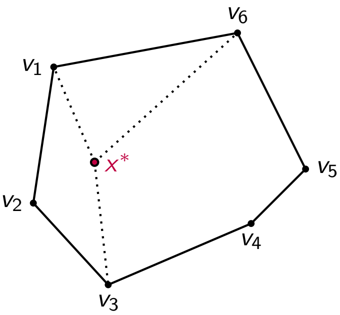


Introduction

Theorem (Carathéodory [1907])

Every point in the convex hull of a set $\mathcal{V} \subset \mathbb{R}^n$ is the convex combination of at most $n+1$ points in $\mathcal{V}$.

- In $\mathbb{R}^2$, every point in $\text{conv}(\mathcal{V})$ is the convex combination of at most 3 points in $\mathcal{V}$
- Can we reduce $n+1$ when we can afford an ϵ-approximation?

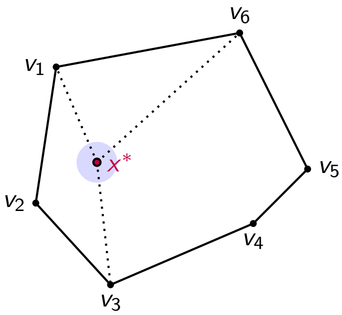


Introduction

Theorem (Carathéodory [1907])

Every point in the convex hull of a set $\mathcal{V} \subset \mathbb{R}^n$ is the convex combination of at most $n+1$ points in $\mathcal{V}$.

- In $\mathbb{R}^2$, every point in $\text{conv}(\mathcal{V})$ is the convex combination of at most 3 points in $\mathcal{V}$
- Can we reduce $n+1$ when we can afford an ϵ-approximation?

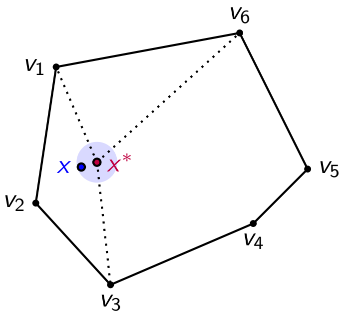


Introduction

Theorem (Carathéodory [1907])

Every point in the convex hull of a set $\mathcal{V} \subset \mathbb{R}^n$ is the convex combination of at most $n+1$ points in $\mathcal{V}$.

- In $\mathbb{R}^2$, every point in conv($\mathcal{V}$) is the convex combination of at most 3 points in $\mathcal{V}$
- Can we reduce $n+1$ when we can afford an ϵ-approximation?

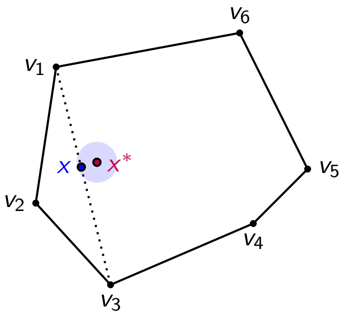


Introduction

Theorem (Carathéodory [1907])

Every point in the convex hull of a set $\mathcal{V} \subset \mathbb{R}^n$ is the convex combination of at most $n + 1$ points in $\mathcal{V}$.

- In $\mathbb{R}^2$, every point in $\text{conv}(\mathcal{V})$ is the convex combination of at most 3 points in $\mathcal{V}$
- Can we reduce $n + 1$ when we can afford an ϵ-approximation?

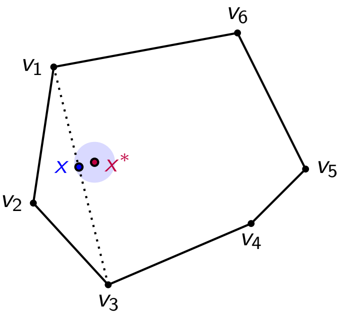


Introduction

Theorem (Carathéodory [1907])

Every point in the convex hull of a set $\mathcal{V} \subset \mathbb{R}^n$ is the convex combination of at most $n+1$ points in $\mathcal{V}$.

- In $\mathbb{R}^2$, every point in $\text{conv}(\mathcal{V})$ is the convex combination of at most 3 points in $\mathcal{V}$
- Can we reduce $n+1$ when we can afford an ϵ-approximation?
- Define the *sparsity* of x as the minimum number of vertices necessary to form x as a convex combination



The approximate Carathéodory problem

Problem

Find $x \in \mathrm{conv}(\mathcal{V})$ with high sparsity satisfying $\|x - x^*\|_p \leqslant \epsilon$.

The approximate Carathéodory problem

Problem

Find $x \in \mathrm{conv}(\mathcal{V})$ with high sparsity satisfying $\|x - x^*\|_p \leqslant \epsilon$.

- Applications in game theory, combinatorial optimization, and machine learning

The approximate Carathéodory problem

Problem

Find $x \in \mathrm{conv}(\mathcal{V})$ with high sparsity satisfying $\|x - x^*\|_p \leqslant \epsilon$.

- Applications in game theory, combinatorial optimization, and machine learning

Theorem

Let $p \geqslant 2$. Then there exists $x \in \mathrm{conv}(\mathcal{V})$ with sparsity $\mathcal{O}(pD_p^2/\epsilon^2)$ satisfying $\|x - x^*\|_p \leqslant \epsilon$, where $D_p = \sup_{v,w \in \mathcal{V}} \|w - v\|_p$.

The approximate Carathéodory problem

Problem

Find $x \in \mathrm{conv}(\mathcal{V})$ with high sparsity satisfying $\|x - x^*\|_p \leqslant \epsilon$.

- Applications in game theory, combinatorial optimization, and machine learning

Theorem

Let $p \geqslant 2$. Then there exists $x \in \mathrm{conv}(\mathcal{V})$ with sparsity $\mathcal{O}(pD_p^2/\epsilon^2)$ satisfying $\|x - x^*\|_p \leqslant \epsilon$, where $D_p = \sup_{v,w \in \mathcal{V}} \|w - v\|_p$.

- This result is independent of the space dimension n

The approximate Carathéodory problem

Problem

Find $x \in \operatorname{conv}(\mathcal{V})$ with high sparsity satisfying $\|x - x^*\|_p \leqslant \epsilon$.

- Applications in game theory, combinatorial optimization, and machine learning

Theorem

Let $p \geqslant 2$. Then there exists $x \in \operatorname{conv}(\mathcal{V})$ with sparsity $\mathcal{O}(pD_p^2/\epsilon^2)$ satisfying $\|x - x^*\|_p \leqslant \epsilon$, where $D_p = \sup_{v,w \in \mathcal{V}} \|w - v\|_p$.

- This result is independent of the space dimension n
- The bound is tight [Mirrokni et al., 2017]

The approximate Carathéodory problem

Problem

Find $x \in \mathrm{conv}(\mathcal{V})$ with high sparsity satisfying $\|x - x^*\|_p \leqslant \epsilon$.

- Applications in game theory, combinatorial optimization, and machine learning

Theorem

Let $p \geqslant 2$. Then there exists $x \in \mathrm{conv}(\mathcal{V})$ with sparsity $\mathcal{O}(pD_p^2/\epsilon^2)$ satisfying $\|x - x^*\|_p \leqslant \epsilon$, where $D_p = \sup_{v,w \in \mathcal{V}} \|w - v\|_p$.

- This result is independent of the space dimension n
- The bound is tight [Mirrokni et al., 2017]
- Probabilistic proofs by Pisier [1981] and Barman [2015]

The approximate Carathéodory problem

Problem

Find $x \in \mathrm{conv}(\mathcal{V})$ with high sparsity satisfying $\|x - x^*\|_p \leqslant \epsilon$.

- Applications in game theory, combinatorial optimization, and machine learning

Theorem

Let $p \geqslant 2$. Then there exists $x \in \mathrm{conv}(\mathcal{V})$ with sparsity $\mathcal{O}(pD_p^2/\epsilon^2)$ satisfying $\|x - x^*\|_p \leqslant \epsilon$, where $D_p = \sup_{v,w \in \mathcal{V}} \|w - v\|_p$.

- This result is independent of the space dimension n
- The bound is tight [Mirrokni et al., 2017]
- Probabilistic proofs by Pisier [1981] and Barman [2015]
- Deterministic proof by Mirrokni et al. [2017] using mirror descent [Nemirovsky and Yudin, 1983] on the dual problem

The approximate Carathéodory problem

Problem

Find $x \in \mathrm{conv}(\mathcal{V})$ with high sparsity satisfying $\|x - x^*\|_p \leqslant \epsilon$.

- Applications in game theory, combinatorial optimization, and machine learning

Theorem

Let $p \geqslant 2$. Then there exists $x \in \mathrm{conv}(\mathcal{V})$ with sparsity $\mathcal{O}(pD_p^2/\epsilon^2)$ satisfying $\|x - x^*\|_p \leqslant \epsilon$, where $D_p = \sup_{v,w \in \mathcal{V}} \|w - v\|_p$.

- This result is independent of the space dimension n
- The bound is tight [Mirrokni et al., 2017]
- Probabilistic proofs by Pisier [1981] and Barman [2015]
- Deterministic proof by Mirrokni et al. [2017] using mirror descent [Nemirovsky and Yudin, 1983] on the dual problem
- Can we solve $\min\limits_{x \in \mathrm{conv}(\mathcal{V})} \|x - x^*\|_p$ by sequentially picking up vertices?

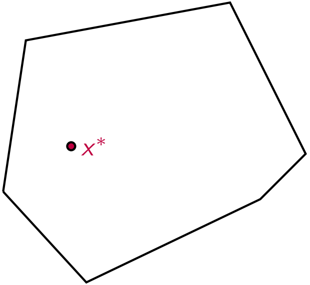
The Frank-Wolfe algorithm

Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

$x_0 \in \mathcal{V}$
for $t = 0$ **to** $T - 1$ **do**
    $v_t \leftarrow \underset{v \in \mathcal{V}}{\arg\min} \langle \nabla f(x_t), v \rangle$
    $x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
end for

$$f(x) = \|x - x^*\|_2^2$$



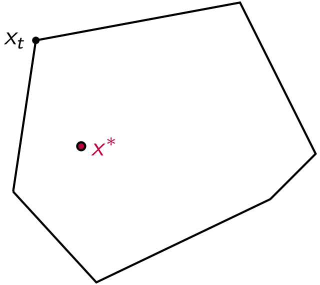
The Frank-Wolfe algorithm

Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \operatorname*{arg\,min}_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
5: **end for**

$$f(x) = \|x - x^*\|_2^2$$



The Frank-Wolfe algorithm

Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$

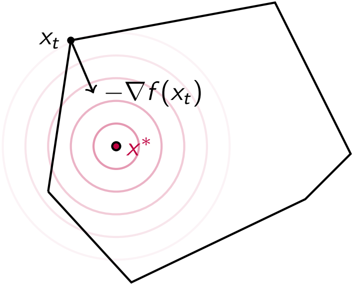
2: **for** $t = 0$ **to** $T - 1$ **do**

3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$

4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$

5: **end for**

$$f(x) = \|x - x^*\|_2^2$$



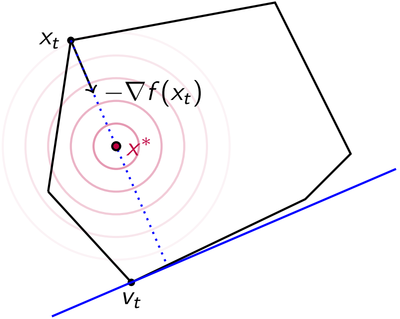
The Frank-Wolfe algorithm

Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \operatorname*{arg\,min}_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
5: **end for**

$f(x) = \|x - x^*\|_2^2$



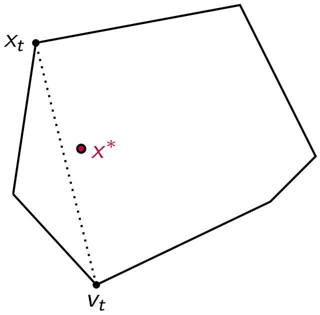
The Frank-Wolfe algorithm

Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
5: **end for**

$$f(x) = \|x - x^*\|_2^2$$



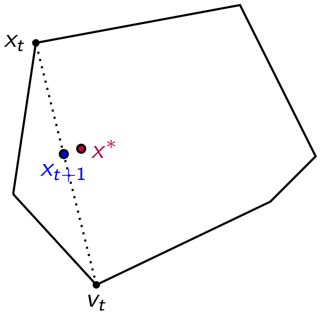
The Frank-Wolfe algorithm

Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
5: **end for**

$$f(x) = \|x - x^*\|_2^2$$



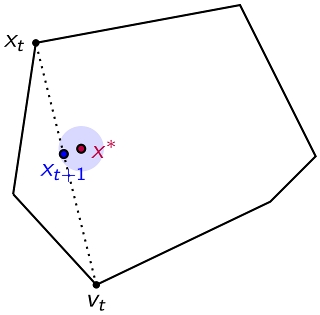
The Frank-Wolfe algorithm

Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \operatorname*{arg\,min}_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
5: **end for**

$$f(x) = \|x - x^*\|_2^2$$



The Frank-Wolfe algorithm

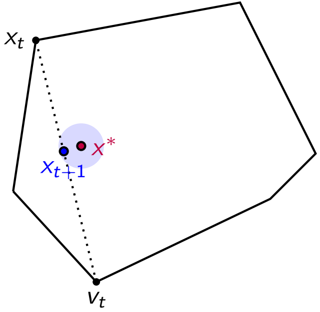
Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

$x_0 \in \mathcal{V}$
for $t = 0$ **to** $T-1$ **do**
    $v_t \leftarrow \underset{v \in \mathcal{V}}{\arg\min} \langle \nabla f(x_t), v \rangle$
    $x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
end for

- FW minimizes f over $\mathrm{conv}(\mathcal{V})$ by sequentially picking up vertices

$$f(x) = \|x - x^*\|_2^2$$



The Frank-Wolfe algorithm

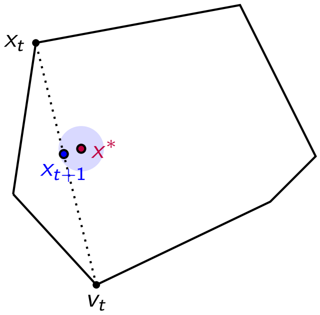
Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \underset{v \in \mathcal{V}}{\arg\min} \langle \nabla f(x_t), v \rangle$
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
5: **end for**

- FW minimizes f over $\text{conv}(\mathcal{V})$ by sequentially picking up vertices
- The final iterate x_T has sparsity at most $T + 1$

$$f(x) = \|x - x^*\|_2^2$$



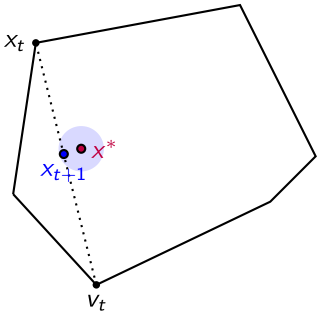
The Frank-Wolfe algorithm

Frank & Wolfe [1956], Levitin & Polyak [1966]

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
5: **end for**

$f(x) = \|x - x^*\|_2^2$



- FW minimizes f over $\mathrm{conv}(\mathcal{V})$ by sequentially picking up vertices
- The final iterate x_T has sparsity at most $T + 1$
- This is like a greedy method for the approximate Carathéodory problem!

The approximate Carathéodory problem via FW

- Apply FW to $f(x) = \|x - x^*\|_p^2$ and count the number of iterations T to achieve ϵ^2-convergence: $\|x_T - x^*\|_p^2 \leqslant \epsilon^2$

The approximate Carathéodory problem via FW

- Apply FW to $f(x) = \|x - x^*\|_p^2$ and count the number of iterations T to achieve ϵ^2-convergence: $\|x_T - x^*\|_p^2 \leqslant \epsilon^2$
- Then x_T has sparsity at most $T + 1$ and satisfies $\|x_T - x^*\|_p \leqslant \epsilon$

The approximate Carathéodory problem via FW

- Apply FW to $f(x) = \|x - x^*\|_p^2$ and count the number of iterations T to achieve ϵ^2-convergence: $\|x_T - x^*\|_p^2 \leqslant \epsilon^2$
- Then x_T has sparsity at most $T + 1$ and satisfies $\|x_T - x^*\|_p \leqslant \epsilon$
- Classical convergence results of FW and variants require smoothness and/or strong convexity of f w.r.t. a norm $\|\cdot\|$

The approximate Carathéodory problem via FW

- Apply FW to $f(x) = \|x - x^*\|_p^2$ and count the number of iterations T to achieve ϵ^2-convergence: $\|x_T - x^*\|_p^2 \leqslant \epsilon^2$
- Then x_T has sparsity at most $T + 1$ and satisfies $\|x_T - x^*\|_p \leqslant \epsilon$
- Classical convergence results of FW and variants require smoothness and/or strong convexity of f w.r.t. a norm $\|\cdot\|$

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

The approximate Carathéodory problem via FW

- Apply FW to $f(x) = \|x - x^*\|_p^2$ and count the number of iterations T to achieve ϵ^2-convergence: $\|x_T - x^*\|_p^2 \leqslant \epsilon^2$
- Then x_T has sparsity at most $T + 1$ and satisfies $\|x_T - x^*\|_p \leqslant \epsilon$
- Classical convergence results of FW and variants require smoothness and/or strong convexity of f w.r.t. a norm $\|\cdot\|$

$$\frac{S}{2}\|y - x\|^2 \leqslant f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant \frac{L}{2}\|y - x\|^2$$

The approximate Carathéodory problem via FW

- Apply FW to $f(x) = \|x - x^*\|_p^2$ and count the number of iterations T to achieve ϵ^2-convergence: $\|x_T - x^*\|_p^2 \leqslant \epsilon^2$
- Then x_T has sparsity at most $T + 1$ and satisfies $\|x_T - x^*\|_p \leqslant \epsilon$
- Classical convergence results of FW and variants require smoothness and/or strong convexity of f w.r.t. a norm $\|\cdot\|$

$$\frac{S}{2}\|y - x\|^2 \leqslant f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant \frac{L}{2}\|y - x\|^2$$

- For $p \geqslant 2$, $f(x) = \|x - x^*\|_p^2$ is smooth w.r.t. $\|\cdot\|_p$ but it is strongly convex only when $p = 2$

The approximate Carathéodory problem via FW

- Apply FW to $f(x) = \|x - x^*\|_p^2$ and count the number of iterations T to achieve ϵ^2-convergence: $\|x_T - x^*\|_p^2 \leqslant \epsilon^2$
- Then x_T has sparsity at most $T + 1$ and satisfies $\|x_T - x^*\|_p \leqslant \epsilon$
- Classical convergence results of FW and variants require smoothness and/or strong convexity of f w.r.t. a norm $\|\cdot\|$

$$\frac{S}{2}\|y - x\|^2 \leqslant f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant \frac{L}{2}\|y - x\|^2$$

- For $p \geqslant 2$, $f(x) = \|x - x^*\|_p^2$ is smooth w.r.t. $\|\cdot\|_p$ but it is strongly convex only when $p = 2$
- Replace strong convexity with a weaker condition: the PL inequality w.r.t. $\|\cdot\|$ [Polyak, 1963, Łojasiewicz, 1963]

$$f(x) - \min_{\mathbb{R}^n} f \leqslant \frac{1}{2\mu}\|\nabla f(x)\|_*^2$$

The approximate Carathéodory problem via FW

- Apply FW to $f(x) = \|x - x^*\|_p^2$ and count the number of iterations T to achieve ϵ^2-convergence: $\|x_T - x^*\|_p^2 \leqslant \epsilon^2$
- Then x_T has sparsity at most $T + 1$ and satisfies $\|x_T - x^*\|_p \leqslant \epsilon$
- Classical convergence results of FW and variants require smoothness and/or strong convexity of f w.r.t. a norm $\|\cdot\|$

$$\frac{S}{2}\|y - x\|^2 \leqslant f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leqslant \frac{L}{2}\|y - x\|^2$$

- For $p \geqslant 2$, $f(x) = \|x - x^*\|_p^2$ is smooth w.r.t. $\|\cdot\|_p$ but it is strongly convex only when $p = 2$
- Replace strong convexity with a weaker condition: the PL inequality w.r.t. $\|\cdot\|$ [Polyak, 1963, Łojasiewicz, 1963]

$$f(x) - \min_{\mathbb{R}^n} f \leqslant \frac{1}{2\mu}\|\nabla f(x)\|_*^2$$

- For $p \geqslant 2$, $f(x) = \|x - x^*\|_p^2$ is $2(p-1)$-smooth and 2-PL w.r.t. $\|\cdot\|_p$

Sparsity bounds via FW convergence rates

Levitin & Polyak [1966], Guélat & Marcotte [1986], Jaggi [2013], Garber & Hazan [2015]

- $p \geqslant 2$
- $\mathcal{C} \subset \mathbb{R}^n$ be a compact convex set
- $\mathcal{V} \subseteq \partial\mathcal{C}$ be the compact set of interest (e.g., $\mathcal{C} = \mathrm{conv}(\mathcal{V})$)
- We want a sparse approximate convex decomposition of $x^* \in \mathrm{conv}(\mathcal{V})$

Sparsity bounds via FW convergence rates

Levitin & Polyak [1966], Guélat & Marcotte [1986], Jaggi [2013], Garber & Hazan [2015]

- $p \geqslant 2$
- $\mathcal{C} \subset \mathbb{R}^n$ be a compact convex set
- $\mathcal{V} \subseteq \partial\mathcal{C}$ be the compact set of interest (e.g., $\mathcal{C} = \mathrm{conv}(\mathcal{V})$)
- We want a sparse approximate convex decomposition of $x^* \in \mathrm{conv}(\mathcal{V})$

Assumptions	**FW rate**	**Sparsity bound**
–	$\dfrac{4(p-1)D_p^2}{t+2}$	$\dfrac{4(p-1)D_p^2}{\epsilon^2} = \mathcal{O}\left(\dfrac{pD_p^2}{\epsilon^2}\right)$
$\mathcal{C}$ is S_p-strongly convex	$\dfrac{\max\{9(p-1)D_p^2, 1152(p-1)^2/S_p^2\}}{(t+2)^2}$	$\mathcal{O}\left(\dfrac{\sqrt{p}D_p + p/S_p}{\epsilon}\right)$
$x^* \in \mathrm{relint}_p(\mathcal{C})$ with radius r_p	$\left(1 - \dfrac{1}{p-1}\dfrac{r_p^2}{D_p^2}\right)^t \epsilon_0$	$\mathcal{O}\left(\dfrac{pD_p^2}{r_p^2}\ln\left(\dfrac{1}{\epsilon}\right)\right)$

Sparsity bounds via FW convergence rates

Levitin & Polyak [1966], Guélat & Marcotte [1986], Jaggi [2013], Garber & Hazan [2015]

- $p \geqslant 2$
- $\mathcal{C} \subset \mathbb{R}^n$ be a compact convex set
- $\mathcal{V} \subseteq \partial\mathcal{C}$ be the compact set of interest (e.g., $\mathcal{C} = \mathrm{conv}(\mathcal{V})$)
- We want a sparse approximate convex decomposition of $x^* \in \mathrm{conv}(\mathcal{V})$

Assumptions	**FW rate**	**Sparsity bound**
–	$\dfrac{4(p-1)D_p^2}{t+2}$	$\dfrac{4(p-1)D_p^2}{\epsilon^2} = \mathcal{O}\left(\dfrac{pD_p^2}{\epsilon^2}\right)$
$\mathcal{C}$ is S_p-strongly convex	$\dfrac{\max\{9(p-1)D_p^2, 1152(p-1)^2/S_p^2\}}{(t+2)^2}$	$\mathcal{O}\left(\dfrac{\sqrt{p}D_p + p/S_p}{\epsilon}\right)$
$x^* \in \mathrm{relint}_p(\mathcal{C})$ with radius r_p	$\left(1 - \dfrac{1}{p-1}\dfrac{r_p^2}{D_p^2}\right)^t \epsilon_0$	$\mathcal{O}\left(\dfrac{pD_p^2}{r_p^2}\ln\left(\dfrac{1}{\epsilon}\right)\right)$

- FW adapts to the geometry of the problem to yield higher sparsity

Fully-Corrective Frank-Wolfe algorithm

Algorithm Frank-Wolfe (FW)

$x_0 \in \mathcal{V}$

for $t = 0$ **to** $T - 1$ **do**

  $v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$

  $x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$

end for

Fully-Corrective Frank-Wolfe algorithm

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
5: **end for**

- Selected vertices in FW may be redundant, can we fix this?

Fully-Corrective Frank-Wolfe algorithm

Algorithm Frank-Wolfe (FW)

$x_0 \in \mathcal{V}$
for $t = 0$ **to** $T - 1$ **do**
  $v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
  $x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
end for

- Selected vertices in FW may be redundant, can we fix this?

Algorithm Fully-Corrective Frank-Wolfe (FCFW)

$x_0 \in \mathcal{V}$
$\mathcal{S}_0 \leftarrow \{x_0\}$
for $t = 0$ **to** $T - 1$ **do**
  $v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
  $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
  $x_{t+1} \leftarrow \arg\min_{\mathrm{conv}(\mathcal{S}_{t+1})} f$
end for

Fully-Corrective Frank-Wolfe algorithm

Algorithm Frank-Wolfe (FW)

1: $x_0 \in \mathcal{V}$
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
4: $\quad x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
5: **end for**

- Selected vertices in FW may be redundant, can we fix this?

Algorithm Fully-Corrective Frank-Wolfe (FCFW)

1: $x_0 \in \mathcal{V}$
2: $\mathcal{S}_0 \leftarrow \{x_0\}$
3: **for** $t = 0$ **to** $T - 1$ **do**
4: $\quad v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
5: $\quad \mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
6: $\quad x_{t+1} \leftarrow \arg\min_{\mathrm{conv}(\mathcal{S}_{t+1})} f$
7: **end for**

Fully-Corrective Frank-Wolfe algorithm

Algorithm Frank-Wolfe (FW)

$x_0 \in \mathcal{V}$
for $t = 0$ **to** $T - 1$ **do**
  $v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
  $x_{t+1} \leftarrow x_t + \gamma_t (v_t - x_t)$
end for

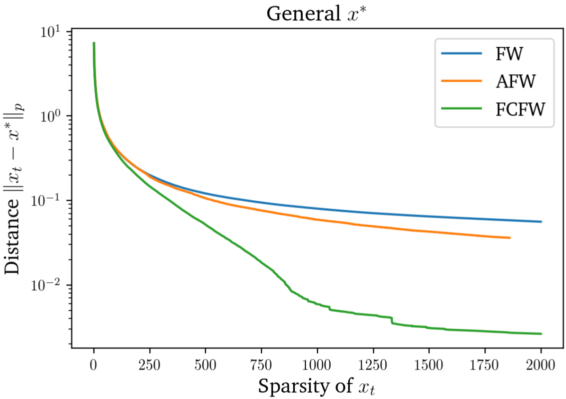
- Selected vertices in FW may be redundant, can we fix this?

Algorithm Fully-Corrective Frank-Wolfe (FCFW)

$x_0 \in \mathcal{V}$
$\mathcal{S}_0 \leftarrow \{x_0\}$
for $t = 0$ **to** $T - 1$ **do**
  $v_t \leftarrow \arg\min_{v \in \mathcal{V}} \langle \nabla f(x_t), v \rangle$
  $\mathcal{S}_{t+1} \leftarrow \mathcal{S}_t \cup \{v_t\}$
  $x_{t+1} \leftarrow \arg\min_{\mathrm{conv}(\mathcal{S}_{t+1})} f$
end for

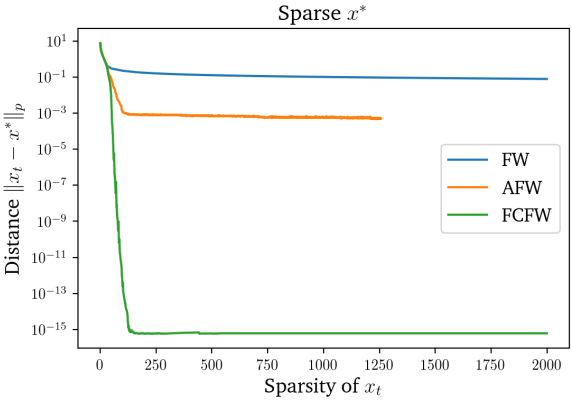
Fully-Corrective Frank-Wolfe algorithm

- We randomly generated 1000 vertices and $x^* \in \mathrm{conv}(\mathcal{V})$
- Here arbitrarily chose $p = 4$



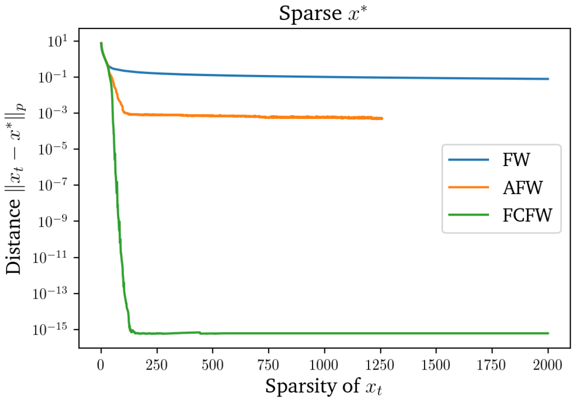
Fully-Corrective Frank-Wolfe algorithm

- Here x^* is generated by a convex combination of only 50 vertices



Fully-Corrective Frank-Wolfe algorithm

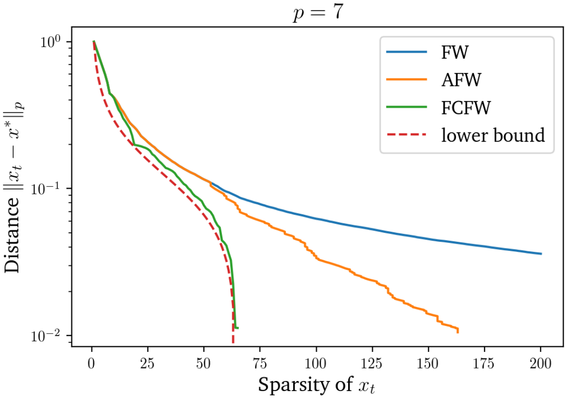
- Here x^* is generated by a convex combination of only 50 vertices



- FCFW obtains an *exact* convex decomposition of x^* once it picks up all its vertices

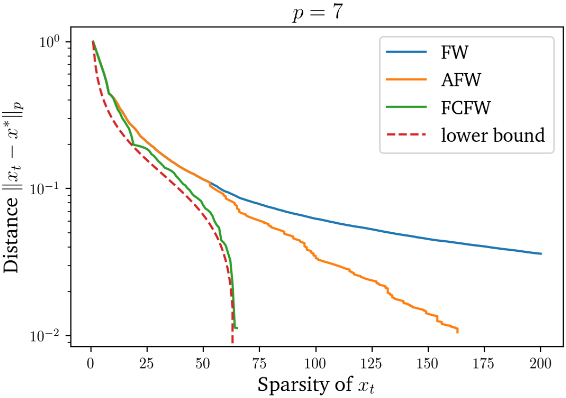
Fully-Corrective Frank-Wolfe algorithm

- FCFW matches the theoretical lower bound!



Fully-Corrective Frank-Wolfe algorithm

- FCFW matches the theoretical lower bound!



- Can we derive a precise convergence rate for FCFW?

Thank you!

https://arxiv.org/pdf/1911.04415.pdf

References

S. Barman. Approximating Nash equilibria and dense bipartite subgraphs via an approximate version of Carathéodory's theorem. *STOC*, 2015.

C. Carathéodory. Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen. *Mathematische Annalen*, 1907.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 1956.

D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. *ICML*, 2015.

E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 1966.

S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Colloques Internationaux du CNRS*, 1963.

V. Mirrokni, R. Paes Leme, A. Vladu, and S. C.-W. Wong. Tight bounds for approximate Carathéodory and beyond. *ICML*, 2017.

A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, 1983.

G. Pisier. Remarques sur un résultat non publié de B. Maurey. *Séminaire d'Analyse Fonctionnelle*, École Polytechnique, 1981.

B. T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 1963.