Limited memory Kelley's Method Converges for Composite Convex and Submodular Objectives

Madeleine Udell

Operations Research and Information Engineering
Cornell University

Song Zhou (Cornell), Swati Gupta (Georgia Tech)

NeurIPS, December 2018

Problem to solve

$$\text{minimize} \quad g(x) + f(x)$$

- $g : \mathbb{R}^n \to \mathbb{R}$ **strongly convex**
- $f : \mathbb{R}^n \to \mathbb{R}$ Lovász extension of submodular function F
  - piecewise linear
  - convex envelope of F
  - generically, exponentially many linear pieces

L-KM solves composite convex + submodular problems whose natural size is **exponential** with **linear memory**.

Submodular optimization background

- **Ground set** $V = \{1, \ n\}$.
- $F : 2^V \to \mathbb{R}$ is **submodular** if for all $A, \ B \subseteq V$,

$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B)$$

- the **base polytope** of F is

$$B(F) = \{w \in \mathbb{R}^n : w(V) = F(V), w(A) \leq F(A), \ \forall A \subseteq V\}$$

- the **Lovász extension** of F is the homogeneous piecewise linear convex function

$$f(x) = \max_{w \in B(F)} w^\top x$$

- linear optimization over $B(F)$ is easy
- $\implies$ evaluating $f(x)$ and $\partial f(x)$ is easy

Original Simplicial Method (OSM) [Bach 2013]

Intuition:

- approximate f with pwl function whose values and (sub)gradients match f at all previous iterates
- minimize approximation to determine the next iterate

Advantages: Finite convergence [Bach 2013]

Drawbacks:

- *Memory.* memory $|\mathcal{V}^{(i)}| = i$ grows with iteration counter i
- *Computation.* subproblem size grows with memory
- *Convergence rate.* no known rate of convergence [Bach 2013]

Limited Memory Kelley's Method ($\mathrm{L\text{-}KM}$)

Algorithm 1 $\mathrm{L\text{-}KM}$ (to minimize $g(x) + f(x)$)

initialize $\mathcal{V} \neq \emptyset$ affinely independent. repeat

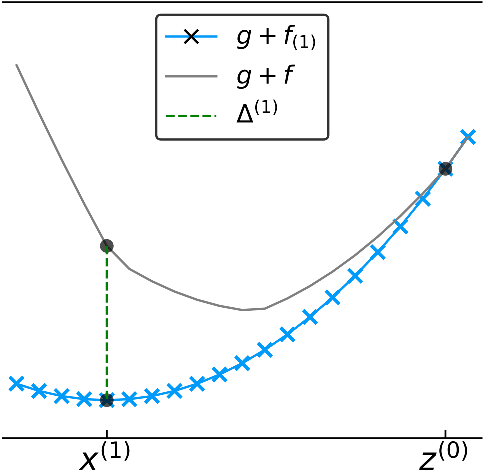
1. define $\hat{f}(x) = \max_{w \in \mathcal{V}} w^\top x$
2. solve subproblem

$$\hat{x} \leftarrow \operatorname{argmin} g(x) + \hat{f}(x)$$

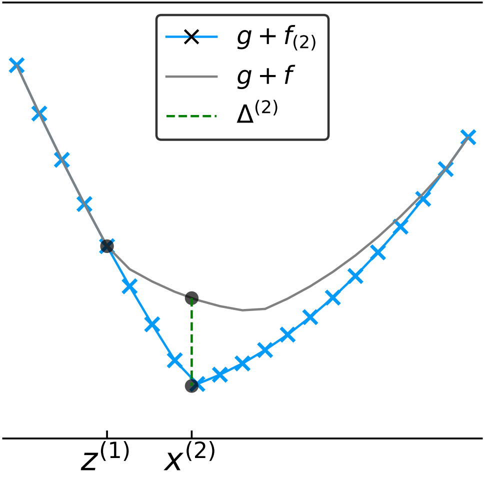
3. compute $v \in \partial f(\hat{x}) = \operatorname{argmax}_{w \in B(F)} \hat{x}^\top w$
4. $\mathcal{V} \leftarrow \{w \in \mathcal{V} : w^\top x = f(\hat{x})\} \cup v$

unlike OSM, $\mathrm{L\text{-}KM}$ drops subgradients $w \in \mathcal{V}$ that are not tight at current iterate

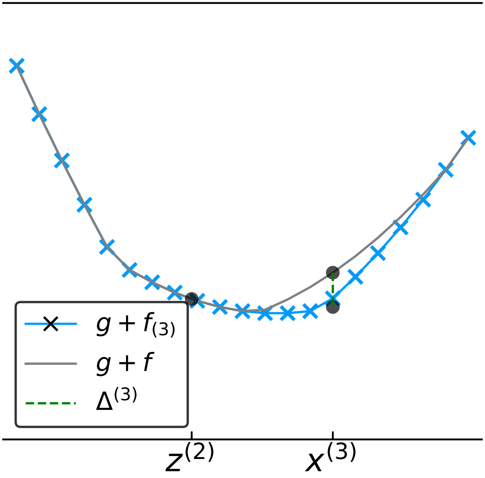
L-KM: example



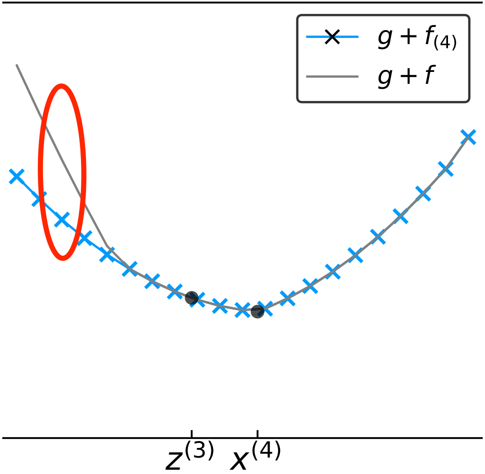
L-KM: example



L-KM: example



L-KM: example



Properties of $\mathrm{L\text{-}KM}$

- **Limited memory:** In $\mathrm{L\text{-}KM}$, for all $i \geq 0$, vectors in $\mathcal{V}^{(i)}$ are affinely independent. Moreover, $|\mathcal{V}^{(i)}| \leq n + 1$.
- **Finite convergence:** When g is strongly convex, $\mathrm{L\text{-}KM}$ converges finitely.
- **Linear convergence:** When g is smooth and strongly convex, the duality gap of $\mathrm{L\text{-}KM}$ and OSM converges linearly to 0.

Limited-memory Fully Corrective Frank Wolfe
L-FCFW

Algorithm 2 L-FCFW (to minimize $-g^*(-y)$ over $y \in B(F)$)

initialize $\mathcal{V} \neq \emptyset$ affinely independent. repeat

1. solve subproblem

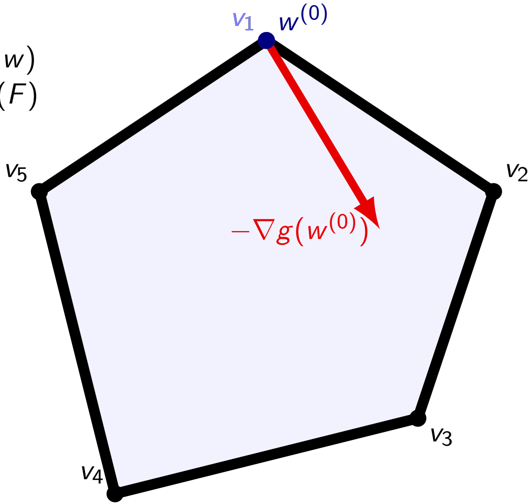
$$\begin{array}{ll} \text{minimize} & -g^*(-y) \\ \text{subject to} & y \in \mathbf{conv}(\mathcal{V}) \end{array}$$

   do convex decomposition of the solution $\hat{y} = \sum_{w \in \mathcal{V}} \lambda_w w$ with $\lambda_w \geq 0$ and $\sum_{w \in \mathcal{V}} \lambda_w = 1$

2. compute gradient $\hat{x} = \nabla(-g^*(-\hat{y}))$
3. solve linear optimization $v = \mathrm{argmax}_{w \in B(F)} \hat{x}^\top w$
4. $\mathcal{V} \leftarrow \{w \in \mathcal{V} : \lambda_w > 0\} \cup v$

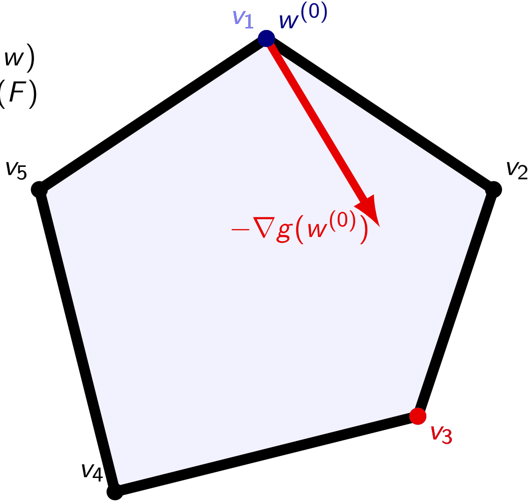
Fully corrective Frank-Wolfe

minimize $\quad -g^*(-w)$
subject to $\quad w \in B(F)$



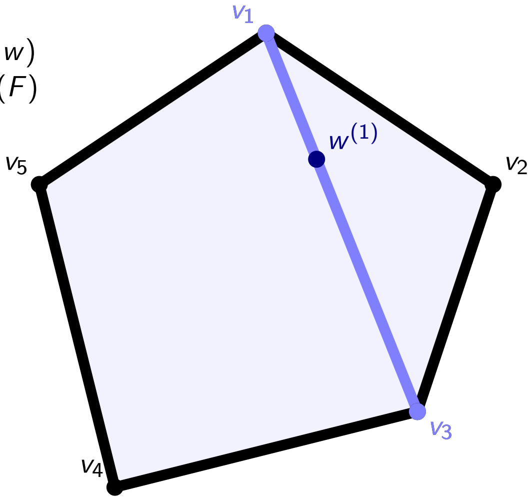
Fully corrective Frank-Wolfe

$$\begin{array}{ll} \text{minimize} & -g^*(-w) \\ \text{subject to} & w \in B(F) \end{array}$$



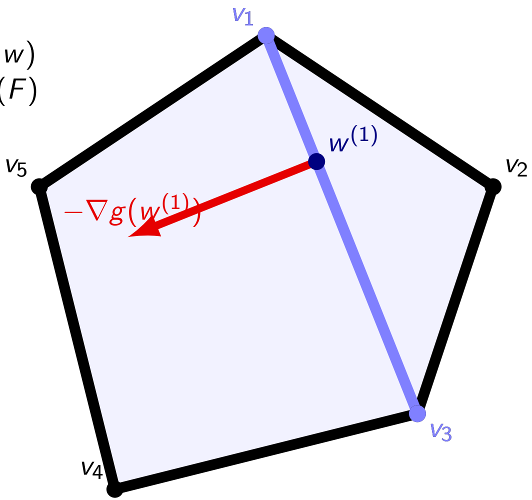
Fully corrective Frank-Wolfe

minimize $\quad -g^*(-w)$
subject to $\quad w \in B(F)$



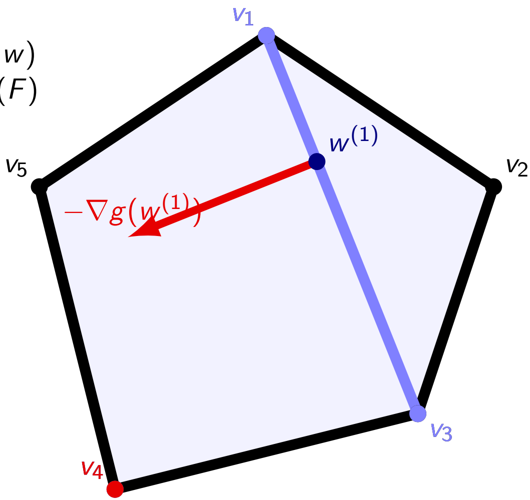
Fully corrective Frank-Wolfe

$$\begin{array}{ll} \text{minimize} & -g^*(-w) \\ \text{subject to} & w \in B(F) \end{array}$$



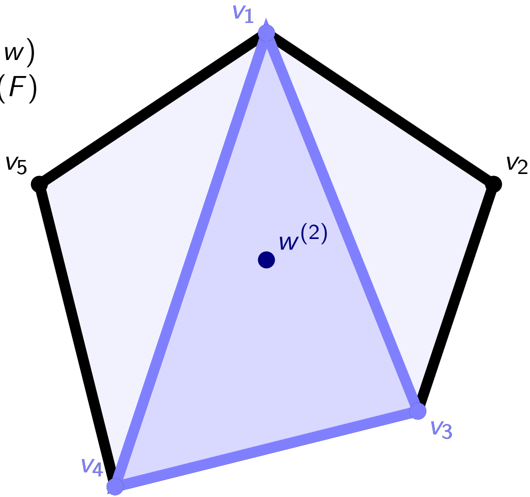
Fully corrective Frank-Wolfe

$$\begin{array}{ll} \text{minimize} & -g^*(-w) \\ \text{subject to} & w \in B(F) \end{array}$$



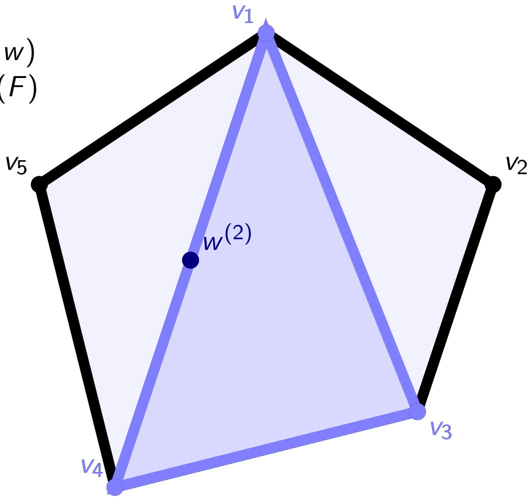
Fully corrective Frank-Wolfe

$$\begin{array}{ll} \text{minimize} & -g^*(-w) \\ \text{subject to} & w \in B(F) \end{array}$$



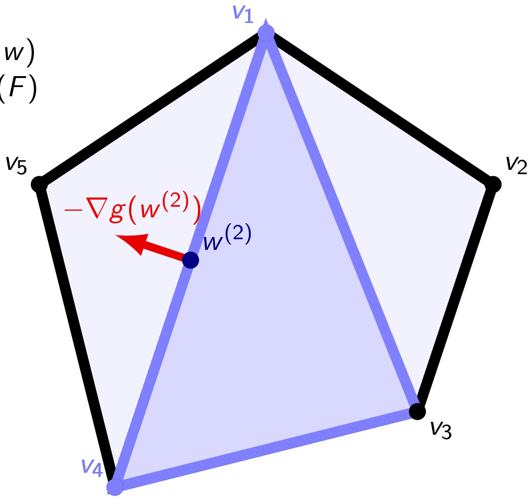
Fully corrective Frank-Wolfe

$$\begin{array}{ll} \text{minimize} & -g^*(-w) \\ \text{subject to} & w \in B(F) \end{array}$$



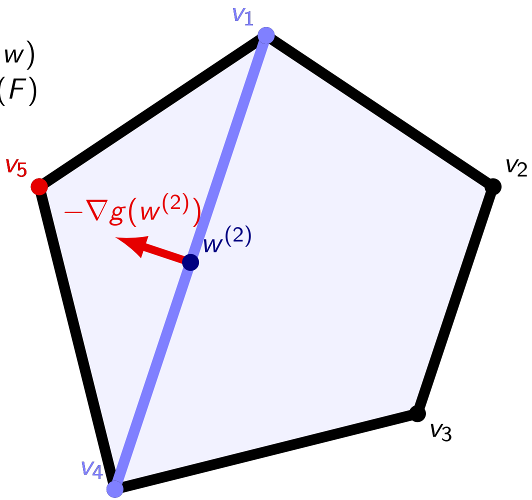
Fully corrective Frank-Wolfe

$$\begin{array}{ll} \text{minimize} & -g^*(-w) \\ \text{subject to} & w \in B(F) \end{array}$$



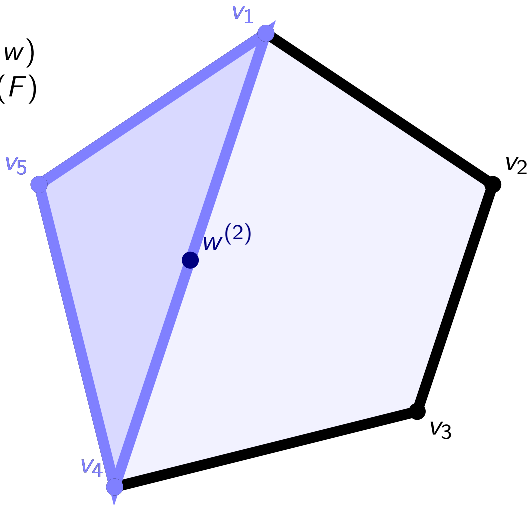
Fully corrective Frank-Wolfe

$$\begin{array}{ll} \text{minimize} & -g^*(-w) \\ \text{subject to} & w \in B(F) \end{array}$$



Fully corrective Frank-Wolfe

minimize $\quad -g^*(-w)$
subject to $\quad w \in B(F)$



Properties of L-FCFW

- **Limited memory**: By Carathéodory's theorem, we can choose $\leq n+1$ active vertices to represent the current iterate.
- **Linear Convergence** [Lacoste-Julien and Jaggi, 2015]: When g is smooth and strongly convex, the duality gap of L-FCFW converges linearly to 0.
- **Duality**: Two algorithms are dual if their iterates solve dual subproblems. If g is smooth and strongly convex and
  - $\mathcal{B}^{(i)} = \{w \in \mathcal{V}^{(i-1)} : \lambda_w > 0\}$, L-FCFW is dual to L-KM.
  - $\mathcal{B}^{(i)} = \mathcal{V}^{(i-1)}$, L-FCFW is dual to OSM.

Summary

$\mathrm{L\text{-}KM}$ solves composite convex + submodular problems whose natural size is **exponential** with **linear memory**.

- S. Zhou, S. Gupta, and M. Udell. Limited Memory Kelley's Method Converges for Composite Convex and Submodular Objectives. NIPS 2018.
- 5–7pm Room 210 Poster #16