kFW: A Frank-Wolfe style algorithm with stronger subproblem oracles

Lijun Ding, Jicong Fan, and Madeleine Udell

Cornell University

Problem setup

Consider optimization problem with the decision x:

$$\begin{aligned} &\text{minimize} \quad f(x) := g(\mathcal{A}x) + \langle c, x\rangle \\ &\text{subject to } x \in \Omega. \end{aligned} \tag{1}$$

- Ω convex and compact with diameter D
- g smooth
- $\mathcal{A}$ a linear map, and c a vector

Applications: LASSO, SVM, matrix completion, phase retrieval, and one-bit matrix completion, etc.

Frank-Wolfe

FW: choose $x_0 \in \Omega$, iterate

1. *Linear Optimization Oracle (LOO):* Find a direction v_t that solves $\min_v \langle \nabla f(x_t), v\rangle$.
2. *Line Search:* Find x_{t+1} that solves $\min_{x=\eta v_t + (1-\eta)x_t, \eta\in[0,1]} f(x)$.

Slow convergence of FW and Zigzag

- FW: slow in both theory and practice, $\mathcal{O}(\frac{1}{t})$ convergence rate.
- Zigzag: cause of slow convergence when when the optimal solution $x_\star \in \partial\Omega$ and is a convex combination of $r_\star$ many extreme points $v_1^\star, \dots, v_{r_\star}^\star \in \Omega$. See Figure 1 for $r_\star = 2$ The grey arrows are the negative gradients $-\nabla f$.

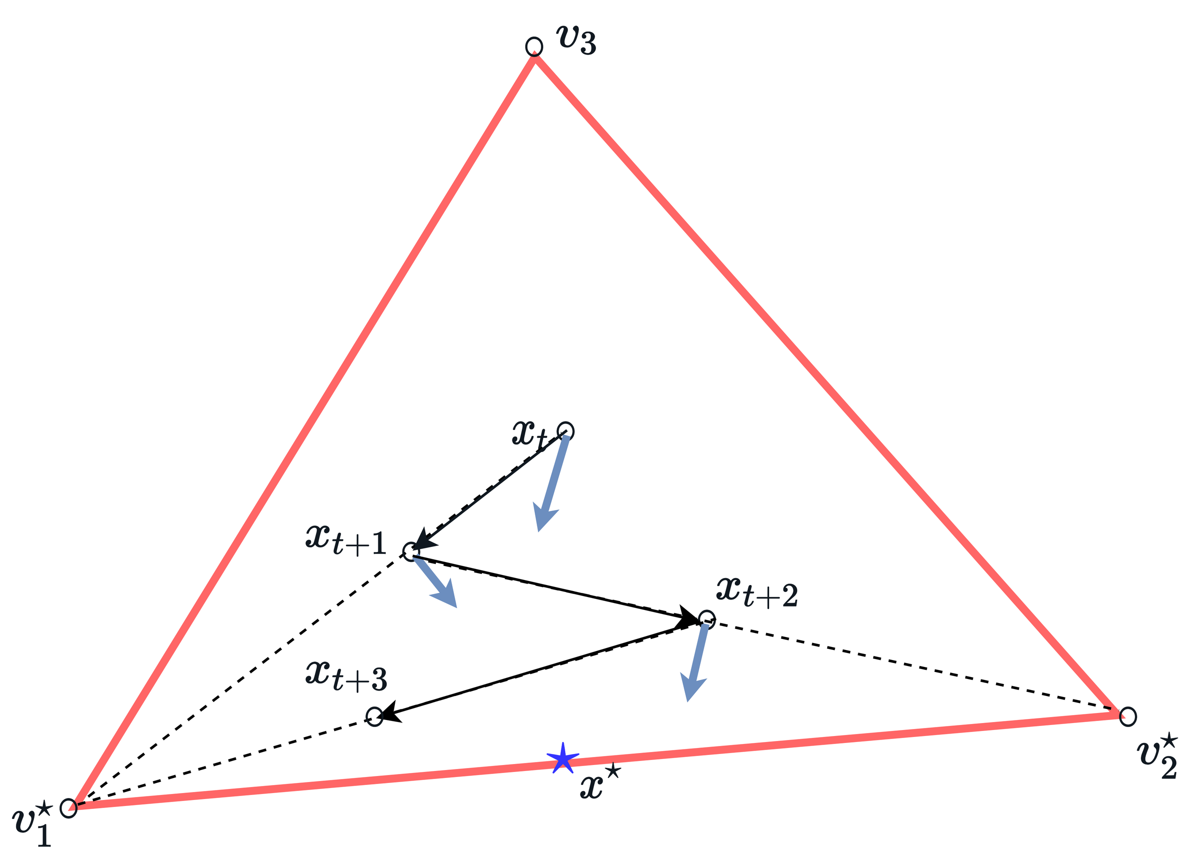

Fig. 1: Zig-Zag: black arrows show trajectory of the iterates. Optimal solution $x_\star$ is a convex combination of $v_1^\star$ and $v_2^\star$, and $r_\star = 2$. The grey arrows are the negative gradients $-\nabla f$.

Our key insight

The sparstity $r_\star$ is small in many applications and $\nabla f(x_\star)$ has the smallest inner product with $v_1^\star, \dots, v_{r_\star}^\star$ among all $v \in \Omega$. Our key insight:

- Compute all extreme points $v_i^\star$ that minimize $\langle \nabla f(x_\star), v\rangle$;
- Solve the smaller problem $\min_{x\in\mathbf{conv}(x_t, v_1^\star, \dots, v_{r_\star}^\star)} f(x)$.

See Figure 2 for an illustration.

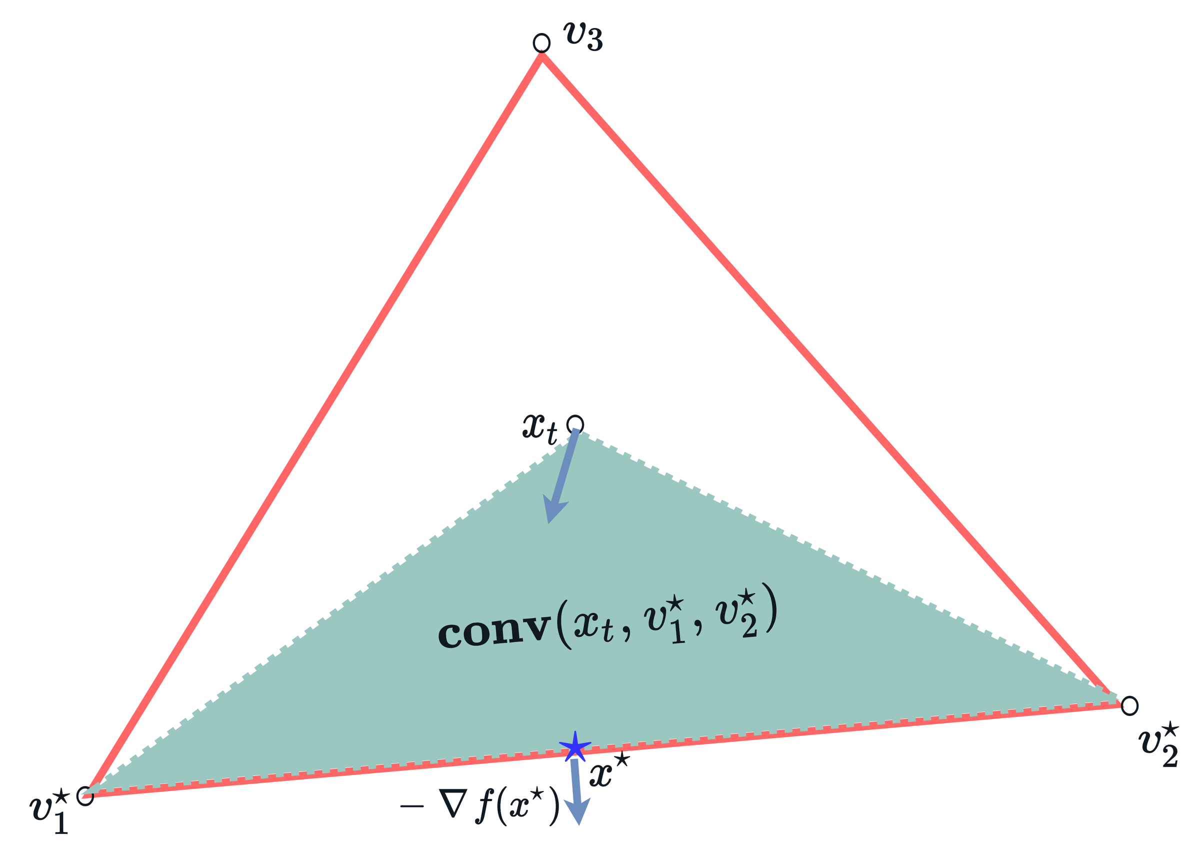

Fig. 2: Optimization over $\mathbf{conv}(x_t, v_1^\star, v_2^\star)$ (green).

k-FW

Inspired by our key insight, we introduce the following two sub-problem oracles for polytope:

- k linear optimization oracle (kLOO): for any $y \in \mathbb{R}^n$, compute the k extreme points $v_1, \dots, v_k$ (k best directions) with the smallest k inner products $\langle v, y\rangle$ among all extreme points v of Ω.
- k direction search (kDS): given input directions $w, v_1, \dots, v_k \in \Omega$, output $x_{k\text{DS}} = \arg\min_{x\in\mathbf{conv}(w,v_1,\dots,v_k)} f(x)$.

kFW simply iterates kL00 and kDS.

- Many polytopes admit efficient kLOO and kDS: probability simplex, flow polytope for directed acyclic graph, matching polytope, matroid, spanning tree polytope, etc.
- kLOO and kDS for nonpolytope is also available! Example includes group norm ball, spetrahedron, and nuclear norm ball.

Theoretical Result

Analytical Conditions

- Sparsity measure $r_\star$: number of extreme points of the smallest face $\mathcal{F}(x_\star)$ containing $x_\star$.
- Strict complementarity (SC) and its measure δ: a unique solution $x_\star \in \partial\Omega$ and $-\nabla f(x_\star) \in$ relint$(N_\Omega(x_\star))$ $N_\Omega(x_\star)$ normal cone). The SC measure is $\delta = \min\{\langle \nabla f(x_\star, v - x_\star\rangle \mid v \notin \mathcal{F}(x_\star), v$ extreme point$\}$.
- γ - quadratic growth (QG): for all $x \in \Omega$, $f(x) - f(x_\star) \geq \gamma\|x - x_\star\|^2$.

Theorem Statement Suppose f is L_f-smooth and convex and Ω is convex compact with diameter D.

- Then for any $k \geq 1$ and for all $t \geq 1$, the iterate x_t in kFW satisfies $f(x_t) - f(x_\star) \leq \frac{2L_f D^2}{t}$.
- Moreover, suppose Problem (1) satisfies strict complementarity and quadratic growth, and $k \geq r_\star$. If the constraint set Ω is a polytope or a unit group norm ball, then the gap $\delta > 0$ and kFW finds $x_\star$ in at most $T+1$ iterations, where T is
$$T = \frac{4L_f^3 D^4}{\gamma\delta^2}. \tag{2}$$
- If the constraint set is the spechedron or the unit nuclear norm ball, the gap $\delta > 0$ and kFW satisfies that for any $t \geq T_1 := \frac{72L_f^3}{\gamma\delta^2}$, $f(X_{t+1}) - f(X_\star) \leq \left(1 - \min\left\{\frac{\gamma}{4L_f}, \frac{\delta}{12L_f}\right\}\right)(f(X_t) - f(X_\star))$.

Numerics

We compare our method kFW with FW, away-step FW (awayFW) , pairwise FW (pairFW), DICG [Garber and Meshi 2016], and blockFW [Allen-Zhu et. al. 2017] for the Lasso, support vector machine (SVM), group Lasso, and matrix completion problems on synthetic data. All algorithms terminate when the relative change of the objective is less than 10^{-6} or after 1000 iterations. As shown in Figure 1, kFW converges in many fewer iterations than other methods. Table 1 shows that kFW also converges faster in wall-clock time, with one exception (blockFW in matrix completion).

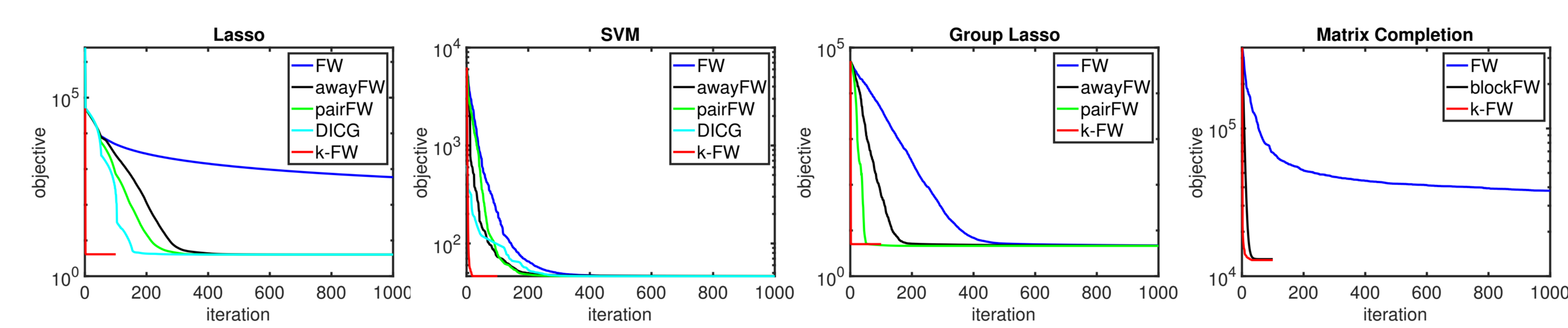

Fig. 3: kFW vs. FW and its variants

Table 1: Computation time (seconds): Sign " -" means the algorithm is not suited to the problem.

	FW	awayFW	pairFW	DICG	blockFW	kFW
Lasso	>14	7	6	10	-	**0.5**
SVM	6	4.5	2.9	2.5	-	**0.6**
Group Lasso	17	6	1.8	-	-	**0.3**
Matrix completion	>180	-	-	-	1.8	4.8