On the Global Linear Convergence of Frank-Wolfe Optimization Variants

Simon Lacoste-Julien

INRIA / École Normale Supérieure, Paris, France


Martin Jaggi

Dept. of Computer Science, ETH Zürich, Switzerland

Summary

- **Frank-Wolfe (FW)** is a popular algorithm for constrained optimization over **combinatorial** objects, but its rate is **sublinear**
- We show the **global linear convergence** for the first time of several variants of FW:
  away-steps FW (AFW), pairwise FW (PFW), fully-corrective FW (FCFW) and Wolfe's min-norm point algorithm (MNP)
- The constant in the rate introduces the `**condition number**' of the constraint set:

$$\frac{1}{\rho_f} \leq \underbrace{\frac{L}{\mu}}_{\text{condition number of } f} \cdot \underbrace{\left(\frac{\text{diam}(\mathcal{M})}{\text{pyr.width}(\mathcal{M})}\right)^2}_{\text{condition number of } \mathcal{M}}$$

Problem Setup

We want to optimize over: $\mathcal{M} = \text{conv}(\mathcal{A})$

$$\min_{\boldsymbol{x}\in\mathcal{M}} f(\boldsymbol{x}) \qquad \mathcal{A} \subseteq \mathbb{R}^d \text{ is a } \textit{finite} \text{ set of } \textit{atoms}$$

with only access to a *linear minimization oracle*:

$$\text{LMO}_{\mathcal{A}}(\boldsymbol{r}) \in \arg\min_{\boldsymbol{x}\in\mathcal{A}} \langle \boldsymbol{r}, \boldsymbol{x}\rangle$$

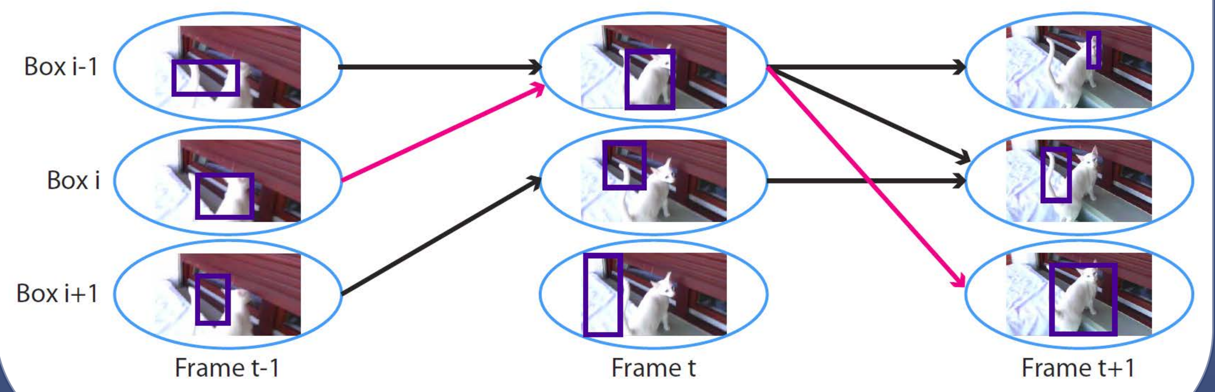
assume f is μ-strongly convex

and ∇f is L-Lipschitz continuous

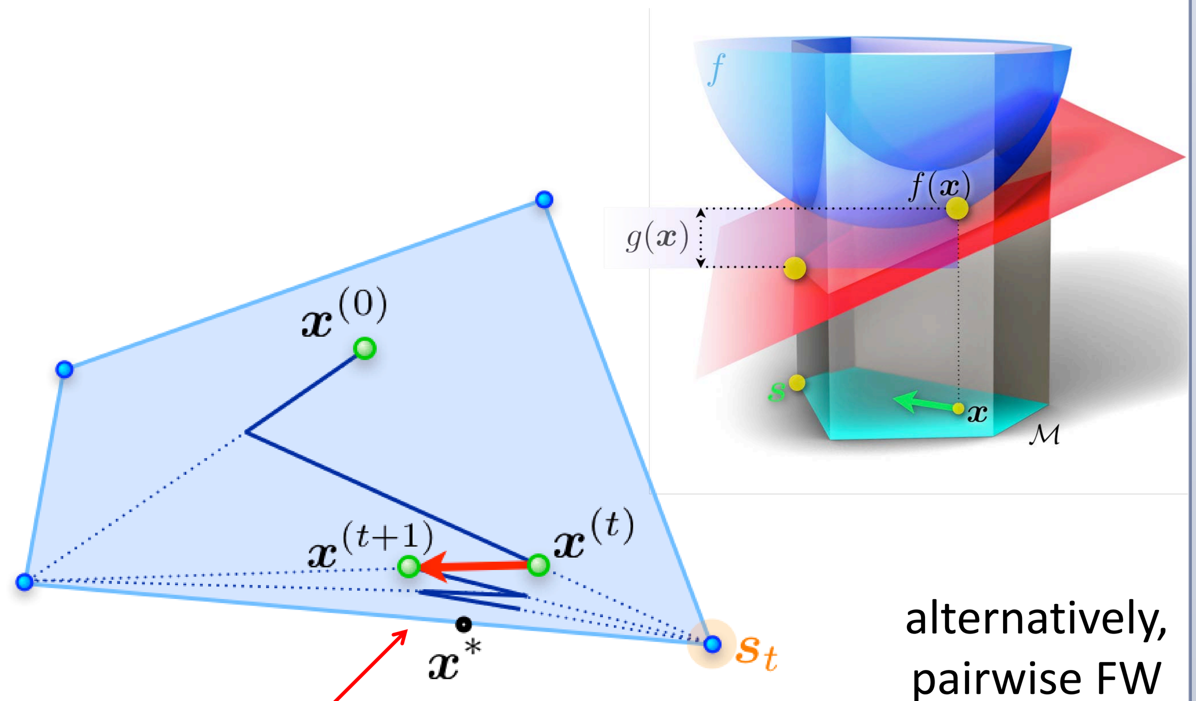
Examples: QP over combinatorial polytopes

1) for submodular function optimization [3], $\mathcal{M}$ is *base polytope*, $\mathcal{A}$ contains indicators on subsets
2) for structured SVM learning [21] or for approximate marginal inference [18] (poster **#48**), $\mathcal{M}$ is *marginal polytope*, $\mathcal{A}$ contains joint assignments in a MRF
3) for tracking [7] or video co-localization [16], $\mathcal{M}$ is *flow polytope*, $\mathcal{A}$ contains integer flows

video co-localization [16 - Joulin et al. ECCV14]:

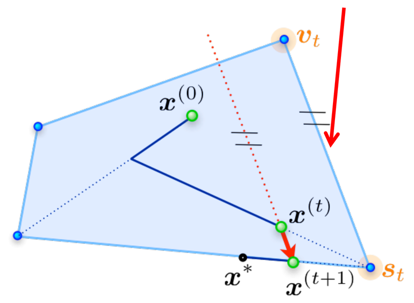


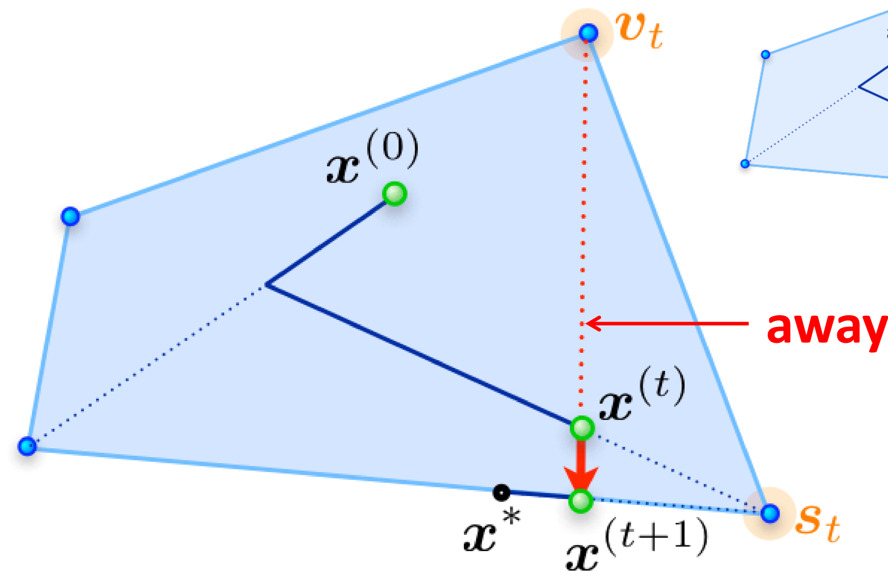
Variants of FW



zig-zagging problem for FW

alternatively, pairwise FW direction





away step fix

Algorithm 1: Away-steps Frank-Wolfe: $\textbf{AFW}(\boldsymbol{x}^{(0)}, \mathcal{A}, \epsilon)$

Let $\boldsymbol{x}^{(0)} \in \mathcal{A}$ and $\mathcal{S}^{(0)} := \{\boldsymbol{x}^{(0)}\}$; we maintain $\boldsymbol{x}^{(t)} = \sum_{\boldsymbol{v}\in\mathcal{S}^{(t)}} \alpha_{\boldsymbol{v}}^{(t)} \boldsymbol{v}$
for $t = 0 \dots T$ **do**
  Let $\boldsymbol{s}_t := \text{LMO}_{\mathcal{A}}(\nabla f(\boldsymbol{x}^{(t)}))$ and $\boldsymbol{d}_t^{\text{FW}} := \boldsymbol{s}_t - \boldsymbol{x}^{(t)}$ *(the FW direction)*
  Let $\boldsymbol{v}_t \in \arg\max_{\boldsymbol{v}\in\mathcal{S}^{(t)}} \langle \nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{v}\rangle$ and $\boldsymbol{d}_t^{\text{A}} := \boldsymbol{x}^{(t)} - \boldsymbol{v}_t$ *(the away direction)*
  if $g_t^{\text{FW}} := \langle -\nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{d}_t^{\text{FW}}\rangle \leq \epsilon$ **then return** $\boldsymbol{x}^{(t)}$
  if $\langle -\nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{d}_t^{\text{FW}}\rangle \geq \langle -\nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{d}_t^{\text{A}}\rangle$ **then**
    $\boldsymbol{d}_t := \boldsymbol{d}_t^{\text{FW}}$, and $\gamma_{\max} := 1$ *(choose the FW direction)*
  else
    $\boldsymbol{d}_t := \boldsymbol{d}_t^{\text{A}}$, and $\gamma_{\max} := \alpha_{\boldsymbol{v}_t}/(1-\alpha_{\boldsymbol{v}_t})$
    (choose the away direction, and maximum feasible step-size)
  end
  Line search: $\gamma_t \in \arg\min_{\gamma\in[0,\gamma_{\max}]} f(\boldsymbol{x}^{(t)} + \gamma \boldsymbol{d}_t)$
  Update $\boldsymbol{x}^{(t+1)} := \boldsymbol{x}^{(t)} + \gamma_t \boldsymbol{d}_t$ *(and weights $\boldsymbol{\alpha}^{(t+1)}$ accordingly)*
  Update $\mathcal{S}^{(t+1)} := \{\boldsymbol{v} \text{ s.t. } \alpha_{\boldsymbol{v}}^{(t+1)} > 0\}$
end

Algorithm 2: Pairwise Frank-Wolfe: $\textbf{PFW}(\boldsymbol{x}^{(0)}, \mathcal{A}, \epsilon)$

... as in Algorithm 1, except using $\boldsymbol{d}_t = \boldsymbol{d}_t^{\text{PFW}} := \boldsymbol{s}_t - \boldsymbol{v}_t$, and $\gamma_{\max} := \alpha_{\boldsymbol{v}_t}$

Fully-corrective FW: at each step, make *correction* over convex hull of visited vertices

Convergence Result

- **THM: Global linear convergence** for the four variants of FW for strongly convex f:
  - away-steps FW (AFW)
  - fully-corrective FW (FCFW)
  - pairwise FW (PFW)
  - min-norm point alg. (MNP)

Let $h_t := f(\boldsymbol{x}^{(t)}) - f(\boldsymbol{x}^*)$ be the suboptimality, then:

$$h_t \leq h_0 \exp\left(-\frac{1}{4}\rho_f\, k(t)\right)$$

where

geometric strong convexity constant **(new!)**

$$\rho_f := \frac{\mu_f \geq \mu(\text{pyr.width}(\mathcal{M}))^2}{C_f \leq L(\text{diam}(\mathcal{M}))^2}$$

curvature constant

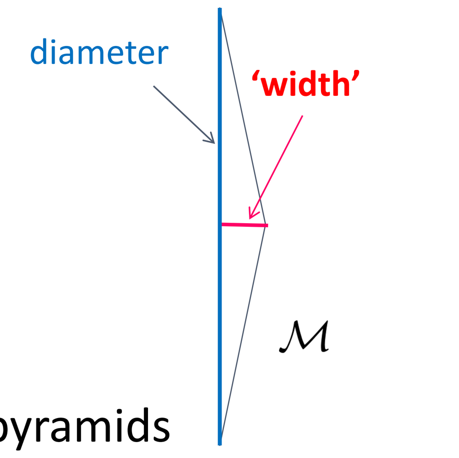
and $k(t)$:= number of 'good steps' after t iterations:

$k(t) = t$	for FCFW;
$k(t) \geq t/2$	for AFW and MNP;
$k(t) \geq t/4\lvert\mathcal{A}\rvert!$	for PFW

properties:

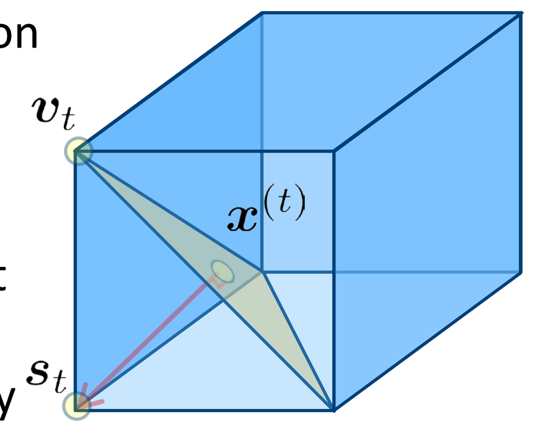
- analysis is affine invariant
- constant bounded away from zero (unlike [12] for AFW)
- first linear convergence result for FCFW and MNP
- can be generalized also to: $f(\boldsymbol{x}) := g(\boldsymbol{A}\boldsymbol{x}) + \langle \boldsymbol{b}, \boldsymbol{x}\rangle$ with g strongly convex
- gives `**condition number**' for domain $\mathcal{M}$:

$$\frac{1}{\rho_f} \leq \underbrace{\frac{L}{\mu}}_{\text{condition number of } f} \cdot \underbrace{\left(\frac{\text{diam}(\mathcal{M})}{\text{pyr.width}(\mathcal{M})}\right)^2}_{\text{condition number of } \mathcal{M}}$$



pyramidal width:

- smallest directional width of pyramids built with **active set** as base, **FW point** as summit, and using a **feasible** direction
- value of $\approx 1/\sqrt{d}$ for both the prob. simplex and unit cube in dimension d
- regular simplex has smallest **cond. number: d/2** giving $O(d\frac{L}{\mu}\log\frac{1}{\epsilon})$ complexity
- cond. num. of cube: d^2



pyr. width(cube) $= 1/\sqrt{3}$

Proof Elements

- as f has L-Lipschitz gradient: let $\boldsymbol{r}_t := -\nabla f(\boldsymbol{x}^{(t)})$

$$f(\boldsymbol{x}^{(t+1)}) \leq f(\boldsymbol{x}^{(t)} + \gamma\boldsymbol{d}_t) \leq f(\boldsymbol{x}^{(t)}) + \gamma\left\langle \nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{d}_t\right\rangle + \frac{\gamma^2}{2}L\|\boldsymbol{d}_t\|^2 \quad \forall\gamma\in[0,\gamma_{\max}]$$

Choose γ to minimize RHS: $h_t - h_{t+1} \geq \frac{\langle \boldsymbol{r}_t, \boldsymbol{d}_t\rangle^2}{2L\|\boldsymbol{d}_t\|^2} = \frac{1}{2L}\langle \boldsymbol{r}_t, \hat{\boldsymbol{d}}_t\rangle^2$ (when $\gamma_{\max}$ is big enough)

- as f is μ-strongly convex: with $\boldsymbol{e}_t := \boldsymbol{x}^* - \boldsymbol{x}^{(t)}$

$$f(\boldsymbol{x}^{(t)} + \gamma\boldsymbol{e}_t) \geq f(\boldsymbol{x}^{(t)}) + \gamma\left\langle \nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{e}_t\right\rangle + \frac{\gamma^2}{2}\mu\|\boldsymbol{e}_t\|^2 \quad \forall\gamma\in[0,1]$$

Use $\gamma = 1$ on LHS; compare with lower bound of RHS: $h_t \leq \frac{\langle \boldsymbol{r}_t, \hat{\boldsymbol{e}}_t\rangle^2}{2\mu}$

Combine to get: (valid for any $\boldsymbol{d}_t$)

$$\boxed{h_t - h_{t+1} \geq \frac{\mu}{L}\frac{\langle \boldsymbol{r}_t, \hat{\boldsymbol{d}}_t\rangle^2}{\langle \boldsymbol{r}_t, \hat{\boldsymbol{e}}_t\rangle^2} h_t \geq \frac{\mu}{L}\langle \hat{\boldsymbol{r}}_t, \hat{\boldsymbol{d}}_t\rangle^2 h_t}$$

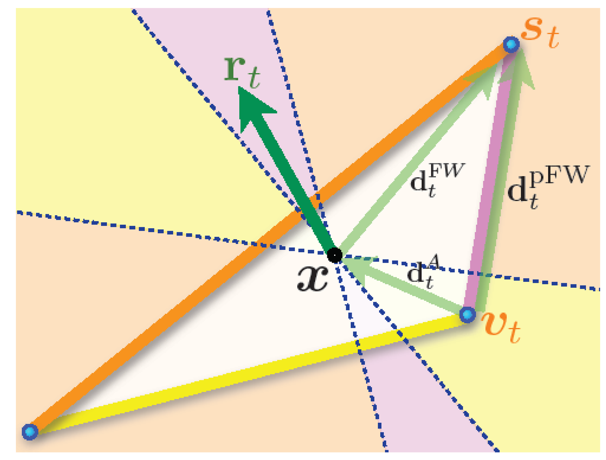
angle between negative gradient and update direction

- **2 Key points:**

$$2\langle \boldsymbol{r}_t, \boldsymbol{d}_t\rangle \geq \langle \boldsymbol{r}_t, \boldsymbol{d}_t^{\text{FW}} + \boldsymbol{d}_t^{A}\rangle = \langle \boldsymbol{r}_t, \boldsymbol{d}_t^{\text{PFW}}\rangle \text{ for AFW}$$

$$\left\langle \hat{\boldsymbol{r}}, \hat{\boldsymbol{d}}^{\text{PFW}}(\hat{\boldsymbol{r}})\right\rangle \geq \frac{\textit{PWidth}(\mathcal{M})}{\text{diam}(\mathcal{M})}$$

(see illustration on right showing possible PFW directions as **r** varies)

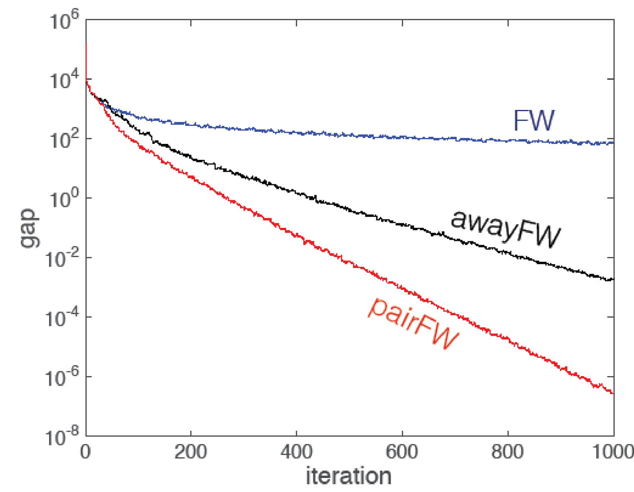


Illustrative Experiments

Lasso regression:

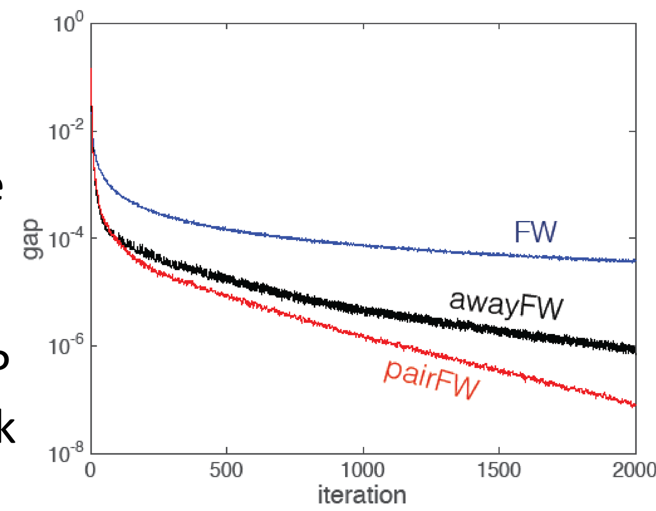
$$\min_{\|\boldsymbol{x}\|_1\leq 20} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$$

- Gaussian matrix $\boldsymbol{A} \in \mathbb{R}^{200\times 500}$
- $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^* + 0.1 \cdot$ Gaussian noise where $\boldsymbol{x}^*$ has 50 entries of ± 1's



Video co-localization [16]:

- QP over flow-polytope
- d = 660
- $\text{LMO}_{\mathcal{A}}$ can be solved using **shortest path** DP algorithm in a network



Empirical tightness of rate: