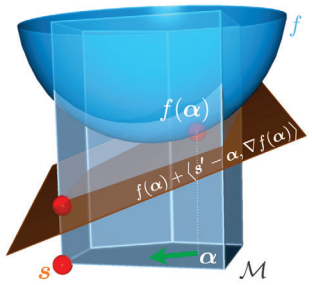
Recent Advances in Frank-Wolfe Optimization

Simon Lacoste-Julien

Université de Montréal



OSL 2017 – Les Houches – April 13th, 2017

Outline

- Frank-Wolfe algorithm review
- global linear convergence of FW optimization variants
  - condition number of domains & pyramidal width
- saddle point Frank-Wolfe

Frank-Wolfe algorithm [Frank, Wolfe 1956]

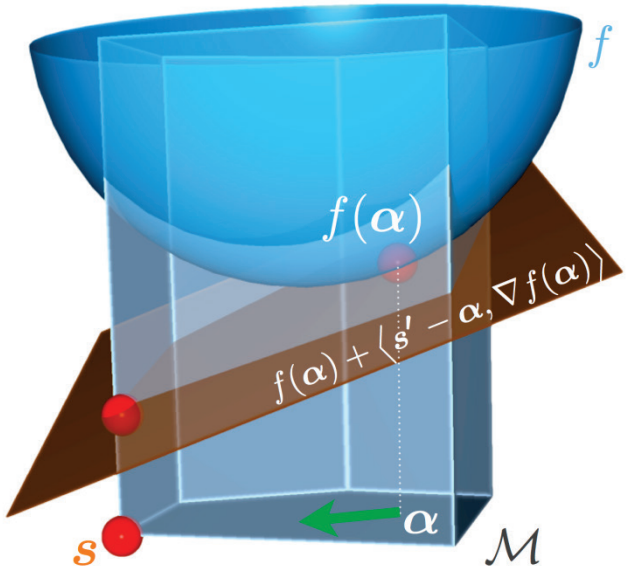
(aka conditional gradient)

- alg. for constrained opt.: $\min_{\alpha \in \mathcal{M}} f(\alpha)$

  where:

  f convex & cts. differentiable

  M convex & compact



- FW algorithm – repeat:

1) Find good feasible direction by minimizing linearization of f :

$$s_{t+1} \in \arg\min_{s' \in \mathcal{M}} \langle s', \nabla f(\alpha_t) \rangle$$

2) Take convex step in direction:

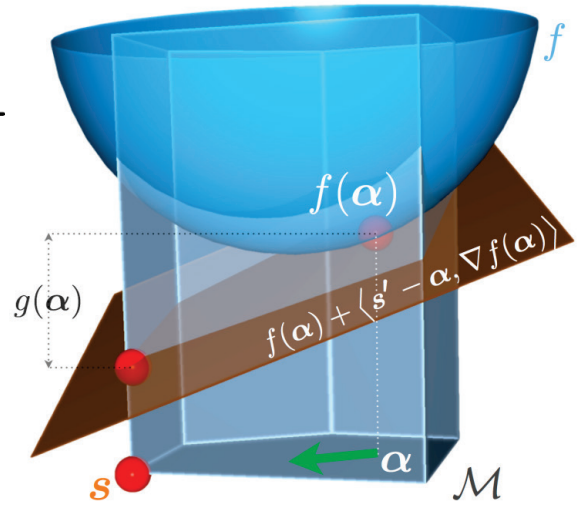
$$\alpha_{t+1} = (1 - \gamma_t)\, \alpha_t + \gamma_t\, s_{t+1}$$

- Properties: O(1/T) rate
  - sparse iterates
  - get duality gap $g(\alpha)$ for free
  - affine invariant
  - rate holds even if linear subproblem solved **approximately**

Frank-Wolfe: properties

- convex steps => convex sparse combo:

$$\boldsymbol{\alpha}_T = \rho_0 \boldsymbol{\alpha}_0 + \sum_{t=1}^{T} \rho_t \boldsymbol{s}_t \quad \text{where} \sum_{t=0}^{T} \rho_t = 1$$



- get duality gap certificate for free

$$g(\boldsymbol{\alpha}) := \max_{\boldsymbol{s}' \in \mathcal{M}} \langle \boldsymbol{s}' - \boldsymbol{\alpha}\,, -\nabla f(\boldsymbol{\alpha}) \rangle = \langle \boldsymbol{s} - \boldsymbol{\alpha}\,, -\nabla f(\boldsymbol{\alpha}) \rangle$$

  - (special case of Fenchel duality gap)
  - also converge as O(1/T)!

$O(1/\sqrt{T})$ for non-convex f
[L.-J. arXiv 2016]

- only need to solve linear subproblem *__approximately__* (additive/multiplicative bound)

- affine invariant!

- numerically stable

see survey [Jaggi ICML 2013]
(also [Lan arXiv 2013])

Why comeback of FW in ML?

- big data -> first order algorithm
- sparse algorithms – e.g. see references in [Locatello et. AISTATS 2017]
- structured constrained sets with cheaper LMOs:

$\mathcal{X}$	Optimization Domain Atoms $\mathcal{A}$	$\mathcal{D} = \text{conv}(\mathcal{A})$	Complexity of one Frank-Wolfe Iteration $\Omega^*_{\mathcal{D}}(y) = \sup_{s \in \mathcal{D}} \langle s, y \rangle$	Complexity
$\mathbb{R}^n$	Sparse vectors	$\|.\|_1$-ball	$\|y\|_\infty$	$O(n)$
$\mathbb{R}^n$	Sign-vectors	$\|.\|_\infty$-ball	$\|y\|_1$	$O(n)$
$\mathbb{R}^n$	ℓ_p-Sphere	$\|.\|_p$-ball	$\|y\|_q$	$O(n)$
$\mathbb{R}^n$	Sparse non-neg. vectors	Simplex Δ_n	$\max_i\{y_i\}$	$O(n)$
$\mathbb{R}^n$	Latent group sparse vectors	$\|.\|_{\mathcal{G}}$-ball	$\max_{g \in \mathcal{G}} \|y_{(g)}\|^*_q$	$\sum_{g \in \mathcal{G}} \|g\|$
$\mathbb{R}^{m \times n}$	Matrix trace norm	$\|.\|_{tr}$-ball	$\|y\|_{op} = \sigma_1(y)$	$\tilde{O}(N_f/\sqrt{\varepsilon'})$ (Lanczos)
$\mathbb{R}^{m \times n}$	Matrix operator norm	$\|.\|_{op}$-ball	$\|y\|_{tr} = \|(\sigma_i(y))\|_1$	SVD
$\mathbb{R}^{m \times n}$	Schatten matrix norms	$\|(\sigma_i(.))\|_p$-ball	$\|(\sigma_i(y))\|_q$	SVD
$\mathbb{R}^{m \times n}$	Matrix max-norm	$\|.\|_{\max}$-ball		$\tilde{O}(N_f(n+m)^{1.5}/\varepsilon'^{2.5})$
$\mathbb{R}^{n \times n}$	Permutation matrices	Birkhoff polytope		$O(n^3)$
$\mathbb{R}^{n \times n}$	Rotation matrices			SVD (Procrustes prob.)
$\mathbb{S}^{n \times n}$	Rank-1 PSD matrices of unit trace	$\{x \succeq 0,\ \text{Tr}(x) = 1\}$	$\lambda_{\max}(y)$	$\tilde{O}(N_f/\sqrt{\varepsilon'})$ (Lanczos)
$\mathbb{S}^{n \times n}$	PSD matrices of bounded diagonal	$\{x \succeq 0,\ x_{ii} \leq 1\}$		$\tilde{O}(N_f\, n^{1.5}/\varepsilon'^{2.5})$

(table from [Jaggi ICML 2013])

On the Global Linear Convergence of Frank-Wolfe Optimization Variants

[L.-J. and Jaggi, NIPS 2015]

joint work with
Martin Jaggi

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Problem setup

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

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

with only access to a *linear minimization oracle*:

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

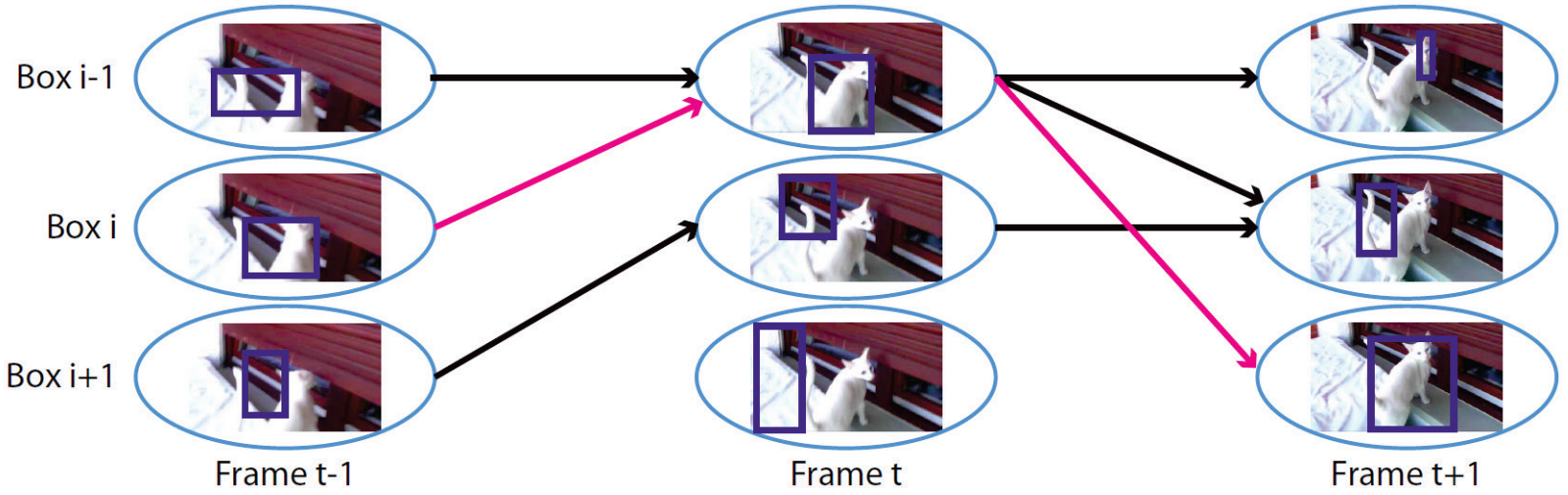
assume f is μ-strongly convex

and ∇f is L-Lipschitz continuous

Examples: QP over combinatorial polytopes

- For tracking [Chari, L.-J. et al. CVPR 15] or video co-localization [Joulin, Tang, Fei-Fei ECCV 14]

$\mathcal{M}$ is *flow polytope*, $\mathcal{A}$ contains integer flows



video co-localization

other examples...

- for structured SVM learning [L.-J., Jaggi et al. ICML 13] or approximate marginal inference [Krishnan, L.-J., Sontag NIPS 15]

$\mathcal{M}$ is *marginal polytope*,
$\mathcal{A}$ contains clique assignments in a MRF

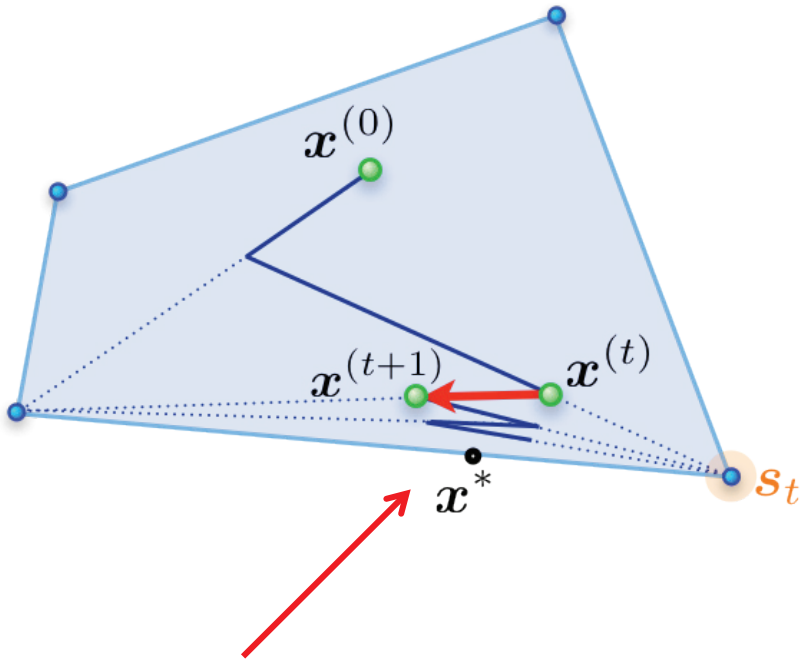
- for submodular function optimization [Bach TML 13]

$\mathcal{M}$ is *base polytope*,

$\mathcal{A}$ has one possible atom
per permutation of ground set

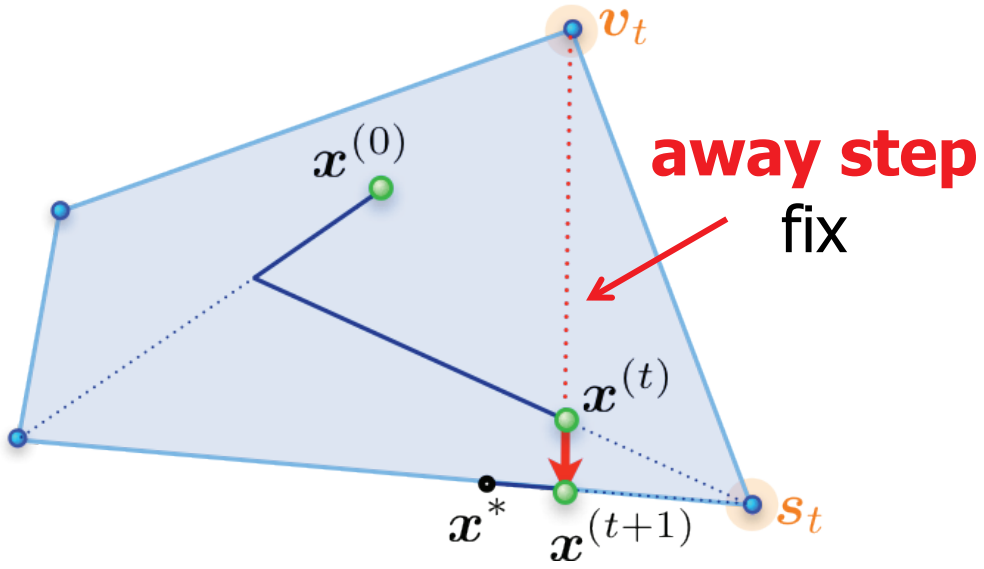
slow convergence of Frank-Wolfe...

standard FW



zig-zagging problem for FW

away-step FW



[Wolfe 1970]
[Guelat & Marcotte 1986]

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

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 $\boldsymbol{s}_t := \mathrm{LMO}_{\mathcal{A}}\left(\nabla f(\boldsymbol{x}^{(t)})\right)$ and $\boldsymbol{d}_t^{\mathrm{FW}} := \boldsymbol{s}_t - \boldsymbol{x}^{(t)}$ *(the FW direction)*

Let $\boldsymbol{v}_t \in \arg\max_{\boldsymbol{v} \in \mathcal{S}^{(t)}} \left\langle \nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{v} \right\rangle$ and $\boldsymbol{d}_t^{\mathrm{A}} := \boldsymbol{x}^{(t)} - \boldsymbol{v}_t$ *(the away direction)*

if $g_t^{\mathrm{FW}} := \left\langle -\nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{d}_t^{\mathrm{FW}} \right\rangle \leq \epsilon$ **then return** $\boldsymbol{x}^{(t)}$

if $\left\langle -\nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{d}_t^{\mathrm{FW}} \right\rangle \geq \left\langle -\nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{d}_t^{\mathrm{A}} \right\rangle$ **then**

$\boldsymbol{d}_t := \boldsymbol{d}_t^{\mathrm{FW}}$, and $\gamma_{\max} := 1$ *(choose the FW direction)*

else

$\boldsymbol{d}_t := \boldsymbol{d}_t^{\mathrm{A}}$, and $\gamma_{\max} := \alpha_{\boldsymbol{v}_t} / (1 - \alpha_{\boldsymbol{v}_t})$

(choose the away direction, and maximum feasible step-size)

end

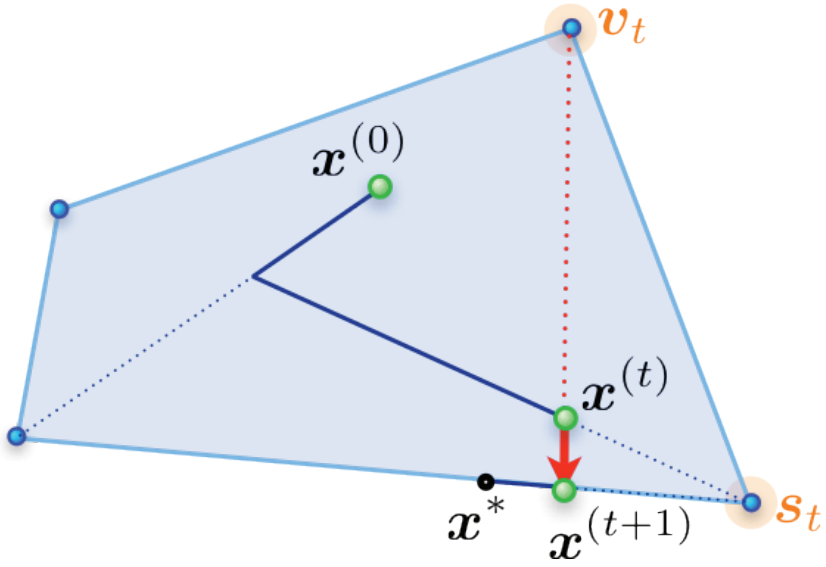
Line search: $\gamma_t \in \arg\min_{\gamma \in [0, \gamma_{\max}]} f\left(\boldsymbol{x}^{(t)} + \gamma \boldsymbol{d}_t\right)$

Update $\boldsymbol{x}^{(t+1)} := \boldsymbol{x}^{(t)} + \gamma_t \boldsymbol{d}_t$ *(and weights $\boldsymbol{\alpha}^{(t+1)}$ accordingly)*

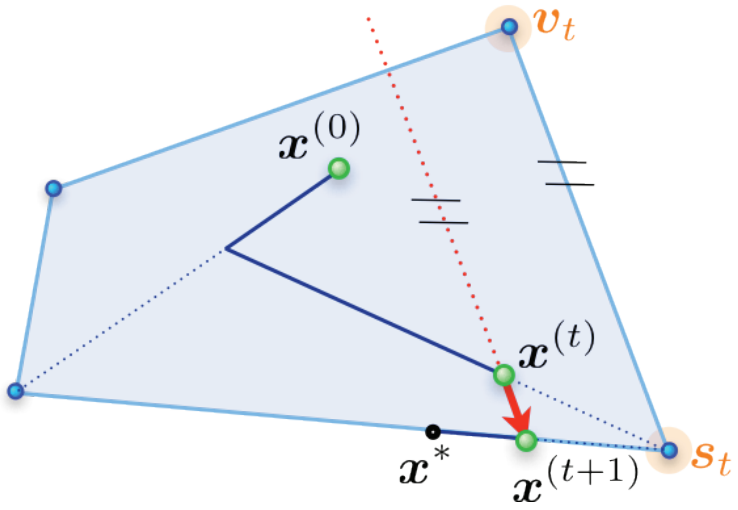
Update $\mathcal{S}^{(t+1)} := \{\boldsymbol{v} \; s.t. \; \alpha_{\boldsymbol{v}}^{(t+1)} > 0\}$

end

other variants:



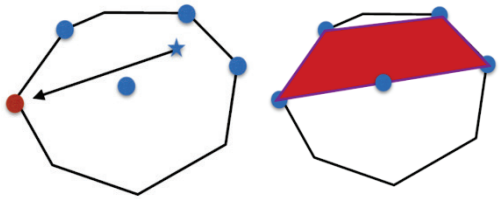
away-step FW



pairwise FW

[MDM 1974]

- fully-corrective FW (FCFW): re-optimize over convex hull of previously found vertices (correction polytope)



[Halloway 1974, Von Hohenbalken 1977, ...]

Previous convergence results

assumption: f is strongly convex (with Lipschitz gradient)

[Wolfe 70, Guélat & Marcotte 86]:

- Frank-Wolfe algorithm converges **linearly** if solution x* is in relative **interior** of M
- Frank-Wolfe **with away steps** converges **linearly** with a constant depending on the distance between x* and the boundary of M in the optimal face containing x*

Problems:

- constant could be arbitrarily close to zero -> not a true linear convergence result
- constant depends on unknown x*
- analysis is **not affine invariant**

(FW alg. is invariant to affine transformations of variables)

Our contribution: [L.-J. & Jaggi NIPS 15, arxiv 13]

- we give an **affine invariant** analysis of the **global linear convergence** of Frank-Wolfe with away steps with constant bounded away from zero:

thm: $f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^*) \leq h_0 \exp(-\frac{1}{8}\rho_f k)$

where:

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

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

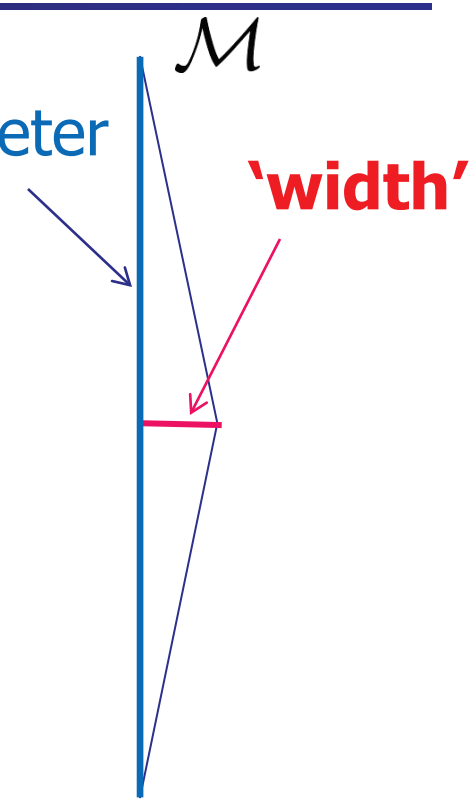
curvature constant

f has L-Lipschitz gradient
and is μ-strongly convex

‘Condition number’ of domain!

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

‘condition number’



eccentricity in dimension d:

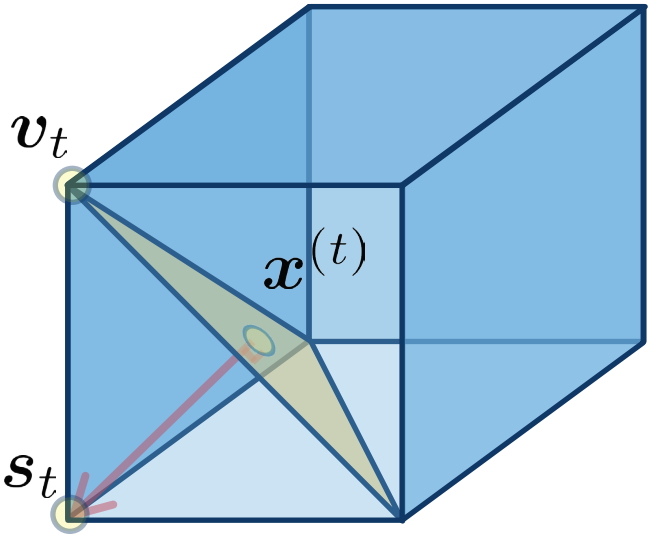
- probability simplex: $O(d)$
- unit cube: $O(d^2)$

Pyramidal width

- smallest directional width of pyramids built with **active set** as base, **FW point** as summit, and using a **feasible direction**
- value of $\approx 1/\sqrt{d}$ in dimension d:
  - prob. simplex
  - unit cube
  - l1-ball
- regular simplex has **smallest cond. number: d/2**
  - this gives a complexity of:

$$O(d\frac{L}{\mu}\log\frac{1}{\epsilon})$$

- unit cube has cond. number: d^2



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

[Pena & Rodriguez arXiv 2015]
shows equivalent to **facial distance:**

$$\min_{F \in \text{faces}(\mathcal{A})} \text{dist}(F, \text{conv}(\mathcal{A} \setminus F))$$

Proof elements [from Guélat & Marcotte 86]

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

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

(suboptimality: $h_t := f(\boldsymbol{x}^{(t)}) - f(\boldsymbol{x}^*)$)

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

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

$$f(\boldsymbol{x}^{(t)} + \gamma \boldsymbol{e}_t) \geq f(\boldsymbol{x}^{(t)}) + \gamma \left\langle \nabla f(\boldsymbol{x}^{(t)}), \boldsymbol{e}_t \right\rangle + \frac{\gamma^2}{2} \mu \|$$

Use $\gamma = 1$ on LHS; compare with lower bound of RHS:

angle between negative gradient and update direction

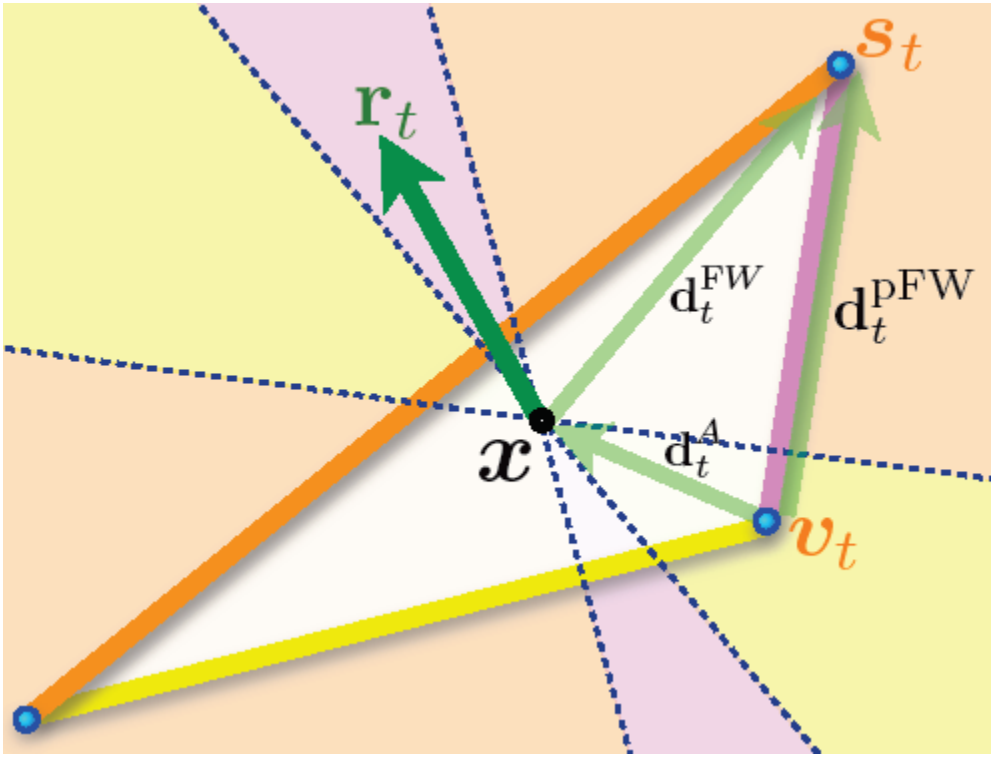
Combine to get: (valid for any $\boldsymbol{d}_t$)

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

2 key insights:

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

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



(illustration showing possible PFW directions as **r** varies)

Important inequality

- key inequality which has been re-used several times:

$$g_t^{\mathrm{PFW}} := \langle -\nabla f(\boldsymbol{x}_t), \boldsymbol{s}_t - \boldsymbol{v}_t \rangle$$

$$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{1}{2\mu_f} (g_t^{\mathrm{PFW}})^2$$

here is an unconstrained analog:

$$\frac{1}{2L} \|\nabla f(\boldsymbol{x}_t)\|^2 \leq f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leq \frac{1}{2\mu} \|\nabla f(\boldsymbol{x}_t)\|^2$$

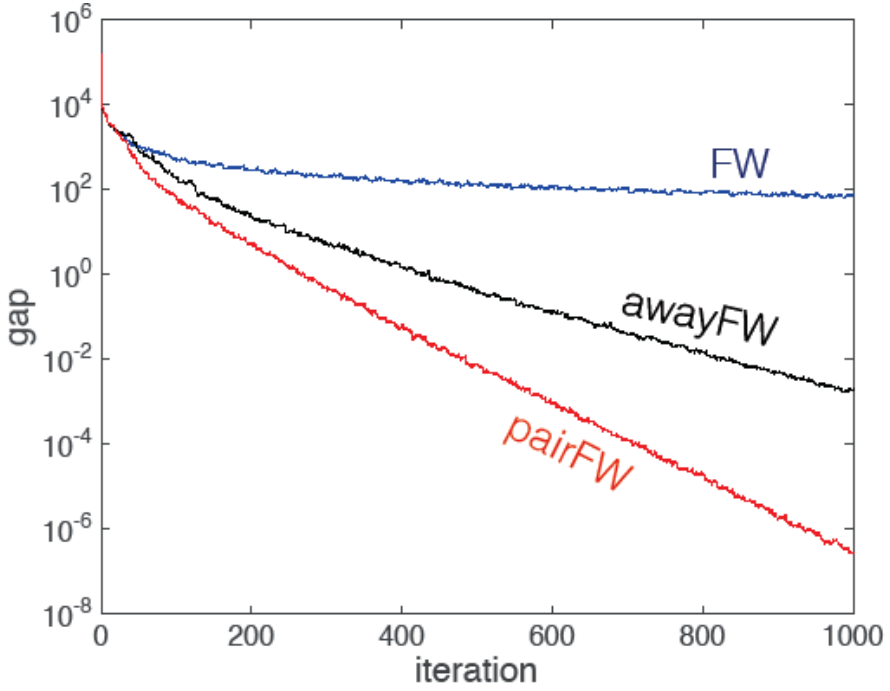
- used for:
  - ADMM + FW: [Yen et al. ICML 2016]
  - bandits [Berthet & Perchet arXiv 2017]
  - saddle pt. FW [Gidel et al. AISTATS 2017] (see 2nd part)
  - etc...

Illustrative experiments

Lasso regression:

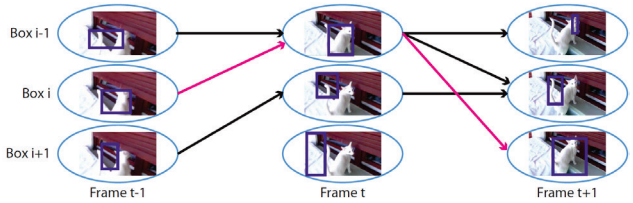
$$\min_{\|\boldsymbol{x}\|_1 \leq 20} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2$$

- 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

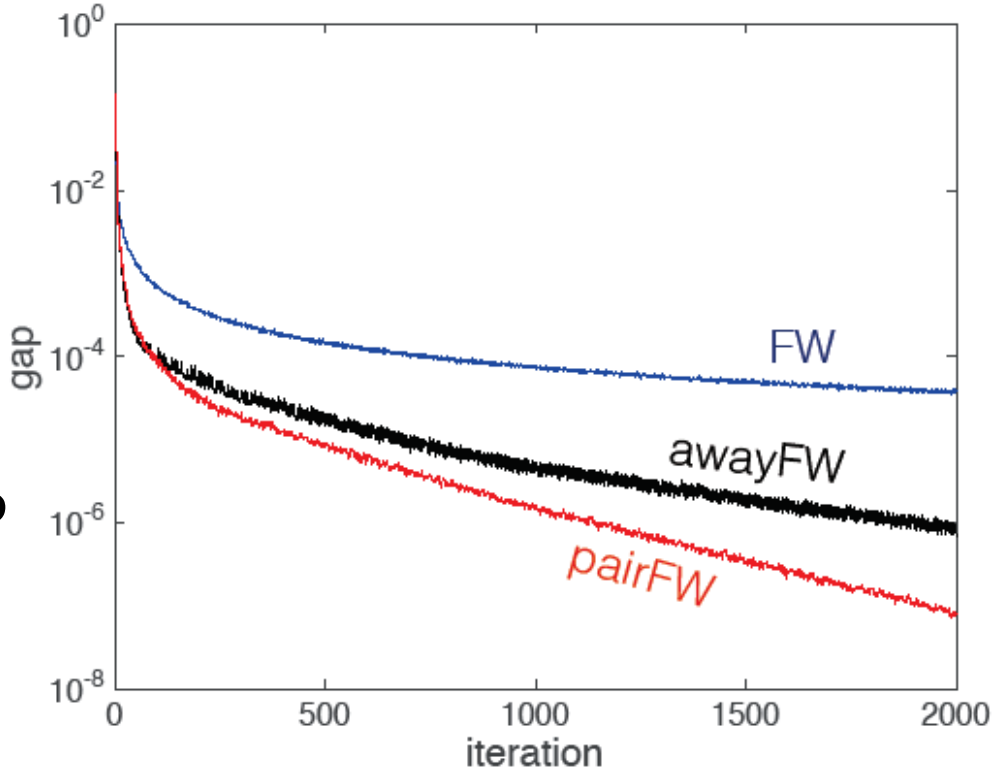


Video co-localization

- problem from [Joulin, Tang, Fei-Fei ECCV 14]

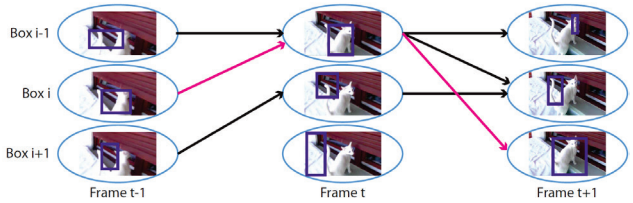


- QP over flow-polytope
- d = 660
- $\mathrm{LMO}_{\mathcal{A}}$ can be solved using **shortest path DP** algorithm over network

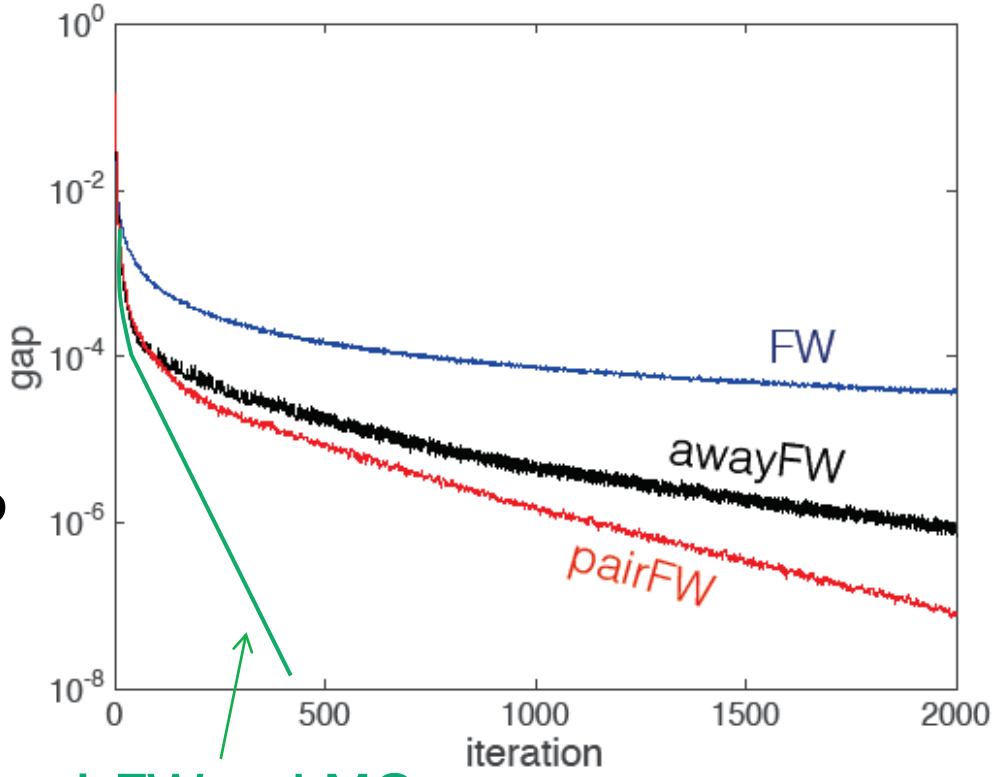


Video co-localization

- problem from [Joulin, Tang, Fei-Fei ECCV 14]



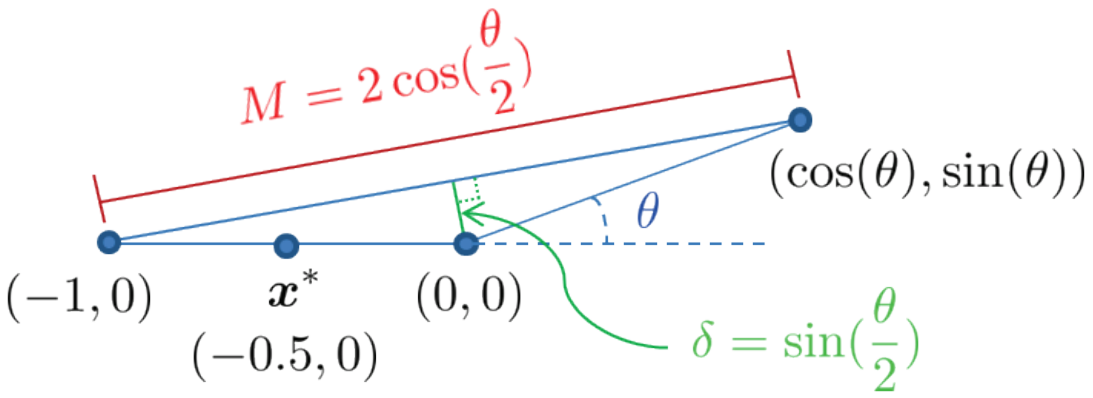
- QP over flow-polytope
- d = 660
- $\mathrm{LMO}_{\mathcal{A}}$ can be solved using **shortest path DP** algorithm over network

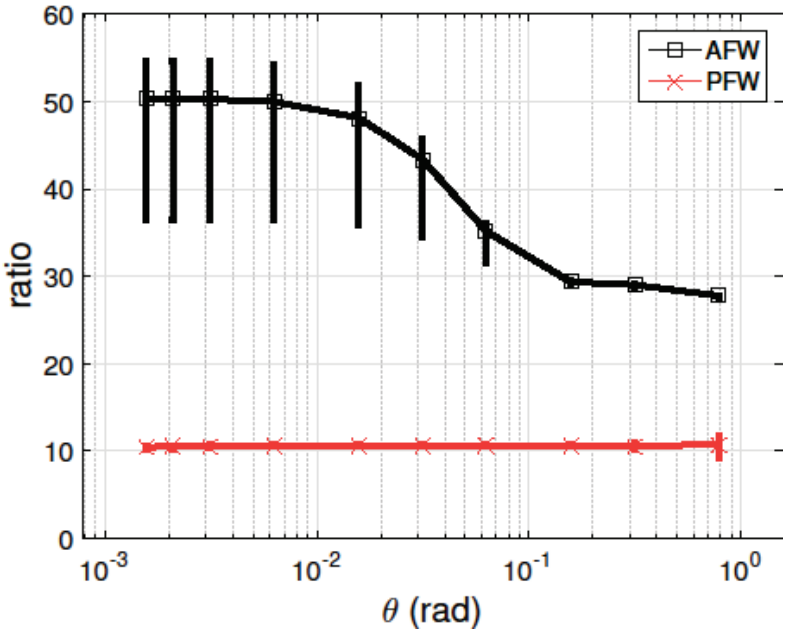


pairFW + LMO away corner

[Garber & Meshi NIPS 2016]

Rate is empirically tight!





Discussion

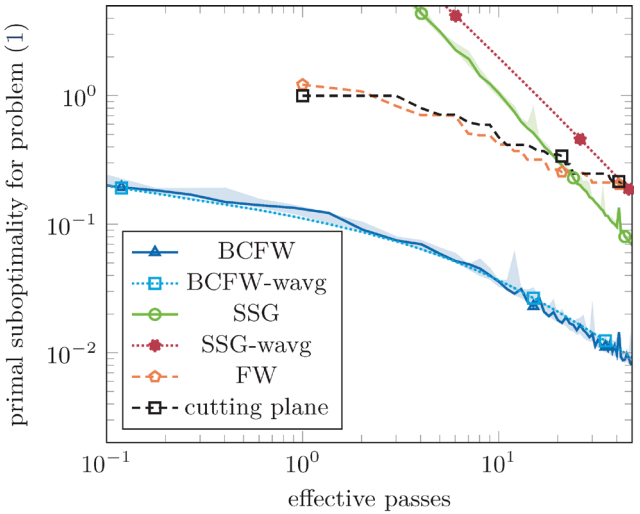
- FW and variants popular in machine learning for optimization over structured polytopes
- Provide first truly global linear convergence rate for a Frank-Wolfe type algorithm *which does not need to compute any constants* (vs. [Garber & Hazan 13])
  - and *analysis is **affine invariant***
  - can bound constant with condition number and purely geometric quantity 'eccentricity' -> **condition number for M**
- give first linear rate for FCFW, PFW and MNP
- extensions: used for ADMM / FW alg.; saddle point FW, etc.
  - reduce dependence to **~dimension of optimal face**?
    - -> YES: [Garber & Meshi NIPS 2016] for special 0-1 polytopes
  - AFW also linear rate for strongly convex sets
    - but general infinite number of atoms -> **still open question**

Other FW extensions / applications

- block-coordinate FW (for structured SVMs)

  [L.-J. et al. ICML 2013]

  AFW -> [Osokin et al. ICML 2016]



- barrier FW

  [Krishnan, L.-J. & Sontag NIPS 2015]

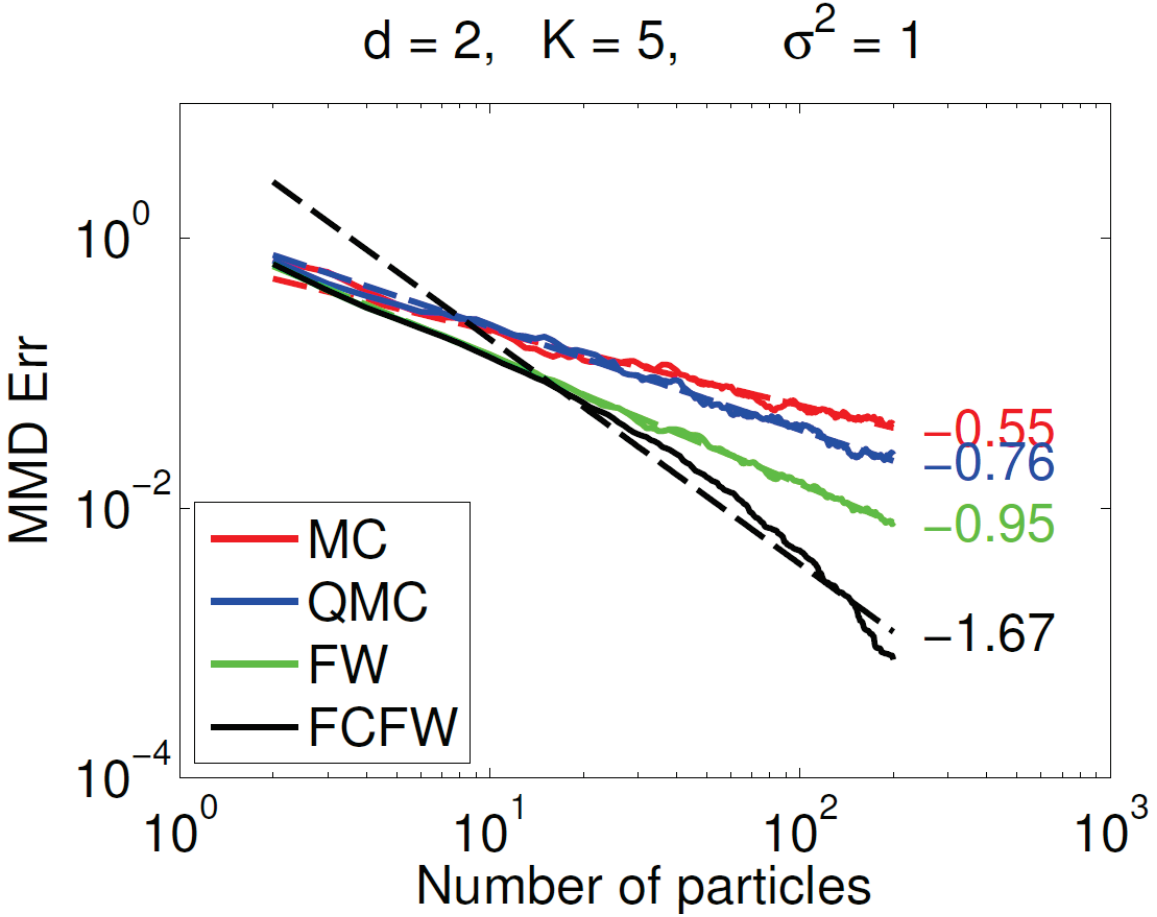
- FW quadrature

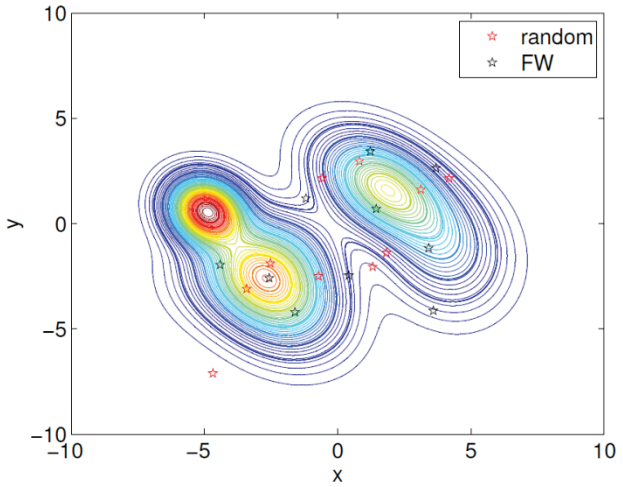
  [Bach, L.-J., Obozinski ICML 2012],

  [L.-J., Lindsten, Bach AISTATS 2015]

FW quadrature

- for a mixture of Gaussians:





[L.-J., Lindsten, Bach, AISTATS 15]

Frank-Wolfe Algorithms for Saddle Point Problems

[Gidel, Jebara & L.-J., AISTATS 2017]

with
Gauthier Gidel

Université de Montréal

Overview

Saddle point problem: solve $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{L}(x, y)$

- want to solve using **only with LMOs**
- approach: extend FW to saddle point problems
- *straightforward* extension but **nontrivial** analysis

- related work:
  - [Lan arXiv 2013] -> use smoothing
  - [He & Harchaoui NIPS 2015] -> approximate projections
  - [Juditsky & Nemirovski MathProg 2016]
    -> VIP transformations

Saddle point and link with variational inequalities

Let $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{X}$ and $\mathcal{Y}$ are convex and compact.

Saddle point problem:	solve $\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{L}(x, y)$

A solution (x^*, y^*) is called a ***Saddle Point***.

- **Necessary *stationary conditions:***

$$\langle x - x^*, \quad \nabla_x \mathcal{L}(x^*, y^*) \rangle \geq 0$$
$$\langle y - y^*, -\nabla_y \mathcal{L}(x^*, y^*) \rangle \geq 0$$

- ***Variational inequality:***

$$\forall z \in \mathcal{X} \times \mathcal{Y} \quad \langle z - z^*, r(z^*) \rangle \geq 0$$

where $(x^*, y^*) = z^*$ and $r(z) = (\nabla_x \mathcal{L}(z), -\nabla_y \mathcal{L}(z))$

- **Sufficient *condition*: *Global solution*** if $\mathcal{L}$ *convex-concave.* $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$

$$x' \mapsto \mathcal{L}(x', y) \text{ is convex} \quad \text{and} \quad y' \mapsto \mathcal{L}(x, y') \text{ is concave.}$$

Motivations

- two-player games: $\min_{x \in \Delta(I)} \max_{y \in \Delta(J)} x^\top M y$
- structured SVM: $\min_{\omega \in \mathbb{R}^d} \lambda \Omega(\omega) + \frac{1}{n} \sum_{i=1}^{n} \underbrace{\max_{y \in \mathcal{Y}_i} (L_i(y) - \langle \omega, \phi_i(y) \rangle)}_{\text{structured hinge loss}}$

  Regularization: penalized $\rightarrow$ constrained.

  $$\min_{\Omega(\omega) \leq \beta} \max_{\alpha \in \Delta(|\mathcal{Y}|)} b^T \alpha - \omega^T M \alpha$$

  Hard to project when:

  - *Structured sparsity* norm (group lasso norm).
  - The output $\mathcal{Y}$ is structured: *exponential* size.

- -> still looking for more: **call for applications!**

Standard methods to solve SP

projected gradient

Simple algorithm to solve Saddle point optimization:

$$\begin{aligned} \boldsymbol{x}^+ &= P_{\mathcal{X}}(\boldsymbol{x} - \gamma \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})) \\ \boldsymbol{y}^+ &= P_{\mathcal{Y}}(\boldsymbol{y} + \gamma \nabla_y \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})) \end{aligned}$$

Can oscillate so use averaging:

$$\frac{1}{T} \sum_{t=1}^{T} \left(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)} \right) \underset{T \to \infty}{\longrightarrow} (\boldsymbol{x}^*, \boldsymbol{y}^*)$$

projected extra-gradient

$$\begin{aligned} \bar{\boldsymbol{x}} &= P_{\mathcal{X}}(\boldsymbol{x} - \gamma \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})) \\ \bar{\boldsymbol{y}} &= P_{\mathcal{Y}}(\boldsymbol{y} + \gamma \nabla_y \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})) \\ \boldsymbol{x}^+ &= P_{\mathcal{X}}(\boldsymbol{x} - \gamma \nabla_x \mathcal{L}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})) \\ \boldsymbol{y}^+ &= P_{\mathcal{Y}}(\boldsymbol{y} + \gamma \nabla_y \mathcal{L}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}})) \end{aligned}$$

More stable and usually **faster**:

$$(\boldsymbol{x}^{(T)}, \boldsymbol{y}^{(T)}) \underset{T \to \infty}{\longrightarrow} (\boldsymbol{x}^*, \boldsymbol{y}^*)$$

SP-FW

Algorithm Saddle point FW algorithm

1: Let $\boldsymbol{z}^{(0)} = (\boldsymbol{x}^{(0)}, \boldsymbol{y}^{(0)}) \in \mathcal{X} \times \mathcal{Y}$
2: **for** $t = 0 \dots T$ **do**
3: Compute $\boldsymbol{r}^{(t)} := \begin{pmatrix} \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \\ -\nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \end{pmatrix}$
4: Compute $\boldsymbol{s}^{(t)} \in \underset{z \in \mathcal{X} \times \mathcal{Y}}{\operatorname{argmin}} \left\langle \boldsymbol{z}, \boldsymbol{r}^{(t)} \right\rangle$
5: Compute $g_t := \left\langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t)}, \boldsymbol{r}^{(t)} \right\rangle$
6: **if** $g_t \leq \epsilon$ **then return** $\boldsymbol{z}^{(t)}$
7: Let $\gamma = \min\left(1, \frac{\nu}{C} g_t\right)$ **or** $\gamma = \frac{2}{2+t}$
8: Update $\boldsymbol{z}^{(t+1)} := (1-\gamma)\boldsymbol{z}^{(t)} + \gamma \boldsymbol{s}^{(t)}$
9: **end for**

proposed by [Hammond 1984] with O(1/t) step-size

30 years old conjecture for polytopes!

Convergence

SP extension of FW with **away step**:

- ***Linear*** rate with ***adaptive*** step size $\gamma_t := \frac{\nu}{2LD^2} g_t$.

$$\min_{s \leq t} g_s \leq \left(1 - \nu^2 \frac{\delta^2}{D^2} \frac{\mu}{2L}\right)^t$$

- ***Sublinear*** rate with ***universal*** step size $\gamma_t := \frac{2}{2+k(t)}$.

Assumptions

Similar assumptions as AFW:

- $\mathcal{L}$ is L-**smooth** and μ-**strongly convex**.
- $\mathcal{X}$ and $\mathcal{Y}$ **polytopes**.

Additional assumption on **bilinearity**:

$$\nu := \frac{1}{2} - \frac{\sqrt{2}\|M\|}{\mu} \frac{D}{\delta} > 0$$

Details on the additional assumption

"*Strong convexity μ big enough compared to the bilinear coupling $\|M\|$*"

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{x}^\top M \boldsymbol{y} - g(\boldsymbol{y}).$$

$$D := \max\{\mathrm{diam}(\mathcal{X}), \mathrm{diam}(\mathcal{Y})\}, \ \delta := \min\{PWidth(\mathcal{X}), PWidth(\mathcal{Y})\}$$

Difficulties for saddle point

Usual **descent Lemma**:

$$h_{t+1} \leq h_t - \underbrace{\gamma_t g_t}_{\geq 0} + \gamma_t^2 \frac{L\|\boldsymbol{d}^{(t)}\|^2}{2}$$

With γ_t small enough the sequence decreases.

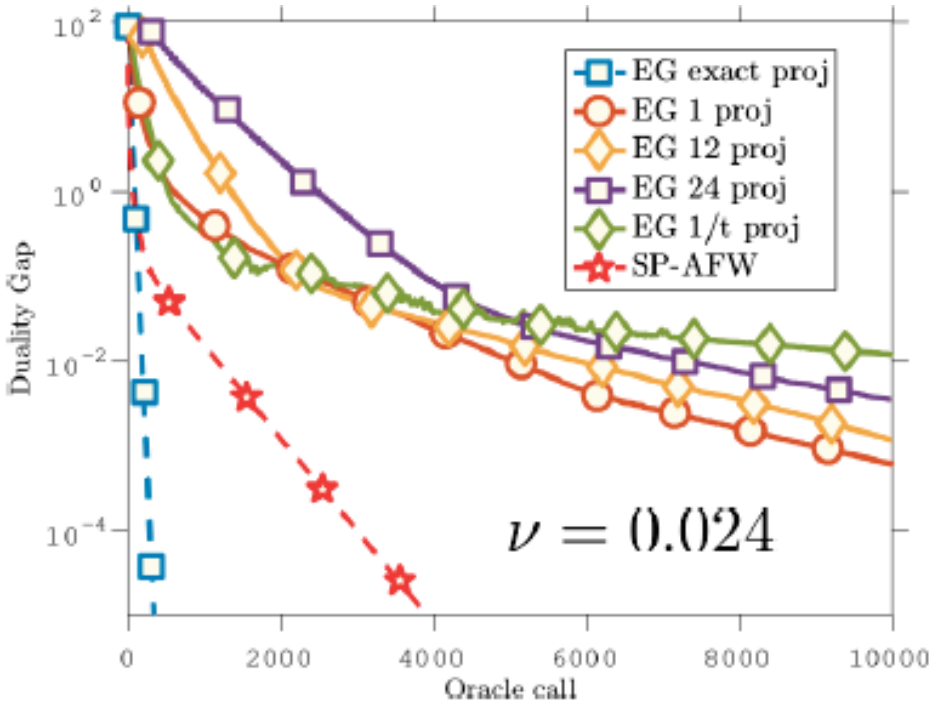
For saddle point problem the Lipschitz gradient property gives

$$\mathcal{L}_{t+1} - \mathcal{L}^* \leq \mathcal{L}_t - \mathcal{L}^* - \underbrace{\gamma_t \left(g_t^{(x)} - g_t^{(y)} \right)}_{\text{arbitrary sign}} + \gamma_t^2 \frac{L\|\boldsymbol{d}^{(t)}\|^2}{2}.$$

- Cannot control the oscillation of the sequence.
- Must introduce other quantities to establish convergence.

Toy experiments

- SP-AFW vs. extragradient with approx. projection



[He & Harchaoui NIPS 2015]

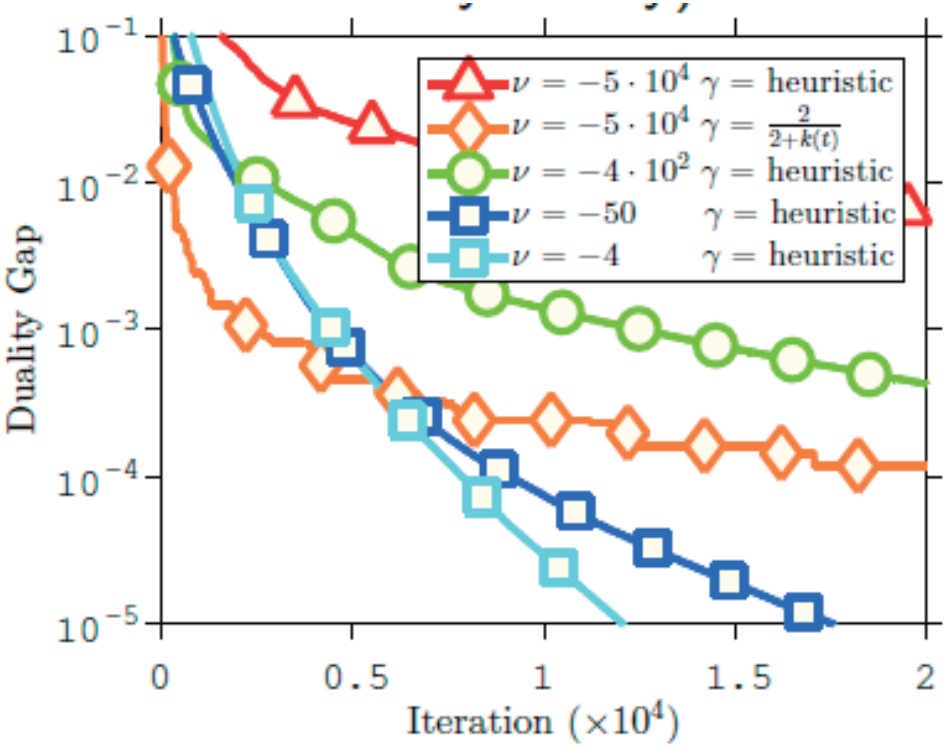
Theoretical: $\gamma_t = \frac{\nu}{C} g_t$

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) := \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + (\boldsymbol{x} - \boldsymbol{x}^*)^\top M(\boldsymbol{y} - \boldsymbol{y}^*) - \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{y}^*\|_2^2$$

$$\mathcal{X} = \mathcal{Y} := [0,1]^d \qquad d = 30 \qquad C := 2LD^2 \qquad L = \mu$$

Toy experiments

- SP-AFW with heuristic step sizes when $\nu < 0$



(not covered by theory)

Heuristic: $\gamma_t = \dfrac{g_t}{C + 2\frac{\|M\|^2 D^2}{\mu}}$

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) := \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + (\boldsymbol{x} - \boldsymbol{x}^*)^\top M(\boldsymbol{y} - \boldsymbol{y}^*) - \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{y}^*\|_2^2$$

$$\mathcal{X} = \mathcal{Y} := [0,1]^d \qquad d = 30 \qquad C := 2LD^2 \qquad L = \mu$$

Discussion

- also linear convergence of SP-FW on product of **strongly convex** sets
- for bilinear objective, Karlin's conjecture [1960] gives $O(1/\sqrt{t})$ rate (only empirical so far)
- more general convergence still open!

Thank you! Any question?

constants...

$$C_f := \sup_{\substack{x,s\in\mathcal{D},\\ \gamma\in[-1,1],\\ y=x+\gamma(s-x)}} \frac{2}{\gamma^2}\Big(f(y)-f(x)-\langle\nabla f(x), y-x\rangle\Big)$$

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

$$\mu_f := \inf_{\substack{x,y\in\mathcal{D}\\ \text{s.t. } \langle\nabla f(x),y-x\rangle<0}} \frac{2}{\gamma^2(x,y)}\Big(f(y)-f(x)-\langle\nabla f(x), y-x\rangle\Big)$$

$$\gamma(x,y) := \frac{\langle\nabla f(x), y-x\rangle}{\big\langle\nabla f(x), s_f(x)-v_f(x)\big\rangle}$$

towards vertex

away vertex

Part I: Adaptive quadrature rule with Frank-Wolfe optimization

- Approximating integrals: $\int_{\mathcal{X}} f(x)p(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)})$

  for **fixed** p, and multiple f's in a RKHS $\mathcal{H}$

  - Random sampling $x^{(i)} \sim p(x)$ yields $O(1/\sqrt{N})$ error
  - Kernel herding [Chen et al. 10] (can) yield $O(1/N)$ error! (like quasi-MC)
    (need finite dim. $\mathcal{H}$)
  - -> generalized to FW optimization [Bach et al. 12] and could even get $O(e^{-cN})$ error

- Trick: run Frank-Wolfe optimization on dummy objective:

  where $\mathcal{M} = \text{cl-conv}(\Phi(\mathcal{X}))$
  is the marginal polytope

  $$\min_{g \in \mathcal{M}} \frac{1}{2}\|g - \mu(p)\|_{\mathcal{H}}^2$$

  and $\mu(p) = \mathbb{E}_{p(x)}\Phi(x)$ is the *mean map*
  representer: $k(x, \cdot) \in \mathcal{H}$

Approx. integrals in RKHS

- Why? Well, controlling **moment discrepancy** $\|\mu(\hat{p}) - \mu(p)\|_{\mathcal{H}}$ is enough to control **error of integrals** in RKHS H :

- Reproducing property: $f \in \mathcal{H} \Rightarrow f(x) = \langle f, \Phi(x) \rangle$
- Define *mean map* : $\mu(p) = \mathbb{E}_{p(x)} \Phi(x)$
- Want to approximate integrals of the form:

$$\mathbb{E}_{p(x)} f(x) = \mathbb{E}_{p(x)} \langle f, \Phi(x) \rangle = \langle f, \mu(p) \rangle$$

- Use weighted sum to get approximated mean: $\hat{p} = \sum_{i=1}^{N} w_t^{(i)} \delta_{x^{(i)}}$

$$\mu(\hat{p}) = \mathbb{E}_{\hat{p}(x)} \Phi(x) = \sum_{i=1}^{N} w^{(i)} \Phi(x^{(i)}) \Rightarrow \mathbb{E}_{\hat{p}(x)} f(x) = \sum_{i=1}^{N} w^{(i)} f(x^{(i)})$$

- Approximation error is then bounded by:

$$|\mathbb{E}_{p(x)} f(x) - \mathbb{E}_{\hat{p}(x)} f(x)| \leq \|f\|_{\mathcal{H}} \, \|\mu(p) - \mu(\hat{p})\|_{\mathcal{H}}$$

FW quadrature

- Run Frank-Wolfe optimization on dummy objective:

$$\min_{g \in \mathcal{M}} \frac{1}{2} \|g - \mu(p)\|_{\mathcal{H}}^2$$

where $\mathcal{M} = \text{cl-conv}\,(\Phi(\mathcal{X}))$ is the marginal polytope

and $\mu(p) = \mathbb{E}_{p(x)} \Phi(x)$ is the *mean map*

representer: $k(x, \cdot) \in \mathcal{H}$

FW-Quad

repeat: input: p

1) FW search:

$$x^{(k+1)} = \arg\min_{x \in \mathcal{X}} \; g_k(x) - \mu(p)(x)$$

2) convex combo:

$$g_{k+1} = (1 - \gamma_k)\, g_k + \gamma_k\, \Phi(x^{(k+1)})$$

e.g. minimum of a difference of mixture of Gaussian bumps!
(for a Gaussian kernel)

output:

$$g_N = \sum_{i=1}^{N} w^{(i)} \Phi(x^{(i)}) = \mu(\widehat{p}_N)$$

- Requirements: can compute $\mu(p)$

  + approx. solve (1) -> use exhaustive search through M random samples from p

  -> **super-samples** selection [Chen et al. 10]