today: • properties of FW
• AFW alg.

properties of FW

1) $f(x_t) - \underbrace{\min_{x \in M} f(x)}_{f^*}$ $\quad O\left(\frac{1}{t}\right)$

2) FW-gap $g_t \geq f(x_t) - f^* \rightarrow$ certificate of subopt.

$\min_{s \leq t} g_s \leq O\left(\frac{1}{t}\right)$ [ie. will stop in $O(\frac{1}{\varepsilon})$ iterations?]

3) $x_t = \rho_0^t x_0 + \sum_{u=1}^{t} \rho_u^t s_{u-1}$ $\rightsquigarrow x_t$ has a t "sparse" expansion in terms of the FW corners $\{s_u\}_{u=0}^{t-1}$

where $\sum_u \rho_u^t = 1$, $\rho_u^t \geq 0$

"sparse methods"
→ popular in ML

→ see later how to apply on dual of SVMstruct

4) there is a $\Omega\left(\frac{1}{t}\right)$ lower bound for FW-like methods for $t \leq d$

5) FW is affine covariant (like Newton's method)

Let $\tilde{M}$ be a new constraint set s.t. $\tilde{M} \xrightarrow{A} M$ (linear transformation & surjective) ie. $M = A\tilde{M}$

define $\tilde{f}(\tilde{x}) \triangleq f(A\tilde{x})$

$$\min_{\tilde{x} \in \tilde{M}} \tilde{f}(\tilde{x}) = \min_{\tilde{x} \in \tilde{M}} f(A\tilde{x}) = \min_{x \in A\tilde{M}} f(x), \quad A\tilde{M} = M$$

$$\boxed{\begin{tikzcd} x_t & \tilde{x}_t \arrow[l, "A"'] \\ f \,\&\, M \arrow[u, "FW"] & \tilde{f} \,\&\, \tilde{M} \arrow[l, "A"'] \arrow[u] \end{tikzcd}}$$

affine covariance

If run FW on $\tilde{f}$ & $\tilde{M}$ to get $\tilde{x}_t$ as iterates then $x_t \triangleq A\tilde{x}_t$ corresponds to running FW on f & M

(modulo tie breaking)

why? → inner product with gradient

$$\tilde{s}_t = \arg\min_{\tilde{s} \in \tilde{M}} \langle \tilde{s}, \nabla_{\tilde{x}} \tilde{f}(\tilde{x}) \rangle \qquad x_t \triangleq A\tilde{x}_t$$

$$\langle \tilde{s}, A^T \nabla_x f(x_t) \rangle$$

$$\langle \underbrace{A\tilde{s}}_{s = A\tilde{s}}, \nabla_x f(x_t) \rangle$$

$$s_t = \arg\min_{s \in M = A\tilde{M}} \langle s, \nabla_x f(x_t) \rangle \implies \underline{s_t = A\tilde{s}_t} \quad \text{(modulo tie breaking)}$$

$$S_t = A\tilde{S}_t \quad \text{set of argmin (ties)}$$

$\implies$ we want an <u>affine invariant</u> analysis

$$\leadsto C_f \leq L_{\|\cdot\|} (\mathrm{diam}_{\|\cdot\|}(M))^2 \text{ for any } \|\cdot\|$$

affine invariant constant (we'll see next class)

6) Convergence of FW for non-convex f

<u>aside:</u> necessary first order condition for constrained optimum

$$\min_{x \in M} f(x) \qquad x^* \text{ is a local min}$$

$$\implies \langle \nabla f(x^*), s - x^* \rangle \geq 0 \quad \forall s \in M$$

"stationary pt." for const. opt. problem

$$\iff \min_{s \in M} \langle \nabla f(x^*), s - x^* \rangle \geq 0$$

$$\iff \underbrace{\max_{s \in M} \langle -\nabla f(x^*), s - x^* \rangle}_{\text{FW-gap}(x^*)} \leq 0$$

(if f & M are convex, then it is a sufficient <u>condition</u> for global min)

FW-gap(x^*) quantify the "non-stationarity"

see Lacoste-Julien 2016 arXiv

$$\boxed{\min_{s \leq t} \text{gap}(x_s) \leq O\left(\frac{1}{\sqrt{t}}\right) \text{ for FW with line search}}$$

"non-convex FW"

when f is L-smooth, M is <u>bounded</u> & convex <u>but</u> f is not nec. convex

$\implies f$ is <u>lower bounded</u>

aside: if f is <u>concave</u>, $O\left(\frac{1}{t}\right)$

14h30

away-step FW

comment: $x_t = \sum_u \rho_u^t s_u$ ("coordinate")

$$x_{t+1} = (1-\gamma_t)x_t + \gamma_t s_t$$

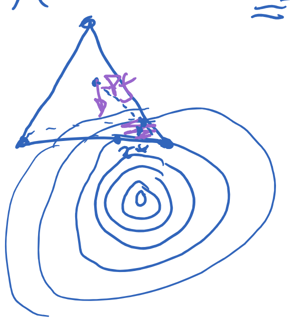
$$= \sum_u \underbrace{\rho_u^t (1-\gamma_t)}_{\rho_u^{t+1}} s_u + \gamma_t s_t$$

ρ_u^{t+1} → previous coordinates shrunk by $(1-\gamma_t)$

* FW step moves mass uniformly away from active set $\{s_u\}_{u=1}^t$ to FW corner s_t

→ unless step-size $\gamma_t = 1$, FW never removes a corner from active set / expansion

M

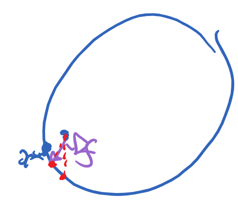


⇒ zig-zag phenomenon close to the boundary on polytopes

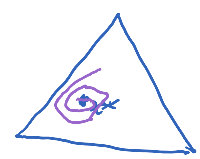
(this is why you get a $\Omega\left(\frac{1}{t}\right)$ rate even if f is μ-strongly convex)

$$\left\langle \frac{-\nabla f(x_t)}{\|\nabla f(x_t)\|}, \frac{d_t}{\|d_t\|} \right\rangle \xrightarrow{t\to\infty} 0$$

* but FW has no problem
  - on "strongly convex sets" ↳ sublevel set of a strongly convex set.



  - when sol'n is in the relative interior of M



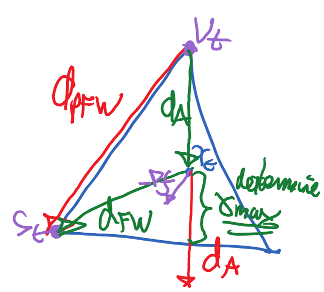
away-step FW fix (solves zig-zagging problem)

in addition to compute FW corner

$$s_t = \underset{s \in M}{\operatorname{argmin}} \langle s, \nabla f(x_t) \rangle$$

also compute the "away corner"

$$v_t = \underset{s \in \text{active set}(x_t)}{\operatorname{argmax}} \langle s, \nabla f(x_t) \rangle$$



$v_t = \arg\max_{s \in \text{active set}(x_t)} \langle s, \nabla f(x_t) \rangle$ $\quad \downarrow d_A$

→ (aside: Wolfe's original alg. used whole $\mathcal{M}$ ⟹ non-convergent alg.)

$d_{FW} = s_t - x_t$

$d_A \triangleq x_t - v_t$

* AFW pick direction best inner-product with $-\nabla f$

i.e. pick d_A if $\underbrace{\langle d_A, -\nabla f(x_t) \rangle}_{\text{"}g_A\text{"}} > \overbrace{\langle d_{FW}, -\nabla f(x_t) \rangle}^{g_{FW}}$

o.w. pick d_{FW}

$g_A = 0$ when $\nabla f(x)$ is $\perp$ active set

* if use d_A, let $\gamma_t = \arg\min_{\gamma \in [0, \gamma_{max}]} f(x_t + \gamma d_A)$

where γ_{max} depends on ρ_i^t coefficients

i.e. suppose $x_t = \sum_u \alpha_u s_u$

$x_{t+1} = x_t + \gamma_t (\overbrace{x_t - v_t}^{d_A})$

$= \sum_u (1+\gamma_t)\alpha_u s_u - \gamma_t v_t$

let α be coeff. of v_t in expansion of x_t $(\alpha > 0)$

$(1+\gamma)\alpha - \gamma \geq 0$

$(1+\gamma_{max})\alpha - \gamma_{max} = 0 \Rightarrow \boxed{\gamma_{max} = \frac{\alpha}{1-\alpha}}$

* when $\gamma_t = \gamma_{max}$

call this a "drop step" → removing v_t from expansion

* when run AFW alg.:

either you maintain same expansion $x_t = \sum_u \rho_u s_u$

or

you have a $\underbrace{\text{feasibility oracle}}_{\text{get } \gamma_{max}}$ + $\underbrace{\text{away-step oracle}}_{\text{get } v_t}$

[see Neurips 2016 paper Meshi & Garber]

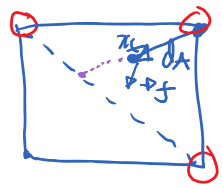
→ assumption needed for convergence ...

↳ see Neurips 2016 paper Meshi & Garber]

↳ they assume $\text{corners}(M) \subseteq \{0,1\}^d$ → assumption needed for convergence result

and $\boxed{M \text{ is described as } Ax = b,\ x \geq 0}$

assumption to run alg



AFW has linear convergence rate on polytopes when f is strongly convex

(vs. FW → $O(\frac{1}{t})$)

(*) combine d_{FW} & d_A → pairwise FW direction

$$d_{PFW} = d_{FW} + d_A = s_t - \cancel{x_t} + \cancel{x_t} - v_t = \underbrace{s_t - v_t}_{d_{PFW}}$$

$$\underbrace{\langle -\nabla f(x_t), d_{PFW} \rangle}_{g_{PFW}} = g_{FW} + g_A \qquad g_{FW} + g_A \leq 2 \max\{\ldots\}$$

during AFW: $g_t = \langle -\nabla f(x_t), d_t \rangle = \max\{g_{FW}, g_A\}$

$$\geq \frac{g_{PFW}}{2}$$