Lecture 16 - FW

Thursday, February 28, 2019 13:39

today: • away-step FW
• FW convergence & properties

away-step FW

comment: $x_t = \sum_u p_u^t s_u$

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

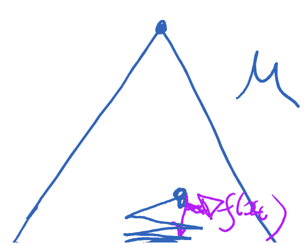
$$= \sum_u \underbrace{p_u^t (1-\gamma_t)}_{p_u^{t+1}} s_u + \gamma_t\, s_t$$

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

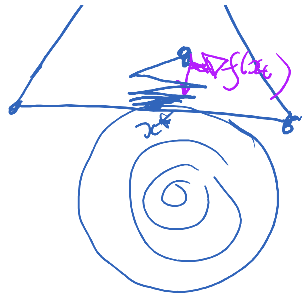
* FW step moves mass uniformly away from active set $\{s_u\}_{u=1}^t$ to FW corner s_t

→ unless step size $\gamma_t = 1$, FW is <u>never</u> removing a corner from its expansion

⇒ Zig-zag phenomenon close to boundary on polytopes (this is why $O(\frac{1}{t})$ rate <u>even</u> if f is μ-strongly convex)



$$\frac{\langle -\nabla f(x_t), d_t \rangle}{\|\nabla f(x_t)\| \|d_t\|} \to 0 \quad \text{as } t \to \infty \Rightarrow \text{slow rate}$$



$\langle -\nabla f(x_t), d_t \rangle \to 0$ as $t \to \infty$

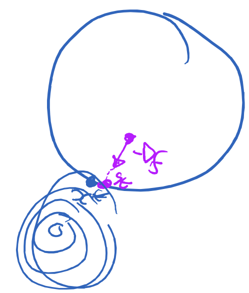
$\|\nabla f(x_t)\| \|d_t\| \Rightarrow$ slow rate

for strongly convex fct.

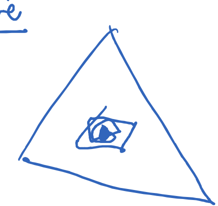
$O(\exp(-\rho t))$ rate

* but FW has no problem → on "strongly convex sets"

↳ sublevel set of a strongly convex fct,



→ when solution is in the relative interior of $\mathcal{M}$



||

<u>away-step FW fix</u>: (solve the zig-zagging problem)

in addition to compute $s_t = \underset{s \in \mathcal{M}}{\arg\min} \langle \nabla f(x_t), s \rangle$

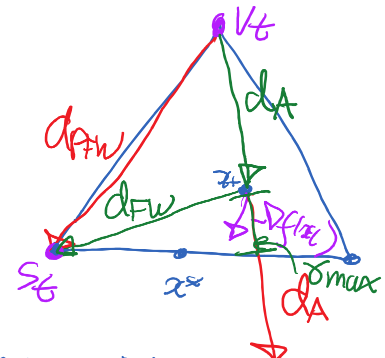
compute $v_t = \underset{s \in \text{active set}(x_t)}{\arg\max} \langle \nabla f(x_t), s \rangle$

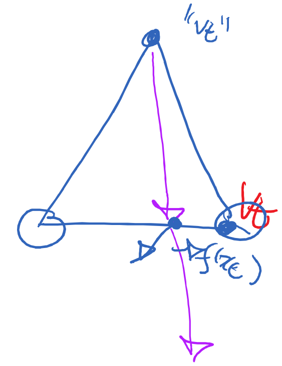
↳ (aside: Wolfe's original alg. used whole $\mathcal{M}$ here ⇒ non-convergent)

$d_{FW} \triangleq s_t - x_t$

$d_A \triangleq x_t - v_t$

g_{FW}





*AFW picks the direction with best inner product

$d_A = x_t - v_t$ $\qquad\qquad$ → non-convergent)

* AFW picks the direction with best inner-product

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 can be zero when $\nabla f(x_t)$ is $\perp$ to active set

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

where γ_{max} depends on α_u^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 coefficient of v_t

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

$$\rightarrow \boxed{\gamma_{max} = \frac{\alpha}{1-\alpha}}$$

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

"drop step" → removed v_t from expansion

(*) when run AFW:

either you maintain some expansion of $x_t = \sum_u \alpha_u s_u$

or

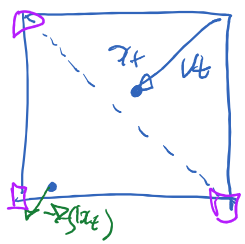
you have a feasibility oracle + away-step oracle

[see NIPS 2016 paper Meshi & Garber]

↳ they assume $\underbrace{\text{corners}(M) \subseteq \{0,1\}^d}_{\text{assumption to prove convergence}}$

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

↳ assumption to run alg.



AFW has <u>linear convergence</u> rate on <u>polytopes</u> when f is strongly convex

14h36

* <u>combining d_{FW} & d_A → pairwise FW</u>

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

$$\langle -\nabla f(x_t), d_{PFW} \rangle = g_{FW} + g_A$$

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

$g_t \geq g_{PFW}/2$

(*) <u>note</u>: if $M = \mathrm{conv}(\mathcal{A})$ where $\mathcal{A}$ is some finite set ("atoms" called)

$$LMO(r): \min_{s \in M = \mathrm{conv}(\mathcal{A})} \langle s, r \rangle = \min_{a \in \mathcal{A}} \langle a, r \rangle$$

↳ lot of application in ML

e.g.: • $\mathcal{A} \to$ integer flows — $LMO \to$ min cost network flow
$\mathrm{conv}(\mathcal{A}) \to$ flow polytope

• $\mathcal{A} \to$ clique assignment in graph — $LMO \to$ max-product or graph-cut alg. for submodular potentials (see later)
$\mathrm{conv}(\mathcal{A}) \to$ marginal polytope

<u>more properties of FW</u>

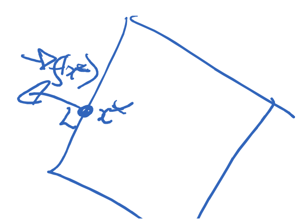
6) <u>Convergence for non-convex f</u>:

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

$\min_{x \in M} f(x)$ $\quad x^*$ global min

$\Rightarrow \langle \nabla f(x^*), s - x^* \rangle \geq 0 \quad \forall s \in M$ "stationary pt" for const. opt. problem

$\Leftrightarrow \min (\quad) \geq 0$



$\Leftrightarrow \min_{s \in M} \big(\qquad \big) \geq 0$

$\Leftrightarrow \underbrace{\max_{s \in M} \langle -\nabla f(x^*), s - x^* \rangle}_{\text{FW-gap}(x^*)} \leq 0$ (sufficient cond. if f & M are convex)

↳ quantify the "non-stationarity"

see L-J. 2016 arxiv

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

"non-convex" FW

for FW with line search and f L-smooth and M bounded & convex but f is not nec. convex

4) affine covariance of FW

let $\tilde{M}$ be a new domain s.t. $\tilde{M} \xrightarrow{A} M$ (linear transf. surjective) i.e. $M = A\tilde{M}$

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

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

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

$$A\tilde{M} = M$$

affine covariance of FW: if run FW on $\tilde{f}$ & $\tilde{M}$ to get $\tilde{x}_t$ iterates

then $x_t \triangleq A\tilde{x}_t$ corresponds to running FW on f & M (modulo tie-breaking) (?)

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

why? $\rightarrow$ inner product is affine invariant $\nabla_{\tilde{x}} \tilde{f}(\tilde{x}) = \nabla_{\tilde{x}} f(A\tilde{x}) = A^T \nabla_x f(x(\tilde{x}))$

$$\tilde{s}_t = \operatorname*{argmin}_{\tilde{s}\in\tilde{M}} \langle \tilde{s}, \nabla_{\tilde{x}} \tilde{f}(\tilde{x}_t)\rangle \qquad x_t = A\tilde{x}_t$$

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

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

$$s_t = \operatorname*{argmin}_{s\in M} \langle s, \nabla_x f(x_t)\rangle \qquad M = A\tilde{M}$$

$$\underline{\underline{s_t = A\tilde{s}_t}} \quad \text{(modulo tie breaking)}$$

$$S_t = A\tilde{S}_t$$

$\Rightarrow$ we want affine <u>invariant analysis</u>

Curvature constant

$$C_f \triangleq \sup_{\substack{\gamma \in [0,1] \\ x, s \in \mathcal{M} \\ x_\gamma = (1-\gamma)x + \gamma s}} \frac{2}{\gamma^2} \Big[f(x_\gamma) - \big(f(x) + \underbrace{\langle \nabla f(x), x_\gamma - x \rangle}_{} \big) \Big]$$

x_γ: potential FW step; γ: step-size

worst-case deviation from linear approximation

this is affine invariant $\Rightarrow$ C_f is aff. inv.

$\rightarrow$ we'll get an affine invariant descent lemma