

Lecture 16 - FW

Thursday, March 12, 2020 13:35

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

away-step FW

comment: $x_t = \sum_u \overset{\text{"coordinate"}}{\beta_u^t} s_u$

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

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

→ previous coordinates shrunk by $(1-\alpha_t)$

* FW step moves mass uniformly away from active set $\sum_{u=1}^k \beta_u^t$ to FW core s_t

→ unless step size $\alpha_t = 1$, FW never removes a coord from its expansion

⇒ zig-zag phenomenon close to boundary on polytopes



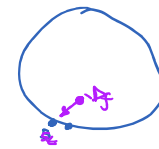
(this is why $\Omega(\frac{1}{\epsilon})$ rate even if f is μ -strongly convex)

$$\left\langle \frac{-\nabla f(x_t)}{\|\nabla f(x_t)\|}, \frac{dx_t}{dt} \right\rangle \rightarrow 0 \text{ as } t \rightarrow \infty$$

⇒ slow rate

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

↳ subset of a strongly convex set.



for strongly convex: $O(\log(1/\epsilon))$ rate

when sol'n is in the relative interior of M



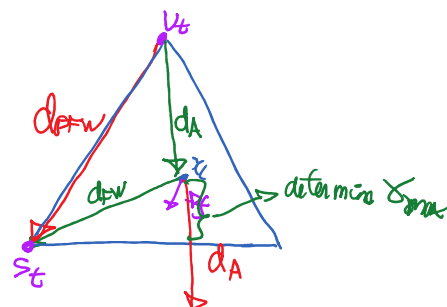
away-step FW fix: (solves zig-zagging problem)

in addition to compute

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

also

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



also

compute $v_t \triangleq \arg \max_{S \in \text{activeset}(x_t)} \langle \nabla f(x_t), S \rangle$



$d_{FW} \triangleq s_t - x_t$

$d_A \triangleq x_t - v_t$

(aside: Wolfe's original dg. used a prob M here \Rightarrow non-convergent dg.)

* AFW picks the direction with best inner-product

i.e. pick d_A if $\langle d_A, -\nabla f(x_t) \rangle > \langle d_{FW}, -\nabla f(x_t) \rangle$

o.w. pick d_{FW} "g*" d_A can be zero when $\nabla f(x_t)$ is \perp to 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 μ^t coefficients

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

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

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

let α be coefficient of v_t

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

$\Rightarrow \gamma_{max} = \frac{\alpha}{1 - \alpha}$

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

"drop step" \rightarrow remove v_t from expansion

⊗ when run AFW:

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

or

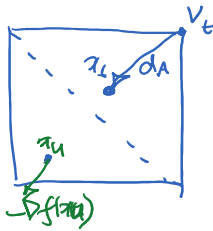
you have a feasibility oracle + away-step oracle

[see NIPS 2016 paper by Meshi & Gonen]

assumption to prove convergence

\hookrightarrow they assume $\text{convex}(M) \subseteq \{0, 1\}^d$





↳ they assume $\text{conv}(M) \subseteq \mathbb{E}_{0,1}^n$

and M is described as $Ax=b$
 $x \geq 0$

↳ assumption to run alg.

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

14h35

* combine d_{FW} & d_A → pairwise FW

$$d_{FW} + d_A = c_t - x_t + x_t - v_t = \underbrace{c_t - v_t}_{d_{FW}}$$

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

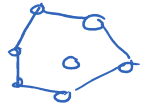
during AFW: $Q_t = \langle -\nabla f(x_t), d_t \rangle = \max \{g_{FW}, g_A\}$

$$Q_t \geq \frac{Q_{FW}}{2}$$

(*) note: if $M = \text{conv}(A)$ where A is some finite set (called "atoms")

$$\text{LMO}(r) : \min_{S \in M = \text{conv}(A)} \langle S, r \rangle = \min_{a \in A} \langle a, r \rangle$$

↳ lot of application in ML



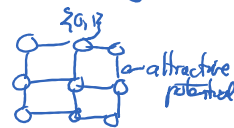
e.g. $A \rightarrow$ integer flows
 $\text{conv}(A) \rightarrow$ flow polytope

LMO → min cost network flow

$A \rightarrow$ clique assignment in graph $(S_{y_c})_{c \in C}$ LMO → max product alg.

$\text{conv}(A) \rightarrow$ marginal polytope

or Graph-ut alg. for submodular potentials (see later)



more properties of FW

(c) convergence for non-convex f:

aside: necessary first order condition for constrained opt.

$$\min_{x \in M} f(x)$$

x^* is a local min

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

'stationary pt.' for const. opt. problem



$$\min_{x \in M} f(x)$$

x^* is a local min

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

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

$$\Leftrightarrow \max_{s \in M} \langle -\nabla f(x^*), s - x^* \rangle \leq 0$$

FW-gap(x^*)

↳ quantify the "non-stationarity"

minimizing pt. for const. opt. problem



(if f and M are convex, this is a sufficient cond.)

see L-J. 2016 contriv

$$\min_{s \in M} \text{gap}(x_s) \leq O\left(\frac{1}{\sqrt{t}}\right) \quad \text{for FW with one search}$$

f is L -smooth and lower bounded
 M is bounded & convex
 but f is not nec. convex

"non-convex" FW

4) affine covariance of FW

let \tilde{M} be a new domain s.t. $\tilde{M} \xrightarrow{A} M$ ie. $M = A\tilde{M}$

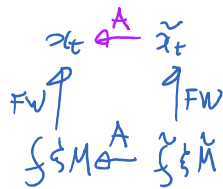
define $\tilde{f}(\tilde{x}) \triangleq f(A\tilde{x})$

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

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)

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

$$\tilde{s}_t = \underset{s \in \tilde{M}}{\text{argmin}} \langle \tilde{s}, \nabla_{\tilde{x}} \tilde{f}(\tilde{x}_t) \rangle \quad x_t \triangleq A\tilde{x}_t$$

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

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

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

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

to find \tilde{A} $\tilde{S}_t = A \tilde{S}_t$

\Rightarrow we want affine invariant analysis $\leadsto C_S \in L_{1,1}(\text{dom}_{1,1}(M))^2$