Lecture 16 - FW & AFW

Tuesday, March 16, 2021 14:19

today: • properties of FW
• AFW alg.

Properties of FW

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

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

$\min_{s \leq t} g_s \leq O(\frac{1}{t})$ [i.e. will stop in $O(\frac{1}{\varepsilon})$ iterations!]

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

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

"sparse method" → popular in ML

→ see later how ~~to~~ apply on dual of SVM struct

4) there is a $\Omega(\frac{1}{t})$ 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 trans & surjective, i.e. $M = A\tilde{M}$

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

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

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

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

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

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

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

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

$$\underset{\text{set of ties}}{S_t} = A\tilde{S}_t$$

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

$$\leadsto \underset{\text{affine invariant constant (we'll see later)}}{C_f} \leq L_{\|\cdot\|}\left(\mathrm{diam}_{\|\cdot\|}(M)\right)^2 \text{ for any } \|\cdot\|$$

6) <u>convergence of FW for non-convex f</u>

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

$\min_{x\in M} f(x)$ $\quad x^*$ is a <u>local min</u>

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

"<u>stationary pt.</u>" for const. opt. problem

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

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

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

$\leftarrow$ quantify the "non-stationarity"

see L.-J. 2016 arXiv

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

for FW with line search

"non-convex FW"

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

15h23

<u>away-step FW</u>

<u>comment</u>: $x_t = \sum_u \underset{\text{"coordinate"}}{\rho_u^t}\, s_u$

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

$$= \sum \underline{\rho_u^t (1-\gamma_t)}\, s_u + \gamma_t s_t$$

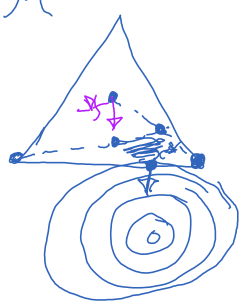
ρ_u^{t+1} → previous coordinates shrunk by $(1-\gamma_t)$

* FW step moves mass uniformly away from active set $\{s_u\}_{u\in...}$ to FW corner s_t

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

⇒ zig-zag phenomenon close to boundary on polytopes

M



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

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

* but FW has no problem

- on "strongly convex" sets
  - → sublevel set of a strongly convex convex set.

⇒ linear convergence rate $O(\exp(-\delta t))$

- when soln is in the relative interior of M

for f is strongly convex ⇒ " linear rate

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

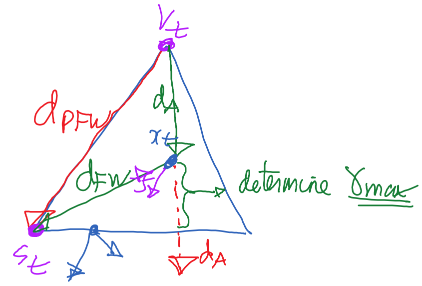
in addition to compute FW corner

$$s_t = \arg\min_{s\in M} \langle s, \nabla f(x_t)\rangle$$

also compute the "away corner"

$$v_t = \arg\max_{s\in \text{activeset}(x_t)} \langle s, \nabla f(x_t)\rangle$$

→ (aside: Wolfe's original alg. used whole M ⇒ non-convergent alg.)



$d_{FW} = s_t - x_t$

$d_A \triangleq x_t - v_t$

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

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

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

o.w. pick d_{FW}

g_A can be zero when $\nabla f(x_t)$ 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_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 coeff. of v_t in expansion for x_t

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

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

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

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

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

* when run AFW algo:

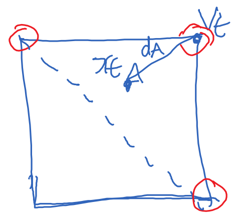
either you maintain some expansion $x_t = \sum_u \rho_u s_u$

or

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

[see NIPS 2016 paper by Meshi & Gerber]

↳ 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 algo.

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

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

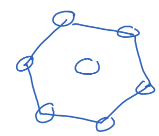
$$\underbrace{\langle -\nabla f(x_t), d_{PFW} \rangle}_{g_{PFW}} = g_{FW} + g_A \qquad g_{FW} + g_A \leq 2 \max\{\dots\}$$

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

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

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

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



$\hookrightarrow$ lot of application in ML where LMO is efficient

e.g.: $\mathcal{A} \leftrightarrow$ integer flows

$\text{conv}(\mathcal{A}) \to$ flow polytope

LMO $\to$ min cost network algorithm

- $\mathcal{A} \to$ clique assignments in graph $(\delta_{y_c})_{c \in \mathcal{C}}$

$\text{conv}(\mathcal{A}) \to$ marginal polytope

LMO $\to$ max product alg.

or

graph cut alg. for Ising model with attractive potential ("Associative Markov network") ($\to$ "submodular potentials") (see later)