Lecture 16 - FW & AFW

Tuesday, March 19, 2024 9:32 AM

today: • properties of FW
• AFW alg.

properties of FW

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

1) $f(x_t) - \min_{x \in M} f(x) \leq O(\frac{1}{t})$
$\triangleq f^*$

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

$\min_{s \leq t} g_s \leq O(\frac{1}{t})$

3) $x_t = \beta_0^t x_0 + \sum_{u=1}^{t} \rho_u^t s_{u-1}$ ⇝ 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

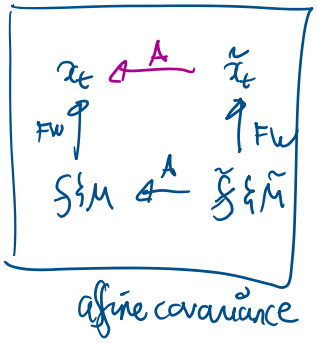
4) there is a $\Omega(\frac{1}{t})$ lower bound for FW-like methods for $t \leq d$

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 & surjective) i.e. $M = A\tilde{M}$

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

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



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 is affine invariant

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

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

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

$x_t \triangleq A\tilde{x}_t$

$\Rightarrow s_t = A\tilde{s}_t$ (modulo tie breaking)

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

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

$\Rightarrow$ we want an affine invariant analysis

$\rightsquigarrow C_f \leq L_{\|\cdot\|} \left(\operatorname{diam}_{\|\cdot\|}(M)\right)^2$ for any $\|\cdot\|$

C_f: affine invariant constant (will see later)

6) Convergence of FW for non-convex f

aside: necessary first order condition for constrained opt. $\min_{x \in M} f(x)$

x^* is a local min $\Rightarrow \langle \nabla f(x^*), s - x^* \rangle \geq 0 \quad \forall s \in M$ (feasible direction)

"stationary pt." for const. opt. problem

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

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

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

(If M & f are convex, then this is a sufficient cond. for global min)

see L.-J. 2016 arXiv

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

"non convex FW"

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

lower bound of f on M

$O\left(\frac{1}{t}\right)$ if f is concave

10/19

away-step FW

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

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

$= \sum_u \rho_u^t (1-\gamma_t) s_u + \gamma_t s_t$

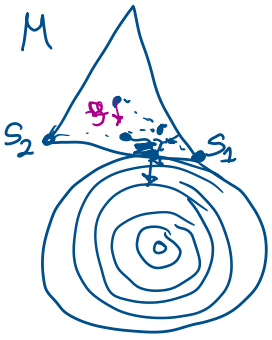
ρ_u^{t+1} → previous coordinates shrunk by $(1-\gamma_t)$

* FW step moves mass uniformly away from active set $\{s_u : \rho_u > 0\}$ 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

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



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

⇒ no linear convergence rate for FW

* but FW has no problem on "strongly convex sets" → subset of a strongly convex set

⇒ linear convergence rate $O(\exp(-\rho t))$ when f is strongly convex

when set'n is the relative interior of M

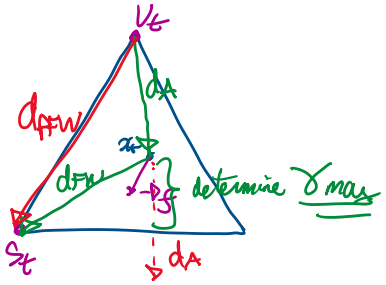
~~Away-Step FW~~: (solves zig-zagging problem)

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"

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



$d_{FW} = s_t - x_t$

$d_A \triangleq x_t - v_t$

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

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

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

o.w. pick d_{FW}

* 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 β_i^t coefficients

ie. suppose $x_t = \sum_i \alpha_i s_i$

d_A

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

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

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

let α be the coeff. of v_t in expansion of x_t

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

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

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

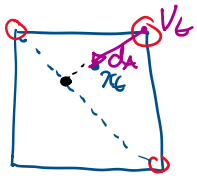
call this a "drop step" → removing v_t from expansion

④ when run AFW alg.:

either you maintain some expansion $x_t = \sum_u \alpha_u s_u$ } this is "storing" by being conservative

or

you have a feasibility oracle (γ_{max}) + away-step oracle (get v_t)



[see NeurIPS 2016 paper by Meshi & Gerber]

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

and M is described as $Ax = b$, $x \geq 0$ ← assumption to run alg.

AFW has linear convergence rate on polytopes when f is μ-strongly convex (vs. FW → $O(\frac{1}{t})$)

⑤ combine d_{FW} & d_A → pairwise FW direction

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

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

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

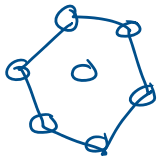
$$g_t \geq \frac{g_{PFW}}{2}$$

⑥ note: if $M = \text{conv}(A)$ where A is some finite set (called "atoms")

$\mathcal{M} = \overline{\text{conv}}(\mathcal{A})$

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

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



↳ lots of applications in ML where LMO is efficient

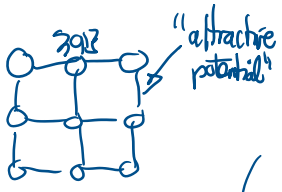
e.g. • $\mathcal{A}$ → integer flows

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

LMO → min cost network flow alg.

• $\mathcal{A}$ → clique assignments on a graph $(\delta_{y_c})_{c \in \mathcal{C}}$

$\text{conv}(\mathcal{A})$ → marginal polytope LMO → max product alg.



or graph cut. alg. for Ising model with attractive potentials

(→ "submodular potentials" (see later))