Lecture 16 - scribbles - Frank-Wolfe

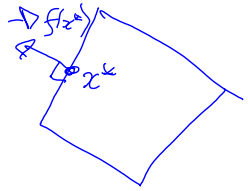
Friday, March 16, 2018 14:27

today: FW convergence & properties
FW for SVMstruct

aside: necessary first order optimality condition for constrained opt.

$\min_{x \in M} f(x)$ x^* global min (i.e. f looks increasing in all feasible directions)

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



"stationarity pt." for const. opt. problem

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

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

(sufficient cond. if f & M are convex)

FW-gap(x^*)

i.e. want FW-gap(x^*) = 0 since gap ≥ 0

quantifying "non-stationarity"

see L-J. 2016 onsite,

$\min_{s \leq t}$ gap(x_s) $\leq O\left(\frac{1}{\sqrt{t}}\right)$ for FW with line search and f L-smooth and M bounded & convex but f not nec. convex

"non-convex FW"

4) affine co-variance of FW

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

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

$x(\tilde{x})$

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

$A\tilde{M} = M$

affine covariance 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)

$$\begin{array}{ccc} x_t & \xleftarrow{A} & \tilde{x}_t \\ \text{FW}\uparrow & & \uparrow\text{FW} \\ f\,\&\,M & \xleftarrow[A]{} & \tilde{f}\,\&\,\tilde{M} \end{array}$$

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

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

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

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

$s = A\tilde{s}$

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

$A\tilde{M}$

only first-order method I know with this property (like Newton which is 2nd order)

⇒ we want affine invariant analysis i.e. $f(x_t) - f(x^*) = f(\tilde{x}_t) - f(\tilde{x}^*)$

this depends on $\langle \nabla f(x), s - x \rangle$ is affine invariant

so C_f is " "

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

- x_γ: potential FW step
- γ: step-size
- $f(x_\gamma) - (f(x) + \langle \nabla f(x), x_\gamma - x \rangle)$: worst-case deviation from linear approximation

by descent lemma, if ∇f is L-Lipschitz i.e. $\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\| \quad \forall x, y \in \mathcal{M}$

$$\|d\|_* \triangleq \sup_{\|x\| \leq 1} \langle d, x \rangle$$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$$

⇒ generalized C-S $\quad |\langle d, x \rangle| \leq \|d\|_* \|x\|$

$$\Rightarrow C_f \leq \sup_{\gamma} \frac{2}{\gamma^2} \left[\frac{L}{2} \underbrace{\|x_\gamma - x\|^2}_{\|\gamma(s-x)\|^2} \right]$$

$$\leq L \sup_{x, s \in \mathcal{M}} \|s - x\|^2$$

$(\|\cdot\|_p)_* = \|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$

$$\operatorname{diam}_{\|\cdot\|}(\mathcal{M}) \triangleq \sup_{x, s} \|x - s\|$$

$$C_f \leq L_{\|\cdot\|} \cdot \operatorname{diam}_{\|\cdot\|}(\mathcal{M})^2$$

- C_f: affine invariant
- $L_{\|\cdot\|}$, $\operatorname{diam}_{\|\cdot\|}(\mathcal{M})^2$: depends on $\|\cdot\|$ and thus A

④ by def. of C_f, we get affine invariant version of descent lemma:

$$f(x_\gamma) \le f(x) + \gamma \langle \nabla f(x), s-x \rangle + \frac{\gamma^2}{2} C_f \qquad \forall \gamma \in [0,1] \quad \forall x, s \in M$$

let $x = x_t$ $s = s_t$ FW corner

$$\langle \nabla f(x_t), s_t - x \rangle = -g_t$$

for FW-step with stepsize γ,

$$(\dagger) \quad \boxed{f(x_\gamma) \le f(x) - \gamma g_t + \frac{\gamma^2}{2} C_f}$$

optimal step-size minimizes the RHS

$$\gamma^* = \min\left\{ \frac{g_t}{C_f}, 1 \right\}$$

adaptive step-size

aff. inv.

$$f(x_{\gamma^*}) \le f(x_t) - \frac{g_t^2}{2C_f}$$

$$\le f(x_t) - \frac{\varepsilon_t^2}{2C_f} \quad \text{where } \varepsilon_t \triangleq f(x_t) - f(x^*) \le g_t$$

15h37

thm: FW alg. with γ_t chosen either $2/(t+2)$, g_t/C_f, line search yields $\varepsilon_t \le \frac{2C_f}{t+2}$ (when f is convex)

proof: let $x_\gamma = x_t + \gamma(s - x_t)$ + apply $(\dagger)$

$$f(x_{t+1}) \le f(x_t) - \gamma g_t + \frac{\gamma^2 C_f}{2} \qquad g_t \ge \varepsilon_t \text{ by convexity}$$

$$\le f(x_t) - \gamma \varepsilon_t + \frac{\gamma^2 C_f}{2}$$

$$\underbrace{f(x_{t+1}) - f^*}_{\varepsilon_{t+1}} \le \underbrace{f(x_t) - f^*}_{\varepsilon_t} - \gamma_t \varepsilon_t + \frac{\gamma_t^2 C_f}{2}$$

$$\boxed{\varepsilon_{t+1} \le (1 - \gamma_t)\varepsilon_t + \frac{\gamma_t^2 C_f}{2}}$$

- See notes last year for cool ODE trick + induction

here, brute force approach to solve the recurrence:

$$\varepsilon_{t+1} \le (1-\gamma_t)\varepsilon_t + \frac{\gamma_t^2 C_f}{2}$$

$$\le (1-\gamma_t)\left[(1-\gamma_{t-1})\varepsilon_{t-1} + \frac{\gamma_{t-1}^2 C_f}{2}\right] + \frac{\gamma_t^2 C_f}{2}$$

$$\le \cdots$$

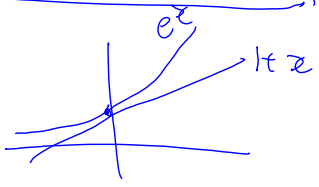
$$\boxed{\varepsilon_{t+1} \le \prod_{s=0}^{t}(1-\gamma_s)\varepsilon_0 + \frac{C_f}{2}\sum_{s=0}^{t}\gamma_s^2\left(\prod_{u=s+1}^{t}(1-\gamma_u)\right)}$$

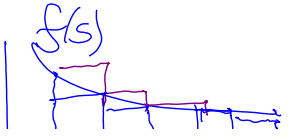
initial condition — Lipschitz part

use $(1+\gamma) \le e^{\gamma} \quad \forall \gamma$

$$(1-\gamma) \le e^{-\gamma}$$

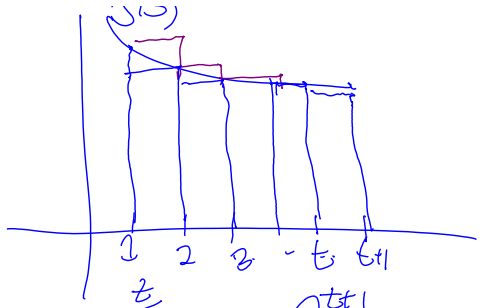
$$\varepsilon_{t+1} \le \varepsilon_0 \exp\left(-\sum_{s=0}^{t}\gamma_s\right) + \frac{C_f}{2}\sum_{s=0}^{t}\gamma_s^2 \exp\left(-\sum_{u=s+1}^{t}\gamma_u\right)$$





$$\varepsilon_{t+1} \le \varepsilon_0 \exp\left(-\sum_{s=0}^{t}\gamma_s\right) + \frac{C_f}{2}\sum_{s=0}^{t}\gamma_s^2 \exp\left(-\sum_{u=s+1}^{t}\gamma_u\right)$$

$\gamma_s \sim \frac{1}{s} \Rightarrow \sum_s \gamma_s \approx \log(t)$



$$\sum_{s=1}^{t} f(s) \ge \int_{s=1}^{t+1} f(s)\,ds$$

$$\exp\left(-\sum_{s=0}^{t}\gamma_s\right) \approx \exp(-\log t) = O\left(\frac{1}{t}\right)$$

$$\exp\left(-\sum_{u=s+1}^{t}\gamma_u\right) \approx \exp(-\log t/s) \approx O\left(\frac{s}{t}\right) \Rightarrow \sum_s \gamma_s^2 \exp(-\cdots)$$

($\gamma_s^2 \to \frac{1}{s^2}$, $\exp(-\cdots) \to \frac{s}{t}$)

$$= \frac{1}{t}\sum_{s}^{t}\frac{1}{s} \approx O\left(\frac{\log t}{t}\right)$$

in fact, if use $\gamma_t = \frac{1}{t+1}$, you get $O\left(\frac{\log t}{t}\right)$

see notes last year for $\gamma_s = \frac{\alpha}{t+\alpha}$ ($O\left(\frac{1}{t}\right)$ for $\alpha \ge 2$) //

* Linear rate for AFW: "linear rate constant"

linear rate: $\varepsilon_{t+1} \le (1-\rho)\varepsilon_t \le \varepsilon_0 (1-\rho)^t \le \varepsilon_0 \exp(-\rho t)$ (Or $\rho \le 1$)

sublinear rate $\varepsilon_t \le O\left(\frac{1}{t^{\text{power}}}\right)$

recall for FW: $\varepsilon_{t+1} \le (1-\gamma_t)\varepsilon_t$ $\gamma_t \approx \frac{g_t}{C_f} \approx \frac{\varepsilon_t}{C_f}$

AFW paper: get $\varepsilon_{t+1} \le \varepsilon_t - \gamma g_t + \frac{\gamma^2}{2} C_f$ $\gamma^* = \frac{g_t}{C_f}$

$\varepsilon_{t+1} \le \varepsilon_t - \frac{g_t^2}{2C_f}$

under some conditions, can show that $a_t^2 \ge \text{lin}\,\varepsilon_t$ ("growth condition")

f is μ-strongly convex

$\varepsilon_{t+1} \le \varepsilon_t - \frac{g_t^2}{2L_g}$ under some conditions, can show that $g_t^2 \ge \frac{\mu_g}{2}\varepsilon_t$ (f is μ_g-strongly convex)

$\Rightarrow \varepsilon_{t+1} \le \left(1 - \frac{\mu_g}{4L_g}\right)\varepsilon_t$

ie. linear rate with $\rho = \frac{\mu_g}{4L_g}$

→ a) FW with L.S. when $x^* \in \text{int}(M)$

b) AFW and M is polytope

Application of FW to structured SVM:

(dual) $\min_{\alpha_i \in \Delta_{|\mathcal{Y}|}} \frac{\lambda \|A\alpha\|^2}{2} - b^T\alpha$

$A\alpha = \frac{1}{\lambda n}\sum_i \sum_y \alpha_i(y)\psi_i(y) = w(\alpha)$

ie. $M = \bigtimes_{i=1}^{n} \Delta_{|\mathcal{Y}_i|}$

Let $\alpha_i^{(0)} = \delta_{y_i}$ (y_i: ground truth) $\Rightarrow w(\alpha^{(0)}) = 0$

FW step: $s_t = \arg\min_{s \in M} \langle s, \nabla f(\alpha_t)\rangle$

$\nabla f(\alpha_t) = \lambda A^T \underbrace{A\alpha_t}_{w_t} - b$ $\quad w_t = A\alpha_t$

$(\nabla f(\alpha_t))_{i,y} = \lambda \psi_i(y)^T w_t - \frac{L_i(y)}{n} = -\frac{1}{n} H_i(y; w_t)$

$\min_{s \in M} \langle s, \nabla f(\alpha_t)\rangle = \min_{\{s_i \in M_i\}} \sum_i \langle s_i, \nabla_i f(\alpha_t)\rangle$

$= \sum_i \min_{s_i \in M_i} \langle s_i, \nabla_i f(\alpha_t)\rangle$

$M_i = \Delta_{|\mathcal{Y}_i|}$

$\min_y \langle \delta_y, \nabla_i f(\alpha_t)\rangle$

$\nabla_{i,y} f(\alpha_t) = -\frac{1}{n} H_i(y; w_t)$

$V_{i,y}(\alpha_t) = \frac{1}{n} H_i(y_i|w_t)$

Thus $S_t = (\hat{S}_i)_{i=1}^n$

$\hat{S}_i = \delta_{\hat{y}_i(w_t)}$ where $\hat{y}_i(w_t) = \underset{\tilde{y} \in \mathcal{Y}_i}{\operatorname{argmax}} H_i(\tilde{y}_i|w_t)$ [less-augmented decoding]

$\Updownarrow$

$\hat{S}_i(y) = \mathbb{1}\{y = \hat{y}_i(w_t)\}$