Lecture 17 - FW convergence

Thursday, March 18, 2021 13:25

today: • convergence of FW alg.
• apply SVMstruct obj.

Curvature constant C_f

curvature constant

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

$x_\gamma = (1-\gamma)x + \gamma s$ ← potential FW step update

this is affine invariant ⇓ C_f is affine invariant

worst case deviation from linear approximate

$*$ by descent lemma, if ∇f is L-Lipschitz

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 \quad \Big[\|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\| \Big]$$

$$\big[|\langle d, x \rangle| \leq \|d\|_* \|x\| \big]$$

$$\Rightarrow C_f \leq \sup \frac{2}{\gamma^2}\frac{L}{2}\|x_\gamma - x\|^2 \qquad x_\gamma \mapsto x + \gamma(s - x)$$

$$\frac{L}{\cancel{2}} \frac{\cancel{2}}{\cancel{\gamma^2}} \|s - x\|^2 \cancel{\gamma^2}$$

$$C_f \leq L \cdot \sup_{x, s \in M} \|s - x\|^2 \qquad \mathrm{diam}_{\|\cdot\|}(M) \triangleq \sup_{x, s \in M} \|x - s\|$$

$$\boxed{C_f \leq L_{\|\cdot\|}\, \mathrm{diam}_{\|\cdot\|}(M)^2} \quad \text{for any } \|\cdot\|$$

affine invariant (C_f) depends on $\|\cdot\|$ ($L_{\|\cdot\|}$)

$(*)$ by def. of C_f, we get affine invariant version of descent lemma

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

let $x = x_t$ and $s = s_t$, FW corner $\quad \langle \nabla f(x_t), s_t - x_t \rangle = -g_t$ ← FW gap

for FW step of size γ

$$(+) \quad \boxed{f(x_\gamma) \leq f(x_t) - \gamma g_t + \frac{\gamma^2}{2} C_f \quad \forall \gamma \in [0,1]}$$

$$(+) \quad \boxed{f(x_\gamma) \leq f(x_t) - \gamma g_t + \frac{\gamma^2}{2} C_f \quad \forall \gamma \in [0,1]}$$

optimize step-size for bound (RHS)

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

this gives you <u>affine invariant</u> adaptive step size

$$f(x_{\gamma^*}) \leq f(x_t) - \frac{g_t^2}{2C_f} \quad \left[\text{when } \frac{g_t}{C_f} \leq 1\right]$$

let $\varepsilon_t \triangleq f(x_t) - f^* \leq g_t$

$$\leq f(x_t) - \frac{\varepsilon_t^2}{2C_f}$$

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

note:
- non-convex f: $\min_{s \leq t} g_s \leq O\left(\frac{1}{\sqrt{t}}\right)$
- f is <u>concave f</u>, <u>$C_f = 0$!?</u> $\min_{s \leq t} g_s \leq O\left(\frac{1}{t}\right)$

<u>proof</u>: let $x_\gamma = x_t + \gamma(s_t - x_t)$ } apply (+)

$$f(x_\gamma) \leq f(x_t) - \gamma g_t + \frac{\gamma^2}{2} C_f \quad \forall \gamma \in [0,1]$$

by convexity, $g_t \geq \varepsilon_t$
$-g_t \leq -\varepsilon_t$

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

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

* see notes 2017 for a cool <u>ODE trick</u> + induction

here, <u>brute force approach</u> to solve recurrence

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

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

ooo

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

initial condition

Lipschitz constant

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

$(1-\gamma) \le \exp(-\gamma)$

loose!

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

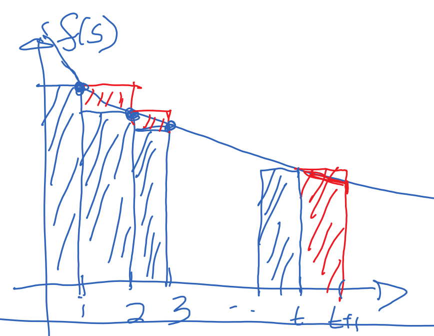
$$\gamma_s \sim \frac{1}{s} \Rightarrow \sum_{s}^{t}\gamma_s \approx \log t$$

$$\exp\Big(-\sum_{s=0}^{t}\gamma_s\Big) \approx \exp(-\log t)$$

$$\exp\big(\log \tfrac{1}{t}\big) \approx O\big(\tfrac{1}{t}\big)$$

$$\exp\Big(-\sum_{u=s+1}^{t}\gamma_u\Big) \approx \exp(-\log t/s) \approx O\big(\tfrac{s}{t}\big)$$

$$\sum_{s} \underbrace{\gamma_s^2}_{\frac{1}{s^2}} \underbrace{\exp\Big(-\sum \cdots\Big)}_{\frac{s}{t}} \sim \frac{1}{t}\sum_{s}^{t}\frac{1}{s} \to O\Big(\frac{\log t}{t}\Big)$$



$$\boxed{\int_{s=1}^{t+1} f(s)\,ds \le \sum_{s=1}^{t} f(s) \le \int_{s=0}^{t} f(s)\,ds}$$

f decreasing

$$\sum_{s=1}^{t}\frac{1}{s} = 1 + \sum_{s=2}^{t}\frac{1}{s} \le 1 + \int_{s=1}^{t}\frac{1}{s}\,ds$$

$$= 1 + [\log s]_1^t$$

$$= 1 + \log t$$

(*) in fact, if use $\gamma_t = \frac{1}{t+1}$, you do get $O\big(\frac{\log t}{t}\big)$ rate

but if $\gamma_t = \frac{2}{t+2}$, here bound says $O\big(\frac{\log t}{t}\big)$;

but a (tighter) direct analysis $O\big(\frac{1}{t}\big)$

see notes in 2017, for $\gamma_t = \frac{\alpha}{t+\alpha}$ ($O(\frac{1}{t})$ for $\alpha \ge 2$)

[telescoping product]

Lecture 11-- 2017/2/20 -- http://www.iro.umontreal.ca/~slacoste/teaching/ift6085/W17/protected/notes/lecture11_scribbles.pdf

14h30

Linear rate for AFW:

"linear rate" constant

$(1-\rho) \leq \exp(-\rho)$

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

(for gradient descent $\rho = \frac{\mu}{L} = \frac{1}{\kappa}$ ← condition #)

sublinear rate: $\varepsilon_t \leq O\left(\frac{1}{t^{\text{some pow}}}\right)$

(*) recall for FW with LS: $\varepsilon_{t+1} \leq \varepsilon_t - \frac{g_t^2}{2C_f}$

[aside: $-g_t^2 \leq -\varepsilon_t^2$]

$\Rightarrow \varepsilon_t \left(1 - \underbrace{\frac{\varepsilon_t}{2C_f}}_{\to 0}\right)$ ⇒ this is why you don't get linear rate

AFW paper, under some conditions

can show that $g_t^2 \geq \frac{\mu_f}{2}\varepsilon_t$

$\Rightarrow \varepsilon_{t+1} \leq \left(1 - \frac{\mu_f}{4C_f}\right)\varepsilon_t$

i.e. linear rate with $\rho = \frac{\mu_f}{4C_f}$

f is μ-strongly convex

a) FW with LS when α^* is in int$(\mathcal{M})$

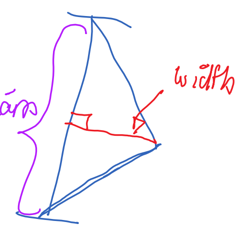
b) AFW and $\mathcal{M}$ is a polytope

$\mu_f \to$ "geometric strong convexity constant"

$\mu_f \geq \mu \cdot \text{pwidth}(\mathcal{M})^2$

linear rate: $\rho = \frac{\mu_f}{4C_f} \geq \frac{\mu\, \text{pwidth}(\mathcal{M})^2}{4L\, \underbrace{\text{diam}(\mathcal{M})^2}}$

$\frac{1}{\kappa_f}$ "$\frac{1}{\text{condition \# of set } \mathcal{M}}$"



FW for SVMstruct dual

dual of SVMstruct: $\min_{\alpha_i \in \Delta_{|\mathcal{Y}_i|}} \frac{\lambda}{2}\|A\alpha\|^2 - b^T\alpha$

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

$A\alpha = \frac{1}{\lambda n}\sum_{i=1}^{n}\sum_{y}\alpha_i(y)\psi_i(y) = w(\alpha)$

let $\alpha_i^{(0)} = \delta_{y^{(i)}}$ $\Rightarrow w(\alpha^{(0)}) = 0$

let $\alpha_i^{(0)} = \delta_{y^{(i)}} \Rightarrow w(\alpha^{(0)}) = 0$

FW step:

$$s_t = \underset{s \in \mathcal{M}}{\operatorname{argmin}} \langle s, \nabla f(\alpha_t) \rangle \qquad \nabla f(\alpha_t) = \lambda A^T \underbrace{A\alpha_t}_{w_t \triangleq w(\alpha_t)} - b \qquad w_t = A\alpha_t$$

$$(\nabla f(\alpha_t))_{i,y} = \cancel{\lambda} \frac{\psi_i(y)^T}{\cancel{\lambda} n} w_t - \frac{l_i(y)}{n}$$

$$= -\frac{1}{n} \underbrace{[l_i(y) - w_t^T \psi_i(y)]}_{H_i(y; w_t)}$$

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

(block-structure)

$$= \sum_i \underbrace{\min_{s_i \in \mathcal{M}_i} \langle s_i, \nabla_i f(\alpha_t) \rangle}_{\min_y \langle \delta_y, \underbrace{\nabla_i f(\alpha_t)}_{\nabla_{i,y} f(\alpha_t) = -\frac{1}{n} H_i(y; w_t)} \rangle}$$

$\alpha_t = \sum_u \alpha_u^t s_u$ $\qquad \mathcal{M}_i = \Delta_{|\mathcal{Y}_i|}$

thus $s_t \triangleq (\hat{s}_i)_{i=1}^n$ where $\hat{s}_i = \delta_{\hat{y}_i(w_t)}$ where $\hat{y}_i(w_t) = \underset{\tilde{y} \in \mathcal{Y}_i}{\operatorname{argmax}} H_i(\tilde{y}; w_t)$

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

[loss-augmented decoding]

$$\alpha^{(t)} \xrightarrow{A} w^{(t)}$$

$$\alpha_i^{(t+1)} = (1-\gamma)\alpha_i^{(t)} + \gamma \hat{s}_i^{(t)}$$

[here, need to maintain active set $\{\delta_{\hat{y}_i^u}\}_{u=0}^t$]

$$\alpha^{(t+1)} = (1-\gamma)\alpha^{(t)} + \gamma s^{(t)}$$

A

$$w^{(t+1)} = (1-\gamma)\underbrace{A\alpha^{(t)}}_{w^{(t)}} + \gamma \underbrace{As^{(t)}}_{\frac{1}{\lambda n}\sum_{i=1}^n \psi_i(\hat{y}_i^{(t)})}$$

you can choose via analytic L.S. on dual obj.

recall primal obj.

$$p(w) = \frac{\lambda \|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^n H_i(w) \qquad H_i(w) = \max_{\tilde{y}} l_i(\tilde{y}) - w^T \psi_i(\tilde{y})$$

$$p'(w_t) = \lambda w_t - \frac{1}{n}\sum_{i=1}^n \psi_i(\hat{y}_i^{(t)})$$

$$w^{(t+1)} = w^{(t)} - \beta p'(w^{(t)})$$

$$= (1-\lambda\beta) w^{(t)} + \beta \frac{1}{n}\sum_{i=1}^n \psi_i(\hat{y}_i^{(t)})$$

$$p'(w_t) = \lambda w_t - \frac{1}{n}\sum_{i=1}^{n} \psi_i(\hat{y}_i^{(t)}) = (1-\lambda\beta)\, w^{(t)} + \beta \sum_{i=1}^{n} \psi_i(\hat{y}_i^{(t)})$$

$\lambda\beta \to \gamma$, $\beta \to \frac{\gamma}{\lambda}$

if we set $\boxed{\beta = \frac{\gamma}{\lambda}}$

then batch subgradient step on primal is equivalent to a batch FW step on dual with $\beta = \frac{\gamma}{\lambda}$ step-size relationship

$w^{(t)} = A\alpha^{(t)}$

FW perspective gives you "adaptive stepsize" for batch subgradient method?