Lecture 17 - FW convergence

Thursday, March 21, 2024 9:39 AM

today: • convergence of FW alg.
• apply SVM struct alg.

Curvature constant C_f

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

- $x_\gamma = (1-\gamma)x + \gamma s$: potential FW step update
- $f(x_\gamma) - (f(x) + \langle \nabla f(x), x_\gamma - x\rangle)$: this is affine invariant
- $\Rightarrow$ C_f is affine invariant

→ worst case duration from linear approximation

* by descent lemma, if ∇f is L-Lipschitz $\rightsquigarrow$ $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x-y\|$

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

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

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

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

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

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

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

- $L_{\|\cdot\|}$: affine invariant
- $\operatorname{diam}_{\|\cdot\|}(M)$: depends on $\|\cdot\|$

for any $\|\cdot\|$

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

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

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

for FW step of size γ

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

optimize step-size for bound (RHS)

$$\gamma^* = \min\{g_t, 1\}$$

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

this gives you affine invariant adaptive step-size

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

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

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

Thm.: FW alg. with γ_t chosen either as $\frac{2}{t+2}$, g_t/C_f, line search

(when f is convex)

yields $\varepsilon_t \le \frac{2C_f}{t+2}$

rate:
- non-convex f: $\min_{s \le t} g_s \le O\left(\frac{1}{\sqrt{t}}\right)$
- f is concave, $C_f = 0$? $\min_{s \le t} g_s \le O\left(\frac{1}{t}\right)$

proof: let $x_\gamma = x_t + \gamma(s_t - x_t)$ & apply (1)

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

by convexity, $g_t \ge \varepsilon_t$

$$-g_t \le -\varepsilon_t$$

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

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

* see notes 2017 for a cool ODE trick + induction proof

here, brute force approach to solve recurrence

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

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

$$\vdots$$

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

ε_0: initial condition; C_f: Lipschitz constant

initial condition Lipschitz constant

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

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

loose?

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

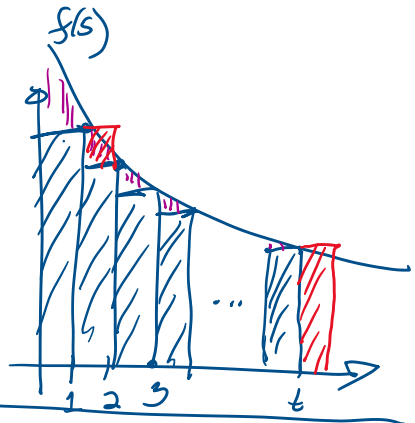
$$\gamma_s \sim \frac{1}{s} \Rightarrow \sum_{s}^{t} \gamma_s \approx \log t$$

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

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

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

$$\sum_{s} \gamma_s^2 \exp\left(-\sum_{u=s+1}^{t}\right) \quad \gamma_s \le \frac{1}{s} \quad \exp \le \frac{s}{t} \longrightarrow \frac{1}{t} \sum_{s}^{t} \frac{1}{s} \longrightarrow O\left(\frac{\log t}{t}\right)$$



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

f is 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 $O(\frac{\log t}{t})$ rate

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

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

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

10h30

linear rate for AFW:

"linear rate" constant $(1-\rho) \le \exp(-\rho)$

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

[for gradient descent $\rho = \frac{\mu}{L} = \frac{1}{\kappa}$ condition #]

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

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

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

$\Rightarrow \varepsilon_t \left(1 - \frac{\varepsilon_t}{2C_f}\right)$ $\to 0$

$\Rightarrow$ this is why you don't rec. get linear rate

AFW paper, under some conditions can show that $g_t^2 \geq \frac{\mu_S}{2} \varepsilon_t$

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

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

f is μ-strongly convex

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

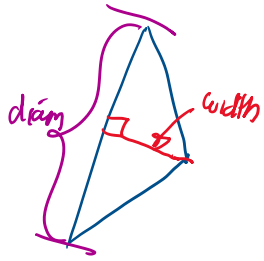
b) AFW and $\mathcal{M}$ is a polytope

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

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

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

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



FW for SVM struct dual

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

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

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

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

FW step:

let $w_t \triangleq w(\alpha_t) = A\alpha_t$

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

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

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

$= -\frac{1}{n} [\ell_i(y) - w_t^T \psi_i(y)]$

$H_i(y; w_t)$

$\min_{s \in \mathcal{M}} \langle s, \nabla f(\alpha_t) \rangle = \min_{s_i \in \mathcal{M}_i, \forall i} \sum_{i=1}^{n} \langle s_i, \nabla_i f(\alpha_t) \rangle$

(block structure)

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

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

$M_i = \Delta_{|\mathcal{Y}_i|}$ $\quad \min_y \langle \delta_y, \nabla_i f(\alpha_t) \rangle$

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

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

$\hat{s}_i(y) = \mathbb{I}\{y = \hat{y}_i\}$ [loss-augmented decoding]

$\alpha_t = \sum_u \alpha_u^+ s_u$

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

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

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

$A\alpha^{(t)} = w^{(t)}$ $\quad A s^{(t)} = \frac{1}{\lambda n} \sum_{i=1}^n \psi_i(\hat{y}_i^{(t)})$

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

recall primal obj:

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

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

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

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

γ $\quad \frac{\gamma}{\lambda}$

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

FW perspective gives you "adaptive step-size" for batch subgradient method!