Lecture 17 - FW convergence

Tuesday, March 17, 2020 14:33

today: - convergence FW alg.
- apply SVM struct obj.

curvature constant C_f

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

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

this affine invariant

worst case deviation from linear approximation

$\Rightarrow C_f$ is affine invariant

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

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

$[\ \|\nabla f(x) - \nabla f(y)\|_* \le L\|x-y\| \quad \forall x,y \in M\ ]$

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

$$\Rightarrow C_f \le \sup \frac{2}{\gamma^2}\left[\frac{L}{2}\|x_\gamma - x\|^2\right], \quad x_\gamma = x + \gamma(s-x)$$

$$\|\gamma(s-x)\|^2$$

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

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

$$C_f \le L_{\|\cdot\|} \cdot \text{diam}_{\|\cdot\|}(M)^2 \quad \text{for any } \|\cdot\|$$

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

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

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

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

FW-gap *

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

for FW step of size γ

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

optimize step-size for bound (RHS)

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

↑ this gives you an affine inv adaptive step-size

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

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

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

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

Note:
non-convex f

$$\min_{s \le t} g_s \le O\left(\frac{1}{\sqrt{t}}\right)$$

concave f, $C_f = 0$

$$\min_{s \le t} g_s \le O\left(\frac{1}{t}\right)$$

proof: let $x_\gamma = x_t + \gamma(s_t - x_t)$ + apply (4)

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

by convexity, $g_t \ge \varepsilon_t$

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

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

* see notes 2017 for a cool ODE trick + induction

here, brute force approach to solve recurrence

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

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

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

initial condition; Lipschitz constant

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

$(1-\gamma) \le e^{-\gamma}$ ← loose?

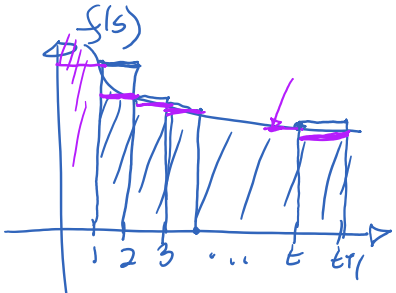
$$\Rightarrow \varepsilon_{t+1} \le \varepsilon_0 \exp\left(-\sum_{s}^{t}\gamma_s\right) + \frac{C_f}{2}\sum_{s=0}^{t}\gamma_s^2 \exp\left(-\sum_{u=s+1}^{t}\gamma_u\right)$$

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

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

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

$\sum_{s}\gamma_s^2 \exp(-\sum\ldots) \approx O\left(\frac{\log t}{t}\right)$ ($\gamma_s^2 \sim \frac{1}{s^2}$, $\exp \sim \frac{s}{t}$)



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

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 $O\left(\frac{\log t}{t}\right)$ rate

but $\gamma_t = \frac{2}{t+2}$, here our bound says $O\left(\frac{\log t}{t}\right)$; but (tighter) analysis $O\left(\frac{1}{t}\right)$

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

15h40

Linear rate for AFW:

"linear rate constant"

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

(for gradient descent $p = \frac{\mu}{L}$)

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

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

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

AFW paper, under some conditions → f is μ-strongly convex

can show $g_t^2 \ge \frac{\mu_S}{2}\varepsilon_t$

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

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

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

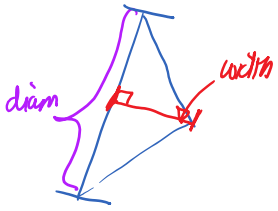
i.e. linear rate with $\rho = \mu_S/4C_f$

μ_S → geometric strong convexity constant

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

linear rate $\rho = \frac{\mu_S}{4C_f} \ge \frac{\mu \cdot \text{pwidth}(\mathcal{M})^2}{4L\,\text{diam}(\mathcal{M})^2}$

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



FW for SVMstruct

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

i.e. $\mathcal{M} = \bigoplus_{i=1}^n \Delta_{|Y_i|}$ $\quad A\alpha = \frac{1}{\lambda n}\sum_{i=1}^n \sum_y \alpha_i(y)\psi_i(y) = w(\alpha)$

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

FW step:

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

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

$= -\frac{1}{n} H_i(y; w_t)$

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

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

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

$M_i = \Delta_{|\mathcal{Y}_i|}$ $\quad \min_y \langle \delta_y, \underbrace{\nabla_i f(w_t)}_{\nabla_{i,y} f(w_t) = -\frac{1}{n} H_i(y; w_t)} \rangle$

thus $s_t \triangleq (\hat{s}_i)_{i=1}^n$ where $\hat{s}_i = \delta_{\hat{y}_i(w_t)}$ where $\hat{y}_i(w_t) \triangleq \operatorname{argmax}_{\tilde{y} \in \mathcal{Y}_i} H_i(\tilde{y}; w_t)$ [loss-augmented decoding] ⊛⊛

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

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

$\alpha_i^{(t+1)} = (1-\gamma)\alpha_i^{(t)} + \gamma \hat{s}_i^{(t)}$ ← [here, need to maintain active set $\{\delta_{\hat{y}_i^{(u)}}\}_{u=0}^t$]

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

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

$\gamma_t^* = \operatorname{argmin}_{\gamma \in [0,1]} f(\alpha^{(t)} + \gamma(s^{(t)} - \alpha^{(t)}))$

recall primal obj. $\quad \max_{\tilde{y}} \ell_i(\tilde{y}) - w^T \psi_i(\tilde{y})$

$p(w) = \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^n H_i(w)$

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

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

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

γ ; $\frac{1}{\lambda n}$

If set $\beta = \frac{\gamma}{\lambda}$

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

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