Lecture 17 - FW convergence

Tuesday, March 12, 2019 14:19

today : • convergence of FW
• apply FW to submodular

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

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

$\left[\|\nabla f(x) - \nabla f(x')\|_* \leq L \|x - x'\| \quad \forall x, x' \in M \right]$

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

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

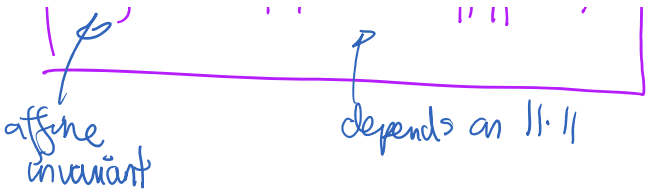
$$\Rightarrow C_f \leq \sup \frac{2}{\gamma^2} \left[\frac{L}{2} \|x_\gamma - x\|^2 \right] \quad x_\gamma = x + \gamma(s - x)$$

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

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

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

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



(*) 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

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

for FW-step of size γ

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

optimize step-size for bound (RHS)

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

this gives 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 $2/t+2$, g_t/C_f, line search (when f is convex) yields $\varepsilon_t \leq \frac{2C_f}{t+2}$

note:

non-convex f

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

concave f, $C_f = 0$

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

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

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

by convexity, $g_t \geq \varepsilon_t$

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

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

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

* see notes 2017 for a cool BOE trick + induction

here, brute force approach to solve the recurrence

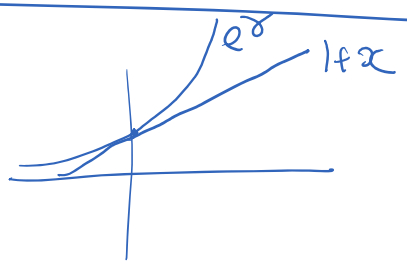
$$\varepsilon_{t+1} \le (1-\gamma_t)\varepsilon_t + \frac{\gamma_t^2 C_f}{2}$$

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

$$\le \cdots$$

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

initial condition — Lipschitz part



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

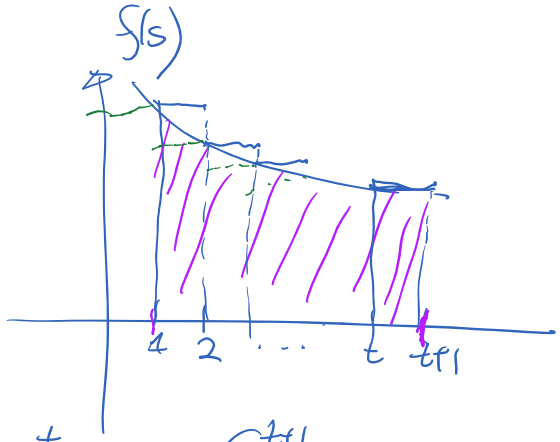
$(1-\gamma) \le e^{-\gamma}$ loose?

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

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

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

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



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

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

$u=s+1$ ... t

$$\sum_s \gamma_s^2 \exp\left(-\sum \ldots\right) \approx O\left(\frac{\log t}{t}\right)$$

$\gamma_s^2 \approx \frac{1}{s^2}$, $\exp(\ldots) \approx \frac{s}{t}$

$$\sum_{s=1}^{t} \frac{1}{s} = 1 + \sum_{s=2}^{t} \frac{1}{s} \leq 1 + \int_{s=1}^{t} \frac{1}{s}\,ds = 1 + [\log s]_1^t = 1 + \log t$$

(*) 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 $\gamma_t = \frac{\alpha}{t+\alpha}$ $\left(O\left(\frac{1}{t}\right) \text{ for } \alpha \geq 2\right)$

15h36

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

Linear rate for AFW:

"linear rate constant"

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

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

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

recall for FW (with LS): $\varepsilon_{t+1} \leq \varepsilon_t - \frac{g_t^2}{2C_f}$ $\left[\text{aside } -g_t^2 \leq \varepsilon_t^2 \quad \varepsilon_t\left(1 - \frac{\varepsilon_t}{2C_f}\right)\right]$

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

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

a) FW with LS, then $x^* \in \text{int}(M)$

b) AFW and M is a polytope

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

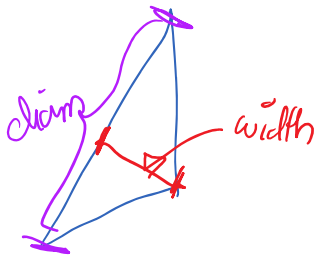
i.e. linear rate with $\rho = \frac{\mu_S}{4C_S}$

$\mu_S \rightarrow$ geometric strong convexity constant

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

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

K_S ← condition # of set M



FW for SVMstruct

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

i.e. $M = \bigoplus_{i=1}^{n} \Delta_{|\mathcal{Y}_i|}$ $\qquad A\alpha = \frac{1}{\lambda n} \sum_{i=1}^{n} \sum_{\tilde{y}} \alpha_i(\tilde{y}) \psi_i(\tilde{y}) = w(\alpha)$

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

FW step:

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

FW step:

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

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

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

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

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

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

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

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

[loss-augmented decoding] ⊗ ⊗

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

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

$\alpha_i^{(t+1)} = (1-\gamma)\alpha^{(t)} + \gamma \hat{s}_i^{(t)}$ [have need to maintain active set:

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

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

A

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

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

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

→ you can choose via analytic L.s on dual objective

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

recall primal obj.

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

$$H_i(w) \downarrow \max_y \ell_i(y) - w^T\psi_i(y)$$

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

(batch) Subgradient method update

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

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

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

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

⊗ FW perspective gives you an 'adaptive step-size' batch subgradient method