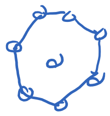
Lecture 17 - FW convergence

Tuesday, March 21, 2023 2:30 PM

today: · convergence of FW alg.
· apply SUMstruct obj.

⊗ note: if $M = \text{conv}(\mathcal{A})$ where $\mathcal{A}$ is some finite set (called "atoms")

$$LMO(r): \min_{s \in M = \text{conv}(\mathcal{A})} \langle s, r \rangle = \min_{a \in \mathcal{A}} \langle a, r \rangle$$



↳ lot of applications in ML where LMO is efficient

e.g. · $\mathcal{A} \to$ integer flows

$\text{conv}(\mathcal{A}) \to$ flow polytope

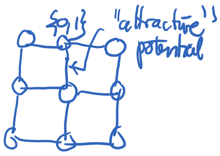
LMO → min cost network flow algorithm

· $\mathcal{A} \to$ clique assignment in graph $(\delta y_c)_{c \in \mathcal{C}}$

$\text{conv}(\mathcal{A}) \to$ marginal polytope

LMO → max product alg.

or



Graph cut alg. for binary Ising model with attractive potential ("Associative Markov network")

(→ "submodular potentials" (see later))

<u>curvature constant C_f</u>

<u>curvature constant</u> $C_f \triangleq \sup_{\substack{\gamma \in [0,1] \\ x, s \in M \\ x_\gamma = (1-\gamma)x + \gamma s}} \frac{2}{\gamma^2}\left[f(x_\gamma) - \underbrace{\left(f(x) + \langle \nabla f(x), x_\gamma - x \rangle\right)}_{\text{this is affine invariant}} \right]$

⇓

C_f is affine invariant

potential FW step update

worst case deviation from linear approximation

✱ by descent lemma, if ∇f is L-Lipschitz $\left[\|\nabla f(x) - \nabla f(y)\|_* \le L \|x - y\| \right]$

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

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

$$\Rightarrow C_f \le \sup \frac{2}{\gamma^2} \frac{L}{2} \|x_\gamma - x\|^2$$

$\hookrightarrow x + \gamma(s - x)$

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

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

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

affine invariant

depends on $\|\cdot\|$

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

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

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

for FW step of size γ

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

optimize step-size for bound (RHS)

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

this gives you an affine-invariant adaptive step-size

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

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

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

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

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

note:

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

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

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

To 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 ODE trick + induction

here, brute force approach to solve recurrence

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

$\vdots$

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

initial condition   Lipschitz constant

use $(1+\gamma) \leq \exp(\gamma) \quad \forall \gamma$

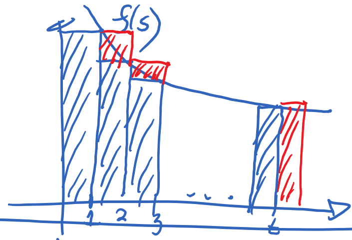
$(1-\gamma) \leq \exp(-\gamma)$

loose!

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

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

15h37



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

f decreasing

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

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

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

$$\sum_{s} \underbrace{\gamma_s^2}_{\frac{1}{s^2}} \underbrace{\exp\left[-\sum \text{''}\right]}_{\frac{s}{t}} \longrightarrow \frac{1}{t} \sum_{s}^{t} \frac{1}{s} \to \frac{\log t}{t}$$

⊛ in fact! if use $\gamma_t = \frac{1}{t+1}$, you do get $O\left(\frac{\log t}{t}\right)$ rate but if $\gamma_t = \frac{2}{t+2}$, here the bound says $O\left(\frac{\log t}{t}\right)$

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

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

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

[telescoping product]

Linear rate of AFW

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

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

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

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

(*) recall for FW with line search:

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

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

$\Rightarrow \varepsilon_t\left(1 - \frac{\varepsilon_t}{2C_f}\right)$, $\frac{\varepsilon_t}{2C_f} \to 0$ $\Rightarrow$ 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 $\rho = \frac{\mu_f}{4C_f}$

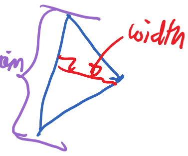
- f is μ-strongly convex
  - a) FW with LS (when x^* is int(M))
  - b) AFW and M is a polytope

μ_f → "geometric strong convexity constant"

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

linear rate : $\rho = \frac{\mu_f}{4C_f} \geq \frac{1}{4}\frac{\mu}{L}\frac{\text{pwidth}(M)^2}{\text{diam}(M)^2}$

$\frac{\mu}{L} \to \frac{1}{\kappa_f}$, $\quad \frac{\text{pwidth}(M)^2}{\text{diam}(M)^2} \to \frac{1}{\text{"condition \# of } M\text{"}}$



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

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

FW step:

$w_t \triangleq w(\alpha_t)$

$$s_t = \underset{s\in\mathcal{M}}{\arg\min} \langle s, \nabla f(\alpha_t)\rangle \qquad \nabla f(\alpha^t) = \lambda A^T A\alpha_t - b$$

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

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

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

$$= \sum_i \min_{s_i\in M_i} \langle s_i, \nabla_i f(\alpha_t)\rangle \qquad \text{(block-separable structure)}$$

$$\alpha_t = \sum_u \alpha_u^t s_u \qquad M_i = \Delta_{|\mathcal{Y}_i|} \qquad \min_{\tilde y} \langle \delta_{\tilde y}, \nabla_i f(\alpha_t)\rangle$$

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

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

$\hat s_i(\tilde y) = \mathbb{1}\{\tilde y = \hat y_i\}$

(*) 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 $\{\delta_{\hat y_i^{(u)}}\}_{u=0}^t$]

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

$$A \downarrow$$

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

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

recall primal obj:

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

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

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

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

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

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 step-size"

- for batch subgradient method :