Lecture 4 - scribbles

Wednesday, January 18, 2017
10:27

today: optimization (of $\hat{J}(w)$)

$$\hat{J}(w) = R(w) + \frac{1}{n}\sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}, w)$$

structured SVM: $R(w) = \frac{\lambda \|w\|^2}{2}$ $\quad S(x,y,w) = \langle w, \varphi(x,y)\rangle$

$$\ell(x^{(i)}, y^{(i)}, w) = \max_{\tilde{y} \in \mathcal{Y}(x^{(i)})} \left[\langle w, \varphi(x^{(i)}, \tilde{y})\rangle + \ell(y^{(i)}, \tilde{y})\right] - \langle w, \varphi(x^{(i)}, y^{(i)})\rangle$$

(structured hinge loss)

let $\ell_i(\tilde{y}) \triangleq \ell(y^{(i)}, \tilde{y})$ $\quad \mathcal{Y}_i \triangleq \mathcal{Y}(x^{(i)})$

$\psi_i(\tilde{y}) \triangleq \varphi(x^{(i)}, y^{(i)}) - \varphi(x^{(i)}, \tilde{y})$

$H_i(w) \triangleq \ell(x^{(i)}, y^{(i)}, w) = \max_{\tilde{y} \in \mathcal{Y}_i} \left[\ell_i(\tilde{y}) - \langle w, \psi_i(\tilde{y})\rangle\right]$

structured SVM objective (non-smooth unconstrained form)

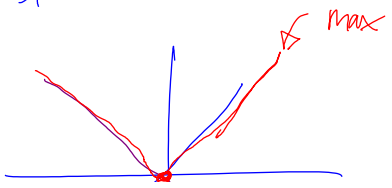
regularization hyperparameter

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

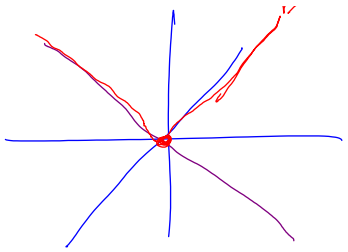
(somewhat akin to $(1 - y_i w^T x_i)_+$)

$$H_i(w) = \max_{\tilde{y} \in \mathcal{Y}_i} \left[\ell_i(\tilde{y}) - \langle w, \psi_i(\tilde{y})\rangle\right]$$

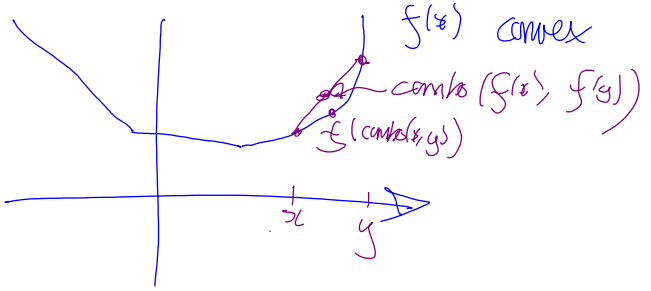
$$|x| = \max\{x, -x\}$$



$|x| = \max \{ x, -x \}$



Convex analysis recap



f is convex

$\Leftrightarrow f(\underbrace{px + (1-p)y}_{\text{cvx-combo}(x,y)}) \leq p f(x) + (1-p) f(y)$

$p \in [0,1]$

subgradient: $v \in \partial f(x)$, "v is a subgradient of f at x"

($\partial f(x)$: subdifferential of f at x)

$\Leftrightarrow \forall y (\in \text{dom} f) \quad f(y) \geq f(x) + \langle v, y - x \rangle$

standard assumptions on f:

smooth f:

- ∇f is L-Lipschitz continuous i.e. $\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|$

  [if f is twice differentiable; you have $L = \max_{x \in \text{dom} f} \|H(f)\|$ (Hessian)]

non-smooth f:

- ∂f is bounded i.e. $\|v\| \leq B$ for all subgradients v of f

strong convexity of f:

f is μ-strongly convex $\Leftrightarrow$ $f(x_1) \geq f(x_2) + \langle \partial f(x_2), x_1 - x_2 \rangle + \frac{\mu}{2} \|x_1 - x_2\|^2 \quad \forall x_1, x_2 \in \text{dom} f$

f is μ-strongly convex $\iff$ $f(y) \geq f(x) + \langle \partial f(x), y-x \rangle + \frac{\mu}{2}\|x-y\|^2 \quad \forall y \in \text{dom} f$

μ: strong convexity constant

$\langle v, y-x \rangle$ for any $v \in \partial f(x)$

[If f is C^2, $\mu = \lambda_{\min}(\text{Hessian}(f))$] (minimum e-value)

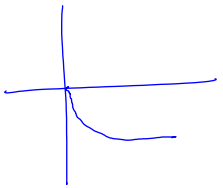
properties: ∇f L-Lipschitz $\Rightarrow$ $f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|y-x\|^2$

f μ-strongly convex $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\mu}{2}\|y-x\|^2$

[question: if $f(x)$ is bounded and convex,

$\overset{?}{\Rightarrow}$ f is L-lipschitz

answer: no; counterexample $f(x) = -\sqrt{x}$ on $[0,1]$



(f is not Lipschitz continuous)]

Landscape of convergence rates: → for first order methods

example:

convergence result: $f(x_k) - f(x^*) \leq \frac{\text{const.}}{\sqrt{k}}$ (non-smooth rate)

function suboptimality

suppose $\text{dist}(x_0, X^*) \leq r_0$ (set of solution)

assumptions	rate deterministic		rate stochastic → $\frac{1}{n}\sum_{i=1}^{n} f_i(x)$	
1) non-smooth $\|\partial f\| \leq B$	$O\left(\frac{Br_0}{\sqrt{k}}\right)$	subgradient method	$O\left(\frac{1}{\sqrt{k}}\right)$	stochastic subgradient method

2) smooth L-Lipschitz	$O\left(\frac{Lr_0^2}{K}\right)$ gradient method	$O\left(\frac{1}{\sqrt{K}}\right)$	SGD
	$O\left(\frac{Lr_0^2}{K^2}\right)$ Nesterov method		
f is μ strongly convex: 3) non-smooth $\|g\| \le B$	$O\left(\frac{B^2}{\mu K}\right)$ subgradient method	$O\left(\frac{B^2}{\mu K}\right)$	$\to \mathbb{E}_\xi f(x,\xi)$
4) smooth L-Lipschitz	$O(\exp(-\frac{\mu}{L}K))$ gradient method	true expectation $O\left(\frac{1}{K}\right)$	SGD
	$O(\exp(-\sqrt{\frac{\mu}{L}}K))$ Nesterov method	but finite sum using SAG, etc..	faster rate $\exp(-\text{something}\cdot K)$ (stochastic averaged gradient method) $\to \frac{1}{n}\sum_{i=1}^n f_i(x)$

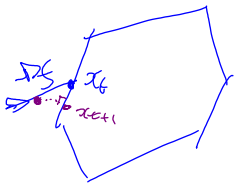
⊗ note: projecting gives same rate → reason is contraction property of projection ie. $\|P_C(y) - P_C(x)\|_2 \le \|y - x\|_2$

e.g. gradient method (unconstrained optimization) $x_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$

projected gradient method for $\min_{x \in C} f(x)$: $x_{t+1} = P_C\left[x_t - \frac{1}{L}\nabla f(x_t)\right]$

Euclidean projection on C ie.

$P_C(y) \triangleq \operatorname{argmin}_{x \in C} \|x - y\|^2$



minimizer

$\|P_C(y) - P_C(x^*)\| = \|P_C(y) - x^*\| \le \|y - x^*\|$

⇒ so unconstrained rates transfer to "projected" versions

Complexity in optimization

→ "work" to get to ε suboptimality $O\left(\frac{1}{\sqrt{k}}\right)$ rate $\rightsquigarrow O\left(\frac{1}{\varepsilon^2}\right)$ # of iterations

If want to compare methods, include also "cost" of iterations

stochastic subgradient method

say want to solve $\min_{x \in C} f(x)$ where $f(x) \triangleq \mathbb{E}_{\xi \sim P}[h(x,\xi)]$

assumptions: 1) projection on C is cheap

2) f is convex in x

3) we have stochastic oracle which gives g_t at time t, a random direction such that

$$\mathbb{E}[g_t \mid x_t] = f'(x_t)$$

(f': (any) subgradient of f)

$\mathbb{E}[g_t \mid \text{"past"}]$

[e.g. if f is differentiable in x and "well behaved"

$\nabla_x h(x, \xi_t)$ for $\xi_t \sim P$ then $\mathbb{E}_{\xi}[\nabla h(x,\xi)] = \nabla f(x)$]

more specifically if $f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x)$

then pick i_t unif. at random $1, \dots, n$

and letting $g_t = \hat{f}'_{i_t}(x_t)$

then $\mathbb{E}[g_t | x_t] = \frac{1}{n} \sum_{j=1}^{n} f'_i(x_t) = f'(x_t)$

4) $\mathbb{E}\|g_t\|^2 \leq B^2$ [finite variance condition]

($\|f'_i(x)\| \leq B$ is sufficient for this then)

algorithm:

- $x_0 \in C$ initialization
- for $t = 0, \ldots, T-1$

  get g_t from oracle

  let $x_{t+1} = P_C[x_t - \gamma_t g_t]$ (step-size)

- output $\hat{x}_T = \sum_{t=0}^{T} \rho_t x_t$ where $\rho_t \propto t$

  "weighted average"

  i.e. $\rho_t = \frac{2t}{T(T+1)}$

pointers:

- book for rates of convergence & lower bound:
  - Nesterov, "Introductory Lectures on Convex Optimization", 2004
- proof for convergence of weighted average stochastic subgradient method:
  - Lacoste-Julien, Schmidt, Bach, "A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method", arXiv:1212.2002
    https://arxiv.org/abs/1212.2002