Lecture 9 - convex optimization

Tuesday, February 5, 2019 14:29

today: finish calibration fct.
convex optimization

(continuation)

$\boxed{\begin{array}{l} \underline{\text{thm. 6}}: \text{ learning complexity} \\ \quad \text{let } \theta^* \text{ minimizes } L(\theta) \text{ with } \|\theta^*\|_{HS} \leq D \\ \text{choosing } n \geq \dfrac{4D^2M^2}{H(\varepsilon)^2} \text{ implies } \mathbb{E}[L(\hat{\theta}^{(n)})] < L(\theta^*) + \varepsilon \end{array}}$

define a meaningful scale

in the paper: we compute D & M & $H(\varepsilon)$ for specific losses ℓ and the quadratic $\mathcal{L}$
to get sample complexity

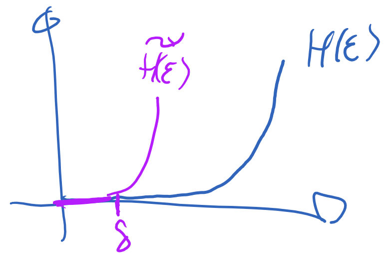
(*) **Moral here**:

- * some losses are harder than others (worst case sample complexity) [0-1 loss is difficult in general]
- * have linked computation to statistical performance in consistency framework
  - ↳ convex surrogate loss
- * could handle dependence on x using RKHS

caveats: • distribution free result (i.e. worst-case over all distributions)

→ still need more theory? (e.g. role of $p(y|x)$)? or other surrogates)

* follow-up paper: inconsistent surrogate losses with computational/statistical advantage NIPS 2018



15h15

PART II: convex optimization

convex surrogate loss

PART II: convex optimization

motivation: $\min_{w} \frac{\lambda \|w\|^2}{2} + \frac{1}{n}\sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}, w)$

convex surrogate loss

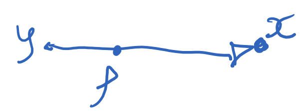
convex analysis recap:

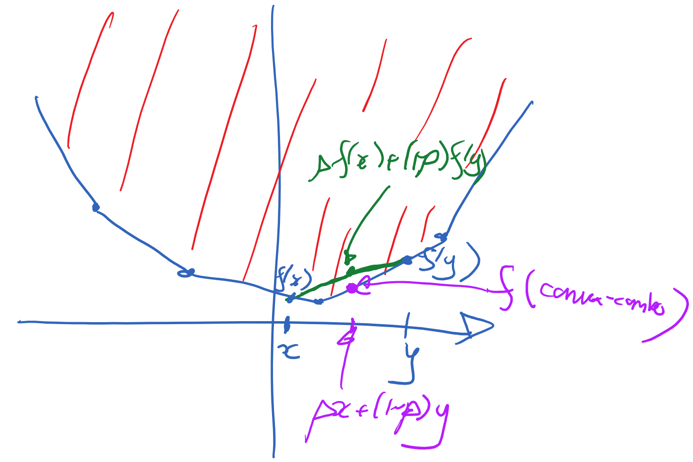
$f : \mathbb{R}^d \to \mathbb{R}$

f is convex $\Leftrightarrow$

$$f(\underbrace{px + (1-p)y}) \le p f(x) + (1-p) f(y) \quad \forall x, y, \; p \in [0,1]$$

convex combination between x & y $\to y + p(x-y)$

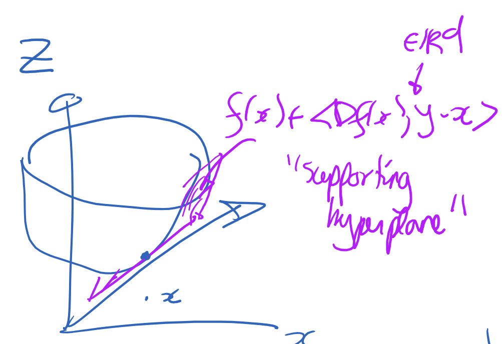




epigraph$(f) \triangleq \{(x, y) : y \ge f(x)\}$, $x \in \mathbb{R}^d$, $y \in \mathbb{R}$

* if f is differentiable at x and convex

$$\Rightarrow f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle \quad \forall y$$



(suppose f is convex)

subdifferential

subgradients of f at x: $v \in \partial f(x)$

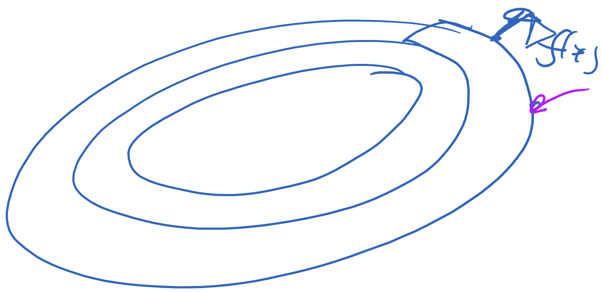
$|x| = \max\{x, -x\}$

(Suppose f is convex)

subdifferential

subgradient v of f at x : $v \in \partial f(x)$

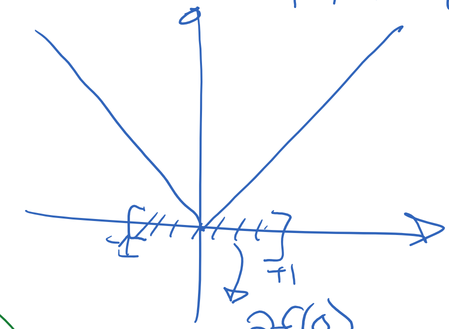
$$\Leftrightarrow \forall y \in dom(f), \quad f(y) \geq f(x) + \langle v, y - x \rangle$$



$\|x\|_1$

$\partial f(x) = conv\{\nabla f(x_1), \nabla f(x_2)\}$

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



when $f(x) = \max_i f_i(x)$ where f_i is differentiable

$$\partial f(x) = conv\{\nabla f_i(x) : i \in \arg\max_j f_j(x)\}$$

Danskin's theorem : https://en.wikipedia.org/wiki/Danskin%27s_theorem

some standard assumptions:

$dom(f) \triangleq \{x \in \mathbb{R}^d : f(x) < \infty\}$ $\quad f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$

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

$\forall x, y \in dom(f)$

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

strong convexity constant

$$\boxed{f \text{ is } \underline{L\text{-smooth}} \quad \text{i.e. } f \text{ has } L\text{-Lipschitz gradient } \forall x \iff \|\nabla f(x) - \nabla f(y)\|_* \le L\|x-y\|}$$

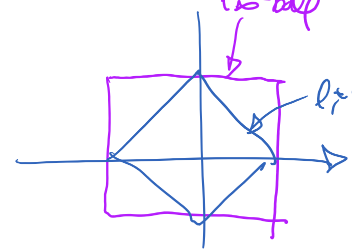
$$\|w\|_* \triangleq \sup_{\|v\|\le 1} \langle w, v\rangle$$

generalized C.-S.

$$\boxed{\langle w, v\rangle \le \|w\|_* \|v\|}$$

$$(\|\cdot\|_p)^* = \|\cdot\|_q \quad \text{where } \frac{1}{p} + \frac{1}{q} = 1$$

$$p=2 \implies q=2$$
$$p=1 \iff q=\infty$$



Fundamental descent lemma:

when ∇f is L-Lipschitz (even if f is not nec. convex)

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

$$* \quad f(\underbrace{x - \gamma\nabla f(x)}_{y_\gamma}) \le f(x) - \gamma\langle \nabla f(x), \nabla f(x)\rangle + \frac{\gamma^2}{2} L \|\nabla f(x)\|^2$$

$$= f(x) - \underbrace{\left[\gamma\left(1 - \frac{\gamma L}{2}\right)\right]}_{>0} \|\nabla f(x)\|_2^2$$

$$>0 \iff \boxed{0 < \gamma < \frac{2}{L}}$$

→ minimize RHS with respect to γ

gives $\boxed{\gamma^* = \frac{1}{}}$ $\qquad f(y_{\gamma^*}) \le f(x) - \frac{1}{} \|\nabla f(x)\|_2^2$

gives $\boxed{\gamma^* = \frac{1}{L}}$ $\quad f(y_{\gamma^*}) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2$

[think of 2nd order Taylor expansion

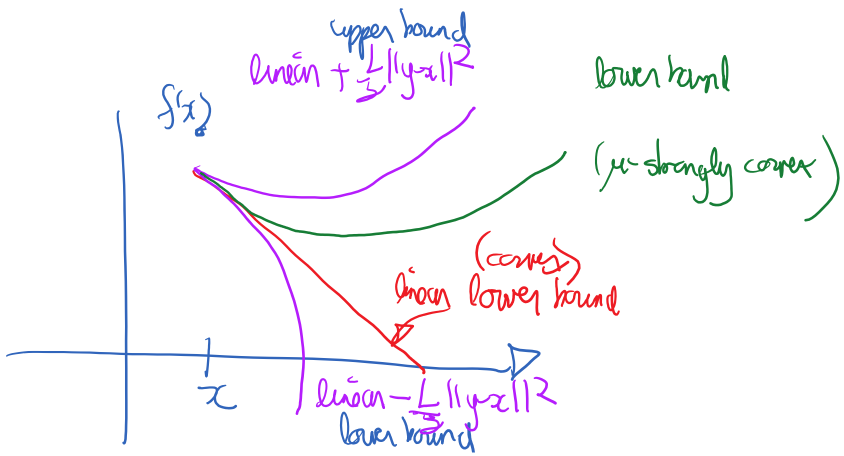
$$f(y) = f(x) + \langle \nabla f(x), y-x \rangle + \frac{1}{2}\int_{\gamma=0}^{1} \langle y-x, H(x+\gamma(y-x))\, y-x \rangle \, d\gamma$$]

top-evalue of H in absolute value $\leq L$

$$\lambda_{min}(H)\|v\|^2 \leq v^T H v \leq \lambda_{max}(H)\|v\|^2$$

or "pro-way" is to use fundamental thm. of calculus

$$f(y) = f(x) + \int_0^1 \underbrace{\frac{d}{d\gamma} f(x+\gamma(y-x))}_{\langle \nabla f(x+\gamma(y-x)),\, y-x \rangle} d\gamma \qquad \text{(see Nesterov)}$$



when f is twice differentiable

$L = \lambda_{max}(H)$ Hessian

x | linear $- \frac{L}{2}\|y-x\|^2$ lower bound

$L = \lambda_{max}(H)$

$\mu = \lambda_{min}(H)$

$$\boxed{f \text{ is } \mu\text{-strongly convex} \iff f - \frac{\mu}{2}\|\cdot\|^2 \text{ is convex}}$$

gradient descent: $x_{t+1} = x_t - \gamma \nabla f(x_t)$ $\quad \gamma = \frac{1}{L}$

a) when f is convex & L-smooth

$$f(x_t) - \underbrace{\min_x f(x)}_{f^*} \leq O\left(\frac{L r_0^2}{t}\right) \quad \text{where } r_0 \geq \text{dist}(x_0, \underbrace{X^*}_{\operatorname{argmin}_x f(x)})$$

"sublinear rate"

↳ [see Nesterov book for $O(\frac{1}{t})$ rate]

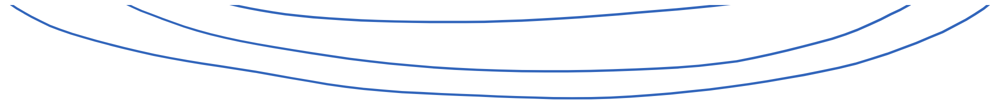
note: no guarantee on $\text{dist}(x_t, X^*)$ (in general)

→ Nesterov lower bounds

b) if f is μ-strongly convex & L-smooth "linear rate"

$$f(x_t) - f(x^*) \leq O\left(\exp\left(-\frac{\mu}{L} t\right)\right)$$

$\frac{L}{\mu} \triangleq$ condition# of f



<u>Newton's method</u>: $x_{t+1} = x_t - \gamma\, H(x_t)^{-1} \nabla f(x_t)$

$$\nabla \nabla f(x_t)$$