Lecture 9 - convex optimization

Tuesday, February 4, 2020 14:30

today : · finish calibration
· start convex opt.

optimization normalization for calibration scr.

setup let $s(x,\tilde{y})$ be of the form $F\theta(x)$ $\theta(x) \in \mathbb{R}^K$

$s(x,\cdot) = F\theta(x)$ $\quad \theta_y(\cdot) \in \mathcal{H}$ ← RKHS

θ_y optimization variables

$\ell(\theta) = \mathbb{E}_{(x,y)\sim p}\, \ell(x,y;\theta)$

$(x^{(t)}, y^{(t)}) \overset{iid}{\sim} p$

run projected SGD (kernelized) on $\ell(\theta)$ i.e. $\theta^{(t+1)} = P_{\mathcal{H}}\left(\theta^{(t)} - \gamma \nabla_\theta \ell(x^{(t)}, y^{(t)}; \theta^{(t)})\right)$

projection

$\mathcal{H}$ = ball of radius D around 0

"$\nabla_\theta \ell(x^{(t)}, y^{(t)}; \theta) = F^T \nabla_s \ell(x^{(t)}, y^{(t)}; s)\, \phi(x^{(t)})^T$"

ϕ feature map of $\mathcal{H}$

$r \times \infty$

$K(x^{(t)}, \cdot)$

convergence result: (thm. 5) if $\|\theta^*\|_{HS} \le D$ and ℓ is convex and differentiable

and if $\mathbb{E}_{(x,y)\sim p} \|\nabla_\theta \ell(x,y;\theta)\|_{HS}^2 \le M^2$

then averaged projected SGD with step-size $\gamma = \frac{2D}{M\sqrt{n}}$

gives $\boxed{\mathbb{E}[\ell(\theta^{[n]})] - \ell(\theta^*) \le \frac{2DM}{\sqrt{n}}}$ (convergence result)

$\theta^{[n]} = \frac{1}{n}\sum_{t=1}^{n} \theta^{(t)}$

thm. 6 learning complexity

let θ^* minimizes $L(\theta)$ with $\|\theta^*\|_{HS} \le D$

choosing $n \ge \frac{4D^2M^2}{H(\epsilon)^2}$ implies $\mathbb{E}[L(\theta^{[n]})] \overset{(5)}{\le} L(\theta^*) + \epsilon$

define a meaningful scale

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

Moral here:
* some losses are harder than others (worst case sample complexity)

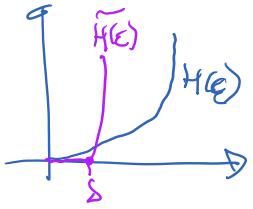
* some losses are harder than others (worst case sample complexity)
[0-1 loss is difficult in general]

* have linked computation & 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 distribution) modulo $\|g\|_{\mathcal{H}} \leq D$ constraint
→ still need more theory? (e.g. role of $p(y|x)$? or other surrogates?)

* follow-up paper: inconsistent surrogate loss with computational/statistical advantage NIPS 2015



15h38

part II: convex optimization

Motivation: $\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} \ell(x^{(i)}, y^{(i)}, w)$ (convex surrogate loss)

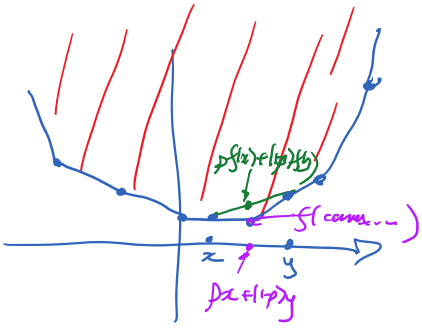
Convex analysis recap:

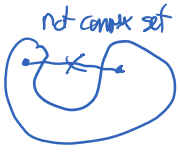
$f: \mathbb{R}^d \to \mathbb{R}$

f is convex $\Leftrightarrow$

$f(px + (1-p)y) \leq pf(x) + (1-p)f(y) \quad \forall x,y$ $p \in [0,1]$

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

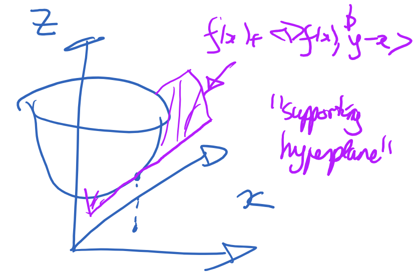




$\text{epigraph}(f) \triangleq \{(x,y) : y \geq f(x)\}$
$x \in \mathbb{R}^d$, $y \in \mathbb{R}$

* if f is differentiable at x and convex

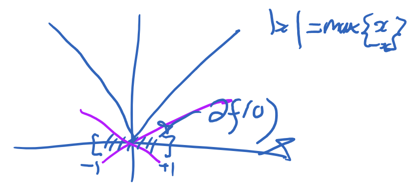
$$\Rightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall y$$

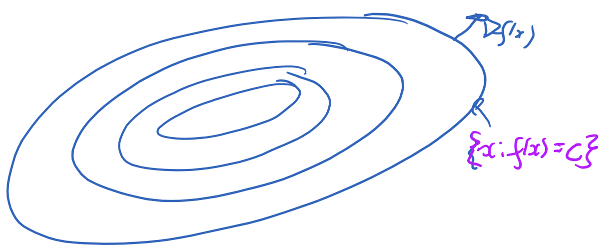


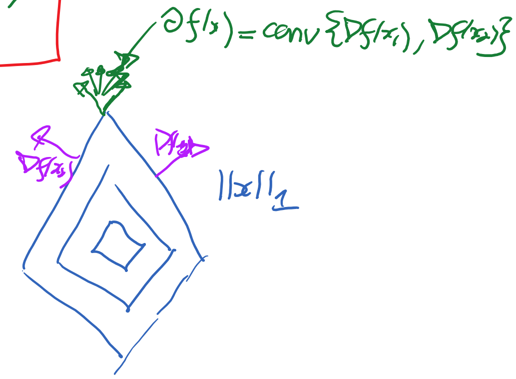
(suppose f is convex)

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

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







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

$$\partial f(x) = \text{conv}\{ \nabla f_{i^*}(x) : i^* \in \arg\max_j f_j(x) \}$$

(Danskin's thm.)

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

Clarke's subdifferential → nice gen. to non-convex fcts.

some standard assumptions :

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

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

(μ: strong convexity constant; $\langle \partial f(x), y-x \rangle$ means $\langle v, y-x \rangle$ for any $v \in \partial f(x)$)

f is L-smooth ie. f has L-Lipschitz gradient $\forall x$ (with ref. norm $\|\cdot\|$)

$$\Longleftrightarrow \|\nabla f(x) - \nabla f(y)\|_* \leq L \|x - y\| \quad \forall x, y$$

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

generalized C-S.

$$\langle w, v \rangle \leq \|w\|_* \|v\|$$

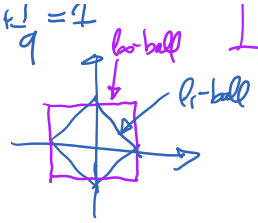
$(\|\cdot\|_p)^* = \|\cdot\|_q$ where $\frac{1}{p} + \frac{1}{q} = 1$

$p = 2 \Rightarrow q = 2$ $\quad$ (ℓ_∞-ball ↔ ℓ_1-ball)

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

$p=2 \Rightarrow q=2$

$p=1 \Rightarrow q=\infty$



$\langle w, v \rangle \leq \|w\|_* \|v\|$