Lecture 9 - convex optimization

Tuesday, February 14, 2023 2:31 PM

today: • finish calibration
• start convex optimization

optimization normalization for calibration function

setup let $s(x,\hat{y})$ be of the form $F\theta(x)$ $\quad \theta(x) \in \mathbb{R}^r$ $\quad \theta \rightsquigarrow r \times D$ (kernel)

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

$\theta(\cdot)$: optimization variables

$\ell(\theta) = \mathbb{E}_{(x,y)\sim\rho}\, \ell(x,y,\theta)$ $\quad$ projection $(x^{(t)}, y^{(t)}) \sim \rho$

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

B: ball of radius D around 0

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

$r \times D$ $\quad \in \mathbb{R}^k$

Φ: feature map of $\mathcal{H}$, $k(x^{(t)}, \cdot)$

$\theta^{(t)} = F^T \sum_t \alpha_t(\cdot)\, \Phi(x^{(t)})^T$ $\quad (r \times D)$

$\theta^{(t)}(x) \longrightarrow \langle \Phi(x^{(t)}), \Phi(x) \rangle$

$= F^T \sum_t \alpha_t(\cdot) \langle \Phi(x^{(t)}), \Phi(x) \rangle$

$\langle \Phi(x^{(t)}), \Phi(x) \rangle = k(x^{(t)}, x)$ $\quad \|\theta\|_{HS}^2 = \sum_{j=1}^{r} \|\theta_j\|_{RKHS}^2$

convergence results (thm. 5)

if $\|\theta^*\|_{HS} \le D$ & ℓ is convex & differentiable

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

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

$\bar{\theta}^{[n]} = \frac{1}{n} \sum_{t=1}^{n} \theta^{(t)}$ gives $\mathbb{E}[\ell(\bar{\theta}^{[n]})] - \ell(\theta^*) \le \frac{2DM}{\sqrt{n}}$

thm. 6 learning complexity

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

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

choosing $n \geq \frac{4D^2M^2}{\tilde{H}(\varepsilon)^2}$ implies $\mathbb{E}[L(\theta^{[n]})] \leq L(\theta^*) + \varepsilon$

offers a meaningful scale

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

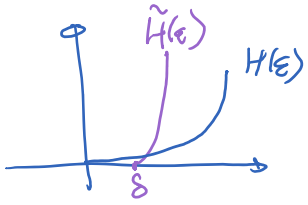
moral here:

* some losses ℓ are harder than others (worst case sample complexity)
  [0-1 loss is difficult in general]
  "too harsh of a loss"
* have linked computation to statistical performance in consistency framework
  → convex surrogate loss
* (non-parametric analysts) could handle dependance on x using RKHS

caveats:

- distribution free result (i.e. worst case over all distributions)
  → still need more theory!
  (e.g. rate $p(y|x)$)
  or other surrogates
  model $\|\theta^*\|_{\mathcal{H}} \leq D$ constraint
  could be big for "bad" p
  [→ no free lunch]

* follow-up: inconsistent surrogate loss with computational/statistical advantages [NeurIPS 2018]



15h18

part II: convex optimization

convex surrogate loss

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

convex analysis recap:

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

f is convex $\Leftrightarrow$

$\Leftrightarrow$ epigraph(f) is convex

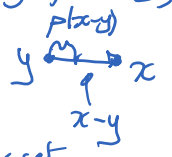
f is convex $\Leftrightarrow$

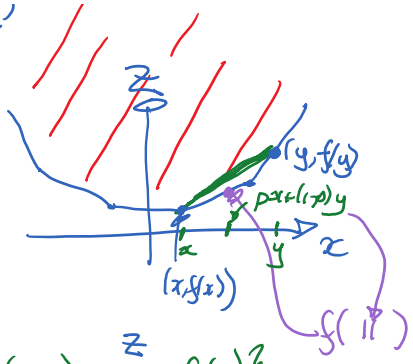
$f(\rho x + (1-\rho) y) \leq \rho f(x) + (1-\rho) f(y) \quad \forall x, y$

$\forall \rho \in [0,1]$

convex combination between x & y

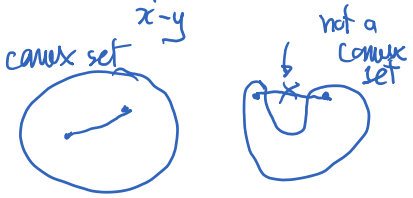
$y + \rho(x-y)$

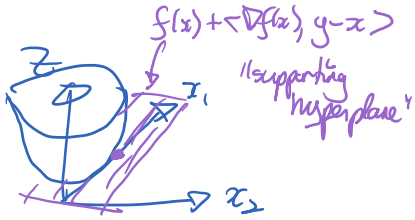




epigraph $f) \triangleq \{(x, y) : y \geq f(x)\}$

$\in \mathbb{R}^d$ $\in \mathbb{R}$





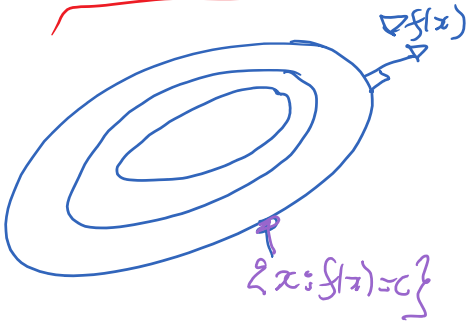
* f is differentiable at x and convex

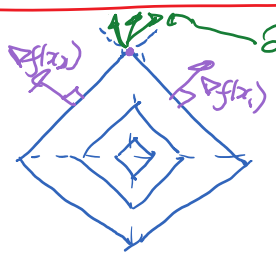
$\Rightarrow f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall y$

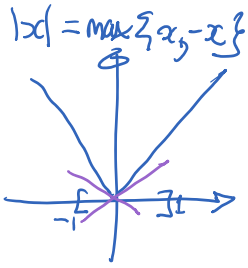
(suppose f is convex) subdifferential

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

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







$\partial |x| \big|_{x=0} = [-1, 1]$

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

$\partial f(x) = \text{conv}\{\nabla f_{i^*}(x) : i^* \in \text{argmax}_i f_i(x)\}$

(Danskin's thm.)

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

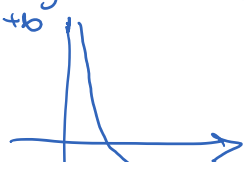
Clarke's subdifferential → nice gen. to non-convex

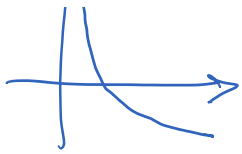
$f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$

"extended reals"

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

e.g. $-\log x$ is convex





$\Rightarrow \min_x f(x) = \min_{x \in \mathrm{dom}(f)} f(x)$

$\mathrm{dom}(f) = \{x \in \mathbb{R} : x > 0\}$

Some standard assumptions

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

μ: strong convexity constant

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

f is μ-strongly convex $\Leftrightarrow f - \frac{\mu}{2}\|\cdot\|^2$ is convex

f is L-smooth i.e. f has L-Lipschitz continuous gradient $\forall x$ (with respect to norm $\|\cdot\|$)

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

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

$\Leftrightarrow$ generalized C.-S.

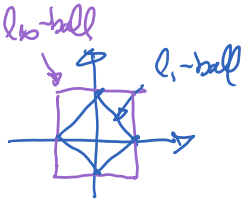
$\langle w, v \rangle \leq \|w\|_* \|v\|$

$(\|\cdot\|_p)_* = \|\cdot\|_q$

where $\frac{1}{p} + \frac{1}{q} = 1$

$p = 2 \Rightarrow q = 2$

$p = 1 \Rightarrow q = \infty$



Fundamental descent lemma:

When ∇f is L-Lipschitz (Lemma holds even if f is not convex)

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

apply lemma with $y = y_\gamma$

$\star$ $f(\underbrace{x - \gamma \nabla f(x)}_{y_\gamma}) \leq f(x) - \gamma \underbrace{\langle \nabla f(x), \nabla f(x) \rangle}_{\|\nabla f(x)\|_2^2} + \frac{L}{2}\gamma^2 \|\nabla f(x)\|^2$

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

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

→ minimize RHS with respect to γ

gives $\boxed{\gamma^* = \frac{1}{L}}$

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

proof intuition of descent lemma:

think of 2nd order Taylor expansion

integral form of the remainder

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

Hessian of f

$\leq L\|y-x\|^2$

f is L-smooth & twice diff. $\Rightarrow$ top e-value of $H \leq L$ in absolute value

$$v^T H v \leq \lambda_{max}(H)\|v\|_2^2$$

(λ_{max} in absolute value)