Lecture 9 - convex optimization

Thursday, February 11, 2021 13:35

today: • finish calibration
• start convex optimization

optimization normalization for calibration functions

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

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

$\in \mathbb{R}^K$ $\quad$ optimization variables

$\ell(\theta) = \mathbb{E}_{(x,y)\sim p}\, \ell(x,y;\theta)$ $\quad$ $(x^{(t)}, y^{(t)}) \overset{iid}{\sim} p$

run projected kernelized SGD on $\ell(\theta)$ i.e. $\theta^{(t+1)} = \underset{\text{projection}}{P_{\text{ball of radius } D \text{ around } 0}}\left(\theta^{(t)} - \gamma \nabla_\theta \ell(x^{(t)}, y^{(t)}; \theta^{(t)})\right)$

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

$r \times \infty$ $\quad$ feature map of $\mathcal{H}$ $\quad$ $k(x^{(t)}, \cdot)$

$\theta^{(t)} = F^T \sum_t \alpha_t \Phi(x^{(t)})^T$

$r \times \infty$ $\quad$ $\theta^{(t)}(x) \rightsquigarrow \langle \Phi(x^{(t)}), \Phi(x) \rangle$ $\quad$ $k(x^{(t)}, x)$

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

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

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 $\|\theta^*\|_{HS} \leq D$

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

define a meaningful scale

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

⊛ moral here:

* some losses 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 losses
* (non-parametric analysis) could handle dependence on x using RKHS

caveats:

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

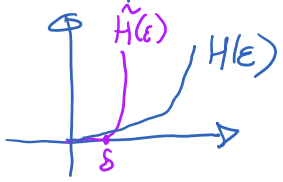
models $\|\theta^*\|_{HS} \leq D$ constraint

→ still need more theory?

could be big for "bad p"
[→ no free lunch]

(e.g. role of $p(y|x)$ or other surrogates?)

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



14:20

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:

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

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

f is convex $\iff$

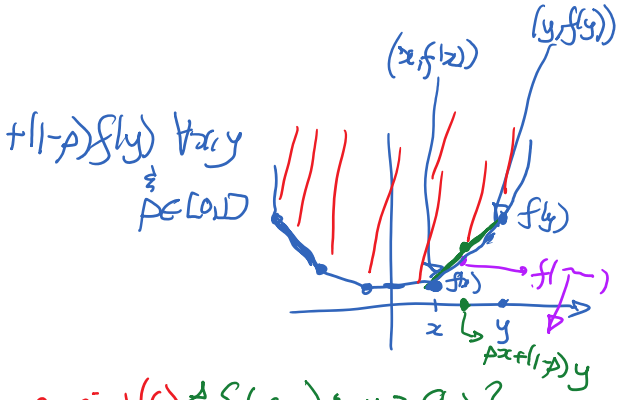
$$f(\rho x + (1-\rho) y) \le \rho f(x) + (1-\rho) f(y) \quad \forall x, y$$

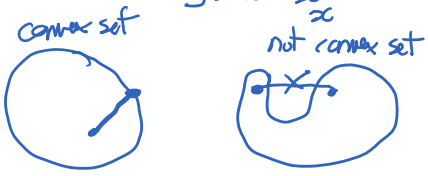
$\rho \in [0,1]$

Convex combination between x & y

$y + \rho(x-y)$

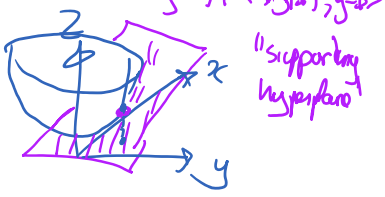
$\rho(y-x)$





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

$x \in \mathbb{R}^d$, $y \in \mathbb{R}$



* If f is differentiable at x and convex

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

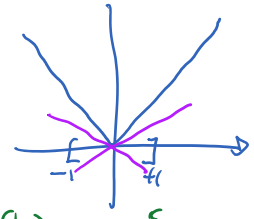
(suppose f is convex)

subdifferential

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

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

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



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

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

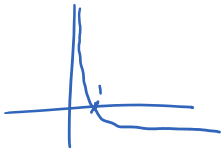
$$\partial f(x) = \text{conv}\{\nabla f_{i^*}(x) : i^* \in \arg\max_j f_j(x)\}$$

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

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

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

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



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

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

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

Some standard assumptions:

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

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

f is μ-strongly convex $\Leftrightarrow$ $f - \frac{\mu}{2}\|\cdot\|_2^2$ is convex

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

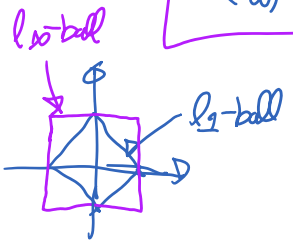
$\Leftrightarrow \|\nabla f(x) - \nabla f(y)\|_* \leq L\|x-y\| \quad \forall x, y$

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

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

$p=2 \Rightarrow q=2$

$p=1 \Rightarrow q=\infty$



Generalized C-S:

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

Fundamental descent lemma:

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

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

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

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

$\geq 0 \Leftrightarrow 0 < \gamma < \frac{2}{L}$

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

$\rightarrow$ 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 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

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

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

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