Lecture 9 - convex optimization

Tuesday, February 13, 2024 9:30 AM

today: - finish calibration
- start convex optimization

optimization normalization for calibration fct.

setup let $s(x,\tilde{y})$ be of the form $F\theta(x)$ $\quad \theta(x) \in \mathbb{R}^r$ (F: $k \times r$)

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

θ_j: optimization variables $\quad \theta_j(x) = \langle \theta_j, \Phi(x) \rangle_{\mathcal{H}}$

$\mathcal{J}(\theta) = \mathbb{E}_{(x,y)\sim P}\, J(x,y,\theta)$ $\quad$ $(x^{(t)}, y^{(t)})$ iid P

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

Π: projection $\quad \Theta_D$: ball of radius D around 0

"∇_θ" $J(x^{(t)}, y^{(t)}; \theta) \approx$

$r \times b$

$F^T \nabla_s J(x^{(t)}, y^{(t)}; s^{(t)})\, \Phi(x^{(t)})^T$

$r \times k$ $\quad \nabla_s J \in \mathbb{R}^k$ $\quad F\theta^{(t)}(x^{(t)})$ $\quad$ $k \times 1$ $\quad$ $1 \times b$ $\quad$ feature map of $\mathcal{H}$ $k(x^{(t)}, \cdot)$

$\theta^{(n)} = F^T \sum_{t=0}^{n-1} \vec{\alpha}_t\, \Phi(x^{(t)})^T$ $\quad$ $\vec{\alpha}_t = -\gamma \nabla_s J(x^{(t)}, y^{(t)}, s^{(t)})$ $\quad$ $k \times 1$

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

convergence result (thm. 5)

If $\|\theta^*\|_{\mathcal{HS}} \le D$ & J is convex and differentiable

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

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 $\quad \mathbb{E}[\mathcal{J}(\bar{\theta}^{[n]})] - \mathcal{J}(\theta^*) \le \frac{2DM}{\sqrt{n}}$

$\|\theta\|^2_{\mathcal{HS}} = \sum_{j=1}^{r} \|\theta_j\|^2_{\mathcal{H}}$

thm. 6 learning complexity

Let θ^* minimize $L(\theta)$ with $\|\theta^*\|_{\mathcal{HS}} \le D$ ← "complexity measure" of θ^*

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

defines a meaningful scale

in the paper: we compute D & M & $\tilde{H}(\epsilon)$ for

defines a meaningful scale

in the paper: we compute D & M & $H(\varepsilon)$ for specific losses ℓ and the quadratic δ 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 losses

* (non-parametric analyses) could handle dependence on x using RKHS

caveats:

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

→ still need more theory?

(eg. role of $p(y|x)$ or other surrogates)

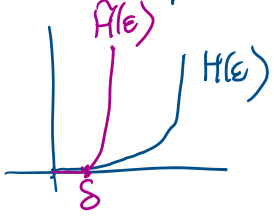
models $\|f^*\|_{HS} \leq D$ constraint

could be big for "bad p"

↳ no free lunch

(related to kernel choice?)

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



$\frac{\varepsilon^2}{C}$

10h17

Part II: Convex optimization

convex surrogate loss

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

convex analysis recap:

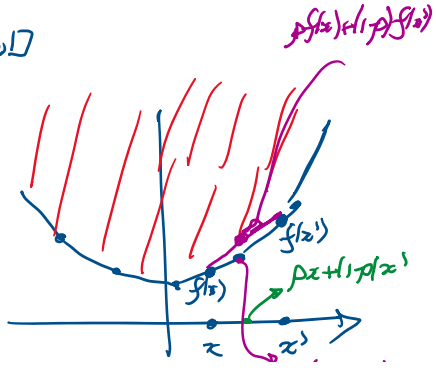
$f: \mathbb{R}^d \to \mathbb{R}$

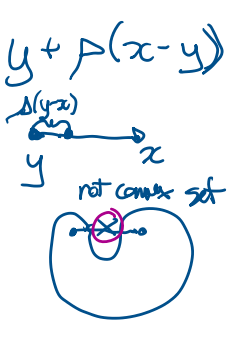
$\forall x, y \in \mathbb{R}^d \ \forall \rho \in [0,1]$

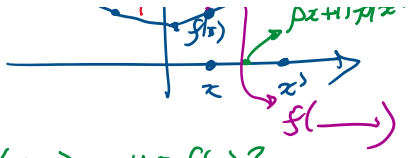
f is convex $\Leftrightarrow$ $f(\rho x + (1-\rho) y) \leq \rho f(x) + (1-\rho) f(y)$

convex combination between x & y

$y + \rho(x - y)$





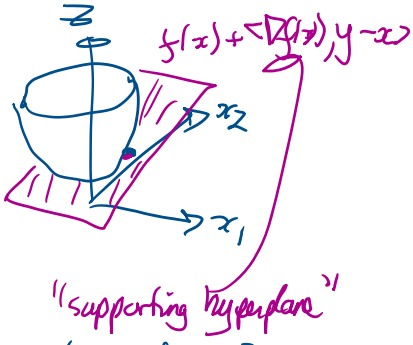


$\text{epigraph}(f) \triangleq \{(x,y) : y \geq f(x)\}$

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

* if f is differentiable at x and convex

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

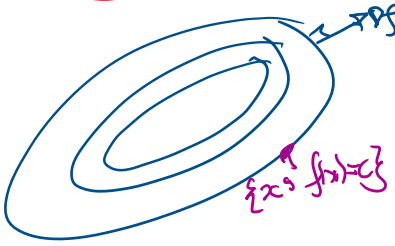


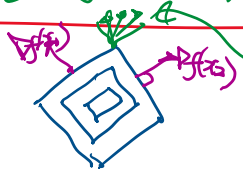
(suppose f is convex)

subdifferential

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

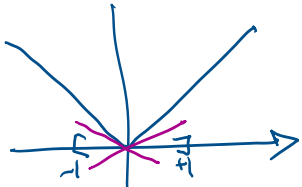
$\Leftrightarrow \forall y \in \text{dom}(f), \quad f(y) \geq f(x) + \langle v, y-x \rangle$





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

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



$\partial f(0) = [-1, +1]$

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

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

(Danskin's thm)

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

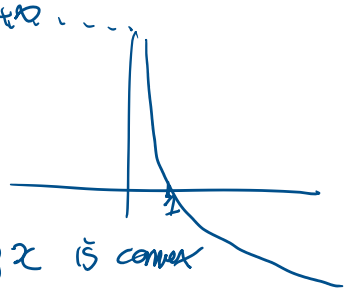
Clarke's subdifferential → nice gen. to non-convex

(limits of gradients for locally Lipschitz → a.e. differentiable)

* Extended reals convention for conv. fct.:

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

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



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 $\iff f(y) \geq f(x) + \langle \partial f(x), y-x \rangle + \frac{\mu}{2}\|y-x\|^2$

(μ: strong convexity constant; $\forall x, y \in \text{dom}(f)$; $\langle v, y-x \rangle$ for any $v \in \partial f(x)$)

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

f is L-smooth ie. f has L-Lipschitz cont. gradient $\forall x \in \text{dom}(f)$ (with respect norm $\|\cdot\|$)

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

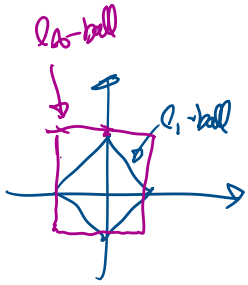
"dual norm" $\|w\|_* \triangleq \sup_{\|v\| \leq 1} \langle w, v \rangle$

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

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 is not convex)

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

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

$= f(x) - \gamma\left(1 - \frac{\gamma L}{2}\right)\|\nabla f(x)\|_2^2$

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

$\rightarrow$ minimize RHS with respect to γ

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

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

$$f(y^*) \le f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2$$

$$f(x) - f(x^*) \le \frac{\|\nabla f(x)\|^2}{2\mu}$$

$$f(x_{t+1}) - f(x^*) \le f(x_t) - f(x^*) - \frac{\mu}{L}\left(f(x_t) - f(x^*)\right)$$

$$\varepsilon_{t+1} \le \left(1 - \frac{\mu}{L}\right)\varepsilon_t \qquad (1-x)^t \le \exp(-xt) \ \forall x$$

$$\varepsilon_t \le \left(1-\frac{\mu}{L}\right)^t \varepsilon_0 \le \exp\left(-\frac{\mu}{L}t\right)\varepsilon_0 \quad \text{"linear rate"}$$

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

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

f is L-smooth $\Rightarrow$ top e-value of $H \le L$

$\hookrightarrow$ twice diff. (in absolute value)

$$v^T H v \le \lambda_{\max}(H)\|v\|$$