Lecture 13: opt: convex $\mathcal{L}$ & Lip H; smooth $\mathcal{L} \circ H$.

Outline for next initialization proofs:

* Recall standard convex opt settings & rates
  * convex Lipschitz & Bounded $\Rightarrow$ $1/\sqrt{t}$ to global
  * smooth $\Rightarrow$ $1/t$ to local opt
  * smooth & convex $\Rightarrow$ $1/t$ to global opt
  * smooth & strongly convex $\Rightarrow$ e^{-t} to global opt

instead consider $\nabla(\mathcal{L} \circ H)$

left column vanish

strategy: make minimal assumptions on H so that we get these rates

+ worst case terms for H.

Bench: recovers, extends, & unifies proofs from the literature

Lemma (from last time: Lipschitz & bounded).
Given convex $\ell:\mathbb{R}^n \to \mathbb{R}$, feature mapping $H:\mathbb{R}^p \to \mathbb{R}^n$, closed convex $U \subseteq \mathbb{R}^p$, and:

$$\varepsilon := \max_{z \in U} \left\| H(z) - \left[H(w_s) + \partial H(w_s)^T (z - w_s)\right] \right\|_\infty$$
(linearization error)

$$A := \max_{s \le t} \|\partial \ell(H(w_s))\|_1 \quad \text{(loss Lipschitz)}$$

$$B := \max_{s \le t} \|\partial H(w_s)\|_{2,\infty} \quad \text{(features Lipschitz)}$$

Then, choosing $\eta := \frac{\eta_0}{\sqrt{t}}$ & arbitrary $z \in U$,

$$\frac{1}{2\eta_0\sqrt{t}}\|w_t - z\|^2 + \min_{s \le t} (\ell \circ H)(w_s) \le \frac{1}{2\eta_0\sqrt{t}}\|w_0 - z\|^2 + \frac{1}{t}\sum_{s \le t} (\ell \circ H)(w_s)$$

$$\le \frac{1}{2\eta_0\sqrt{t}}\|w_0 - z\|^2 + \underbrace{(\ell \circ H)(z)}_{\le R} + \frac{A\varepsilon}{2\eta_0} + \frac{\eta_0 A^2 B^2}{2\sqrt{t}}.$$

(interp: training error $\le$ $\underbrace{z}_{x_{opt}}$'s training error $+ \frac{1}{\sqrt{t}}$)

Examples: $F(x; V) = \sum_i a_i \sigma(v_i^T x) = a^T \sigma(Vx)$

$$H(w) = \begin{bmatrix} y_1 F(x_1; w) \\ \vdots \\ y_n F(x_n; w) \end{bmatrix}$$

$$\|\partial H(w)\|_{2,\infty}^2 = \left\| \begin{bmatrix} | & & | \\ y_1 \partial F(x_1; w) & \cdots & y_n \partial F(x_n; w) \\ | & & | \end{bmatrix} \right\|_{2,\infty}^2 \le \max_i \|\partial F(x_i; w)\|_2^2$$

$$\le \max_i \sum_i a_i^2 \left\| \sigma'(v_i^T x_i) x_i \right\| \le \max_i \sum_i \frac{1}{n} = 1.$$

$$\varepsilon = \frac{O\left(B^{4/3} + B\ln\left(\frac{1}{\delta}\right)^{1/4}\right)}{n^{1/6}}; \quad \text{choose } n = B^8 t^3 \ln\left(\frac{1}{\delta}\right)^2$$

where $B = \sup_{a \in U, b \in U} \|a - b\|$

$$\Rightarrow \quad \varepsilon \le \frac{1}{\sqrt{t}}$$

$$\Rightarrow \min_{s \le t} (\ell \circ H)(w_s) \le (\ell \circ H)(z) + O\left(\frac{1}{\sqrt{t}}\right).$$

"step size" or "learning rate"

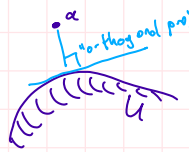
$$w_{s+1} := \Pi_U\left(w_s - \eta \, \partial\left(\ell \circ H\right)(w_s)\right)$$

feature mapping

convex loss

$$\Pi_U(a) := \operatorname{argmin}_{v \in U} \left\{ \tfrac{1}{2}\|a - v\|^2 : v \in U \right\}.$$

"orthogonal projection"



Examples:

* H = identity: $\varepsilon = 0$,

$H(w) = w \Rightarrow \partial H(w) = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$

$\|\partial H(w)\|_{2,\infty} = \|1 \cdots 1\|_\infty = 1.$

* linear: $H(w) = Zw = \begin{bmatrix} - y_1 x_1^T - \\ \vdots \\ - y_n x_n^T - \end{bmatrix} w$, $\partial H(w) = \begin{bmatrix} | & & | \\ y_1 x_1 & \cdots & y_n x_n \\ | & & | \end{bmatrix}$

$\varepsilon = 0$.

$\Rightarrow \|\partial H(w)\|_{2,\infty} \le \max_i \|x_i\|_1 \le 1.$

* Smooth & Lipschitz activations, e.g., $\sigma(z) = \frac{1}{1 + \exp(-z)}$

By same calculation: $\|\partial H(w_s)\| \le 1$,

and $\varepsilon \le \frac{B^2}{\sqrt{n}} \Rightarrow n = t B^4$

Loose ends

Proof of Lemma. We didn't quite finish; recall MD lemma

$$\|w_t - z\|^2 \le \sum_{s\le t} \lambda_s \langle g_s, z - w_s\rangle + \sum_{s\le t} \eta^2 \|g_s\|^2 + \|w_0 - z\|^2$$

where $g_s \in \partial(\ell \circ H)(w_s)$

where $\|g_s\|_2 = \|\partial H(w_s)\, \partial L(H(w_s))\|_2 = \left\| \sum_{i=1}^n \partial H(w_s)_{:,i}\, \partial \ell(H(w_s))_i \right\|$

$$\le \sum_{i=1}^n |\partial \ell(H(w_s))_i| \cdot \|\partial H(w_s)_{:,i}\|_2 \le \|\partial \ell(H(w_s))\|_1 \cdot \max_i \|\partial H(w_s)_{:,i}\|_2.$$

$$\max_i \|\partial H(w_s)_{:,i}\|_2 = \|\partial H(w_s)\|_{2,\infty}$$

Loose end #2: projection. Where it appears: $\|w_{s+1} - z\|^2 \overset{\text{MD lemma}}{=} \|\Pi_U(w_s - \eta g) - z\|^2$

$$\le \|w_s - \eta g - z\|^2 \quad ??$$

$$= \|w_s - z\|^2 + (\text{cross and } g \text{ terms})$$

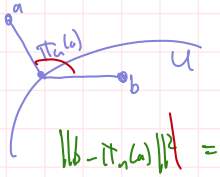
Proposition. If $U \subseteq \mathbb{R}^d$ is closed convex, $\forall a \in \mathbb{R}^d$, $\forall b \in U$, $\|\Pi_U(a) - b\| \le \|a - b\|$.

Proof. By first-order optimality conditions applied to $\operatorname{argmin}\{\frac{1}{2}\|a - v\|^2 : v \in U\}$,

then $\forall b \in U$, $\langle a - \Pi(a), b - \Pi(a)\rangle \le 0$

$$\| \qquad \|$$

$$\langle a - b + b - \Pi_U(a),\ b - \Pi_U(a)\rangle$$



$$\|b - \Pi_U(a)\|^2 = \langle b - \Pi(b),\ b - \Pi_U(a)\rangle \le \langle b - a,\ b - \Pi_U(a)\rangle$$

$$\le \|b - a\| \cdot \|b - \Pi_U(a)\|$$

Refinement to lemma so that $n \ll t$ is possible

Lemma. Same setup as previous convex + Lipschitz lemma, except additionally $\|\partial \ell(H(w_s))\|_1 \le C \cdot \ell(H(w_s))$ (true with $C=1$ for logistic).

Then,

$$\frac{1}{\eta t}\|w_t - z\|^2 + \frac{1}{t}\sum_{s<t}\left(1 - \left(C\eta + \frac{\eta}{2}AB^2\right)\right)(\ell\circ H)(w_s) \le \frac{1}{\eta t}\|w_0 - z\|^2 + (\ell\circ H)(z).$$

Remark. Suppose $\|x\| \le 1 \Rightarrow B^2 = 1$. Suppose logistic: $A > 1$, $C = 1$.

pick $\eta = 1$, want $\varepsilon \le \frac{1}{4}$; e.g. $n = 2^{24} B^8 \ln(1/\delta)^2$.

$$\Rightarrow \frac{2}{t}\|w_t - z\|^2 + \frac{1}{t}\sum_{s<t}(\ell\circ H)(w_s) \le \frac{2}{t}\|w_0 - z\|^2 + 4(\ell\circ H)(z).$$

(approach only makes sense if $(\ell\circ H)(z) = O\left(\frac{1}{t}\right)$; whereby rhs is $O\left(\frac{1}{t}\right)$.)

Proof comment: same proof as before, but use $\|\partial\ell(H(w_s))\|_1 \le C\ell(H(w_s))$ & collect terms.

Remark.
Point of this page: $\frac{1}{t}$ rate and width $\ll t$ possible.

Next time.

Suppose $\ell \circ H = \hat{R}$ is β-smooth (i.e., $\|\nabla \hat{R}(a) - \nabla \hat{R}(b)\| \le \beta \|a - b\|$).

$\Rightarrow$ GD with $\eta = \frac{1}{\beta}$ satisfies

$$\min_{s \le t} \|\nabla \hat{R}(w_s)\|^2 \le 2\beta \left(\frac{f(w_0) - f(w_t)}{t} \right)$$

interp: come close to local opt

If $\hat{R} = \ell \circ H$ & same setup as before

$\Rightarrow$ $\frac{1}{t}$-rks rate

very common ("optimality for nonconvex problems")