Lecture 16: end of strong convexity; weak implicit regularization

Remark: (failure of unification.)
Literature has incomparable results across settings even within near interpolation (e.g., smooth activation vs ReLU, least squares vs logistic).

Recall Lipschitz & bounded setting, which in strongest form required

$$\|\partial \ell\| \le B \cdot \ell \quad \text{(holds for logistic)}.$$

Plan:

* Today is lecture 6/6 of near-int opt.
* Next week: Clarke differential & GF
* Outside NTK ("margins" and other stuff).

* At least 3 Gen lectures including 1 on double descent & interpolation.

---

* HW3 (6b) Tuesday, typed notes with it, 1 "coding" problem

Theorem. Suppose $\hat{R}$ β-smooth (2° H), $\mathcal{L}$ is λ-sc, $\sigma_{LB} := \min_{s \le t} \sigma_n(H(w_s)) > 0$

define $\bar{y} := \arg\min_{\bar{y}} \mathcal{L}(\bar{y})$, and $\eta \le \frac{1}{\beta}$. Then

$$\hat{R}(w_t) - \mathcal{L}(\bar{y}) \le (\hat{R}(w_0) - \mathcal{L}(\bar{y})) \exp(-t\eta\lambda\sigma_{LB}^2).$$

Proof. (last time, based on $\hat{R}(w_{s+1}) \le \hat{R}(w_s) - \frac{\eta}{2}\|\partial\hat{R}(w_s)\|^2 \le \hat{R}(w_s) - \frac{\eta}{2}\|\partial H(w_s)\|_2^2 \cdot \|\partial\mathcal{L}\|_2^2$ //

Issues: β is time varying: $\|\partial\hat{R}(a) - \partial\hat{R}(b)\|^2 \overset{?}{\le} \beta\|a-b\|$

$= \|\partial H(a)\partial\mathcal{L}(H(a)) - \partial H(b)\partial\mathcal{L}(H(b))\|$.

σ_{LB} depends on algorithm

Lemma. Suppose $\|\partial H(a) - \partial H(w_0)\| \le \frac{1}{2}\sigma_n(H(w_0))$ $\forall \|a - w_0\| \le B$.

Then $\sigma_1(\partial H(a)) \le \frac{3}{2}\sigma_1(\partial H(w_0))$, and $\sigma_n(\partial H(a)) \ge \frac{1}{2}\sigma_n(H(w_0))$.

[Reasonable if ∂H smooth: $\|\partial H(a) - \partial H(w_0)\| \le \beta_H\|a - w_0\|$, so we assumed $\|a - w_0\| \le \frac{\sigma_n(H(w_0))}{2\beta_H}$]

Proof. UB: $\sigma_1(\partial H(a)) = \|\partial H(a)\|_2 \le \|\partial H(w_0)\|_2 + \|\partial H(a) - \partial H(w_0)\|_2$

$\le \sigma_1(\partial H(w_0)) + \frac{1}{2}\sigma_n(H(w_0)) \le \frac{3}{2}\sigma_n(\partial H(w_0))$

LB: $J_0 := \partial H(w_0)$, $J := \partial H(a)$. Then

$$\sigma_n(J)^2 = \lambda_{min}(JJ^T) = \min_{\|v\|=1}(v^T JJ^T v) = \min_{\|v\|\le 1} v^T(J - J_0 + J_0)(J - J_0 + J_0)^T v$$

(Variational form for singular values. Recall $J \in \mathbb{R}^{n \times p}$)

$$= \min_{\|v\|=1} v^T J_0 J_0^T v + 2v^T(J - J_0)J_0^T v + v^T(J - J_0)(J - J_0)^T v$$

$$\ge \min_{\|v\|=1} \|J_0^T v\|^2 - 2\|(J - J_0)^T v\| \cdot \|J_0^T v\| + \|(J - J_0)^T v\|^2$$

$$= \min_{\|v\|=1} (\|J_0^T v\| - \|(J - J_0)^T v\|)^2 \ge (\tfrac{1}{2}\sigma_n(J_0))^2.$$ //

(where $\|J_0^T v\| \ge \sigma_n(J_0)$ and $\|(J - J_0)^T v\| \le \|J - J_0\|_2 \cdot \|v\| \le \frac{1}{2}\sigma_n(J_0)$)

$$\min_{\|v\|=1} (\|Jv\|)^2$$

$$\|(J - J_0 + J_0)v\| \le \|Jv\| + \|J - J_0\|$$

$$\|A\| = \|A - B + B\| \le \|A - B\| + \|B\|$$

$$\Rightarrow \|B\| \ge \|A\| - \|A - B\|$$

($B \leftarrow Jv$, $A \leftarrow J_0 v$, $A - B \leftarrow J_0 v$, $J v$)

Easier version: (Thanks!)

$$\sigma_n(J)^2 = \min_{\|v\|\le 1} v^T J J^T v = \left(\min_{\|v\|\le 1} \|Jv\|\right)^2 \ge \left(\min_{\|v\|\le 1} \|J_0 v\| - \|(J - J_0)v\|\right)^2$$

$$\ge \left(\min_{\|v\|\le 1} \|J_0 v\| - \|v\| \cdot \|J - J_0\|\right)^2$$

$$\ge \left(\min \|J_0 v\| - \frac{\sigma_1(J_0)}{2}\right)^2 = \left(\frac{\sigma_1(J_0)}{2}\right)^2$$ //

(In the typed notes: show how to control $\|w_t - z\|$ for this strongly convex setting.)

Weak implicit regularization: GD automatically finds good predictors & stays close to initialization.

Theorem (from before): Assume $w_{s+1} := \Pi_U\left(w_s - \eta \partial \hat{R}(w_s)\right)$, let $z \in U$ be given &

$$\varepsilon := \max_s \max_{s \in t} \left| H(z)_i - (H(w_s)_i - \langle \partial H(w_s)_{:i}, w_s - z \rangle \right|,$$

$$\|\partial \ell(w_s)\| \le A\, \ell(w_s) \quad \forall s \qquad \text{(true for Logistic with } A=1\text{)}$$

$$B := \max_{s \in t} \|\partial H(w_s)\|_{2,\infty} \qquad (\le 1 \text{ for shallow ReLU}).$$

$$\Rightarrow \frac{1}{2t\eta}\|w_t - z\|^2 + \frac{1}{t}\left(1 - \varepsilon - \frac{n}{2} A B^2\right) \sum_{s \in t} \hat{R}(w_s) \le \frac{1}{2t\eta}\|w_0 - z\|^2 + \hat{R}(z).$$

Goal: show don't need to project.

Why: (a) no projection in practice (b) some other settings where projection isn't an option.

Pick t, & pick z & radius R so that near optimal $\hat{R}(z) \le \frac{R^2}{2t}$

and near initialization $\|w_0 - z\| \le \frac{R}{2}$

Recall for ReLU: $\varepsilon \le \frac{1}{m^{1/6}}\left(4R^{4/3} + 2R\ln(1/\delta)^{1/4}\right)$,

$\Rightarrow$ $m \ge \left(\frac{1}{4}\left[4R^{4/3} + 2R\ln(1/\delta)^{1/4}\right)\right.$, then $\varepsilon \le \frac{1}{4}$

$\Rightarrow (1 - \varepsilon - \frac{n}{2}AB^2) \ge \frac{1}{2}$.

Apply theorem with $U := \left\{ \|w - w_0\|^2 \le (1+4\eta)R^2 : w \in \mathbb{R}^p \right\}$.

Suppose $\exists s$ with $w_s - \eta \nabla \hat{R}(w_s) \notin U$, then

$$\frac{1}{2t\eta}\|w_{s+1} - z\|^2 \le \frac{1}{2t\eta}\|z - w_0\|^2 + \hat{R}(z) \le \frac{1}{2t\eta}\left(\frac{R}{2}\right)^2 + \frac{R^2}{2t} \le \frac{R^2}{t}$$

Note also $\|w_{s+1} - w_0\|^2 = \|w_{s+1} - z\|^2 + 2\langle w_{s+1} - z, z - w_0\rangle + \|z - w_0\|^2$

$$\le 2\|w_{s+1} - z\|^2 + 2\|z - w_0\|^2$$

$$\Rightarrow \|w_{s+1} - w_0\|^2 \le 2\|w_{s+1} - z\|^2 + 2\|z - w_0\|^2 \le 4\eta R^2 + \frac{R^2}{2}$$

$\Rightarrow$ did not exit U

s was arbitrary $\Rightarrow$ projection never encountered $\Rightarrow \max_{s \in t} \|w_s - z\|^2 \le (1+4\eta)R^2$