Lecture 27: even looser Rademacher bounds!

Theorem (Bartlett–Mendelson '02) $\sigma_i(0)=0$, σ_i is coordinate-wise ρ-Lipschitz

$$\mathrm{URad}\left(\{x \mapsto \sigma_L(W_L\sigma_{L-1}(W_{L-1}\cdots\sigma_1(W_1x)\cdots)) : \|W_i^T\|_{1,\infty}\le B\}_{|X}\right)$$
$$\le \|X\|_{2,\infty}\sqrt{2\ln d}\,(2\rho B)^L.$$

Rem (looseness). Proof doesn't allow $(W_i - W_i(0))$, therefore $\|W_i^T\|_{1,\infty} \underset{\approx}{} \sqrt{m}$ at init; super loose.

Proof. Define $\mathcal{F}_i :=$ "output of nodes in layer i"

$$= \begin{cases} \mathcal{F}_0 := \{x\mapsto x_j : j\in\{1,\dots,d\}\} \\ \mathcal{F}_{i+1} := \left\{x\mapsto \sigma_{i+1}\left(\sum_{j=1}^m v_j g_j(x)\right) : m\ge 0,\ g_j\in\mathcal{F}_i,\ \|v\|_1\le B\right\}\end{cases}$$

IH: $\mathrm{URad}(\mathcal{F}_{i|X}) \le (2\rho B)^i\|X\|_{2,\infty}\sqrt{2\ln d}$.

Base case: $\mathrm{URad}(\mathcal{F}_{0|X}) \le \left(\sup_{j\in\{1,\dots,d\}}\|X_{:j}\|_2\right)\sqrt{2\ln|\mathcal{F}_{0|X}|} = \|X\|_{2,\infty}\sqrt{2\ln d}$
$= (2\rho B)^0\|X\|_{2,\infty}\sqrt{2\ln d}$.

Massart finite lemma

Ind. step. $\mathrm{URad}(\mathcal{F}_{i+1|X}) = \mathrm{URad}\left(\{x\mapsto\sigma(\sum_{j=1}^m v_j g_j(x)) : m\ge 0,\ \|v\|_1\le B,\ g_j\in\mathcal{F}_i\}_{|X}\right)$

$\le \rho\,\mathrm{URad}\left(\{x\mapsto\sigma(B\sum_{j=1}^m v_j g_j(x)) : m\ge 0,\ \|v\|_1\le 1,\ g_j\in\mathcal{F}_i\}_{|X}\right)$

$\le \rho B\cdot\mathrm{URad}\left(\mathrm{conv}(\mathcal{F}_i\cup-\mathcal{F}_i)_{|X}\right)$

$\overset{\text{conv}}{=} \rho B\cdot\mathrm{URad}(\mathcal{F}_{i|X}\cup-\mathcal{F}_{i|X})$

$\overset{\text{union rule}}{\le} 2\rho B\,\mathrm{URad}(\mathcal{F}_{i|X})$

using $\sigma_i(0)=0$

$\overset{\text{IH}}{\le} (2\rho B)^{i+1}\|X\|_{2,\infty}\sqrt{2\ln d}$. //

Remarks (1) Source of looseness (a) not adapted to GD (searches over too many functions) (b) worst-cases between layers.

(2) Union rule typically misstated, & source of bugs in papers.

Original defn URad was $\mathrm{URad}_{|\cdot|}(V) = \mathrm{URad}(V\cup -V)$
$= \mathbb{E}_\varepsilon\sup_{u\in V}|\langle\varepsilon,u\rangle|$

But $\mathrm{URad}_{|\cdot|}\left(\bigcup_i V_i\right) \le \sum_i\mathrm{URad}_{|\cdot|}(V_i)$.

Union holds URad under some conditions.

(3) $\mathrm{URad}(\mathrm{conv}(V)) = \mathbb{E}_\varepsilon\sup_{u\in\mathrm{conv}V}\langle\varepsilon,u\rangle = \mathbb{E}_\varepsilon\sup_{k\ge0}\sup_{\alpha\in\Delta_k}\sup_{u_1}\dots\sup_{u_k}\langle\varepsilon,\sum_{i=1}^k\alpha_i u_i\rangle$

$= \mathbb{E}_\varepsilon\sup_{k\ge0}\sup_{\alpha\in\Delta_k}\sum_{i=1}^k\alpha_i\boxed{\sup_{u_i\in V}\langle\varepsilon,u_i\rangle}$

$\sup_{u\in V}\langle\varepsilon,u\rangle$

$= \mathbb{E}_\varepsilon\sup_{u\in V}\langle\varepsilon,u\rangle\left[\underbrace{\sup_{k\ge0}\sup_{\alpha\in\Delta_k}\sum_i\alpha_i}_{1}\right]$

$= \mathrm{URad}(V)$. //

Theorem (Golowich–Rakhlin–Shamir '18). $\sigma_i(0)=0$, σ_i coordinate-wise & 1-Lip

[check my lecture notes if need 1-homo]

$$\mathrm{URad}\left(\{x \mapsto \sigma_L(W_L \cdots \sigma_1(W_1 x)\cdots) : \|W\|_F \le B\}_{|X}\right) \le B^L \|X\|_F \left(1+\sqrt{2L\ln 2}\right).$$

Proof ideas.

know this variant of Lipschitz peeling lemma:

Lemma (Talagrand–Ledoux). $\vec{\ell}:\mathbb{R}^n\to\mathbb{R}^n$, $\ell(0)=0$, ℓ coordinate-wise ρ-Lip

$$\mathbb{E}_\varepsilon \sup_{u\in V} \exp\left(\sum \varepsilon_i \ell_i(u_i)\right) \le \mathbb{E}_\varepsilon \sup_{u\in V} \exp\left(\rho\sum \varepsilon_i u_i\right)$$

Proof. tricky case analysis.

$$X_0 = \begin{bmatrix} -x_1^T- \\ \vdots \\ -x_n^T- \end{bmatrix} \in \mathbb{R}^{n\times d}, \qquad X_i^w = \sigma_i(X_{i-1} W_{i-1}^T)$$

$$\mathrm{URad}(\mathcal{F}_{|X}) = \mathbb{E}_\varepsilon \sup_w \varepsilon^T X_L^w$$

$$= \mathbb{E}_\varepsilon \frac{1}{t} \ln \sup_w \exp(t\, \varepsilon^T X_L^w)$$

$$\le \frac{1}{t} \ln \mathbb{E}_\varepsilon \sup_w \exp(t\, \varepsilon^T X_L^w).$$

① induction: $\mathbb{E}_\varepsilon \sup_w \exp(t\, \|\varepsilon^T X_L^w\|) \underset{\text{many steps}}{\le} \mathbb{E} \sup_w \exp(t B^L \|\varepsilon^T X_0\|)$.

② $\|\varepsilon^T X_0\|$ is $\|X\|_F^2$-subGaussian & $\mathbb{E}\|\varepsilon^T X_0\|_2 \le \|X\|_F$.

③ pick $t := \sqrt{\dfrac{2\ln 2}{B^{2L}\|X\|_F}}$.

Theorem (★) (Bartlett – Foster – Telgarsky '17) $\sigma_i(0)=0$, ρ_i-Lip, not coordinate-wise

$$\mathrm{URad}\left(\left\{ x \mapsto \sigma_L(W_L \cdots \sigma_1(W_1 x) \cdots) : \begin{array}{l} \|(W_i - W_i(0))^T\|_{2,1} \le b_i \\ \|W_i\|_2 \le s_i \end{array} \right\}_{|X} \right)$$

$$\le \tilde{O}\left(\|X\|_F \left(\prod_i \rho_i s_i\right) \left(\sum_i \left(\frac{b_i}{s_i}\right)^{2/3} \right)^{3/2} \right).$$

Proof remarks.

① uses "covering numbers":

$C(u) \approx u$, $|C|$ small

$$\mathbb{E}_\varepsilon \sup_{u \in V} \langle \varepsilon, u \rangle = \mathbb{E}_\varepsilon \sup_{u\in V} \left(\langle \varepsilon, u - C(u) \rangle + \langle \varepsilon, C(u) \rangle \right)$$

$$+ \sqrt{n} \sup_{u \in V} \|u - C(u)\| + \underbrace{\mathrm{URad}(C)}_{\sqrt{\ln |C|}}.$$

② (★) has not been proved with Rademacher complexity.

Theorem (VC bound) σ ReLU

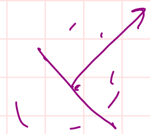
$$\mathrm{URad}\left(\left\{x \mapsto \mathrm{sgn}\left(\sigma_L(W_L \cdots \sigma_1(W_1 x + b_1) + \cdots b_L)\right) : \begin{matrix} p \text{ parameters,} \\ L \text{ layers} \end{matrix}\right\}_{|X}\right)$$

$$\leq \sqrt{2n(1+VC(\cdots))\ln(n+1)}$$

where $VC(\cdots) \leq 6\, pL \ln(pL)$.

Proof remarks:

* $\mathrm{URad}(\mathrm{sgn}(\mathcal{F})_{|X}) \leq \left(\underbrace{\sup_{u \in \mathrm{sgn}(\mathcal{F}_{|X})} \|u\|_2}_{\sqrt{n}}\right) \sqrt{2n \ln \underbrace{|\mathrm{sgn}(\mathcal{F})_{|X}|}_{\leq 2^n}}$, $\{\pm 1\}^n$



$Sh(\mathcal{F}; X)$

$$Sh(\mathcal{F}; n) = \sup_{|X| \leq n} |\mathrm{sgn}(\mathcal{F})_{|X}|$$ "shatter coefficient"

$$VC(\mathcal{F}) := \sup\{i \geq 0 : Sh(\mathcal{F}; i) = 2^i\}$$

$$Sh(\mathcal{F}; n) \leq 1 + n^{1+VC(\mathcal{F})}$$

* Activation matters ($\exists$ a convex-concave monotone & bounded activation ϕ s.t. $VC(\{x \mapsto \phi(W_L \cdots \phi(W_1 x) \cdots\}) = \infty$)

* Specific ReLU comments:

$$x \mapsto \sigma_L(W_L \cdots \sigma_1(W_1 x) \cdots) \quad S_L^w \odot (W_L \cdots (S_2^w \odot (W_2 (S_1^w \odot (W_1 x)))))$$

For a fixed $S_L^w \ldots S_1^w$, then this is linear in x

L-degree polynomial in w (Warren '62) } hard to prove.

$\Rightarrow$ need to count $(S_L^w \ldots S_1^w)$.

recursively refine partitions $P_1 \ldots P_L$

where $\forall S \in P_{i+1}\ \exists S' \in P_i$ s.t. $S \subseteq S'$