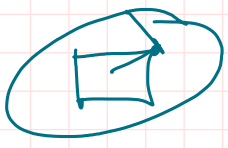
Lecture 26: more Rademacher complexity

Definition. $\mathrm{URad}(V) = \mathbb{E}_\varepsilon \sup_{u \in V} \langle \varepsilon, u \rangle$, $V \subseteq \mathbb{R}^n$

Typical usage: $G|_S = \{(g(z_1), \dots, g(z_n)) : g \in G\} \subseteq \mathbb{R}^n$

$\varepsilon_i \overset{iid}{\sim} \mathrm{Unif}(\{-1, +1\})$

alternatively $\gamma_i \sim \mathcal{N}(0,1)$ "Gaussian complexity"



Theorem (core Rademacher bound). Given G with $g(z) \in [a,b]$ a.s., w/ $p \geq 1-\delta$ $\forall g \in G$

$$\mathbb{E} g(z) \leq \frac{1}{n}\sum_i g(z_i) + \frac{2}{n}\mathrm{URad}(G|_z) + 3(b-a)\sqrt{\frac{\ln 2/\delta}{2n}}$$

Lemma. $\mathrm{URad}(\{w \mapsto w^T x : \|w - w_0\| \leq R\}|_X) \leq \|X\|_F \cdot R$.

Rem. Regularization $\Rightarrow$ good generalization. (sufficient but not necessary.)

Proof.

$$\mathbb{E}_\varepsilon \sup_{\|w-w_0\| \leq R} \sum_i \varepsilon_i w^T x_i$$

$$= \mathbb{E}_\varepsilon \sup_{\|w-w_0\| \leq R} \sum_i \varepsilon_i (w - w_0 + w_0)^T x_i$$

$$= \underbrace{\mathbb{E}_\varepsilon \sum_i w_0^T x_i}_{0} + \mathbb{E}_\varepsilon \sup_{\|w-w_0\| \leq R} \sum_i \varepsilon_i (w - w_0)^T x_i$$

$v := w - w_0$

$$= \mathbb{E}_\varepsilon \sup_{\|v\| \leq R} \langle v, \sum_i \varepsilon_i x_i \rangle = \mathbb{E}_\varepsilon R \cdot \|\sum_i \varepsilon_i x_i\|$$

$$\leq R\sqrt{\mathbb{E}_\varepsilon \|\sum_i \varepsilon_i x_i\|^2} = R\sqrt{\mathbb{E}_\varepsilon \left(\sum_i \varepsilon_i^2 \|x_i\|^2 + \sum_{i \neq j} \varepsilon_i \varepsilon_j x_i^T x_j\right)}$$

$$= R\|X\|_F.$$

Rem. essentially tight, $\mathrm{URad}(\{w \mapsto w^T x\}_X) \geq \frac{1}{\sqrt{2}}\|X\|_F \cdot R$ "Khintchine inequality".

Lemma. $\vec{\ell} : \mathbb{R}^n \to \mathbb{R}^n$, where ℓ_i is e-Lipschitz, then $\mathrm{URad}(\vec{\ell} \circ V) \leq e \cdot \mathrm{URad}(V)$.

Proof idea. Core idea: main Rademacher is proved via "symmetrization", modern proof of this lemma de-symmetrizes then re-symmetrizes. //

Lemma (Massart finite lemma). $\mathrm{URad}(V) \leq \left(\sup_{u \in V} \|u\|_2\right)\sqrt{2 \ln |V|}$.

Proof idea. Subgaussian tricks on $(\varepsilon_1 \cdots \varepsilon_n)$. //

Remark.

① Agrees with union bound:

$$\sup_{g \in G}\left(\mathbb{E} g(z) - \frac{1}{n}\sum_i g(z_i)\right) \leq 3(b-a)\sqrt{\frac{\ln 2/\delta}{2n}} + \frac{2}{n}\underbrace{\mathrm{URad}(G|_x)}_{\simeq \sqrt{\ln |G|_x|}}$$

② Many generalization bounds are proved via discretization + Massart lemma. //

Example (logistic regression).

logistic loss $\ell(z) = \ln(1 + \exp(-z))$

Suppose $\|w\| \leq R$, $\|x\| \leq 1$ a.s.

$\Rightarrow$ ① ℓ is 1-Lip ② $\ell(y w^T x) \in [0, \ln 2 R]$ a.s. ③ $\|X\|_F \leq \sqrt{n}$

w/ $p \geq 1-\delta$, $\forall \|w\| \leq R$,

$$\mathop{\mathbb{E}}_{x,y} \ell(w^T x y) \leq \frac{1}{n}\sum_i \ell(w^T x_i y_i) + \frac{2}{n}\mathrm{URad}\left(\left\{\vec{\ell} \circ \xi : x \mapsto w^T x : \|w\| \leq R\right\}|_S\right) + 3(b-a)\sqrt{\frac{\ln 2/\delta}{2n}}$$

($b - a = 1 + R$)

$\ell_i(f(x_i)) = \ell(y_i f(x_i))$

$$\Rightarrow \mathrm{URad}\left(\left(\vec{\ell} \circ \{w \mapsto \cdots\}\right)|_S\right) \leq 1 \cdot \mathrm{URad}(\{w \mapsto x^T \cdots\}) \leq \|X\| \cdot R = R\sqrt{n}.$$

$$\leq \frac{1}{n}\sum_i \ell(w^T x_i y_i) + \frac{2R}{\sqrt{n}} + \frac{3(1+R)}{\sqrt{2n}}\sqrt{\ln \frac{2}{\delta}}.$$

//

Four deep network Rademacher bounds.

① [Bartlett-Mendelson '02.] σ: coordinate-wise c-Lipschitz, $\sigma(0)=0$

$$\mathrm{URad}\left(\left\{x \mapsto \sigma_L(W_L \sigma_{L-1} \cdots \sigma_1(W_1 x)\cdots) : \|W^i\|_{1,\infty} \le B\right\}|_X\right) \le \|X\|_{2,\infty} (2cB)^L \sqrt{2\ln d}.$$

② [Golowich-Rakhlin-Shamir '18.] σ: coordinate-wise 1-Lipschitz, homogeneous

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

③ [Bartlett-Foster-Telgarsky '17.] σ: c-Lipschitz

$$\mathrm{URad}\left(\left\{x \mapsto \sigma_L(W_L \cdots \sigma_1(W_1 x)\cdots) : \begin{array}{l}\|(W_i - W_i(0))^T\|_{2,1} \le b_i \\ \|W_i\|_2 \le s_i\end{array}\right\}|_X\right) \le \tilde{O}\left(\|X\|_F \left(\prod_{i=1}^L s_i c\right)\left(\sum_{i=1}^L \left(\frac{b_i}{s_i}\right)^{2/3}\right)^{3/2}\right).$$

④ [Bartlett - '18.] ("VC dim bound")

$$\mathrm{URad}\left(\left\{x \mapsto \mathrm{sgn}\left(\sigma_L(W_L \cdots \sigma_1(W_1 x)\cdots)\right) : \begin{array}{l}p \text{ parameters} \\ L \text{ layers}\end{array}\right\}|_X\right) \le 30\sqrt{np \cdot L \cdot (\ln pL)(\ln n)}$$

(ReLU $\subseteq \{\pm 1, \mathrm{Id}\}^n$)

Remarks.

* All have strawman mechanisms (i.e., add garbage to weight matrices)
* All are "vacuous".
* Community efforts:
  * Identify "phenomena"
  * 2-layer case
  * Apply restrictions (e.g., Wei-Ma "Lipschitz Augmentation"; no ReLU interesting)
  * Data & algorithmic assumptions