18. Rademacher Complexity

Last class

- Interested in controlling the uniform deviation

$$\sup_{f \in \mathcal{F}} \left| R(f) - \hat{R}(f) \right|$$

- We give probabilistic upper bound on this stochastic quantity, by:
  - Controlling the mean, via Rademacher complexity
  - Controlling the residual, via McDiarmid's inequality

- Introduced the notion of Rademacher complexity:
  - Given a bounded set $V \subseteq \mathbb{R}^n$, the RC is defined as

$$\mathfrak{R}(V) := \frac{1}{n} \mathbb{E}_{\varepsilon} \sup_{v \in V} \langle \vec{\varepsilon}, v \rangle, \qquad \varepsilon_i \sim \text{Unif}(\{\pm 1\})$$

Last class

- Showed the symmetrization bound:

Theorem (Symmetrization).

We have

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(R(f) - \hat{R}(f) \right) \leq 2 \cdot \mathbb{E} \mathfrak{R}(\ell_{\mathcal{F}}(Z^n))$$

where the set $\ell_{\mathcal{F}}(Z^n)$ denotes the set of length-n sequences

$$\ell_{\mathcal{F}}(Z^n) = \left\{ \left(\ell_f(Z_1), \ldots, \ell_f(Z_n) \right), \;\middle|\; f \in \mathcal{F} \right\}$$

- Thus, we want to upper-bound the RC instead

Today

- Focus on two things:
  - Details about “residual control via McDiarmid”
  - Properties of the RC
    - Basic algebra
    - Finite class lemma
    - Contraction principle

- **Next week.** Analyze the RC of neural nets

Residual control

- We want to control the residual:

$$\sup_{f\in\mathcal{F}}\left|R(f)-\hat{R}(f)\right| - \mathbb{E}\sup_{f\in\mathcal{F}}\left|R(f)-\hat{R}(f)\right|$$

- **Recall.** McDiarmid's inequality states the following

Theorem (McDiarmid).

Let $f(\cdot)$ have the bounded difference property, i.e.,

$$|f(x_1,\cdots,x_{i-1},x_i,x_{i+1},\cdots,x_n)-f(x_1,\cdots,x_{i-1},\tilde{x}_i,x_{i+1},\cdots,x_n)|\le c_i, \qquad \forall\ldots$$

Then, for independent $X_1,\ldots,X_n$, we have

$$\mathbf{Pr}[f-\mathbb{E}f\ge\epsilon]\le\exp\left(\frac{2\epsilon^2}{\sum c_i^2}\right)$$

Residual control

- Thus, it suffices to check that whether the following quantity has a bounded difference

$$g(Z^n) = \sup_{f \in \mathcal{F}} \left| R(f) - \hat{R}(f) \right|$$

  - That is, we should check whether this quantity satisfies

$$\left| g(z_{1:n}) - g(z_{1:i-1}, \tilde{z}_i, z_{i+1:n}) \right| \leq c_i$$

- Any volunteer? 🙋 (assume that the loss is bounded between [0,1])

Basic properties of RC

- Now, let's move on to prove some basic properties of the Rademacher complexity

$$\mathfrak{R}(V) := \frac{1}{n}\mathbb{E}_{\varepsilon} \sup_{v \in V} \langle \vec{\varepsilon}, v \rangle, \qquad \varepsilon_i \sim \text{Unif}(\{\pm 1\})$$

Property#1.

$$\mathfrak{R}(\{v\}) = 0$$

Basic properties of RC

$$\mathfrak{R}(V) := \frac{1}{n}\mathbb{E}_{\varepsilon} \sup_{v \in V} \langle \vec{\varepsilon}, v \rangle, \qquad \varepsilon_i \sim \text{Unif}(\{\pm 1\})$$

Property#2.

$$\mathfrak{R}(U + \{v\}) = \mathfrak{R}(U)$$

Basic properties of RC

$$\mathfrak{R}(V) := \frac{1}{n}\mathbb{E}_{\varepsilon} \sup_{v \in V} \langle \vec{\varepsilon}, v \rangle, \qquad \varepsilon_i \sim \text{Unif}(\{\pm 1\})$$

Property#3.

$$\mathfrak{R}(U + V) = \mathfrak{R}(U) + \mathfrak{R}(V)$$

Basic properties of RC

$$\mathfrak{R}(V) := \frac{1}{n}\mathbb{E}_{\varepsilon} \sup_{v \in V} \langle \vec{\varepsilon}, v \rangle, \qquad \varepsilon_i \sim \text{Unif}(\{\pm 1\})$$

Property#4.

$$\mathfrak{R}(c \cdot U) = |c| \cdot \mathfrak{R}(U)$$

Basic properties of RC

$$\mathfrak{R}(V) := \frac{1}{n}\mathbb{E}_{\varepsilon} \sup_{v \in V} \langle \vec{\varepsilon}, v \rangle, \qquad \varepsilon_i \sim \text{Unif}(\{\pm 1\})$$

Property#5.

If $U \subseteq V$, then $\mathfrak{R}(U) \leq \mathfrak{R}(V)$

Finite class lemma

- Now, recall that we had a very simple bound, whenever the hypothesis space was finite
  - Had the dependency $\sqrt{\log k/n}$
  - Proved with the union bound

- Turns out that RC has a similar result
  - We do not lose anything by UB-ing with RC

Lemma (Finite Class Lemma).

Let $|V| = k$, and let $L := \max_{v \in V} \|v\|_2$. Then, we have

$$\mathfrak{R}(V) \leq \frac{L\sqrt{\log k}}{n}$$

Proof sketch

$$\mathfrak{R}(V) \leq \frac{L\sqrt{\log k}}{n}$$

- Select some $t > 0$. Then, proceed as:

$$\exp\left(t \cdot \mathbb{E} \sup_{v \in V} \langle \varepsilon, v \rangle\right) \leq \mathbb{E} \exp\left(\sup_{v \in V} \langle \varepsilon, t \cdot v \rangle\right) \leq \mathbb{E} \sum_{v \in V} \exp(\langle \varepsilon, t \cdot v \rangle)$$

$$= \sum_{v \in V} \prod_{i=1}^{n} \mathbb{E} \exp(t \cdot \varepsilon_i \cdot v_i)$$

$$\leq \sum_{v \in V} \prod_{i=1}^{n} \exp(t^2 v_i^2 / 2)$$

$$= \sum_{v \in V} \exp(t^2 \|v\|_2^2 / 2) \qquad = k \cdot \exp(t^2 L^2 / 2)$$

- On both sides, take log and divide by t. Then, optimize over t

Contraction principle

- Another very useful property of RC is the contraction principle
  - Allow us to “peel off” compositions of functions

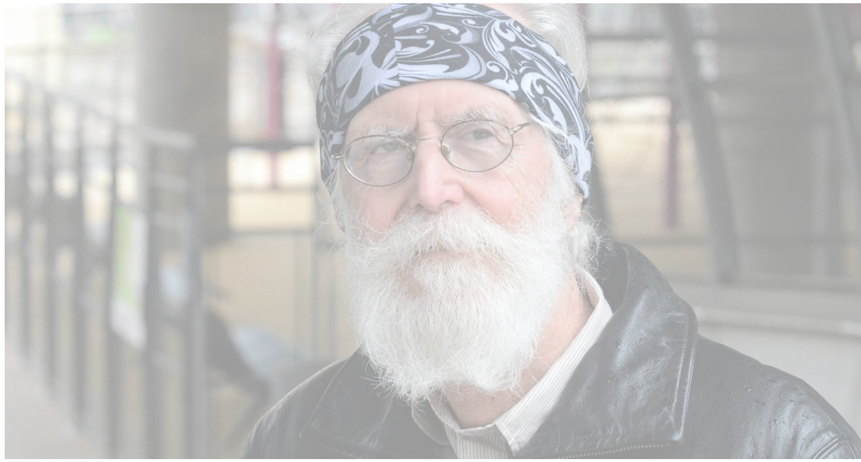
Lemma (Contraction principle).

Let V be a bounded subset of $\mathbb{R}^d$, and let $\phi_i(\cdot) : \mathbb{R} \to \mathbb{R}$ be an M-Lipschitz function. Then,

$$\mathfrak{R}(\phi \circ V) \leq M \cdot \mathfrak{R}(V)$$

- **Proof Sketch.** Let’s make some simplifications…
  - Assume WLOG that $M = 1$
  - Note that we can introduce one ϕ_i at a time.
    - We assume that we only have ϕ_1, and show

$$\mathfrak{R}((\phi_1, \mathrm{Id}, \cdots, \mathrm{Id}) \circ V) \leq \mathfrak{R}(V)$$



Proof sketch

Want-to-show: $\mathfrak{R}((\phi_1, \mathrm{Id}, \cdots, \mathrm{Id}) \circ V) \leq \mathfrak{R}(V)$

- First, investigate the RHS:

$$\begin{aligned}
\mathfrak{R}(V) &= \frac{1}{2} \cdot \frac{1}{n} \mathbb{E} \sup_{v \in V} \Big(v_1 + \langle \varepsilon_{2:n}, v_{2:n} \rangle \Big) + \frac{1}{2} \cdot \frac{1}{n} \mathbb{E} \sup_{v \in V} \Big(- v_1 + \langle \varepsilon_{2:n}, v_{2:n} \rangle \Big) \\
&= \frac{1}{2n} \mathbb{E} \left[\sup_{v \in V} \Big(v_1 + \langle \varepsilon_{2:n}, v_{2:n} \rangle \Big) + \sup_{v \in V} \Big(- v_1 + \langle \varepsilon_{2:n}, v_{2:n} \rangle \Big) \right] \\
&= \frac{1}{2n} \mathbb{E} \left[\sup_{v, \tilde{v} \in V} \Big(| v_1 - \tilde{v}_1 | + \langle \varepsilon_{2:n}, v_{2:n} + \tilde{v}_{2:n} \rangle \Big) \right]
\end{aligned}$$

- Likewise, the LHS can be written as:

$$\mathfrak{R}((\phi_1, \mathrm{Id}, \cdots, \mathrm{Id}) \circ V) = \frac{1}{2n} \cdot \mathbb{E} \left[\sup_{v, \tilde{v} \in V} \Big(| \phi(v_1) - \phi(\tilde{v}_1) | + \langle \varepsilon_{2:n}, v_{2:n} + \tilde{v}_{2:n} \rangle \Big) \right]$$

Next up

- RC of linear models