17. Uniform Convergence

Concentration of Measure

- **Last class.** For a single function f, we have

$$|R(f) - \hat{R}(f)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}, \qquad \text{w.p.} 1 - \delta$$

  - Concentration of measures
    - Markov
    - Chebyshev
    - Chernoff
    - Hoeffding
    - McDiarmid
    - Bernstein

Concentration of Measure

$$|R(f) - \hat{R}(f)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}, \qquad \text{w.p. } 1-\delta$$

- **Problem.** True for a fixed f, but not for f chosen post-hoc
    - To see this, let us first recap the ERM

- **ERM.** Given the data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we solve the optimization:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))}_{:= \hat{R}(f)}$$

- By doing so, we hope to achieve a near-optimal hypothesis such that

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) \approx 0$$

Concentration of Measure

- Suppose that

$$f^* = \arg\min_{f \in \mathcal{F}} R(f)$$

- Then, we have:

$$\begin{aligned} R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) = R(\hat{f}) - R(f^*) & \\ = \left[R(\hat{f}) - \hat{R}(\hat{f})\right] + \left[\hat{R}(\hat{f}) - \hat{R}(f^*)\right] + \left[\hat{R}(f^*) - R(f^*)\right] & \\ \leq \left[R(\hat{f}) - \hat{R}(\hat{f})\right] + \left[\hat{R}(f^*) - R(f^*)\right] & \end{aligned}$$

- **Problem.** The first term is random, and chosen post-hoc
    - A bound that works for a single f is not good enough

Example

- To see this, consider the following example:
  - Suppose that we observe all training data

$$(x_1, y_1), \ldots, (x_n, y_n)$$

  - Then, we construct the function

$$f(x) = \sum_{i=1}^{n} y_i \cdot \mathbf{1}[x = x_i]$$

  - **Problem.** This will never generalize
    - only return 0 on unseen data!

Uniform deviation

- A classic way to handle this stochasticity is via uniform deviation
  - That is, we upper-bound as:

$$R(\hat{f}) - \inf_{f \in \mathcal{F}} R(f) \leq \left[R(\hat{f}) - \hat{R}(\hat{f})\right] + \left[\hat{R}(f^*) - R(f^*)\right]$$

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

  - The goal will be to get a probabilistic upper bound on this quantity, i.e.,

$$\mathbf{Pr}\left[\sup_{f \in \mathcal{F}} \left|R(f) - \hat{R}(f)\right| > \epsilon\right] \leq \delta$$

Finite case

- This is easy to do, whenever our hypothesis space is finite

Proposition (Finite class).

Suppose that we have $\mathscr{F} = \{f_1, f_2, \ldots, f_k\}$. Then, with probability at least $1 - \delta$, the following holds:

$$\sup_{f \in \mathscr{F}} |R(f) - \hat{R}(f)| \leq \sqrt{\frac{\log(2k/\delta)}{2n}} \leq \sqrt{\frac{\log(k)}{2n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

- Compare this with the Hoeffding's theorem for a single function

$$|R(f) - \hat{R}(f)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}, \qquad \text{w.p. } 1 - \delta$$

  - We have an extra $\sqrt{\log k/n}$ term.

Proof Sketch

- Simply a consequence of the union bound + Hoeffding

- Proceed as:

$$\mathbf{Pr}\left[\sup_{f \in \mathscr{F}} |R(f) - \hat{R}(f)| > \epsilon\right] = \mathbf{Pr}\left[|R(f_1) - \hat{R}(f_1)| > \epsilon \quad \textbf{or} \quad \cdots \quad \textbf{or} \quad |R(f_k) - \hat{R}(f_k)| > \epsilon\right]$$

$$\leq \mathbf{Pr}\left[|R(f_1) - \hat{R}(f_1)| > \epsilon\right] + \cdots + \mathbf{Pr}\left[|R(f_k) - \hat{R}(f_k)| > \epsilon\right]$$

$$\leq k \cdot \left(2\exp(-n\epsilon^2)\right)$$

- Thus, we have the first claim.

- The second claim follows from the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$

Handling infinite classes

- Now we have the bound

$$\max_{i \in [k]} |\hat{R}(f_i) - R(f_i)| \leq \sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2n}}, \qquad \text{w.p. } 1-\delta$$

- **Problem.** For neural nets, we know that $|\mathcal{F}| = \infty$
    - We treat weights as continuous parameters

- **Vague idea.** Select some representative functions $f_1, \ldots, f_k$, so that

$$\sup_{f \in \mathcal{F}} \inf_{i} \|f(x) - f_i(x)\| \leq \epsilon?$$

Rademacher Complexity

- For infinite hypothesis space, we'll use a quantity that is called Rademacher complexity

- **Spoiler.** RC will provide an upper bound on the expected value of the uniform deviation
    - Here, the expectation is taken over the randomness of the training data

    - In particular, we will show that:

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

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

$$\leq \text{(Rademacher Complexity bounds)} + \text{(Concentration of Measure bounds)}$$

Rademacher Complexity

- To formalize everything, we'll first define the Rademacher random variable

Definition (Rademacher Random Variable).

The Rademacher random variable ε is a binary random variable, with

$$\mathbf{Pr}[\varepsilon = +1] = \mathbf{Pr}[\varepsilon = -1] = \frac{1}{2}$$

Definition (Rademacher Random Vector).

The Rademacher random vector $\vec{\varepsilon} \in \mathbb{R}^n$ is a random vector, with entries consisting of n independent Rademacher random variables.

Rademacher Complexity

Definition (Rademacher Average).

Given a bounded set $V \subseteq \mathbb{R}^n$, define the Rademacher average of V as

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

- Also known as "Rademacher complexity"
- We will also define a notation for the unnormalized quantity

$$\tilde{\mathfrak{R}}(V) := \mathbb{E} \sup_{v \in V} \langle \vec{\varepsilon}, v \rangle$$

- **Note.** Supremum is inside the expectation — Given some random $\vec{\varepsilon}$, we find the best-fitting v
    - If V is rich, we expect a large $\mathfrak{R}(V)$
    - If V is not diverse, we expect a small $\mathfrak{R}(V)$

Rademacher Complexity

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

- **Example.** Consider the case $n = 2$, and let

$$V_1 = \{(+1, +1), (+1, -1), (-1, +1), (-1, -1)\}$$
$$V_2 = \{v \mid v = (t, t), \quad t \in [-1, +1]\}$$

  - Then, we have

$$\mathfrak{R}(V_1) =$$

$$\mathfrak{R}(V_2) =$$

  - On the other hand, $|V_1| = 4$ and $|V_2| = \infty$

Motivation for RC

- Before formally proving the theorem, let me a hand-wavy explanation on:

  "why random binary can be useful for measuring generalization"

- Suppose that we have $2n$ data at hand.

$$Z_1, \ldots, Z_n, \qquad Z_{n+1}, \ldots, Z_{2n}$$

  - Here, $Z = (X, Y)$
  - **First half.** Used for training

$$\frac{1}{n}\sum_{i=1}^{n} \ell_f(Z_i) = \hat{R}(f)$$

  - **Second half.** Used for approximating the test error

$$\frac{1}{n}\sum_{i=1}^{n} \ell_f(Z_i) = \hat{R}(f)$$

Motivation for RC

- If we consider a sequence

$$\vec{\varepsilon} = (\underbrace{+1, \ldots, +1}_{n \text{ entries}}, \underbrace{-1, \ldots, -1}_{n \text{ entries}})$$

- Then, the generalization gap can be written as:

$$\hat{R}(f) - R(f) \approx \frac{1}{n} \sum_{i=1}^{2n} \varepsilon_i \cdot \ell_f(Z_i) = \frac{1}{n} \langle \varepsilon_{1:n}, \ell_f(Z_{1:n}) \rangle$$

  - Rademacher r.v.s determine whether a sample is on the training side or the test side

Symmetrization

- This intuition is formalized in the following theorem.

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

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

Proof Sketch

- First, we consider "ghost samples" drawn independently from Z^n

$$(Z'_1, \ldots, Z'_n)$$

- Then, we have:

$$\mathbb{E}_{Z^n} \sup_{f \in \mathcal{F}} \left(R(f) - \hat{R}(f)\right) \leq \mathbb{E}_{Z^n} \mathbb{E}_{Z'^n} \sup_{f \in \mathcal{F}} \left(\hat{R}'(f) - \hat{R}(f)\right)$$

  - Here, $\hat{R}'$ denotes the empirical risk w.r.t. $Z'_1, \ldots, Z'_n$ — i.e., the ghost samples

- Now, it suffices to show that

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

Proof Sketch

Want-to-show: $\mathbb{E}_{Z^n}\mathbb{E}_{Z'^n} \sup_{f \in \mathcal{F}} \left(\hat{R}'(f) - \hat{R}(f)\right) \leq 2 \cdot \mathbb{E}\mathfrak{R}(\ell_f(Z^n))$

- Take a closer look at the LHS:

$$\mathbb{E}_{Z^n}\mathbb{E}_{Z'^n} \sup_{f \in \mathcal{F}} \left(\hat{R}'(f) - \hat{R}(f)\right) = \frac{1}{n}\mathbb{E}_{Z^n}\mathbb{E}_{Z'^n} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} \ell_f(Z_i') - \ell_f(Z_i) \right)$$

- We know that $\ell_f(Z_i') - \ell_f(Z_i)$ has a symmetric distribution.
    - Thus, we have

$$\ell_f(Z_i') - \ell_f(Z_i) \quad \overset{d}{=} \quad \varepsilon(\ell_f(Z_i') - \ell_f(Z_i))$$

    - In other words, we have

$$\mathbb{E}_{Z^n}\mathbb{E}_{Z'^n} \sup_{f \in \mathcal{F}} \left(\hat{R}'(f) - \hat{R}(f)\right) = \frac{1}{n}\mathbb{E}_{\varepsilon^n}\mathbb{E}_{Z^n}\mathbb{E}_{Z'^n} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} \varepsilon_i\left(\ell_f(Z_i') - \ell_f(Z_i)\right) \right)$$

Proof Sketch

Want-to-show: $$\frac{1}{n}\mathbb{E}_{\varepsilon^n}\mathbb{E}_{Z^n}\mathbb{E}_{Z'^n}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\varepsilon_i\big(\ell_f(Z_i') - \ell_f(Z_i)\big)\right) \le 2\cdot\mathbb{E}\mathfrak{R}(\ell_f(Z^n))$$

- Now, note that $\sup(X+Y) \le \sup(X) + \sup(Y)$
  - Thus, we have:

$$\frac{1}{n}\mathbb{E}_{\varepsilon^n}\mathbb{E}_{Z^n}\mathbb{E}_{Z'^n}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\varepsilon_i\big(\ell_f(Z_i') - \ell_f(Z_i)\big)\right)$$

$$\le \frac{1}{n}\mathbb{E}_{\varepsilon^n}\mathbb{E}_{Z'^n}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\varepsilon_i\cdot\ell_f(Z_i')\right) + \frac{1}{n}\mathbb{E}_{\varepsilon^n}\mathbb{E}_{Z^n}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}-\varepsilon_i\cdot\ell_f(Z_i)\right)$$

- By the symmetricity of ε, we have:

$$= \frac{2}{n}\mathbb{E}_{\varepsilon^n}\mathbb{E}_{Z^n}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{n}\varepsilon_i\cdot\ell_f(Z_i)\right)$$

Next up

- Residual control via McDiarmid
- Analysis on RC