19. RC of Simple Nets

Recap

- Our goal is to prove generalization bounds

- With probability at least $1-\delta$, we have (roughly)

$$\sup_{f\in\mathcal{F}}\left(R(f)-\hat{R}(f)\right)\le 2\cdot\mathbb{E}\mathfrak{R}(\ell_{\mathcal{F}}(Z^n))+\sqrt{\frac{\log(2/\delta)}{2n}}$$

  - Here, the Rademacher complexity is:

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

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

Today

- Give elementary generalization bounds for neural networks
  - That is, want to upper bound:

$$\mathbb{E}\mathfrak{R}(\ell_{\mathscr{F}}(Z^n))$$

    - for the cases:

$$\ell_f(z; w) = \ell(y, f(x; W_{1:d}))$$

$$f(x; W_{1:L}) = W_L \circ \sigma \circ W_{L-1} \circ \cdots \circ \sigma \circ W_1 x$$

- As the first step, we'll look at the linear model

$$f(x; w) = w^\top x$$

Logistic regression

- Consider a logistic regression with bounded weights & data
    - Bounded data

$$\|X\|_2 \leq M, \quad Y \in \{+1, -1\}$$

    - Logistic loss

$$\ell(y, f(x)) = \log\big(1 + \exp(-y \cdot f(x))\big)$$

    - Bounded function space

$$\mathcal{F} = \left\{ x \mapsto w^\top x \;\middle|\; w \in \mathbb{R}^d, \quad \|w\|_2 \leq B \right\}$$

Logistic regression

- First, we claim that we can "peel off" the loss function

Lemma.

$$\mathfrak{R}(\ell_{\mathcal{F}}(Z^n)) \leq \mathfrak{R}(\mathcal{F}(X^n))$$

- **Proof idea.** Recall the "contraction principle"
    - Let V be a bounded subset of $\mathbb{R}^n$, and let $\phi_i(\,\cdot\,) : \mathbb{R} \to \mathbb{R}$ be an M-Lipschitz function. Then,

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

    - Show that for $y \in \{+1, -1\}$, the following function is 1-Lipschitz

    $$\phi(a) = \log(1 + \exp(-y \cdot a))$$

Logistic regression

- Now, our target of analysis is:

$$\mathfrak{R}(\mathscr{F}(x^n)) = \frac{1}{n}\mathbb{E} \sup_{\|w\|_2 \leq B} \left(\sum_{i=1}^{n} \varepsilon_i \cdot w^\top x_i \right)$$

- This is usually a headache:
  - We expect something that behaves $\sim 1/\sqrt{n}$
  - That is, we expect

$$\mathbb{E} \sup_{\|w\|_2 \leq B} \left(\sum_{i=1}^{n} \varepsilon_i \cdot w^\top x_i \right) \sim \sqrt{n}$$

  - Naïve approaches, e.g., Cauchy-Schwarz, is doomed.

Logistic regression

- In fact, we have the following bound.

Proposition.

$$\mathbb{E} \sup_{\|w\|_2 \leq B} \left(\sum_{i=1}^{n} \varepsilon_i \cdot w^\top x_i \right) \leq B \cdot \sqrt{\sum_{i=1}^{n} \|x_i\|^2}$$

- Not a bound that involves the number of parameters!
- Tight
    - consult Khinchine's inequality

Proof sketch

$$\mathbb{E} \sup_{\|w\|_2 \le B} \left(\sum_{i=1}^{n} \varepsilon_i \cdot w^\top x_i \right) \le B \cdot \sqrt{\sum_{i=1}^{n} \|x_i\|^2}$$

- First, remove supremum:

$$\mathbb{E} \sup_{\|w\|_2 \le B} \left(\sum_{i=1}^{n} \varepsilon_i \cdot w^\top x_i \right) = B \cdot \mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i \cdot x_i \right\|$$

- Then, apply the Jensen's inequality

$$\mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i \cdot x_i \right\| \le \sqrt{\mathbb{E} \left\| \sum_{i=1}^{n} \varepsilon_i \cdot x_i \right\|^2}$$

- Analyze the cross terms, and confirm they are zero.

Logistic regression

- As a corollary, we have:

Corollary.

$$\mathbb{E}\mathfrak{R}(\ell_{\mathscr{F}}(Z^n)) \leq \frac{B \cdot \sqrt{\mathrm{Var}(X)}}{\sqrt{n}} \leq \frac{BM}{\sqrt{n}}$$

- Thus, we have a generalization bound of order $1/\sqrt{n}$
- If we train a lot, then B can be large:
    - Longer training —> Can overfit

Logistic regression — a variant

- Suppose that we have a 1-norm constraint on the weights.

Proposition.

$$\mathbb{E} \sup_{\|w\|_1 \leq B} \left(\sum_{i=1}^{n} \varepsilon_i \cdot w^\top x_i \right) \leq B \cdot \max_i \|x_i\|_\infty \cdot \sqrt{\frac{\log 2d}{n}}$$

- **Proof idea.** Try yourself ;)

Two-layer net

- Consider a slightly different version: Regression with two-layer net
  - Bounded data

$$\|x\|_2 \le 1, \quad |y| \le 1$$

  - Squared loss

$$\ell(y, f(x)) = (y - f(x))^2$$

  - Bounded function space

$$\mathcal{F} = \left\{ x \mapsto w^\top \sigma(Ux) \;\middle|\; w \in \mathbb{R}^m, U \in \mathbb{R}^{m\times d} \quad \|w\|_2 \le B_w, \|u_j\|_2 \le B_u \forall j \in [m] \right\}$$

Two-layer net

- Similarly, begin by peeling off the loss function

Lemma.

$$\mathfrak{R}(\ell_{\mathscr{F}}(Z^n)) \leq 4 \cdot \mathfrak{R}(\mathscr{F}(X^n))$$

- **Proof idea.** Again, inspect the Lipschitz constant of $a \mapsto \|y - a\|^2$

Two-layer net

- Now, we can show the following bound

Proposition.

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} \varepsilon_i \cdot f(x_i) \right) \leq 2 B_w B_u \sqrt{mn}$$

- Unfortunately, we have $\sqrt{m}$
    - Dependent on the number of hidden layer neurons

Proof sketch

- Begin by peeling off the second layer

$$\begin{aligned}
\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} \varepsilon_i \cdot f(x_i) \right) &= \mathbb{E} \sup_{\|w\|_2 \leq B_w} \sup_{g \in \mathcal{G}} \left(\sum_{i=1}^{n} \varepsilon_i \cdot w^\top g(x_i) \right) \\
&= \mathbb{E} \sup_{\|w\|_2 \leq B_w} \sup_{g \in \mathcal{G}} w^\top \left(\sum_{i=1}^{n} \varepsilon_i \cdot g(x_i) \right) \\
&= B_w \cdot \mathbb{E} \sup_{g \in \mathcal{G}} \left\| \sum_{i=1}^{n} \varepsilon_i \cdot g(x_i) \right\|_2 \\
&\leq B_w \sqrt{m} \cdot \mathbb{E} \sup_{g \in \mathcal{G}} \left\| \sum_{i=1}^{n} \varepsilon_i \cdot g(x_i) \right\|_\infty \\
&= B_w \sqrt{m} \cdot \mathbb{E} \sup_{U : \|u_j\|_2 \leq B_u} \max_{j \in [m]} \left| \sum_{i=1}^{n} \varepsilon_i \cdot \sigma(u_j^\top x_i) \right|
\end{aligned}$$

Proof sketch

$$
\begin{aligned}
B_w\sqrt{m}\cdot\mathbb{E}\sup_{U:\|u_j\|_2\le B_u}\max_{j\in[m]}\left|\sum_{i=1}^{n}\varepsilon_i\cdot\sigma(u_j^\top x_i)\right| &= B_w\sqrt{m}\cdot\mathbb{E}\sup_{\|u\|_2\le B_u}\left|\sum_{i=1}^{n}\varepsilon_i\cdot\sigma(u^\top x_i)\right| \\
&\le 2\cdot B_w\sqrt{m}\cdot\mathbb{E}\sup_{\|u\|_2\le B_u}\sum_{i=1}^{n}\varepsilon_i\cdot\sigma(u^\top x_i) \\
&\le 2\cdot B_w\sqrt{m}\cdot\mathbb{E}\sup_{\|u\|_2\le B_u}\sum_{i=1}^{n}\varepsilon_i\cdot u^\top x_i \\
&\le 2\cdot B_wB_u\sqrt{mn}
\end{aligned}
$$

Remarks

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\sum_{i=1}^{n} \varepsilon_i \cdot f(x_i) \right) \leq 2 B_w B_u \sqrt{mn}$$

- The factor $\sqrt{m}$ came from $\|\cdot\|_2 \to \|\cdot\|_\infty$, and then coming back to $\|\cdot\|_2$
- Thus, for depth-L nets, we will have the dependency: $(2\sqrt{\text{width}})^{\text{depth}}$
    - But is this true?

Depth-independent bound

Theorem 14.2.

Consider a ReLU net of form

$$x \mapsto \sigma(W_L\sigma(W_{L-1}\cdots\sigma(W_1 x)\cdots), \qquad \|W_i\|_F \le B$$

Then, we have

$$\mathfrak{R}(\mathscr{F}(Z^n)) \le B^L \|X\|_F \left(1 + \sqrt{2L\log(2)}\right)$$

- Sadly, won't prove today
- **Proof idea.** Use the log-exponential trick

$$\mathbb{E}\sup = \mathbb{E}\log\exp\sup \le \log\mathbb{E}\exp\sup$$

  - Handle everything inside the log

Next up

- Covering number bounds