Lecture 20. (Sketch.)

- Project proposal meetings this Wednesday, November 14! Must attend to receive full points!

1. Rademacher recap.

Concentration controlled one function at a time. To control many functions, out main tool is (unnormalized) Rademacher complexity:

$$\mathrm{URad}(V) := \mathbb{E} \sup_{u \in V} \langle \epsilon, u \rangle, \qquad \mathrm{Rad}(V) := \frac{1}{n}\mathrm{URad}(V).$$

Given data $S := (Z_1, \ldots, Z_n)$ and functions $\mathcal{F}$, define vectors

$$\mathcal{F}_{|S} := \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}.$$

Our main generalization tool involves $\mathrm{URad}(\mathcal{F}_{|S})$, and is a consequence of our two symmetrization lemmas and McDiarmid's inequality.

Theorem. Let $\mathcal{F}$ be given with $f(z) \in [a, b]$ a.s.. With probability $\geq 1 - \delta$,

$$\sup_{f \in \mathcal{F}} \mathbb{E}f - \hat{\mathbb{E}}_n f \leq \frac{2}{n}\mathrm{URad}(\mathcal{F}_{|S}) + 3(b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Remarks.

- Recall that other treatments use Rad, we'll use $\mathrm{URad} = \mathrm{Rad}/n$.
- Some classical texts provide a variety of Generalization bounds which all require custom symmetrization arguments. Instead, we'll prove everything using Rademacher complexity (and the preceding Theorem). This is a standard approach, but is not perfect: some things seem to be hard or even (as far as we know) impossible to prove with Rademacher complexity. (An example of a thing that is hard are cases where the preceding theorem should scale with $1/n$ rather than $1/\sqrt{n}$.)

Remarks (continued).

- A quick note on interpretation: if V is more expressive/complicated, then it can fit more of the random signs, and $\mathrm{URad}(V)$ is larger.
- Some sanity checks.

$$\mathrm{URad}(\{u\}) = \mathbb{E}_\epsilon \langle \epsilon, u \rangle = 0,$$

$$\mathrm{URad}(\{(-1, \ldots, -1), (+1, \ldots, +1)\}) = \mathbb{E}_\epsilon \left| \sum_i \epsilon_i \right| = \Theta(\sqrt{n}),$$

$$\mathrm{URad}(\{-1, +1\}^n) = n.$$

2. Linear predictors.

Theorem. Collect sample $S := (x_1, \ldots, x_n)$ into rows of $X \in \mathbb{R}^{n \times d}$.

$$\mathrm{URad}(\{x \mapsto \langle w, x\rangle : \|w\|_2 \le B\}_{|S}) \le B\|X\|_F.$$

Proof. Fix any $\vec\epsilon \in \{-1, +1\}^n$. Then

$$\sup_{\|w\| \le B} \sum_i \epsilon_i \langle w, x_i\rangle = \sup_{\|w\| \le B} \left\langle w, \sum_i \epsilon_i x_i \right\rangle = B\left\|\sum_i \epsilon_i x_i\right\|.$$

We'll bound this norm with Jensen's inequality (only inequality in whole proof!):

$$\mathbb{E}\left\|\sum_i \epsilon_i x_i\right\| = \mathbb{E}\sqrt{\left\|\sum_i \epsilon_i x_i\right\|^2} \le \sqrt{\mathbb{E}\left\|\sum_i \epsilon_i x_i\right\|^2}.$$

To finish,

$$\mathbb{E}\left\|\sum_i \epsilon_i x_i\right\|^2 = \mathbb{E}\left(\sum_i \|\epsilon_i x_i\|^2 + \sum_{i,j}\langle \epsilon_i x_i, \epsilon_j x_j\rangle\right) = \mathbb{E}\sum_i \|x_i\|^2 = \|X\|_F^2.$$

Remark. We used exactly one inequality: everywhere else we had an equality! Indeed, the bound is tight: we can get a lower bound with Khintchine-Kahane ($\mathbb{E}_\epsilon\|\sum_i \epsilon_i x_i\|_2 \ge C\|X\|_F$).

3. Lipschitz composition.

Lemma. Let $\ell : \mathbb{R}^n \to \mathbb{R}^n$ be a vector of univariate L-lipschitz functions. Then $\mathrm{URad}(\ell \circ V) \le L\,\mathrm{URad}(V)$.

Proof. The idea of the proof is to "de-symmetrize" and get a difference of coordinates to which we can apply the definition of L.

(See next page.)

Proof (continued).

$$\begin{aligned}
\mathrm{URad}(\ell \circ V) &= \mathbb{E}\sup_{u \in V}\sum_i \epsilon_i \ell_i(u_i) \\
&= \frac{1}{2}\mathbb{E}_{\epsilon_{2:n}} \sup_{u,w \in V}\left(\ell_1(u_1) - \ell_1(w_1) + \sum_{i=2}^n \epsilon_i(\ell_i(u_i) + \ell_i(w_i))\right) \\
&\le \frac{1}{2}\mathbb{E}_{\epsilon_{2:n}} \sup_{u,w \in V}\left(L|u_1 - w_1| + \sum_{i=2}^n \epsilon_i(\ell_i(u_i) + \ell_i(w_i))\right) \\
&= \frac{1}{2}\mathbb{E}_{\epsilon_{2:n}} \sup_{\substack{u,w \in V \\ u_1 \ge w_1}}\left(L(u_1 - w_1) + \sum_{i=2}^n \epsilon_i(\ell_i(u_i) + \ell_i(w_i))\right) \\
&= \mathbb{E}_\epsilon \sup_{u \in V}\left(Lu_1 + \sum_{i=2}^n \epsilon_i \ell_i(u_i)\right).
\end{aligned}$$

Other coordinates follow by the same procedure.

We'll overload composition notation:

$$(\ell \circ f) = ((x,y) \mapsto \ell(-yf(x))),$$
$$\ell \circ \mathcal{F} = \{\ell \circ f : f \in \mathcal{F}\}.$$

Corollary. Suppose ℓ is L-lipschitz and $\ell \circ \mathcal{F} \in [a,b]$ a.s.. With probability $\geq 1-\delta$, every $f \in \mathcal{F}$ satisfies

$$\mathcal{R}_\ell(f) \leq \widehat{\mathcal{R}}_\ell(f) + \frac{2L}{n}\mathrm{URad}(\mathcal{F}_{|S}) + 3(b-a)\sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Proof. Use the lipschitz composition lemma with

$$\begin{aligned}|\ell(-y_i f(x_i) - \ell(-y_i f'(x_i))| &\leq L|-y_i f(x_i) + y_i f'(x_i))| \\ &\leq L|f(x_i) - f'(x_i))|.\end{aligned}$$

Example (logistic regression).

Suppose $\|w\| \leq B$ and $\|x_i\| \leq 1$, and the loss is the 1-Lipschitz logistic loss $\ell_{\log}(z) := \ln(1+\exp(z))$. Note $\ell(\langle w, yx\rangle) \geq 0$ and $\ell(\langle w, yx\rangle) \leq \ln(2) + \langle w, yx\rangle \leq \ln(2) + B$.

Combining the main Rademacher bound with the Lipschitz composition lemma and the Rademacher bound on linear predictors, with probability at least $1-\delta$, every $w \in \mathbb{R}^d$ with $\|w\| \leq B$ satisfies

$$\begin{aligned}\mathcal{R}_\ell(w) &\leq \widehat{\mathcal{R}}_\ell(w) + \frac{2}{n}\mathrm{URad}((\ell \circ \mathcal{F})_{|S}) + (\ln(2)+B)\sqrt{\ln(2/\delta)/(2n)} \\ &\leq \widehat{\mathcal{R}}_\ell(w) + \frac{2B\|X\|_F}{n} + (\ln(2)+B)\sqrt{\ln(2/\delta)/(2n)} \\ &\leq \widehat{\mathcal{R}}_\ell(w) + \frac{2B + (B+\ln(2))\sqrt{\ln(2/\delta)/2}}{\sqrt{n}}.\end{aligned}$$

Remarks.

- (Average case vs worst case.) Here we replaced $\|X\|_F$ with the looser $\sqrt{n}$.
- This bound scales as the SGD logistic regression bound proved via Azuma, despite following a somewhat different route (Azuma and McDiarmid are both proved with Chernoff bounding method; the former approach involves no symmetrization, whereas the latter holds for more than the output of an algorithm).
- It would be nice to have an "average Lipschitz" bound rather than "worst-case Lipschitz"; e.g., when working with neural networks and the ReLU, which seems it can kill off many inputs! But it's not clear how to do this. Relatedly: regularizing the gradient is sometimes used in practice?

Remarks (continued).

- The Lipschitz composition rule is more complicated with the absolute value form of Rademacher complexity. The easiest proof I know invokes the one here as a lemma:

$$\mathbb{E}_\epsilon \sup_{u \in V} |\langle \epsilon, \ell \circ v\rangle| = \mathrm{URad}\left((\ell \circ V) \cup (-\ell \circ V)\right).$$