Lecture 25.

- Project presentations next week.
  - 2 slides! 5 minutes!
  - You'll need to upload your slides some time December 12; I'll post details on piazza.
  - Presentation is **for class**, not **for me**!
  - Written part due December 20.

1. Covering and Rademacher bounds for deep networks.

Theorem. Let $\mathcal{F}$ denote a network architecture with L layers, ρ-Lipschitz activations with $\sigma(0) = 0$, and $\|a\|_1 \leq W$ for all node weights a. Then

$$\mathrm{URad}(\mathcal{F}_{|S}) \leq \|X\|_{2,\infty}(2\rho W)^L\sqrt{2\ln(d)}.$$

Proof.

Let $\mathcal{F}_i$ denote functions computed by nodes in layer i. It'll be shown by induction that

$$\mathrm{URad}((\mathcal{F}_i)_{|S}) \leq \|X\|_{2,\infty}(2\rho W)^i\sqrt{2\ln(d)}.$$

Base case ($i = 0$):

$$\begin{aligned}\mathrm{URad}((\mathcal{F}_i)_{|S}) &= \mathrm{URad}\left(\{x \mapsto x_i : i \in \{1,\ldots,d\}\}_{|S}\right)\\ &\leq \left(\max_i \|(x_1)_i,\ldots,(x_n)_i\|_2\right)\sqrt{2\ln(d)}\\ &= \|X\|_{2,\infty}\sqrt{2\ln d} = \|X\|_{2,\infty}(2\rho W)^0\sqrt{2\ln d}.\end{aligned}$$

Proof (inductive step; some steps use a big Rademacher lemma from Lecture 21).

Since $0 = \sigma(\langle 0, F(x)\rangle) \in \mathcal{F}_{i+1}$,

$$\begin{aligned}\mathrm{URad}((\mathcal{F}_{i+1})_{|S}) &= \left(\{x \mapsto \sigma(Wg(x)) : g \in \mathrm{conv}(-\mathcal{F}_i \cup \mathcal{F}_i)\}_{|S}\right)\\ &\leq \rho W\mathrm{URad}\left(-(\mathcal{F}_i)_{|S} \cup (\mathcal{F}_i)_{|S}\right)\\ &\leq 2\rho W\mathrm{URad}\left((\mathcal{F}_i)_{|S}\right)\\ &\leq (2\rho W)^{i+1}\|X\|_{2,\infty}\sqrt{2\ln d}.\end{aligned}$$

Remarks.

- This bound depends on a Lipschitz constant wrt $\|\cdot\|_\infty$; getting a bound with $\|\cdot\|_2$ incurs other factors, but can also get rid of the 2^L; see the work of Neyshabur-Tomioka-Srebro, Bartlett-Foster-Telgarsky, Golowich-Rakhlin-Shamir, Barron-Klusowski.
- The best lower bound is roughly what you get by writing a linear function as a deep network $\ddot\frown$.

Remark (another approach).

- As in the ReLU VC proof, let $X_0 = X^\top$ be a data matrix with examples as columns, and $X_i = \sigma_i(W_iX_{i-1})$ denote the output of layer i. To build a cover $\hat{X}_{i+1}$ of X_{i+1},

$$\begin{aligned}\|X_{i+1} - \widehat{X}_{i+1}\|_F &\leq \rho\|W_iX_i - \widehat{W}_i\widehat{X}_{i+1}\|_F\\ &\leq \rho\|W_iX_i - W_i\widehat{X}_{i+1}\|_F + \rho\|W_i\widehat{X}_i - \widehat{W}_i\widehat{X}_{i+1}\|_F\\ &\leq \rho\|W_i\|_2\|X_i - \widehat{X}_{i+1}\|_F + \rho\|W_i\widehat{X}_i - \widehat{W}_i\widehat{X}_{i+1}\|_F,\end{aligned}$$

  where the first term may be handled by induction, and the second by a per-layer covering number. This argument appears in Anthony-Bartlett's book, and is also used in the Bartlett-Foster-Telgarsky "spectrally-normalized" bound.
- This proof does not "coordinate" the layers in any way, and has two exponential dependences on layers L: a product of Lipschitz constants $\prod_{i\leq L}\rho\|W_i\|_2$, and because it must product the per-layer covers (second term) together.

2. Complexity of Lipschitz functions.

Note that this bound, while scaling with the Lipschitz constant of the networks, is much better than the Rademacher complexity of arbitrary Lipschitz functions.

Theorem. Let data $S = (x_1, \ldots, x_n)$ be given with $R := \max_{i,j} \|x_i - x_j\|_\infty$. Let $\mathcal{F}$ denote all ρ-Lipschitz functions from $[-R, +R]^d \to [-B, +B]$ (where Lipschitz is measured wrt $\|\cdot\|_\infty$). Then the **improper** covering number $\widetilde{\mathcal{N}}$ satisfies

$$\ln \widetilde{\mathcal{N}}(\mathcal{F}, \epsilon, \|\cdot\|_u) \leq \max\left\{0, \left\lceil \frac{4\rho(R+\epsilon)}{\epsilon} \right\rceil^d \ln \left\lceil \frac{2B}{\epsilon} \right\rceil \right\}.$$

Remark.

- Exponential in dimension!

Proof.

- Suppose $B > \epsilon$, otherwise can use the trivial cover $\{x \mapsto 0\}$.
- Subdivide $[-R-\epsilon, +R+\epsilon]^d$ into $\left(\frac{4(R+\epsilon)\rho}{\epsilon}\right)^d$ cubes of side length $\epsilon/2\rho$; call this U.
- Subdivide $[-B, +B]$ into intervals of length ϵ, thus $2B/\epsilon$ elements; call this V.
- Our candidate cover $\mathcal{G}$ is the set of all piecewise constant maps from $[-R-\epsilon, +R+\epsilon]^d$ to $[-B, +B]$ discretized according to U and V, meaning

$$|\mathcal{G}| \leq \left\lceil \frac{2B}{\epsilon} \right\rceil^{\left\lceil \frac{4(R+\epsilon)\rho}{\epsilon} \right\rceil^d}.$$

Proof (continued).

To show this is an improper cover, given $f \in \mathcal{F}$, choose $g \in \mathcal{G}$ by proceeding over each $C \in U$, and assigning $g_{|C} \in V$ to be the closest element to $f(x_C)$, where x_C is the midpoint of C. Then

$$\begin{aligned}
\|f - g\|_u &= \sup_{C \in U} \sup_{x \in C} |f(x) - g(x)| \\
&\leq \sup_{C \in U} \sup_{x \in C} \left(|f(x) - f(x_C)| + |f(x_C) - g(x)|\right) \\
&\leq \sup_{C \in U} \sup_{x \in C} \left(\rho \|x - x_C\|_\infty + \frac{\epsilon}{2}\right) \\
&\leq \sup_{C \in U} \sup_{x \in C} \left(\rho(\epsilon/(4\rho)) + \frac{\epsilon}{2}\right) \leq \epsilon
\end{aligned}$$