Lecture 22. (Sketch.)

- Homework 2 due Wednesday.
- Today and on Wednesday, we'll discuss VC bound for neural networks. These bounds have a bad reputation as "loose", "impractical", vacuous. so why are we studying them?
  - They reveal and are sensitive to some interesting structure in networks (the total possible number of activation patterns).
  - *Before* we "worst-case-ify" the bounds and have $\ln \text{Sh}(\mathcal{F}_{|S})$, it *seems* they could somehow be made average-case-y and tighter, though I don't know how yet...

1. VC Theory recap.

A few definitions:

$$\begin{aligned}
\text{sgn}(U) &:= \{(\text{sgn}(u_1), \ldots, \text{sgn}(u_n)) : u \in V\}, \\
\text{Sh}(\mathcal{F}_{|S}) &:= \left|\text{sgn}(\mathcal{F}_{|S})\right|, \\
\text{Sh}(\mathcal{F}; n) &:= \sup_{|S| \le n} \left|\text{sgn}(\mathcal{F}_{|S})\right|, \\
\text{VC}(\mathcal{F}) &:= \sup\{i \in \mathbb{Z}_{\ge 0} : \text{Sh}(\mathcal{F}; i) = 2^i\}.
\end{aligned}$$

Theorem ("VC Theorem"). With probability at least $1 - \delta$, every $f \in \mathcal{F}$ satisfies

$$\mathcal{R}_z(\text{sgn}(f)) \le \widehat{\mathcal{R}}_z(\text{sgn}(f)) + \frac{2}{n}\text{URad}(\text{sgn}(\mathcal{F}_{|S})) + 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

and

$$\begin{aligned}
\text{URad}(\text{sgn}(\mathcal{F}_{|S})) &\le \sqrt{2n \ln \text{Sh}(\mathcal{F}_{|S})}, \\
\ln \text{Sh}(\mathcal{F}_{|S}) &\le \ln \text{Sh}(\mathcal{F}; n) \le \text{VC}(\mathcal{F}) \ln(n+1).
\end{aligned}$$

2. VC dimension of linear predictors.

Theorem. Define $\mathcal{F} := \left\{x \mapsto \text{sgn}(\langle a, x\rangle - b) : a \in \mathbb{R}^d, b \in \mathbb{R}\right\}$ ("linear classifiers"/"affine classifier"/ "linear threshold function (LTF)"). Then $\text{VC}(\mathcal{F}) = d + 1$.

Remarks.

- By Sauer-Shelah, $\text{Sh}(\mathcal{F}; n) \le n^{d+1} + 1$. Anthony-Bartlett chapter 3 gives an exact equality; only changes constants of $\ln \text{VC}(\mathcal{F}; n)$.
- Let's compare to Rademacher:

$$\begin{aligned}
\text{URad}(\text{sgn}(\mathcal{F}_{|S})) &\le \sqrt{2nd \ln(n+1)}, \\
\text{URad}(\mathbb{R}x \mapsto \langle w, x\rangle : \|w\| \le R\}_{|S}) &\le R\|X_S\|_F,
\end{aligned}$$

where $\|X_S\|_F^2 = \sum_{x \in S} \|x\|_2^2 \le n \cdot d \cdot \max_{i,j} x_{i,j}$. One is scale-sensitive (and suggests regularization schemes), other is scale-insensitive.

Proof of lower bound $\text{VC}(\mathcal{F}) \ge d + 1$.

- Suffices to show $\exists S := \{x_1, \ldots, x_{d+1}\}$ with $\text{Sh}(\mathcal{F}_{|S}) = 2^{d+1}$.
- Choose $S := \{\mathbf{e}_1, \ldots, \mathbf{e}_d, (0, \ldots, 0)\}$.

Given any $P \subseteq S$, define (a, b) as

$$a_i := 2 \cdot \mathbb{1}[\mathbf{e}_i \in P] - 1, \qquad b := \frac{1}{2} - \mathbb{1}[0 \in P].$$

Then

$$\begin{aligned}
\text{sgn}(\langle a, \mathbf{e}_i\rangle - b) &= \text{sgn}(2\mathbb{1}[\mathbf{e}_i \in P] - 1 - b) = 2\mathbb{1}[\mathbf{e}_i \in P] - 1, \\
\text{sgn}(\langle a, 0\rangle - b) &= \text{sgn}(2\mathbb{1}[0 \in P] - 1/2) = 2\mathbb{1}[0 \in P] - 1,
\end{aligned}$$

meaning this affine classifier labels S according to P, which was an arbitrary subset.

Proof (of upper bound $\mathrm{VC}(\mathcal{F}) < d+2$).

- Consider any $S \subseteq \mathbb{R}^d$ with $|S| = d+2$.
- By *Radon's Lemma* (proved on next page), there exists a partition of S into nonempty (P, N) with $\mathrm{conv}(P) \cap \mathrm{conv}(N)$.
- Label P as positive and N as negative. Given any affine classifier, it can not be correct on all of S (and thus $\mathrm{VC}(\mathcal{F}) < d+2$): either it is incorrect on some of P, or else it is correct on P, and thus has a piece of $\mathrm{conv}(N)$ and thus $x \in N$ labeled positive.

Theorem (Radon's Lemma). Given $S \subseteq \mathbb{R}^d$ with $|S| = d+2$, there exists a partition of S into nonempty (P, N) with $\mathrm{conv}(P) \cap \mathrm{conv}(S) \neq \emptyset$.

Proof. Let $S = \{x_1, \ldots, x_{d+2}\}$ be given, and define $\{u_1, \ldots, u_{d+1}\}$ as $u_i := x_i - x_{d+2}$, which must be linearly dependent:

- Exist scalars $(\alpha_1, \ldots, \alpha_{d+1})$ and a j with $\alpha_j := -1$ so that

$$\sum_i \alpha_i u_i = -u_j + \sum_{i \neq j} \alpha_i u_i = 0;$$

- thus $x_j - x_{d+2} = \sum_{\substack{i \neq j \\ i < d+2}} \alpha_i (x_i - x_{d+2})$ and $0 = \sum_{i<d+2} \alpha_i x_i - x_{d+2} \sum_{i<d+2} \alpha_i =: \sum_j \beta_j x_j$, where $\sum_j \beta_j = 0$ and not all β_j are zero.

Proof (continued).

Set $P := \{i : \beta_i > 0\}$, $N := \{i : \beta_i \leq 0\}$; where neither set is empty.

Set $\beta := \sum_{i \in P} \beta_i - \sum_{i \in N} \beta_i > 0$.

Since $0 = \sum_i \beta_i x_i = \sum_{i \in P} \beta_i x_i + \sum_{i \in N} \beta_i x_i$, then

$$\frac{0}{\beta} = \sum_{i \in P} \frac{\beta_i}{\beta} x_i + \sum_{i \in N} \frac{\beta_i}{\beta} x_i$$

and the point $z := \sum_{i \in P} \beta_i x_i / \beta = \sum_{i \in N} \beta_i x_i / (-\beta)$ satisfies $z \in \mathrm{conv}(P) \cap \mathrm{conv}(N)$.

Remarks.

- Generalizes Minsky-Papert "xor" construction from lecture 2.
- Indeed, the first appearance I know of shattering/VC was in approximation theory, the papers of Warren and Shapiro, and perhaps it is somewhere in Kolmogorov's old papers.

3. VC dimension of LTF networks.

Consider iterating the previous construction, giving an "LTF network": a neural network with activation $z \mapsto \mathbb{1}[z \geq 0]$.

We'll analyze this by studying output of all nodes. To analyze this, we'll study not just the outputs, but the behavior of all nodes.

Definition.

- Given a sample S of size n and an LTF network with m nodes (in any topologically sorted order), define activation matrix $A := \text{Act}(S; W := (a_1, \ldots, a_m))$ where A_{ij} is the output of node j on input i, with fixed network weights W.
- Let $\text{Act}(S; \mathcal{F})$ denote the set of activation matrices with architecture fixed and weights W varying.

Remarks.

- Since last column is the labeling, $|\text{Act}(S; \mathcal{F})| \geq \text{Sh}(\mathcal{F}_{|S})$.
- Act seems a nice complexity measure, but it is hard to estimate given a single run of an algorithm (say, unlike a Lipschitz constant).
- We'll generalize Act to analyze ReLU networks.

Theorem.

For any LTF architecture $\mathcal{F}$ with p parameters,

$$\text{Sh}(\mathcal{F}; n) \leq |\text{Act}(S; \mathcal{F})| \leq (n+1)^p.$$

When $p \geq 12$, then $\text{VC}(\mathcal{F}) \leq 6p\ln(p)$.

Proof.

- Topologically sort nodes, let $(p_1, \ldots, p_m)$ denote numbers of respective numbers of parameters (thus $\sum_i p_i = p$).
- Proof will iteratively construct sets $(U_1, \ldots, U_m)$ where U_i partitions the weight space of nodes $j \leq i$ so that, within each partition cell, the activation matrix does not vary.
- The proof will show, by induction, that $|U_i| \leq (n+1)^{\sum_{j \leq i} p_j}$. This completes the proof of the first claim, since $\text{Sh}(\mathcal{F}_{|S}) \leq |\text{Act}(\mathcal{F}; S)| = |U_m|$.
- For convenience, define $U_0 = \{\emptyset\}$, $|U_0| = 1$; the base case is thus $|U_0| = 1 = (n+1)^0$.

Proof (inductive step).

Let $j \geq 1$ be given; the proof will now construct U_{j+1} by refining the partition U_j.

- Fix any cell C of U_j; as these weights vary, the activation is fixed, thus the input to node $j+1$ is fixed for each $x \in S$.
- Therefore, on this augmented set of n inputs (S with columns of activations appended to each example), there are $(n+1)^{p_{j+1}}$ possible outputs via Sauer-Shelah and the VC dimension of affine classifiers with p_{j+1} inputs.
- In other words, C can be refined into $(n+1)^{p_{j+1}}$ sets; since C was arbitrary,

$$|U_{j+1}| = |U_j|(n+1)^{p_{j+1}} \leq (n+1)^{\sum_{l \leq j+1} p_l}.$$

Proof (VC dimension bound).

It ermains to bound the VC dimension via this Shatter bound:

$$
\begin{aligned}
& \mathrm{VC}(\mathcal{F}) < n \\
\Longleftarrow & \forall i \geq n \centerdot \mathrm{Sh}(\mathcal{F}; i) < 2^i \\
\Longleftarrow & \forall i \geq n \centerdot (i+1)^p < 2^i \\
\Longleftrightarrow & \ \forall i \geq n \centerdot p \ln(i+1) < i \ln 2 \\
\Longleftrightarrow & \ \forall i \geq n \centerdot p < \frac{i \ln(2)}{\ln(i+1)} \\
\Longleftarrow & p < \frac{n \ln(2)}{\ln(n+1)}
\end{aligned}
$$

If $n = 6p \ln(p)$,

$$
\begin{aligned}
\frac{n \ln(2)}{\ln(n+1)} \geq \frac{n \ln(2)}{\ln(2n)} &= \frac{6p \ln(p) \ln(2)}{\ln 12 + \ln p + \ln \ln p} \\
&\geq \frac{6p \ln p \ln 2}{3 \ln p} > p.
\end{aligned}
$$

Remarks.

- Had to do handle $\forall i \geq n$ since VC dimension is defined via sup; one can define funky $\mathcal{F}$ where Sh is not monotonic in n.
- Lower bound is $\Omega(p \ln m)$; see Anthony-Bartlett chapter 6 for a proof. This lower bound however is for a specific fixed architecture!
- Other VC dimension bounds: ReLU networks have $\widetilde{\mathcal{O}}(pL)$, sigmoid networks have $\widetilde{\mathcal{O}}(p^2 m^2)$, and there exists a convex-concave activation which is close to sigmoid but has VC dimension ∞.
- Matching lower bounds exist for ReLU, not for sigmoid; but even the "matching" lower bounds are deceptive since they hold for a *fixed* architecture of a given number of parameters and layers.