## Lecture 23. (Sketch.)

- In class we also discussed a recent paper to highlight the role of random initialization in neural networks; I'm not including notes on that...

## 1. VC dimension of ReLU networks.

Today's ReLU networks will predict with

$$x \mapsto A_L\sigma_{L-1}\left(A_{L-1}\cdots A_2\sigma_1(A_1x + b_1) + b_2 \cdots + b_{L-1}\right) + b_L,$$

where $A_i \in \mathbb{R}^{d_i \times d_{i-1}}$ and $\sigma_i : \mathbb{R}^{d_i \to d_i}$ applies the ReLU $z \mapsto \max\{0, z\}$ coordinate-wise.

**Convenient notation:** collect data as rows of matrix $X \in \mathbb{R}^{n\times d}$, and define

$$X_0 := X^\top \qquad Z_0 := \text{all 1s matrix,}$$
$$X_i := A_i(Z_{i-1} \odot X_{i-1}) + b_i 1_n^\top, \quad X_i := \mathbb{1}[X_i \geq 0],$$

where $(Z_1, \ldots, Z_L)$ are the activation matrices.

**Theorem** (See also Bartlett-Harvey-Liaw-Mehrabian Theorem 6).

Let fixed ReLU network $\mathcal{F}$ be given with $p = \sum_{i=1}^L p_i$ parameters, $L$ layers, $m = \sum_{i=1}^L m_i$ nodes. Let examples $(x_1, \ldots, x_n)$ be given and collected into matrix $X$. There exists a partition $U_L$ of the parameter space satisfying:

- Fix any $C \in U_L$. As parameters vary across $C$, activations $(Z_1, \ldots, Z_L)$ are fixed.
- $\text{Sh}(\mathcal{F}; n) \leq |\{Z_L(C) : C \in U_L\}| \leq |U_L| \leq (12nL)^{pL}$, where $Z_L(C)$ denotes the sign pattern in layer $L$ for $C \in U_L$.
- If $pL^2 \geq 72$, then $\text{VC}(\mathcal{F}) \leq 6pL\ln(pL)$.

- As with LTF networks, the prove inductively constructs partitions of the weights up through layer $i$ so that the activations are fixed across all weights in each partition cell.
- Consider a fixed cell of the partition, whereby the activations are fixed zero-one matrices. As a function of the *inputs*, the ReLU network is now *an affine function*; as a function of the *weights* it is *multilinear* or rather *a polynomial of degree L*.
- Consider again a fixed cell and some layer $i$; thus $\sigma(X_i) = Z_i \odot X_i$ is a matrix of polynomials of degree $i$ (in the weights). If we can upper bound the number of possible signs of $A_{i+1}(Z_i \odot X_i) + b_i 1_n^\top$, then we can refine our partition of weight space and recurse. For that we need a bound on sign patterns of polynomials, as on the next slide.

**Theorem** (Warren '68; see also Anthony-Bartlet Theorem 8.3).
Let $F$ denote functions $x \mapsto f(x; w)$ which are $r$-degree polynomials in $w \in \mathbb{R}^p$. If $n \geq p$, then $\mathrm{Sh}(\mathcal{F}; n) \leq 2({}^{2enr}/_{p})^p$.

**Remark.** Proof is pretty intricate, and omitted. It relates the VC dimension of $F$ to the zero sets $Z_i := \{w \in \mathbb{R}^p : f(x; w) = 0\}$, which it controls with an application of Bezout's Theorem. The zero-counting technique is also used to obtain an exact Shatter coefficient for affine classifiers.

**Proof** (of ReLU VC bound).

We'll inductively construct partitions $(U_0, \ldots, U_L)$ where $U_i$ partitions the parameters of layers $j \leq i$ so that for any $C \in U_i$, the activations $Z_j$ in layer $j \leq i$ are fixed for all parameter choices within $C$ (thus let $Z_j(C)$ denote these fixed activations).

The proof will proceed by induction, showing $|U_i| \leq (12nL)^{pi}$.

**Base case** $i = 0$: then $U_0 = \{\emptyset\}$, $Z_0$ is all ones, and $|U_0| = 1 \leq (12nL)^{pi}$.

**Proof** (inductive step).

- Fix $C \in S_i$ and $(Z_1, \ldots, Z_i) = (Z_1(C), \ldots, Z_i(C))$.
- Note $X_{i+1} = A_{i+1}(Z_i \odot X_i) + b_i 1_n^\top$ is polynomial (of degree $i+1$) in the parameters since $(Z_1, \ldots, Z_i)$ are fixed.
- Therefore

  $$\left|\{\mathbb{1}[X_{i+1} \geq 0] : \text{params} \in C\}\right| \leq \mathrm{Sh}(i+1 \text{ deg poly};\ m_i \cdot n \text{ functions})$$
  $$\leq 2\left(\frac{2enm_{i+1}}{\sum_{j \leq i} p_j}\right)^{\sum_{j \leq i+1} p_j} \leq (12nL)^p.$$

  [ Technical comment: to apply the earlier shatter bound for polynomials, we needed $n \cdot m_{i+1} \geq \sum_j p_j$; but if (even more simply) $p \geq nm_{i+1}$, we can only have $\leq 2^{nm_{i+1}} \leq 2^p$ activation matrices anyway, so the bound still holds. ]
- Therefore carving $U_i$ into pieces according to $Z_{i+1} = \mathbb{1}[X_{i+1} \geq 0]$ being fixed gives

  $$|U_{i+1}| \leq |U_i|(12nL)^p \leq (12nL)^{p(i+1)}.$$

**Proof** (VC bound).

As with LTF networks,

$$\begin{aligned}
\mathrm{VC}(\mathcal{F}) < n &\Longleftarrow \forall i \geq n \centerdot \mathrm{Sh}(\mathcal{F}; i) < 2^i \\
&\Longleftarrow \forall i \geq n \centerdot (12iL)^{pL} < 2^i \\
&\Longleftrightarrow \forall i \geq n \centerdot pL \ln(12iL) < i \ln 2 \\
&\Longleftrightarrow \forall i \geq n \centerdot pL < \frac{i \ln 2}{\ln(12iL)} \\
&\Longleftarrow pL < \frac{n \ln 2}{\ln(12nL)}
\end{aligned}$$

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

$$\begin{aligned}
\frac{n \ln 2}{\ln(12nL)} &= \frac{6pL\ln(pL)\ln(2)}{\ln(72pL^2\ln(pL))} = \frac{6pL\ln(pL)\ln(2)}{\ln(72) + \ln(pL^2) + \ln\ln(pL)} \\
&\geq \frac{6pL\ln(pL)\ln(2)}{\ln(72) + \ln(pL^2) + \ln(pL) - 1} \geq \frac{6\ln(pL)\ln(2)}{3\ln(pL^2)} \\
&= 2pL\ln 2 > pL.
\end{aligned}$$

**Remarks.**

- If ReLU is replaced with a degree $r \geq 2$ piecewise polynomial activation, have $r^i$-degree polynomial in each cell of partition, and shatter coefficient upper bound scales with $L^2$ not $L$. The lower bound in this case still has $L$ not $L^2$; it's not known where the looseness is.
- Lower bounds are based on digit extraction, and for each pair $(p, L)$ require a fixed architecture.