Lecture 21

Sample Complexity of Neural Networks

Sasha Rakhlin

Nov 19, 2019

In this lecture we will discuss a couple of ways to analyze sample complexity of neural networks.

First result is concerned with the VC dimension of neural networks.

(Bartlett et al '17): fix an architecture of a neural network with L layers, W parameters, and ReLU activation. Then VC dimension of the collection of functions is $O(WL \log W)$.

See above paper for a more general statement, other activation functions, etc.

Since typical VC bounds scale with $\frac{\text{VCdim}(\mathcal{F})}{n}$, the bound is vacuous whenever $WL \log W$ exceeds n.

However, as we saw earlier, VC dimension is only a loose upper bound on the generalization performance. We saw this on the example of Perceptron, where the dimension could be taken to be infinite, yet inverse margin (or ℓ_2 norm of the separating hyperplane) can determine sample complexity.

In particular, we saw that for Perceptron we could analyze (with the help of the margin bound) the Rademacher averages of the class

$$\mathcal{F}_{\text{lin}} = \{x \mapsto \langle w, x \rangle : \|w\| \leq 1\}.$$

We saw that

$$\widehat{\mathcal{R}}_n(\mathcal{F}_{\text{lin}}) \leq \frac{1}{\sqrt{n}}$$

irrespective of dimensionality, assuming $\|x_i\| \leq 1$.

Fix an architecture of a neural network. Recall our notation:

$$f_W(x) = W^L \sigma\left(W^{L-1} \sigma\left(\dots \sigma\left(W^1 x\right)\dots\right)\right)$$

where we abbreviate $W = (W^1, \dots, W^L)$.

Just as in the case of $\mathcal{F}_{lin}$, we would like to define a "ball" in the space of neural networks:

$$\left\{f_W : \text{compl}(W^1, \dots, W^L) \le 1\right\}$$

for some notion of complexity compl.

Note: many tuples $(W^1, \dots, W^L)$ lead to the same function f_W. Example: take ReLU activation, scale one layer up by 100, another down by 100. Function does not change. This is "invariance" to a transformation. There are many transformations that leave the function intact. We would like to make sure compl does not assign different values of complexity to different sets of parameters if they lead to same function.

Example: take Frobenius norm of all the layers:

$$\mathrm{compl}(W) = \sum_{j=1}^{L} \left\| W^j \right\|_F$$

since this is a natural "generalization" of the corresponding Euclidean norm for $\mathcal{F}_{\mathtt{lin}}$. Unfortunately, this measure does not capture the scaling invariance of the layers. However, a product of Frobenius norms would reflect the invariance (though it may not reflect many other invariances)

$$\mathrm{compl}(W) = \prod_{j=1}^{L} \left\| W^j \right\|_F$$

Of course, it is not at all clear that the Rademacher averages of a unit ball defined with respect to this complexity is non-vacuous. Remember that we relied heavily on linearity of functions to analyze $\widehat{\mathscr{R}}_n(\mathcal{F}_{\mathtt{lin}})$.

Norms based on spectrum

Before we start, some other norms of a $d_1 \times d_2$ matrix A:

Operator norm (or, spectral norm, or 2-norm) of a matrix A:

$$\|A\| = \sigma_{max}(A) = \sqrt{\lambda_{max}(A^*A)}$$

and can also be written as

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

General Schatten norms:

$$\|A\|_p = \left(\sum_{i=1}^{\min(d_1,d_2)} \sigma_i^p \right)^{1/p}$$

The $p = 2$ case coincides with the Frobenius norm.

The $p = 1$ case is termed nuclear norm, or trace norm, or Ky Fan norm:

$$\|A\|_{nuc} = \sum_{i=1}^{\min(d_1,d_2)} \sigma_i = \text{trace}(\sqrt{A^*A}).$$

Entrywise norms

Sum of ℓ_2 norms of columns:

$$\|A\|_{2,1} = \sum_{j=1}^{d_2} \|A_{\cdot,j}\| = \sum_{j=1}^{d_2} \left(\sum_{i=1}^{d_1} A_{i,j}^2 \right)^{1/2}$$

Maximum ℓ_2 norm of columns:

$$\|A\|_{2,\infty} = \max_{j=1 \ldots d_2} \|A_{\cdot,j}\|$$

For general $p, q \geq 1$,

$$\|A\|_{p,q} = \left(\sum_{j=1}^{d_2} \left(\sum_{i=1}^{d_1} |A_{i,j}|^p \right)^{q/p} \right)^{1/q}$$

Sample complexity by recursive peeling

Define the class of L-layer neural networks (with fixed architecture) recursively as

$$\mathcal{F}_i = \left\{ x \mapsto \sum_{j=1}^{d_{i-1}} w_j \sigma(f_j(x)) : f_j \in \mathcal{F}_{i-1}, \|w\|_1 \leq B_i \right\}$$

where d_i is number of hidden units in ith layer. Think of this class as a class of real valued functions implementable as a linear combination of lower-level functions using ℓ_1-bounded combination. We assume that σ is 1-Lipschitz. Base class $\mathcal{F}_1$ is some class (e.g. linear functions computed by first layer). Think of w as any row of W^i.

Claim:

$$\widehat{\mathscr{R}}_n(\mathcal{F}_i) \leq 2B_i \widehat{\mathscr{R}}_n(\mathcal{F}_{i-1})$$

Proof:

$$
\begin{aligned}
n\widehat{\mathscr{R}}_n(\mathcal{F}_i) &= \mathbb{E}_\epsilon \max_{\|w\|_1 \le B_i, f_j \in \mathcal{F}_{i-1}} \sum_{t=1}^{n} \epsilon_t \sum_j w_j \sigma(f_j(x_t)) \\
&= \mathbb{E}_\epsilon \max_{\|w\|_1 \le B_i, f_j \in \mathcal{F}_{i-1}} \sum_j w_j \sum_{t=1}^{n} \epsilon_t \sigma(f_j(x_t)) \\
&\le \mathbb{E}_\epsilon \max_{\|w\|_1 \le B_i, f_j \in \mathcal{F}_{i-1}} \|w\|_1 \max_j \left| \sum_{t=1}^{n} \epsilon_t \sigma(f_j(x_t)) \right|
\end{aligned}
$$

where the last step is by the Cauchy-Schwartz inequality. The last expression is at most

$$
B_i \mathbb{E}_\epsilon \max_{f \in \mathcal{F}_{i-1}} \left| \sum_{t=1}^{n} \epsilon_t \sigma(f(x_t)) \right|
$$

which is at most

$$
2B_i \mathbb{E}_\epsilon \max_{f \in \mathcal{F}_{i-1}} \sum_{t=1}^{n} \epsilon_t \sigma(f(x_t)).
$$

By contraction property this is at most $2B_i \widehat{\mathscr{R}}_n(\mathcal{F}_{i-1})$, as claimed.

Let's fix

$$\mathcal{F}_1 = \{x \mapsto \langle w, x \rangle : \|w\|_1 \leq B_1\}$$

and assume further that $\|x\|_\infty \leq 1$. We have seen that in this case

$$\widehat{\mathscr{R}}_n(\mathcal{F}_1) \leq c\sqrt{\frac{\log d}{n}}$$

Putting everything together (Bartlett and Mendelson '03):

The Rademacher averages of $\mathcal{F}_L$, the class of L-hidden-layer neural networks with rows of weight matrices W^i bounded by B_i in ℓ_1 norm (that is, $\|(W^i)^\top\|_{1,\infty} \leq B_i$) is

$$O\left(2^L \cdot \prod_{i=1}^{L} B_i \cdot \sqrt{\frac{\log d}{n}}\right)$$

- Pros: no (explicit) dependence on number of units, only on the size of weights (similar to Perceptron case in spirit).
- Cons: $\|W^i\|_{1,\infty}$ may be large. Exponential dependence on depth.

Is exponential dependence on depth unavoidable?

Consider a *thin* neural network $f(x) = w^L \sigma\left(\dots \sigma(w^1 x)\dots\right)$ with $w^1 \in \mathbb{R}^{1\times d}$ and all $w^j \in \mathbb{R}_{\geq 0}$ for $j > 1$ be nonnegative numbers. Take σ to be ReLU. Then by positive homogeneity of ReLU,

$$f(x) = \prod_{j>1} w^j \cdot \langle w^1, x\rangle$$

Clearly, in this trivial case there is no exponential dependence on depth. Question is whether this dependence can be avoided beyond this trivial case. The next positive example is diagonal matrices. Beyond that?

Some of the existing results

Ignoring logarithmic factors and constants: generalization error bounded by

(Neyshabur and Srebro '15):

$$\prod_{j=1}^{L} \left\| W^j \right\|_F \cdot 2^L \cdot \frac{1}{\sqrt{n}}$$

(Neyshabur et al '17):

$$\prod_{j=1}^{L} \left\| W^j \right\| \cdot L \cdot \sqrt{h \sum_{j=1}^{L} \frac{\left\| W^j \right\|_F^2}{\left\| W^j \right\|^2}} \cdot \frac{1}{\sqrt{n}}$$

(Bartlett et al '17):

$$\prod_{j=1}^{L} \left\| W^j \right\| \cdot \left(\sum_{j=1}^{L} \left(\frac{\left\| W^j \right\|_{2,1}}{\left\| W^j \right\|} \right)^{2/3} \right)^{3/2} \cdot \frac{1}{\sqrt{n}}$$

Here h is network width.

(Golowich et al '18)

$$\tilde{O}\left(\min\left\{\prod_{j=1}^{L}\|W_j\|_F \cdot \frac{1}{n^{1/4}}\ ,\ \prod_{j=1}^{L}\|W_j\|_F \cdot \sqrt{\frac{L}{n}}\right\}\right).$$

- first regime: Independent of network depth or width
- second regime: mild $\sqrt{L}$ dependence on depth

This is as close to the Perceptron analogue as we can get at this point.

One can show (see Golowich et al) that some dependence on width of network is necessary if one only controls Schatten-p norm for $p > 1/2$. At $p = 2$ (Frobenius norm), one can avoid depth and width dependence.