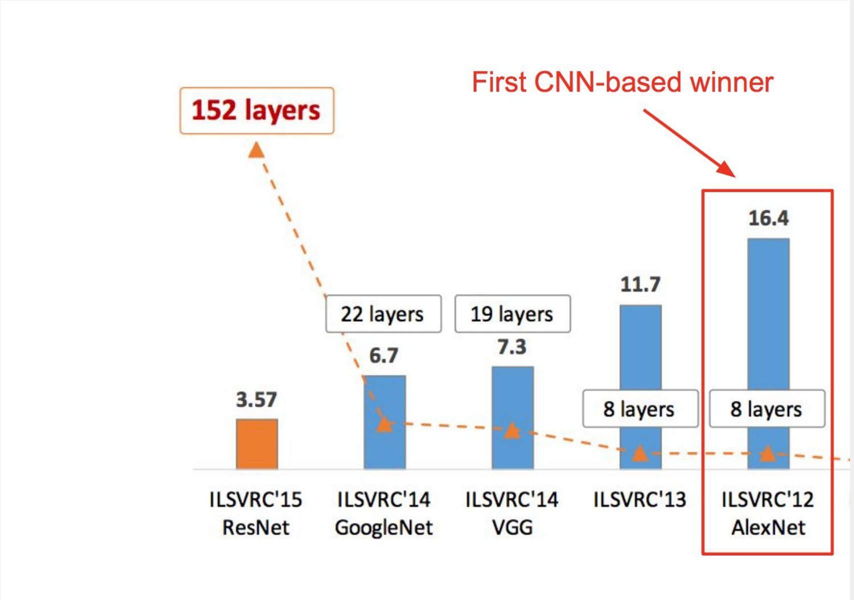
Size-Independent Sample Complexity of Neural Networks

Saurabh Mathur

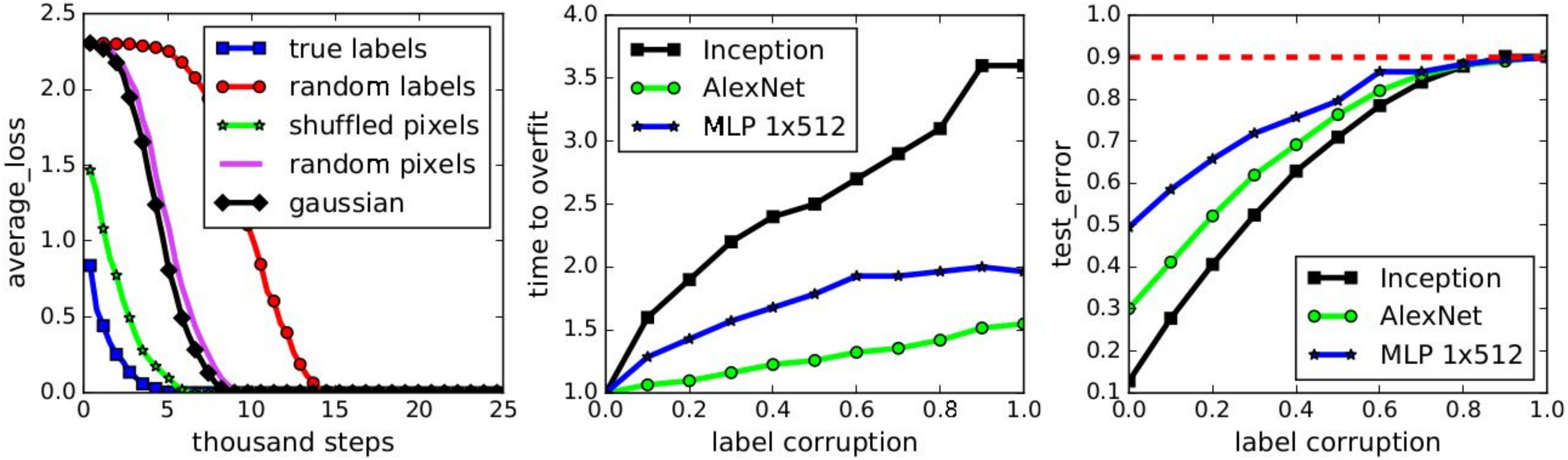
Authors: Noah Golowich, Alexander Rakhlin, Ohad Shamir

Contributions : Bounds on Rademacher complexity

1. Exponential to polynomial-in-depth bound for general neural network.
2. Depth-independent bound for case where weight norm is constrained.



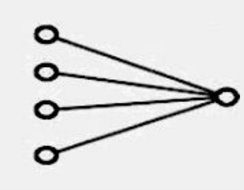
Network	#params
AlexNet	35K
VGG16	138M
GoogleNet	5M
ResNet	25M



average_loss
thousand steps
true labels
random labels
shuffled pixels
random pixels
gaussian
time to overfit
label corruption
Inception
AlexNet
MLP 1x512
test_error
label corruption
Inception
AlexNet
MLP 1x512

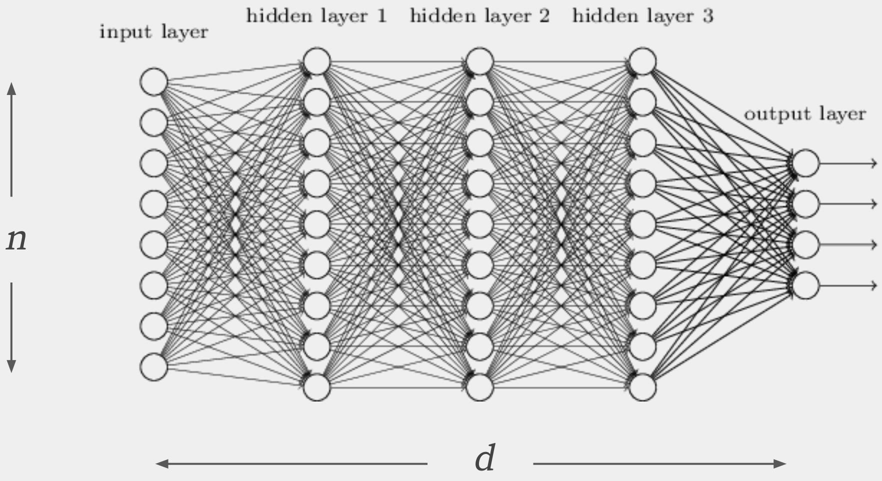
Can we bound the R(generalization error) for neural networks?

Simplest Case



$$\mathcal{O}\left(\sqrt{\frac{B^2}{m}}\right)$$

$$\|w\| \leq B$$



$$x \mapsto W_d \sigma_d(\ldots \sigma_2(W_2 \sigma_1(W_1 \boldsymbol{x}))\ldots)$$

$$\sigma : z \mapsto max\{z, 0\}$$

VC bounds

$$\sqrt{\frac{n^2 d^2}{m}}$$

VC bounds

$$\sqrt{\frac{n^2 d^2}{m}}$$

Lipschitz bounds

$$\frac{\prod_{j=1}^{d} \|W_j\|_{op}}{m^{1/n}}$$

VC bounds

$$\sqrt{\frac{n^2 d^2}{m}}$$

Lipschitz bounds

$$\frac{\prod_{j=1}^{d} \|W_j\|_{op}}{m^{1/n}}$$

Scale-sensitive bounds (e.g. [NS15])

$$2^d \sqrt{\frac{\prod_{j=1}^{d} \|W_j\|_F^2}{m}}$$

VC bounds

$$\sqrt{\frac{n^2 d^2}{m}}$$

Lipschitz bounds

$$\frac{\prod_{j=1}^{d} \|W_j\|_{op}}{m^{1/n}}$$

Scale-sensitive bounds (e.g. [NS15])

$$2^d \sqrt{\frac{\prod_{j=1}^{d} \|W_j\|_F^2}{m}}$$

$$\sqrt{\frac{d^2 n \left(\prod_{j=1}^{d} \|W_j\|_{op}^2\right) \sum_{j=1}^{d} \frac{\|W_j\|_F^2}{\|W_j\|^2}}{m}}$$

[NS17]

VC bounds

$$\sqrt{\frac{n^2 d^2}{m}}$$

Lipschitz bounds

$$\frac{\prod_{j=1}^{d} \|W_j\|_{op}}{m^{1/n}}$$

Scale-sensitive bounds (e.g. [NS15])

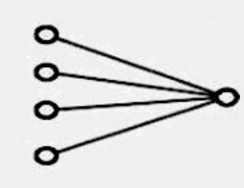
$$2^d \sqrt{\frac{\prod_{j=1}^{d} \|W_j\|_F^2}{m}}$$

$$\sqrt{\frac{d^2 n \left(\prod_{j=1}^{d} \|W_j\|_{op}^2\right) \sum_{j=1}^{d} \frac{\|W_j\|_F^2}{\|W_j\|^2}}{m}}$$ [NS17]

$$\sqrt{\frac{\left(\prod_{j=1}^{d} \|W_j\|_{op}^2\right) \left(\sum_{j=1}^{d} \left(\frac{\|W_j\|_{2,1}}{\|W_j\|_{op}}\right)^{2/3}\right)^3}{m}}$$ [BFT17]

Can we make our bounds independent of depth?

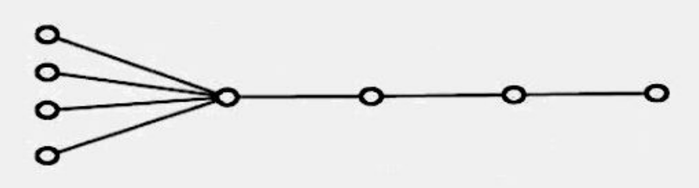
Simplest Case



$$\mathcal{O}\left(\sqrt{\frac{B^2}{m}}\right)$$

$$\|w\| \leq B$$

Thin Case



$$\mathcal{O}\left(\sqrt{\frac{B^2}{m}}\right)$$

$$\|w\| \leq B$$

Lower Bound

$$\Omega(\sqrt{\frac{B^2 n}{m}})$$

$$\prod_{j=1}^{d} \|W_j\|_{op} \leq B$$

$$\Omega(\sqrt{\frac{B^2 n^{max\{0,\frac{1}{2}-\frac{1}{p}\}}}{m}})$$

$$\prod_{j=1}^{d} \|W_j\|_{p-schatten} \leq B$$

Upper Bound

$$\mathcal{O}\left(min\left\{\frac{B}{m^{1/4}}, \sqrt{\frac{dB^2}{m}}\right\}\right)$$

$$\prod_{j=1}^{d} \|W_j\|_F \leq B$$

Interesting Tricks

1. **Log Sum Exp**
2. Eliminating depth-dependence using product of p-schatten norms

$$m \cdot \hat{\mathcal{R}}_m(\mathcal{H}_d) = \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}_d} \sum_i \varepsilon_i h(\mathbf{x}_i) \right]$$

Introducing a parameter $\lambda > 0$,

$$\begin{aligned} m \cdot \hat{\mathcal{R}}_m(\mathcal{H}_d) &= \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}_d} \sum_i \varepsilon_i h(\mathbf{x}_i) \right] \\ &= \frac{1}{\lambda} \log \exp \left(\lambda \cdot \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{h \in \mathcal{H}_d} \sum_i \epsilon_i h(\mathbf{x}_i) \right) \end{aligned}$$

Introducing a parameter $\lambda > 0$,

$$
\begin{aligned}
m \cdot \hat{\mathcal{R}}_m(\mathcal{H}_d) &= \mathbb{E}_{\epsilon}\left[\sup_{h \in \mathcal{H}_d} \sum_i \varepsilon_i h(\mathbf{x}_i)\right] \\
&= \frac{1}{\lambda} \log \exp\left(\lambda \cdot \mathbb{E}_{\epsilon} \sup_{h \in \mathcal{H}_d} \sum_i \epsilon_i h(\mathbf{x}_i)\right) \\
&\leq \ \frac{1}{\lambda} \log\left(\mathbb{E}_{\epsilon} \sup_{h \in \mathcal{H}_d} \exp\left(\lambda \sum_i \epsilon_i h(\mathbf{x}_i)\right)\right)
\end{aligned}
$$

Introducing a parameter $\lambda > 0$,

$$
\begin{aligned}
m \cdot \hat{\mathcal{R}}_m(\mathcal{H}_d) &= \mathbb{E}_{\boldsymbol{\epsilon}} \left[\sup_{h \in \mathcal{H}_d} \sum_i \varepsilon_i h(\mathbf{x}_i) \right] \\
&= \frac{1}{\lambda} \log \exp \left(\lambda \cdot \mathbb{E}_{\boldsymbol{\epsilon}} \sup_{h \in \mathcal{H}_d} \sum_i \epsilon_i h(\mathbf{x}_i) \right) \\
&\leq \frac{1}{\lambda} \log \left(\mathbb{E}_{\boldsymbol{\epsilon}} \sup_{h \in \mathcal{H}_d} \exp \left(\lambda \sum_i \epsilon_i h(\mathbf{x}_i) \right) \right)
\end{aligned}
$$

Using a contraction lemma variant and peeling as before:

$$
\frac{1}{\lambda} \log \left(2^d \mathbb{E}_{\epsilon} \exp \left(\lambda \prod_i C_{\mathcal{W}_i} \cdot f(\mathbf{x}_1, \ldots, \mathbf{x}_m) \right) \right)
$$

Theorem

If $\prod_{j=1}^{d} \|W_j\|_F \leq B$*, generalization error is*

$$\mathcal{O}\left(\sqrt{\frac{dB^2}{m}}\right)$$

Theorem

If $\prod_{j=1}^{d} \|W_j\|_{1,\infty} \leq B$*, generalization error is*

$$\mathcal{O}\left(\sqrt{\frac{(d + \log(n)) \cdot B^2}{m}}\right)$$

Interesting Tricks

1. Log Sum Exp
2. **Eliminating depth-dependence using product of p-schatten norms**

We can eliminate depth dependence using

- A bound for a network of depth r << d
- Composed with univariate Lipschitz function

The r^{th} layer is replaced with its Rank-1 approximation.

(under some weak assumptions) we can modify network with $\prod_{j=1}^{d} \|W_j\|_p \leq B$ by replacing one of first r matrices by rank-1:

$$\mathbf{x} \mapsto W_d \sigma(W_{d-1} \ldots W_k \sigma(\ldots \sigma(W_1 \mathbf{x}) \ldots) \ldots)$$

$\approx$

$$\mathbf{x} \mapsto \underbrace{W_d \sigma(W_{d-1} \ldots s\mathbf{u}}_{\text{Univariate Lipschitz func.}} \underbrace{\mathbf{v}^\top \sigma(\ldots \sigma(W_1 \mathbf{x}) \ldots) \ldots)}_{\text{Depth} \leq r \text{ network}}$$

- r trades-off approximation and statistical complexity

Theorem

If $\prod_{j=1}^{d} \|W_j\|_F \leq B$, *generalization error is*

$$\tilde{\mathcal{O}}\left(B \cdot \min\left\{\frac{\log(B/\Gamma)}{m^{1/4}}, \sqrt{\frac{d}{m}}\right\}\right)$$

Theorem (Depth-Independent Version of BFT17)

If $\prod_{j=1}^{d} \|W_j\|_{op} \leq B$, $\prod_{j=1}^{d} \|W_j\|_p \leq B_p$ *and* $\max_j \frac{\|W_j\|_{2,1}}{\|W_j\|} \leq L$, *generalization error is*

$$\tilde{\mathcal{O}}\left(BL \cdot \min\left\{\frac{(\log(B_p/\Gamma))^{\frac{1}{\frac{3}{2}+p}}}{m^{\frac{1}{2+3p}}}, \sqrt{\frac{d^3}{m}}\right\}\right)$$