Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks

Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, Nathan Srebro [1]

Benjamin Dubois-Taine

Dec 11th, 2019

The University of British Columbia

Brief Recap of this Semester

Why does deep learning work ?

Why does deep learning work ? So far : geometry of minima, implicit regularization of SGD, etc..

Today

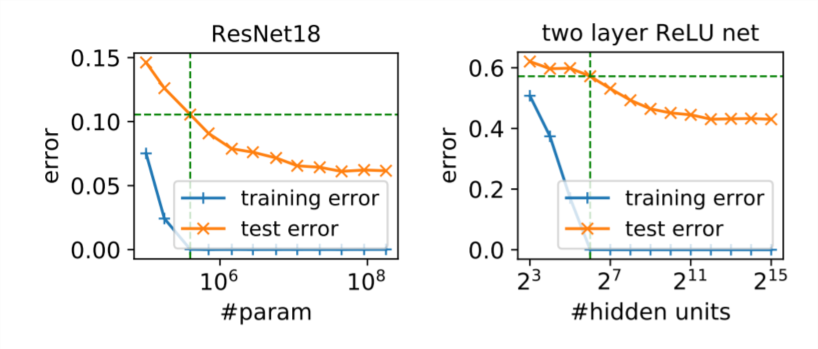
Why does deep learning work ? Today: Bounds on Generalization error.

Today

Why does deep learning work ? Today: Bounds on Generalization error.
More specifically, bounds on the complexity of a certain class of neural networks.

Motivation behind use of two-layer ReLU network

- Study restricted to two-layer ReLU neural networks
- Following experiment on CIFAR-10



Notation

Consider two-layer fully connected ReLU networks with input dimension d, output dimension c, and number of hidden units h.
Prediction function is

$$f_{V,U} : \mathbb{R}^d \to \mathbb{R}^c$$
$$f_{V,U}(x) = V[Ux]_+$$

with $x \in \mathbb{R}^d$, $U \in \mathbb{R}^{h \times d}$, $V \in \mathbb{R}^{c \times h}$.

Loss Function

Margin operator

$$\mu : \mathbb{R}^c \times [c] \to \mathbb{R}$$
$$\mu(f(x), y) = f(x)[y] - \max_{i \neq y} f(x)[i]$$

Loss Function

Margin operator

$$\mu : \mathbb{R}^c \times [c] \to \mathbb{R}$$
$$\mu(f(x), y) = f(x)[y] - \max_{i \neq y} f(x)[i]$$

Ramp loss

$$\ell_\gamma(f(x), y) = \begin{cases} 0 & \mu(f(x), y) > \gamma \\ \mu(f(x), y)/\gamma & \mu(f(x), y) \in [0, \gamma] \\ 1 & \mu(f(x), y) < 0 \end{cases}$$

Expected margin and empirical estimate

Ramp loss

$$\ell_\gamma(f(x), y) = \begin{cases} 0 & \mu(f(x), y) > \gamma \\ \mu(f(x), y)/\gamma & \mu(f(x), y) \in [0, \gamma] \\ 1 & \mu(f(x), y) < 0 \end{cases}$$

Expected margin loss of f

$$L_\gamma(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell_\gamma(f(x), y)\right]$$

Expected margin and empirical estimate

Ramp loss

$$\ell_\gamma(f(x), y) = \begin{cases} 0 & \mu(f(x), y) > \gamma \\ \mu(f(x), y)/\gamma & \mu(f(x), y) \in [0, \gamma] \\ 1 & \mu(f(x), y) < 0 \end{cases}$$

Expected margin loss of f

$$L_\gamma(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell_\gamma(f(x), y)\right]$$

Empirical estimate of expected margin loss

$$\hat{L}_\gamma(f) = \frac{1}{m}\sum_{i=1}^{m} \ell_\gamma(f(x_i), y_i)$$

Expected margin and empirical estimate

Ramp loss

$$\ell_\gamma(f(x), y) = \begin{cases} 0 & \mu(f(x), y) > \gamma \\ \mu(f(x), y)/\gamma & \mu(f(x), y) \in [0, \gamma] \\ 1 & \mu(f(x), y) < 0 \end{cases}$$

Expected margin loss of f

$$L_\gamma(f) = \mathbb{E}_{(x,y)\sim D}\left[\ell_\gamma(f(x), y)\right]$$

Empirical estimate of expected margin loss

$$\hat{L}_\gamma(f) = \frac{1}{m}\sum_{i=1}^{m} \ell_\gamma(f(x_i), y_i)$$

Write $L_0(f)$ and $\hat{L}_0(f)$ for expected risk and training error respectively.

Main Contributions

- Proved tighter bounds on the expected risk $L_0(f)$
- Empirically showed that it is the only known upper bound that decreases with the number of hidden units

Rademacher Complexity

Definition The Rademacher Complexity of a class $\mathcal{H}$ of functions with respect to the training set $\mathcal{S} = \{z_i\}_{i=1}^m$ is defined as

$$\mathcal{R}_{\mathcal{S}}(\mathcal{H}) = \frac{1}{m}\mathbb{E}_{\sigma\sim\{\pm1\}^m}\Big[\sup_{f\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i f(z_i)\Big]$$

Rademacher Complexity

Definition The Rademacher Complexity of a class $\mathcal{H}$ of functions with respect to the training set $\mathcal{S} = \{z_i\}_{i=1}^m$ is defined as

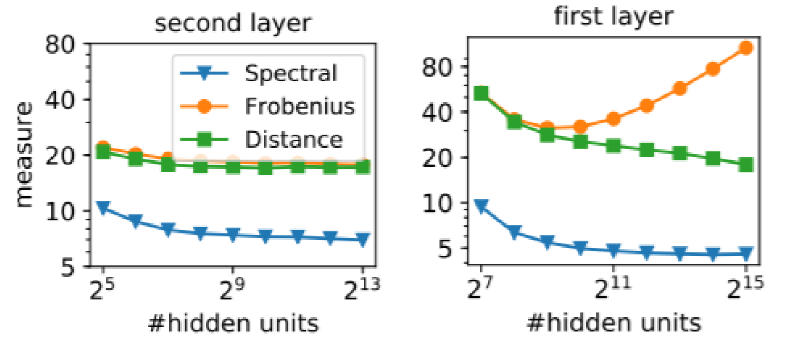
$$\mathcal{R}_{\mathcal{S}}(\mathcal{H}) = \frac{1}{m}\mathbb{E}_{\sigma\sim\{\pm1\}^m}\Big[\sup_{f\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i f(z_i)\Big]$$

Theorem: For a function class $\mathcal{H}$, with probability $1-\delta$, we have for any $f \in \mathcal{H}$

$$L_0(f) \leq \hat{L}_\gamma(f) + 2\mathcal{R}_S(\ell_\gamma \circ \mathcal{H}) + \sqrt{\frac{\ln(2/\delta)}{2m}}$$

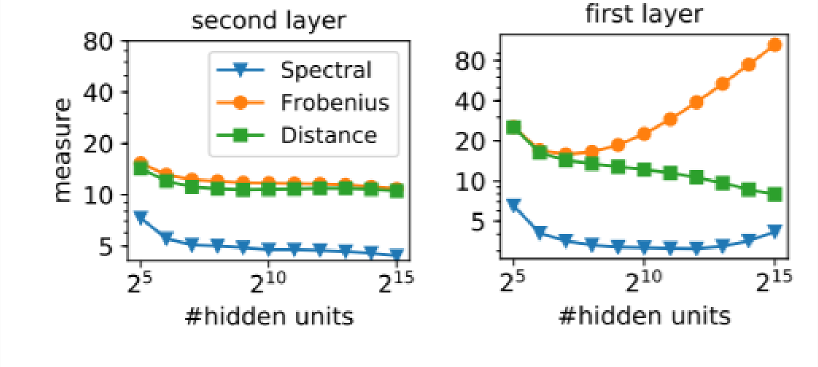
Reducing the class size : Empirical Investigation

Trained two-layer ReLU networks on SVHN



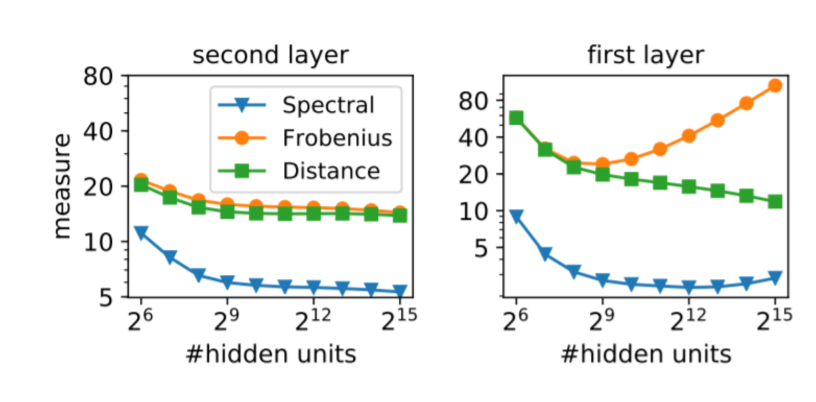
Reducing the class size : Empirical Investigation

Trained two-layer ReLU networks on MNIST



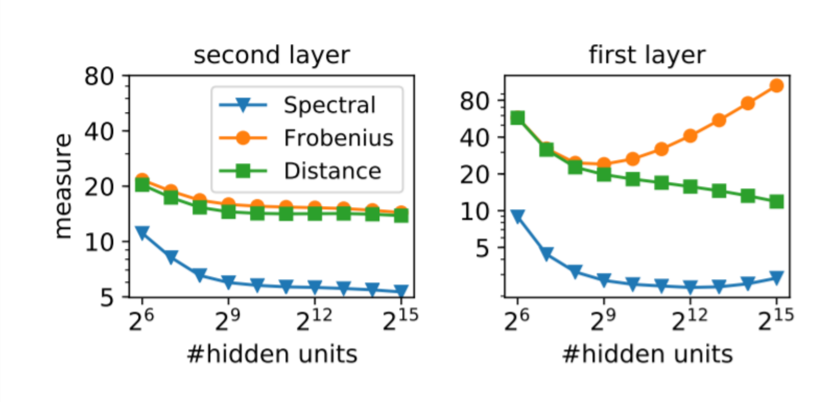
Reducing the class size : Empirical Investigation

Trained two-layer ReLU networks on CIFAR-10



Reducing the class size : Empirical Investigation

Trained two-layer ReLU networks on CIFAR-10



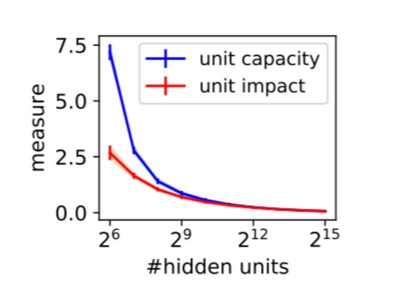
leads to defining

unit capacity: $||u_i - u_i^0||_2$

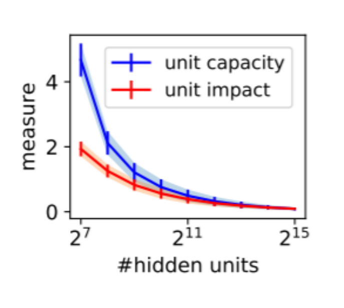
unit impact: $||v_i||_2$

Reducing the class size : Empirical Investigation

unit capacity: $||u_i - u_i^0||_2$
unit impact: $||v_i||_2$



(a) CIFAR-10



(b) SVHN

Reducing the function class size

These results lead to the definition of the following set of parameters

$$\mathcal{W} = \{(V, U) \mid V \in \mathbb{R}^{c \times h}, U \in \mathbb{R}^{h \times d}, ||v_i||_2 \leq \alpha_i, ||u_i - u_i^0||_2 \leq \beta_i\}$$

Reducing the function class size

These results lead to the definition of the following set of parameters

$$\mathcal{W} = \{(V, U) \mid V \in \mathbb{R}^{c \times h}, U \in \mathbb{R}^{h \times d}, ||v_i||_2 \leq \alpha_i, ||u_i - u_i^0||_2 \leq \beta_i\}$$

New function class

$$\mathcal{F}_{\mathcal{W}} = \{f(x) = V[Ux]_+ \mid (V, U) \in \mathcal{W}\}$$

Bounding the Rademacher complexity

Theorem: Given a training set $S = \{x_i\}_{i=1}^m$ and $\gamma > 0$, we have the following bound on the Rademacher complexity

$$\begin{aligned}\mathcal{R}_S(\ell_\gamma \circ \mathcal{F}_\mathcal{W}) &\leq \frac{2\sqrt{2c}+2}{\gamma m} \sum_{j=1}^{h} \alpha_j(\beta_j ||X||_F + ||u_j^0 X||_2) \\ &\leq \frac{2\sqrt{2c}+2}{\gamma m} ||\alpha||_2 \left(||\beta||_2 \sqrt{\frac{1}{m}\sum_{i=1}^{m} ||x_i||_2^2} + \sqrt{\frac{1}{m}\sum_{i=1}^{m} ||U^0 x_i||_2^2} \right)\end{aligned}$$

Proof Idea

Lemma 9 (Rademacher Decomposition). *Given a training set $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^m$ and $\gamma > 0$, Rademacher complexity of the class $\mathcal{F}_\mathcal{W}$ defined in equations (5) and (4) is bounded as follows:*

$$\mathcal{R}_\mathcal{S}(\ell_\gamma \circ \mathcal{F}_\mathcal{W}) \leq \frac{2}{\gamma m} \sum_{j=1}^{h} \mathop{\mathbb{E}}_{\boldsymbol{\xi}_i \in \{\pm 1\}^c, i \in [m]} \left[\sup_{\|\mathbf{v}_j\|_2 \leq \alpha_j} \sum_{i=1}^{m} (\rho_{ij} + \beta_j \|\mathbf{x}_i\|_2) \langle \boldsymbol{\xi}_i, \mathbf{v}_j \rangle \right]$$
$$+ \frac{2}{\gamma m} \sum_{j=1}^{h} \mathop{\mathbb{E}}_{\boldsymbol{\xi} \in \{\pm 1\}^m} \left[\sup_{\left\|\mathbf{u}_j - \mathbf{u}_j^0\right\|_2 \leq \beta_j} \sum_{i=1}^{m} \xi_i \alpha_j \langle \mathbf{u}_j, \mathbf{x}_i \rangle \right].$$

Bound on the generalization error

Theorem: For any $h \geq 2$, $\gamma > 0$, $\delta \in (0, 1)$, and $U^0 \in \mathbb{R}^{h \times d}$, with probability $1 - \delta$ over the choice of the training set $S = \{x_i\}_{i=1}^m$, for any $f(x) = V[Ux]_+$, we have

$$\begin{aligned} L_0(f) &\leq \hat{L}_\gamma(f) + O\left(\frac{\sqrt{c}||V||_F(||U - U^0||_F||X||_F + ||U^0X||_F)}{\gamma m} + \sqrt{\frac{h}{m}} \right) \\ &\leq \hat{L}_\gamma(f) + O\left(\frac{\sqrt{c}||V||_F(||U - U^0||_F + ||U^0||_2)\sqrt{\frac{1}{m}\sum_{i=1}^m ||x_i||_2^2}}{\gamma m} + \sqrt{\frac{h}{m}} \right) \end{aligned}$$

Comparison with other Capacity Bounds

#	Reference	Measure
(1)	Harvey et al. [9]	$\tilde{\Theta}(dh)$
(2)	Bartlett and Mendelson [3]	$\tilde{\Theta}\left(\|\mathbf{U}\|_{\infty,1} \|\mathbf{V}\|_{\infty,1}\right)$
(3)	Neyshabur et al. [20], Golowich et al. [7]	$\tilde{\Theta}\left(\|\mathbf{U}\|_F \|\mathbf{V}\|_F\right)$
(4)	Bartlett et al. [4], Golowich et al. [7]	$\tilde{\Theta}\left(\|\mathbf{U}\|_2 \|\mathbf{V}-\mathbf{V}_0\|_{1,2} + \|\mathbf{U}-\mathbf{U}_0\|_{1,2} \|\mathbf{V}\|_2\right)$
(5)	Neyshabur et al. [23]	$\tilde{\Theta}\left(\|\mathbf{U}\|_2 \|\mathbf{V}-\mathbf{V}_0\|_F + \sqrt{h}\,\|\mathbf{U}-\mathbf{U}_0\|_F \|\mathbf{V}\|_2\right)$
(6)	Theorem 2	$\tilde{\Theta}\left(\|\mathbf{U}_0\|_2 \|\mathbf{V}\|_F + \left\|\mathbf{U}-\mathbf{U}^0\right\|_F \|\mathbf{V}\|_F + \sqrt{h}\right)$

Table 1: Comparison with the existing generalization measures presented for the case of two layer ReLU networks with constant number of outputs and constant margin.

Comparison with other Capacity Bounds

#	Reference	Measure
(1)	Harvey et al. [9]	$\tilde{\Theta}(dh)$
(2)	Bartlett and Mendelson [3]	$\tilde{\Theta}\left(\left\Vert\mathbf{U}\right\Vert_{\infty,1}\left\Vert\mathbf{V}\right\Vert_{\infty,1}\right)$
(3)	Neyshabur et al. [20], Golowich et al. [7]	$\tilde{\Theta}\left(\left\Vert\mathbf{U}\right\Vert_{F}\left\Vert\mathbf{V}\right\Vert_{F}\right)$
(4)	Bartlett et al. [4], Golowich et al. [7]	$\tilde{\Theta}\left(\left\Vert\mathbf{U}\right\Vert_{2}\left\Vert\mathbf{V}-\mathbf{V}_0\right\Vert_{1,2}+\left\Vert\mathbf{U}-\mathbf{U}_0\right\Vert_{1,2}\left\Vert\mathbf{V}\right\Vert_{2}\right)$
(5)	Neyshabur et al. [23]	$\tilde{\Theta}\left(\left\Vert\mathbf{U}\right\Vert_{2}\left\Vert\mathbf{V}-\mathbf{V}_0\right\Vert_{F}+\sqrt{h}\left\Vert\mathbf{U}-\mathbf{U}_0\right\Vert_{F}\left\Vert\mathbf{V}\right\Vert_{2}\right)$
(6)	Theorem 2	$\tilde{\Theta}\left(\left\Vert\mathbf{U}_0\right\Vert_{2}\left\Vert\mathbf{V}\right\Vert_{F}+\left\Vert\mathbf{U}-\mathbf{U}^0\right\Vert_{F}\left\Vert\mathbf{V}\right\Vert_{F}+\sqrt{h}\right)$

Table 1: Comparison with the existing generalization measures presented for the case of two layer ReLU networks with constant number of outputs and constant margin.

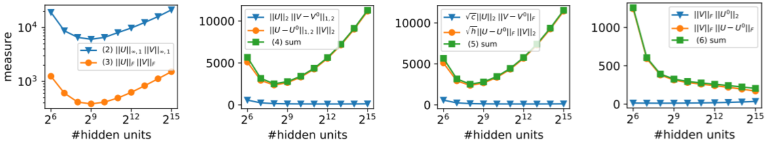
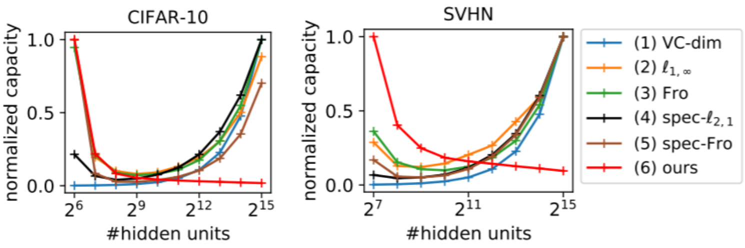


Figure 3: Behavior of terms presented in Table 1 with respect to the size of the network trained on CIFAR-10.

Comparison with other Capacity Bounds

#	Reference	Measure
(1)	Harvey et al. [9]	$\tilde{\Theta}(dh)$
(2)	Bartlett and Mendelson [3]	$\tilde{\Theta}\left(\|\mathbf{U}\|_{\infty,1}\|\mathbf{V}\|_{\infty,1}\right)$
(3)	Neyshabur et al. [20], Golowich et al. [7]	$\tilde{\Theta}\left(\|\mathbf{U}\|_F\|\mathbf{V}\|_F\right)$
(4)	Bartlett et al. [4], Golowich et al. [7]	$\tilde{\Theta}\left(\|\mathbf{U}\|_2\|\mathbf{V}-\mathbf{V}_0\|_{1,2}+\|\mathbf{U}-\mathbf{U}_0\|_{1,2}\|\mathbf{V}\|_2\right)$
(5)	Neyshabur et al. [23]	$\tilde{\Theta}\left(\|\mathbf{U}\|_2\|\mathbf{V}-\mathbf{V}_0\|_F+\sqrt{h}\|\mathbf{U}-\mathbf{U}_0\|_F\|\mathbf{V}\|_2\right)$
(6)	Theorem 2	$\tilde{\Theta}\left(\|\mathbf{U}_0\|_2\|\mathbf{V}\|_F+\left\|\mathbf{U}-\mathbf{U}^0\right\|_F\|\mathbf{V}\|_F+\sqrt{h}\right)$

Table 1: Comparison with the existing generalization measures presented for the case of two layer ReLU networks with constant number of outputs and constant margin.

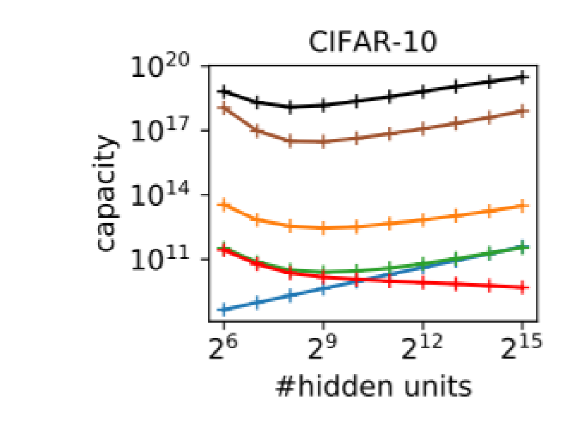


Lower Bound

Under a certain set of assumptions, the upper bound on the Rademacher complexity given is actually tight.

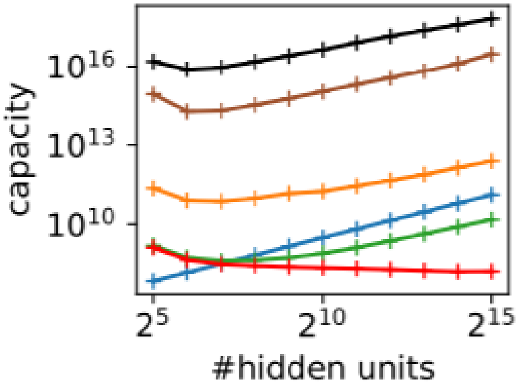
Future Work

Although this bound is the only one that decreases with the size of the network, it is still very loose, i.e. larger than the number of training examples.



Future Work

Although this bound is the only one that decreases with the size of the network, it is still very loose, i.e. larger than the number of training examples.



Future Work

- Get tighter bounds
- Extend those results for deeper networks
- Reduce the class size even more by choice of hyperparameters and optimization algorithms.

Any questions?

Thank you!

References

[1] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks. *CoRR*, abs/1805.12076, 2018. URL `http://arxiv.org/abs/1805.12076`.