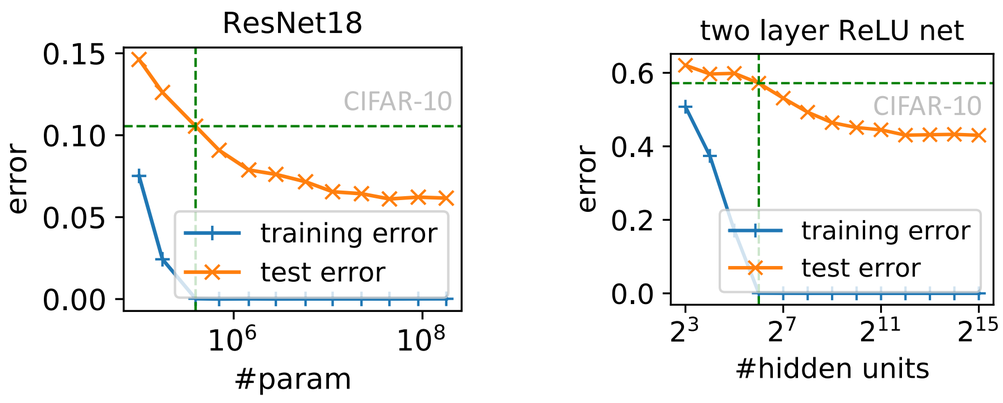
The Role of Over-Parametrization in Generalization of Neural Networks

Behnam Neyshabur NYU | **Zhiyuan Li** Princeton | **Srinadh Bhojanapalli** Google | **Yann LeCun** NYU | **Nathan Srebro** TTI-Chicago

Empirical observation:

Over-parametrization helps generalization



Current complexity measures ↑ with over-parametrization ☹

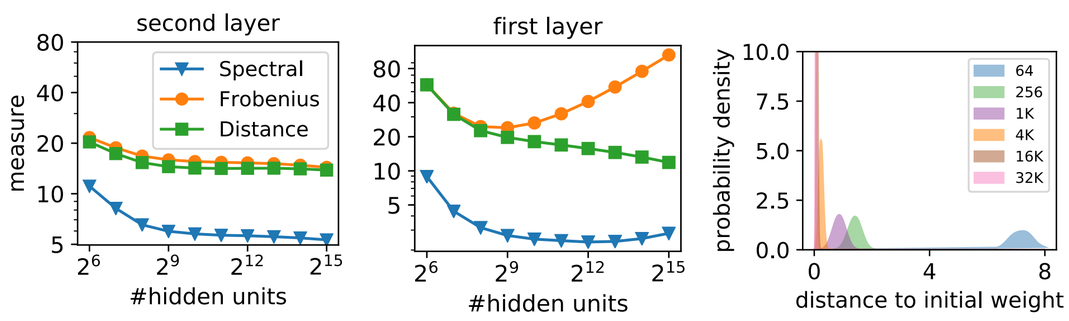
A generalization bound that ↓ with over-parametrization? ☺

Why is this important and useful?

- One of the key mysteries behind success of deep learning
- Understanding the real cause of generalization can be used to design better models and optimization algorithms

An Empirical Investigation

Properties of learned two layer ReLU networks:

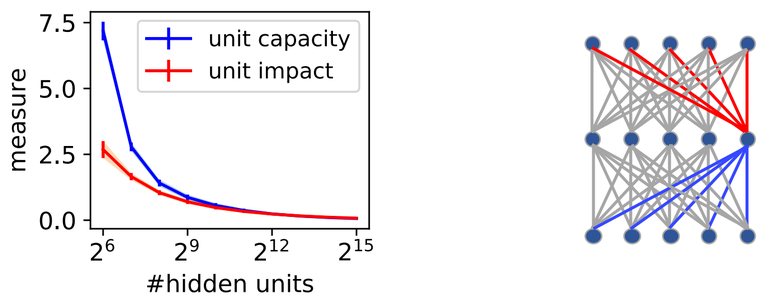


Distance: Euclidean distance from initialization.

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

$$\mathcal{W} = \left\{ (\mathbf{V},\mathbf{U}) \mid \mathbf{V} \in \mathbb{R}^{c\times h}, \mathbf{U} \in \mathbb{R}^{h\times d}, \|\mathbf{v}_i\| \le \alpha_i, \left\|\mathbf{u}_i - \mathbf{u}_i^0\right\|_2 \le \beta_i \right\}$$

unit impact ↑ unit capacity ↑



Generalization Bound

Rademacher complexity bound:

$$\mathcal{R}_{\mathcal{S}}(\ell_\gamma \circ \mathcal{F}_{\mathcal{W}}) \le \frac{2\sqrt{2c}+2}{\gamma m} \sum_{j=1}^{h} \alpha_j \left(\beta_j \|\mathbf{X}\|_F + \left\|\mathbf{u}_j^0 \mathbf{X}\right\|_2 \right)$$

Generalization bound:

For any m, w.p. > 1- δ over the training set:

$$L_0(f) \le \hat{L}_\gamma(f) + \tilde{O}\left(\frac{\sqrt{c}\|\mathbf{V}\|_F \left(\left\|\mathbf{U}-\mathbf{U}^0\right\|_F + \left\|\mathbf{U}^0\right\|_2\right) \sqrt{\frac{1}{m}\sum_{i=1}^m \|\mathbf{x}_i\|_2^2}}{\gamma\sqrt{m}} + \sqrt{\frac{h}{m}} \right).$$

c: #classes h: #hidden units d: input dim m: #samples γ: margin

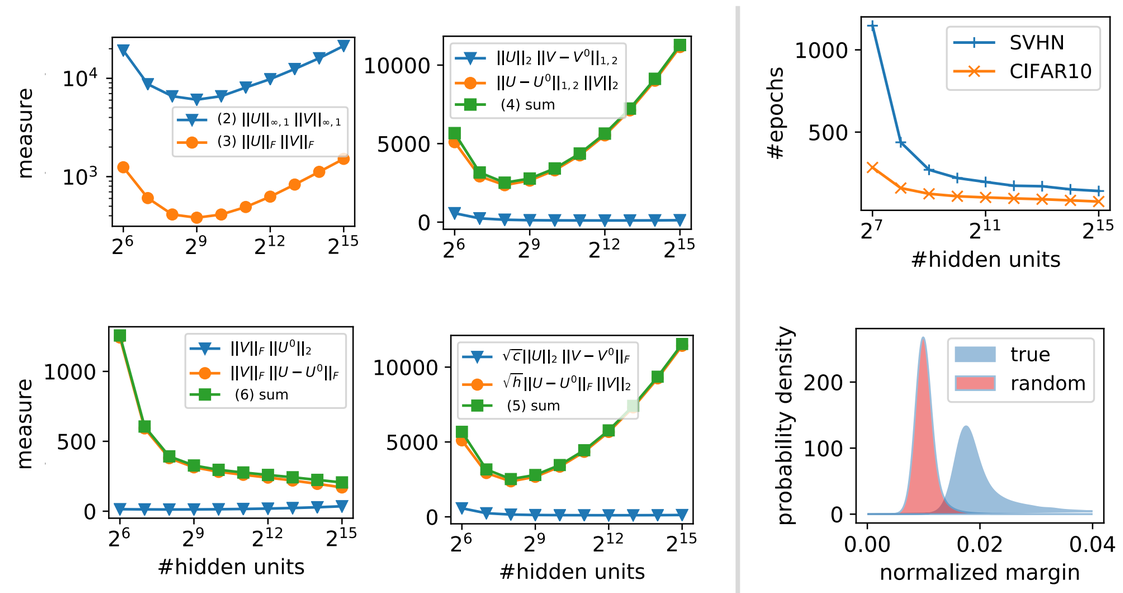
- U^0 is the initialization / any fixed matrix.
- Rademacher complexity: $\mathcal{R}_{\mathcal{S}}(\mathcal{H}) = \frac{1}{m} \mathop{\mathbb{E}}_{\xi\sim\{\pm1\}^m} \left[\sup_{f\in\mathcal{H}} \sum_{i=1}^m \xi_i f(x_i) \right]$
- Margin ramp loss: $\ell_\gamma(f(\mathbf{x}),y) = \begin{cases} 0 & \mu(f(\mathbf{x}),y) > \gamma \\ \mu(f(\mathbf{x}),y)/\gamma & \mu(f(\mathbf{x}),y) \in [0,\gamma] \\ 1 & \mu(f(\mathbf{x}),y) < 0. \end{cases}$

Evaluation

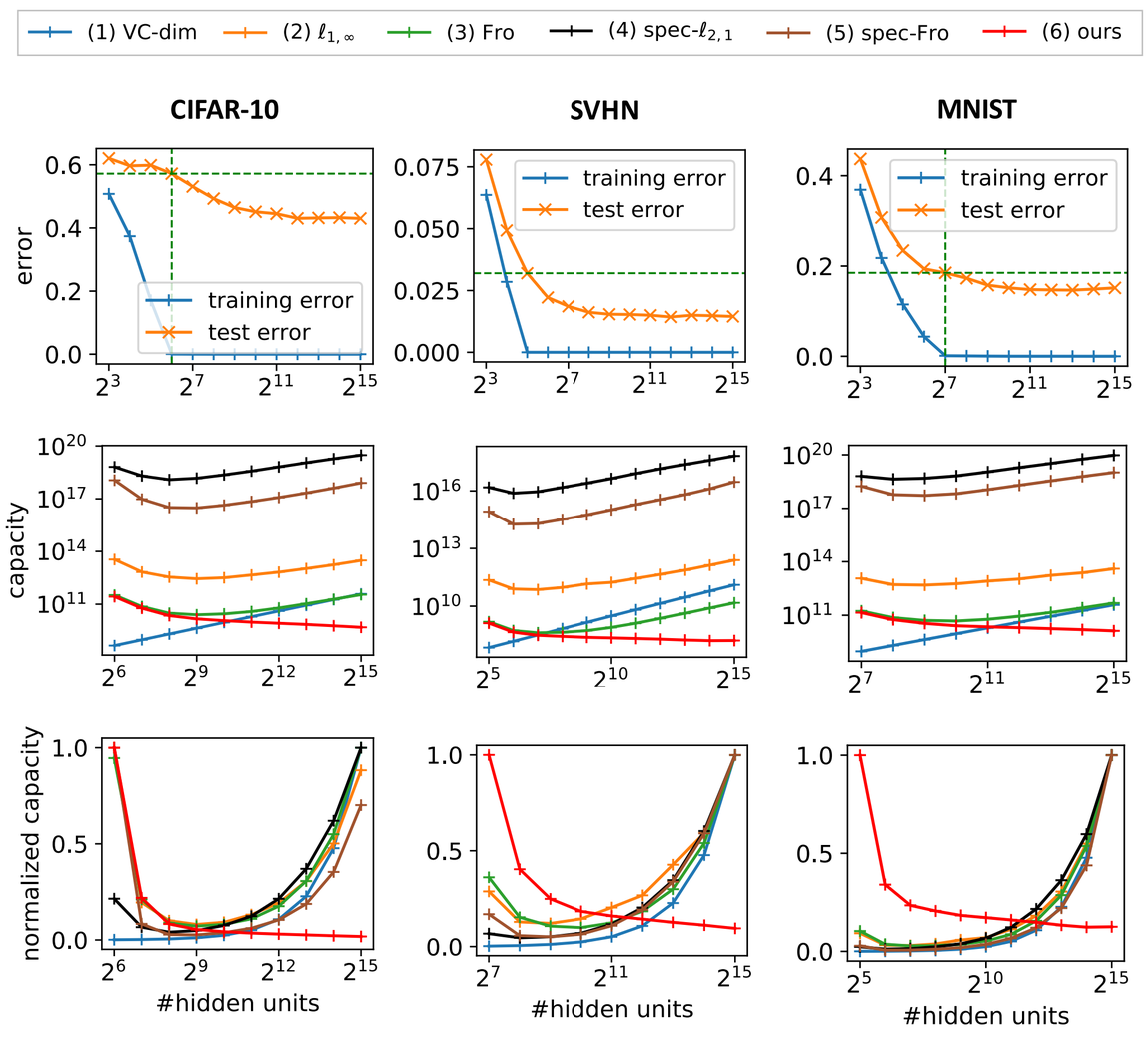
Comparing generalization measures:

#	Reference	Measure
(1)	Harvey et al. 2017	$\tilde{O}(dh)$
(2)	Bartlett and Mendelson 2002	$\tilde{O}\left(\|\mathbf{U}\|_{\infty,1} \|\mathbf{V}\|_{\infty,1}\right)$
(3)	Neyshabur et al. 2015	$\tilde{O}\left(\|\mathbf{U}\|_F \|\mathbf{V}\|_F\right)$
(4)	Bartlett et al. 2017	$\tilde{O}\left(\|\mathbf{U}\|_2 \|\mathbf{V}-\mathbf{V}_0\|_{1,2} + \|\mathbf{U}-\mathbf{U}_0\|_{1,2} \|\mathbf{V}\|_2\right)$
(5)	Neyshabur et al. 2018	$\tilde{O}\left(\|\mathbf{U}\|_2 \|\mathbf{V}-\mathbf{V}_0\|_F + \sqrt{h}\|\mathbf{U}-\mathbf{U}_0\|_F \|\mathbf{V}\|_2\right)$
(6)	ours	$\tilde{O}\left(\|\mathbf{U}_0\|_2 \|\mathbf{V}\|_F + \left\|\mathbf{U}-\mathbf{U}^0\right\|_F \|\mathbf{V}\|_F + \sqrt{h}\right)$

Behavior of different norms with over-parametrization:



Comparing capacity bounds:



Code available at https://github.com/bneyshabur/over-parametrization

Matching Lower Bound

Consider $\mathcal{W}' = \left\{ (\mathbf{V},\mathbf{U}) \mid \mathbf{V} \in \mathbb{R}^{1\times h}, \mathbf{U} \in \mathbb{R}^{h\times d}, \|\mathbf{v}_j\| \le \alpha_j, \left\|\mathbf{u}_j - \mathbf{u}_j^0\right\|_2 \le \beta_j, \|\mathbf{U}-\mathbf{U}^0\|_2 \le \max_{j\in h} \beta_j \right\}$

For any $d = h \le m, \{\alpha_j,\beta_j\}_{j=1}^h \subset \mathbb{R}^+, \mathbf{U}_0 \in \mathbb{R}^{h\times d}$ there exists $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^m \subset \mathbb{R}^d$ s.t.

$$\mathcal{R}_{\mathcal{S}}(\mathcal{F}_{\mathcal{W}}) \ge \mathcal{R}_{\mathcal{S}}(\mathcal{F}_{\mathcal{W}'}) = \Omega\left(\frac{\sum_{j=1}^h \alpha_j \|\mathbf{u}_j^0 \mathbf{X}\|_2}{m} \right).$$

For any $d = h \le m, \{\alpha_j,\beta_j\}_{j=1}^h \subset \mathbb{R}^+, \mathbf{U}_0 = 0$ there exists $\mathcal{S} = \{\mathbf{x}_i\}_{i=1}^m \subset \mathbb{R}^d$ s.t.

$$\mathcal{R}_{\mathcal{S}}(\mathcal{F}_{\mathcal{W}}) \ge \mathcal{R}_{\mathcal{S}}(\mathcal{F}_{\mathcal{W}'}) = \Omega\left(\frac{\sum_{j=1}^h \alpha_j \beta_j \|\mathbf{X}\|_F}{m} \right).$$

- ✓ Matches both terms in the upper bound.
- ✓ First lower bound that is higher than Lipschitz of the network.
- ✓ Improves by a factor of $\sqrt{h}$ over Bartlett et al. 2017.
- ✓ $\sqrt{h}$ gap between linear vs non-linear networks.