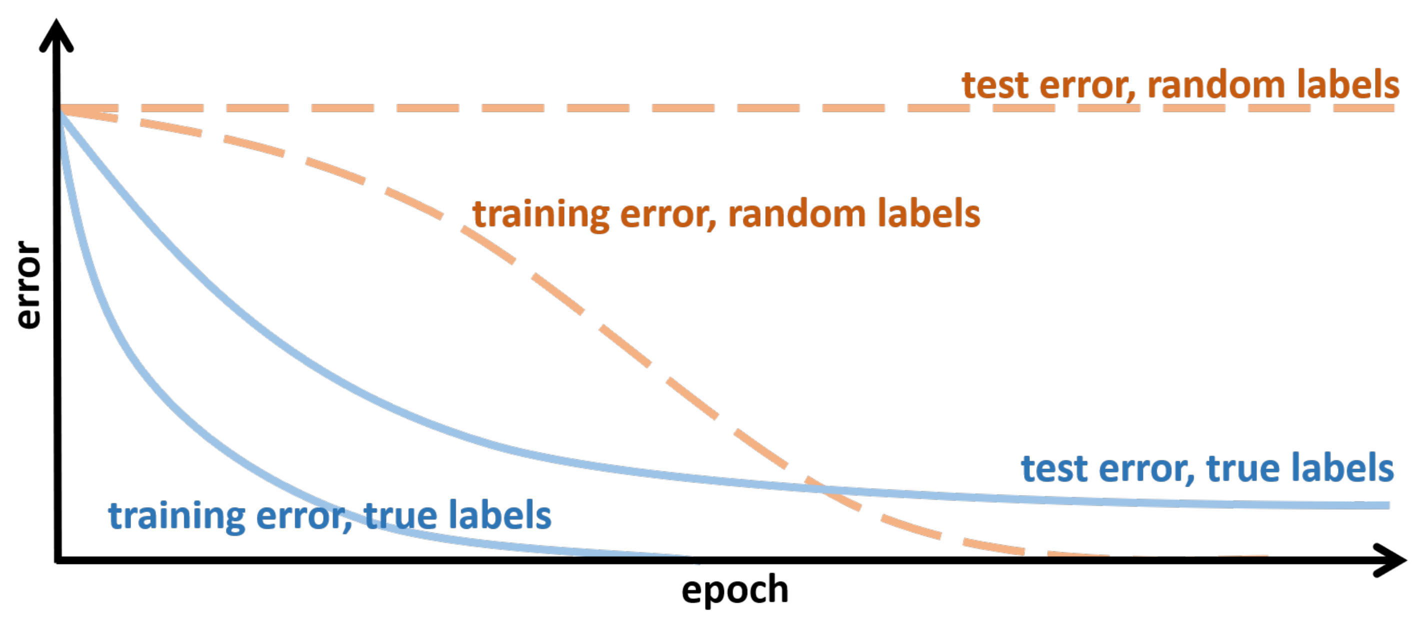
Generalization Bounds of Stochastic Gradient Descent for Wide and Deep Neural Networks

Yuan Cao and **Quanquan Gu**

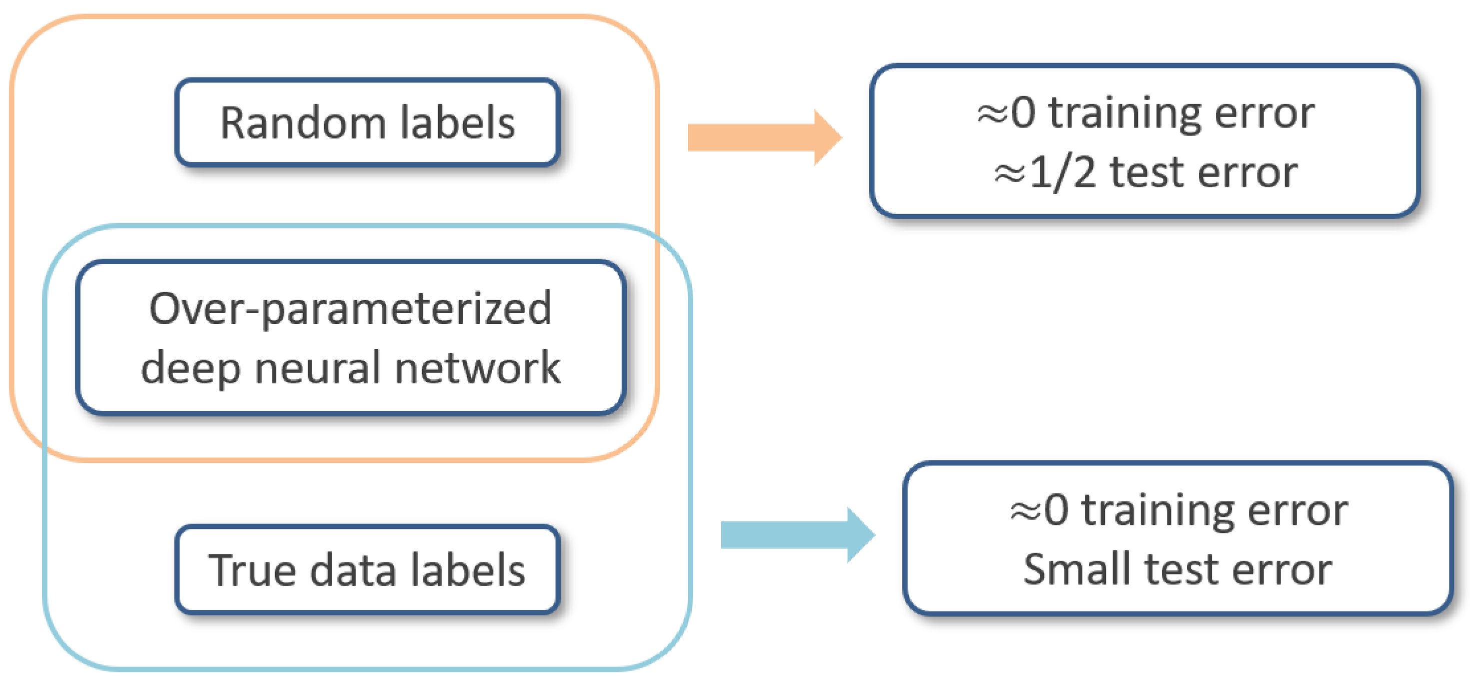
Department of Computer Science, University of California, Los Angeles

Over-parameterization in Deep Learning

- **An empirical observation** (Zhang etal. 2017; Bartlett et al. 2017; Neyshabur et al. 2018; Arora et al. 2019)



Questions We Aim to Answer



Why can extremely wide neural networks generalize?

What data can be learned by deep and wide neural networks?

Learning Over-parameterized DNNs

- Fully connected neural network with width m:

$$f_{\mathbf{W}}(\mathbf{x}) = \sqrt{m} \cdot \mathbf{W}_L \sigma(\mathbf{W}_{L-1} \cdots \sigma(\mathbf{W}_1 \mathbf{x}) \cdots)),$$

- $\sigma(\cdot)$ is the ReLU activation function: $\sigma(t) = \max(0, t)$.
- Suppose that $(\mathbf{x}, y) \sim \mathcal{D}$, and for simplicity, $\|\mathbf{x}\|_2 = 1$.
- $L_{(\mathbf{x}_i, y_i)}(\mathbf{W}) = \ell[y_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)]$, $\ell(z) = \log(1 + \exp(-z))$.

Algorithm SGD for DNNs starting at Gaussian initialization

$\mathbf{W}_l^{(0)} \sim N(0, 2/m)$, $l \in [L-1]$, $\mathbf{W}_L^{(0)} \sim N(0, 1/m)$
for $i = 1, 2, \ldots, n$ **do**
  Draw $(\mathbf{x}_i, y_i)$ from $\mathcal{D}$.
  Update $\mathbf{W}^{(i)} = \mathbf{W}^{(i-1)} - \eta \cdot \nabla_{\mathbf{W}} L_{(\mathbf{x}_i, y_i)}(\mathbf{W}^{(i-1)})$.
end for
Output: Randomly choose $\widehat{\mathbf{W}}$ uniformly from $\{\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(n-1)}\}$.

Generalization Bound, NTRF

For any $R > 0$, if $m \geq \widetilde{\Omega}(\mathrm{poly}(R, L, n))$, then with high probability, SGD returns $\widehat{\mathbf{W}}$ that satisfies

$$\mathbb{E}\big[L_{\mathcal{D}}^{0-1}(\widehat{\mathbf{W}})\big] \leq \inf_{f \in \mathcal{F}(R)} \left\{ \frac{4}{n} \sum_{i=1}^{n} \ell[y_i \cdot f(\mathbf{x}_i)] \right\} + \widetilde{\mathcal{O}}\left(\frac{LR}{\sqrt{n}}\right),$$

where $\mathcal{F}(\mathbf{W}^{(0)}, R)$ is the Neural Tangent Random Feature (NTRF) function class:

$$\mathcal{F}(R) = \big\{ f_{\mathbf{W}^{(0)}}(\cdot) + \langle \nabla f_{\mathbf{W}^{(0)}}(\cdot), \mathbf{W} \rangle : \|\mathbf{W}_l\|_F \leq R m^{-1/2} \big\}.$$

Test error of DNNs $\leq$ Training loss of NTRF $+\ \widetilde{\mathcal{O}}(n^{-1/2})$.

Generalization Bound. NTK

Let $\lambda_0 = \lambda_{\min}(\mathbf{\Theta}^{(L)})$. If $m \geq \widetilde{\Omega}(\mathrm{poly}(L, n, \lambda_0^{-1}))$, then with high probability, SGD returns $\widehat{\mathbf{W}}$ that satisfies

$$\mathbb{E}\big[L_{\mathcal{D}}^{0-1}(\widehat{\mathbf{W}})\big] \leq \widetilde{\mathcal{O}}\left[L \cdot \inf_{\widetilde{y}_i y_i \geq 1} \sqrt{\frac{\widetilde{\mathbf{y}}^\top (\mathbf{\Theta}^{(L)})^{-1} \widetilde{\mathbf{y}}}{n}} \right].$$

where $\mathbf{\Theta}^{(L)}$ is the neural tangent kernel (Jacot et al. 2018) Gram matrix:

$$\mathbf{\Theta}_{i,j}^{(L)} := \lim_{m \to \infty} m^{-1} \langle \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_j) \rangle.$$

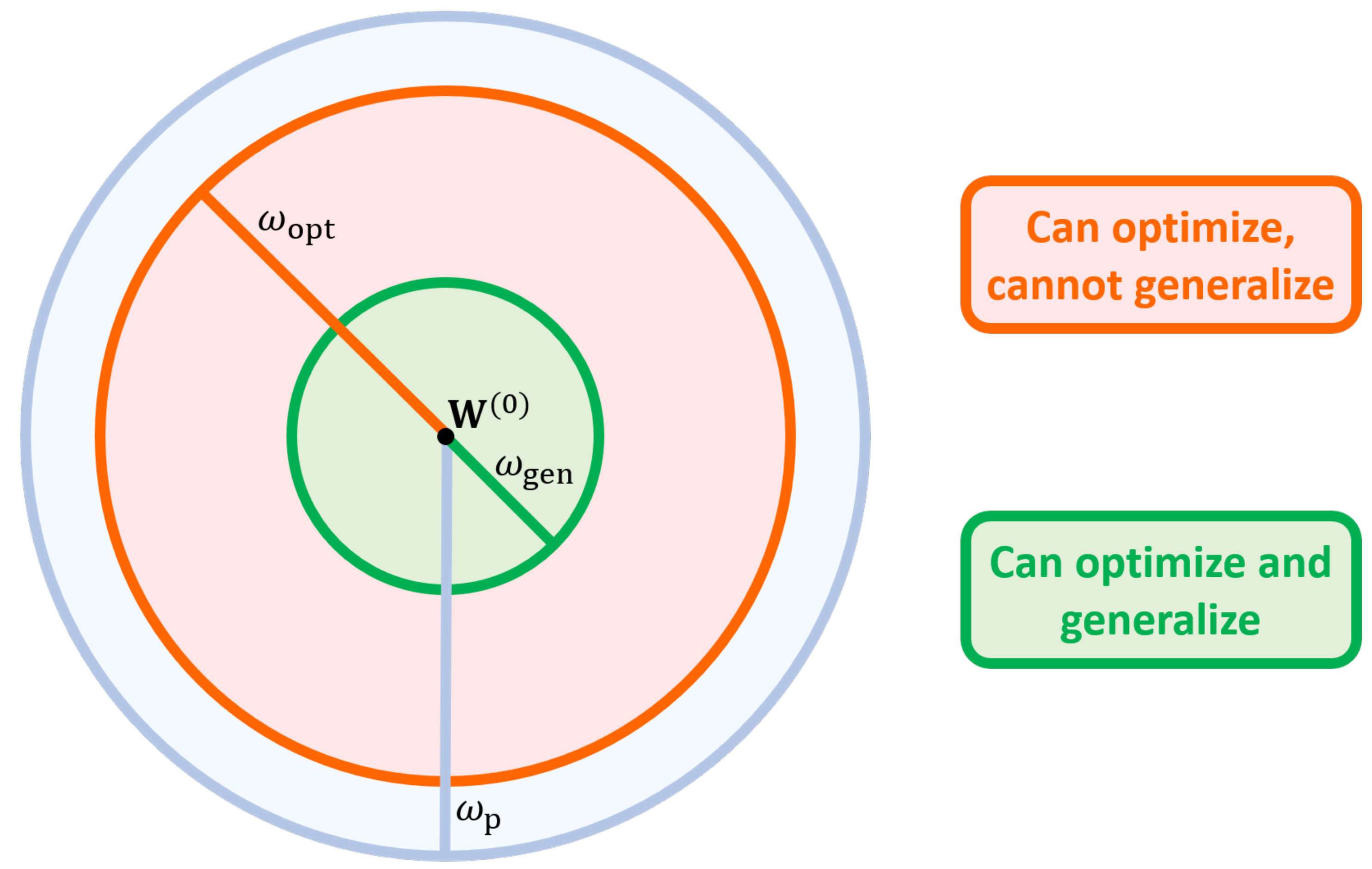
The "classifiability" of the underlying data distribution $\mathcal{D}$ can also be measured by the quantity $\inf_{\widetilde{y}_i y_i \geq 1} \sqrt{\widetilde{\mathbf{y}}^\top (\mathbf{\Theta}^{(L)})^{-1} \widetilde{\mathbf{y}}}$.

Discussion

- **Connection between the two bounds**
  - DNN competes with the best function in $\mathcal{F}(\widetilde{\mathcal{O}}(1))$.
  - $R = \inf_{\widetilde{y}_i y_i \geq 1} \sqrt{\widetilde{\mathbf{y}}^\top (\mathbf{\Theta}^{(L)})^{-1} \widetilde{\mathbf{y}}}$ guarantees small training loss of the NTRF function class.
- **Extremely wide neural networks can generalize**
  - The generalization bounds do not increase with m.
- **Quantification of the "classifiability" of $\mathcal{D}$**
  - For random labels, $\inf_{\widetilde{y}_i y_i \geq 1} \sqrt{\widetilde{\mathbf{y}}^\top (\mathbf{\Theta}^{(L)})^{-1} \widetilde{\mathbf{y}}} \gg \sqrt{n}$.
  - For "good" data, $\inf_{\widetilde{y}_i y_i \geq 1} \sqrt{\widetilde{\mathbf{y}}^\top (\mathbf{\Theta}^{(L)})^{-1} \widetilde{\mathbf{y}}} = \widetilde{\mathcal{O}}(1)$.
- **"Neural tangent kernel regime"**
  - $\sqrt{\widetilde{\mathbf{y}}^\top (\mathbf{\Theta}^{(L)})^{-1} \widetilde{\mathbf{y}}}$ is the NTK-induced RKHS norm of the kernel regression classifier on $\{(\mathbf{x}_i, \widetilde{y}_i), i \in [n]\}$.

Discussion Cont'd

- $\mathcal{B}(\mathbf{W}^{(0)}, \omega) := \{\mathbf{W} : \|\mathbf{W}_l - \mathbf{W}_l^{(0)}\|_F \leq \omega, l \in [L]\}$.



- For $\mathbf{W} \in \mathcal{B}(\mathbf{W}^{(0)}, \omega_{\mathrm{p}})$, $\omega_{\mathrm{p}} = \widetilde{\mathcal{O}}(1)$, neural networks enjoy good properties.
- As long as $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ when $y_i \neq y_j$, SGD converges with trajectory length $\omega_{\mathrm{opt}} \leq \widetilde{\mathcal{O}}(\mathrm{poly}(n) \cdot m^{-1/2})$.
- Under stronger data distribution assumptions, SGD converges with trajectory length $\omega_{\mathrm{gen}} \leq \widetilde{\mathcal{O}}(m^{-1/2})$

Key Ingredients for the Proof

- Deep ReLU networks are *almost linear* in terms of their parameters in a small neighbourhood around random initialization

$$f_{\mathbf{W}'}(\mathbf{x}_i) \approx f_{\mathbf{W}}(\mathbf{x}_i) + \langle \nabla f_{\mathbf{W}}(\mathbf{x}_i), \mathbf{W}' - \mathbf{W} \rangle.$$

- $L_{(\mathbf{x}_i, y_i)}(\mathbf{W})$ is *Lipschitz continuous* and *almost convex*

$$\|\nabla_{\mathbf{W}_l} L_{(\mathbf{x}_i, y_i)}(\mathbf{W})\|_F \leq O(\sqrt{m}),\ l \in [L],$$
$$L_{(\mathbf{x}_i, y_i)}(\mathbf{W}') \gtrsim L_{(\mathbf{x}_i, y_i)}(\mathbf{W}) + \langle \nabla_{\mathbf{W}} L_{(\mathbf{x}_i, y_i)}(\mathbf{W}), \mathbf{W}' - \mathbf{W} \rangle.$$

Optimization for Lipschitz and (almost) convex functions
$+$
Online-to-batch conversion

Applicable to general loss functions:
$\ell(\cdot)$ is convex/Lipschitz/smooth
$\Rightarrow L_{(\mathbf{x}_i, y_i)}(\mathbf{W})$ is (almost) convex/Lipschitz/smooth