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

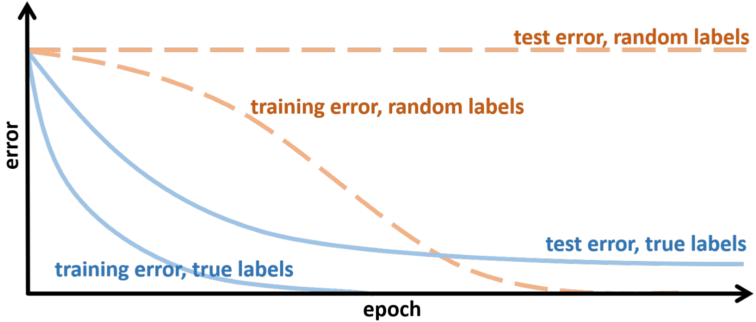
Yuan Cao and Quanquan Gu

Computer Science Department

UCLA

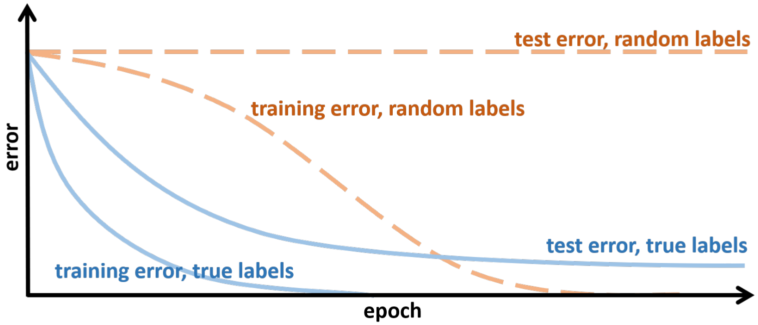
Learning Over-parameterized DNNs

Empirical observation on extremely wide deep neural networks (Zhang et al. 2017; Bartlett et al. 2017; Neyshabur et al. 2018; Arora et al. 2019)



Learning Over-parameterized DNNs

Empirical observation on extremely wide deep neural networks (Zhang et al. 2017; Bartlett et al. 2017; Neyshabur et al. 2018; Arora et al. 2019)



- 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: $\sigma(t) = \max(0, t)$.
- $L_{(\mathbf{x}_i, y_i)}(\mathbf{W}) = \ell[y_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)]$, $\ell(z) = \log(1 + \exp(-z))$.

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: $\sigma(t) = \max(0, t)$.
- $L_{(\mathbf{x}_i, y_i)}(\mathbf{W}) = \ell[y_i \cdot f_{\mathbf{W}}(\mathbf{x}_i)]$, $\ell(z) = \log(1 + \exp(-z))$.

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 Bounds for DNNs

Theorem

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

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

where

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

Generalization Bounds for DNNs

Theorem

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

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

where

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

Neural Tangent Random Feature (NTRF) model

Generalization Bounds for DNNs

Corollary

Let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ *and* $\lambda_0 = \lambda_{\min}(\boldsymbol{\Theta}^{(L)})$. *If* $m \geq \widetilde{\Omega}\big(\mathrm{poly}(L, n, \lambda_0^{-1})\big)$, *then with high probability, SGD returns* $\widehat{\mathbf{W}}$ *that satisfies*

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

where $\boldsymbol{\Theta}^{(L)}$ is the neural tangent kernel (Jacot et al. 2018) Gram matrix.
$\boldsymbol{\Theta}_{i,j}^{(L)} := \lim_{m \to \infty} m^{-1} \langle \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_j) \rangle$.

Generalization Bounds for DNNs

Corollary

Let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ *and* $\lambda_0 = \lambda_{\min}(\boldsymbol{\Theta}^{(L)})$. *If* $m \geq \widetilde{\Omega}\big(\mathrm{poly}(L, n, \lambda_0^{-1})\big)$, *then with high probability, SGD returns* $\widehat{\mathbf{W}}$ *that satisfies*

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

where $\boldsymbol{\Theta}^{(L)}$ is the neural tangent kernel (Jacot et al. 2018) Gram matrix.
$\boldsymbol{\Theta}_{i,j}^{(L)} := \lim_{m \to \infty} m^{-1} \langle \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_i), \nabla_{\mathbf{W}} f_{\mathbf{W}^{(0)}}(\mathbf{x}_j) \rangle$.

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

Overview of the Proof

Key observations

- 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*

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

Overview of the Proof

Key observations

- 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

Overview of the Proof

Key observations

- 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*

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

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

Summary

- Generalization bounds for wide DNNs that do not increase in network width.
- A random feature model (NTRF model) that naturally connects over-parameterized DNNs with NTK.
- A quantification of the "classifiability" of data: $\inf_{\widetilde{y}_i y_i \geq 1} \sqrt{\widetilde{\mathbf{y}}^\top (\mathbf{\Theta}^{(L)})^{-1} \widetilde{\mathbf{y}}}$.
- A clean and simple proof framework for neural networks in the "NTK regime" that is applicable to various problem settings.

Summary

- Generalization bounds for wide DNNs that do not increase in network width.
- A random feature model (NTRF model) that naturally connects over-parameterized DNNs with NTK.
- A quantification of the "classifiability" of data: $\inf_{\widetilde{y}_i y_i \geq 1} \sqrt{\widetilde{\mathbf{y}}^\top (\boldsymbol{\Theta}^{(L)})^{-1} \widetilde{\mathbf{y}}}$.
- A clean and simple proof framework for neural networks in the "NTK regime" that is applicable to various problem settings.

Thank you!

Poster #141