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

Learning Over-parameterized DNNs


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 $m$:
  \[ f_W(x) = \sqrt{m} \cdot W_L \sigma(W_{L-1} \cdots \sigma(W_1x) \cdots) \].

- $\sigma(\cdot)$ is the ReLU activation function: $\sigma(t) = \max(0, t)$.

- $L(x_i, y_i)(W) = \ell[y_i \cdot f_W(x_i)]$, $\ell(z) = \log(1 + \exp(-z))$. 
Learning Over-parameterized DNNs

- Fully connected neural network with width $m$:
  $$f_W(x) = \sqrt{m} \cdot W_L \sigma(W_{L-1} \cdots \sigma(W_1 x) \cdots)).$$

- $\sigma(\cdot)$ is the ReLU activation function: $\sigma(t) = \max(0, t)$.

- $L(x_i, y_i)(W) = \ell[y_i \cdot f_W(x_i)]$, $\ell(z) = \log(1 + \exp(-z))$.

**Algorithm** SGD for DNNs starting at Gaussian initialization

\[
\begin{align*}
W_l^{(0)} &\sim N(0, 2/m), \ l \in [L - 1], \ W_L^{(0)} \sim N(0, 1/m) \\
\text{for } i = 1, 2, \ldots, n \text{ do} \\
&\text{Draw } (x_i, y_i) \text{ from } D. \\
&\text{Update } W^{(i)} = W^{(i-1)} - \eta \cdot \nabla W L(x_i, y_i)(W^{(i-1)}) . \\
\text{end for} \\
\text{Output:} \text{ Randomly choose } \hat{W} \text{ uniformly from } \{W^{(0)}, \ldots, W^{(n-1)}\}.
\]
Generalization Bounds for DNNs

**Theorem**

For any $R > 0$, if $m \geq \tilde{\Omega}\left(\text{poly}(R, L, n)\right)$, then with high probability, SGD returns $\hat{W}$ that satisfies

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

where

$$
\mathcal{F}(W^{(0)}, R) = \{ f_{W^{(0)}}(\cdot) + \langle \nabla w f_{W^{(0)}}(\cdot), W \rangle : \|W_l\|_F \leq R \cdot m^{-1/2}, l \in [L] \}.
$$
Generalization Bounds for DNNs

Theorem

For any \( R > 0 \), if \( m \geq \tilde{\Omega}(\text{poly}(R, L, n)) \), then with high probability, SGD returns \( \hat{W} \) that satisfies

\[
\mathbb{E}[L_D^{-1}(\hat{W})] \leq \inf_{f \in \mathcal{F}(W^{(0)}, R)} \left\{ \frac{4}{n} \sum_{i=1}^{n} \ell[y_i \cdot f(x_i)] \right\} + O \left[ \frac{LR}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \right],
\]

where

\[
\mathcal{F}(W^{(0)}, R) = \left\{ f_{W^{(0)}}(\cdot) + \langle \nabla w f_{W^{(0)}}(\cdot), W \rangle : \|W_l\|_F \leq R \cdot m^{-1/2}, l \in [L] \right\}.
\]
Corollary

Let \( y = (y_1, \ldots, y_n)^\top \) and \( \lambda_0 = \lambda_{\min}(\Theta^{(L)}) \). If \( m \geq \tilde{\Omega}(\text{poly}(L, n, \lambda_0^{-1})) \), then with high probability, SGD returns \( \hat{W} \) that satisfies

\[
\mathbb{E}[L_D^{0^{-1}}(\hat{W})] \leq \tilde{O}\left[ L \cdot \inf_{\tilde{y} \in \mathcal{Y}} \sqrt{\frac{\tilde{y}^\top (\Theta^{(L)})^{-1} \tilde{y}}{n}} \right] + O\left[ \sqrt{\frac{\log(1/\delta)}{n}} \right].
\]

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

\( \Theta^{(L)}_{i,j} := \lim_{m \to \infty} m^{-1} \langle \nabla w f_{W(0)}(x_i), \nabla w f_{W(0)}(x_j) \rangle. \)
Corollary

Let \( \mathbf{y} = (y_1, \ldots, y_n)^\top \) and \( \lambda_0 = \lambda_{\min}(\Theta^{(L)}) \). If \( m \geq \tilde{\Omega}(\text{poly}(L, n, \lambda_0^{-1})) \), then with high probability, SGD returns \( \hat{\mathbf{W}} \) that satisfies

\[
\mathbb{E}[L_D^{0-1}(\hat{\mathbf{W}})] \leq \tilde{O} \left[ L \cdot \inf_{\tilde{y}_i \tilde{y}_i \geq 1} \sqrt{\tilde{y}^\top (\Theta^{(L)})^{-1} \tilde{y}} \right] + O \left[ \sqrt{\log(1/\delta) / n} \right].
\]

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

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

The “classifiability” of the underlying data distribution \( \mathcal{D} \) can also be measured by the quantity \( \inf_{\tilde{y}_i \tilde{y}_i \geq 1} \sqrt{\tilde{y}^\top (\Theta^{(L)})^{-1} \tilde{y}}. \)
Overview of the Proof

Key observations

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

\[ f_{W'}(x_i) \approx f_W(x_i) + \langle \nabla f_W(x_i), W' - W \rangle. \]

- \( L(x_i, y_i)(W) \) is *Lipschitz continuous* and *almost convex*

\[
\|\nabla_w L_{(x_i, y_i)}(W)\|_F \leq O(\sqrt{m}), \quad l \in [L],
\]

\[
L_{(x_i, y_i)}(W') \succeq L_{(x_i, y_i)}(W) + \langle \nabla_w L_{(x_i, y_i)}(W), W' - W \rangle.
\]
Overview of the Proof

Key observations

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

\[ f_{W'}(x_i) \approx f_W(x_i) + \langle \nabla f_W(x_i), W' - W \rangle. \]

▶ \( L(x_i, y_i)(W) \) is *Lipschitz continuous* and *almost convex*

\[
\| \nabla W_l L(x_i, y_i)(W) \|_F \leq O(\sqrt{m}), \; l \in [L],
\]

\[ L(x_i, y_i)(W') \gtrsim L(x_i, y_i)(W) + \langle \nabla W L(x_i, y_i)(W), W' - W \rangle. \]

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

\[ f_{W'}(x_i) \approx f_W(x_i) + \langle \nabla f_W(x_i), W' - W \rangle. \]

▶ \( L(x_i, y_i)(W) \) is Lipschitz continuous and almost convex

\[
\|\nabla_w L(x_i, y_i)(W)\|_F \leq O(\sqrt{m}), \ l \in [L],
\]

\[ L(x_i, y_i)(W') \gtrsim L(x_i, y_i)(W) + \langle \nabla_w L(x_i, y_i)(W), W' - W \rangle. \]

Applicable to general loss functions:

\( \ell(\cdot) \) is convex/Lipschitz/smooth

\[ \Rightarrow L(x_i, y_i)(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: \( \inf \tilde{y}_i \tilde{y}_i \geq 1 \sqrt{\tilde{y}^\top (\Theta(L))^{-1} \tilde{y}}. \)
- 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: \( \inf_{\tilde{y}_i, y_i \geq 1} \sqrt{\tilde{y}^\top (\Theta(L))^{-1} \tilde{y}} \).
▶ A clean and simple proof framework for neural networks in the “NTK regime” that is applicable to various problem settings.

Thank you!

Poster #141