Algorithm-Dependent Generalization Bounds for Overparameterized Deep Residual Networks

Spencer Frei and Yuan Cao and Quanquan Gu
Department of Statistics and Department of Computer Science, UCLA

Background

- Residual connections common in modern NNs: but theoretical justifications lacking.
- Fewer parameters, better generalization observed empirically in many residual architectures.

Problem Description

- Input \((x, y) \in \mathbb{R}^d \times \{-1, 1\}\), binary classification under cross-entropy loss \(\ell(z) := \log(1 + \exp(-z))\).
- \(f_W(x)\) is output of \(L + 1\) hidden layer residual network, \(x_{L+1} = W_{L+1}^T x_L\), \(f_W(x) = \sigma \circ (\sigma \circ \cdots \circ \sigma \circ W_1^T x)\).
- Layer weights \(W_l \in \mathbb{R}^{m_l \times m_l}\) trained by G.D.:
  \[
  W_l^{(t+1)} = W_l^{(t)} - \eta \nabla W_l L_S(W_1, \ldots, W_L),
  L_S(W) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_W(x_i)),
  E_S(W) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_W(x_i)) = \text{surrogate error}.
  \]

Assumptions

- Gaussian initialization: \(W_1^{(0)} \sim_{iid} \mathcal{N}(0, 2/m_l)\).
- Separability by random feature model: \(\mathbb{P}_\mathcal{D}(f(x) = \mathbb{E}_x \mathbb{E}_{u \sim \mathcal{N}(0,1)}[\sigma(u^T x)] \parallel \|\sigma(\cdot)\|_{\infty} \leq 1\), such that \(y \cdot f(x) \geq \gamma > 0\) for all \((x, y) \in \mathcal{D}\).
- Normalized input data: \(\|x\|_2 = 1 \forall x\).
- Widths of same order: \(m_{L+1} = \Theta(m_L)\), denote smallest layer width as \(m = m_L \wedge m_{L+1}\).
- Residual scaling: \(\theta = 1/\Theta(L)\).

Main Theorems

- Weights stay close to init. and bound for surrogate error: denote \(\tau\)-neighborhood of init. by \(\mathcal{W}(\tau) := \{W_1, \ldots, W_{L+1} : \|W_l - W_l^{(0)}\|_F \leq \tau \forall l\}\).
  Let \(\tilde{O}\) hide logarithmic terms in \(L, n, \delta^{-1}\).
- If \(\tau = \tilde{O}(\gamma^{-1})\), \(\eta = \tilde{O}(\gamma^{-1} \epsilon - \gamma^{-1} m^{-1})\), and \(K_\eta = \tilde{O}(\tau \gamma^{-1})\), then provided \(m \geq \tilde{O}(\tau \gamma^{-1} \delta^{-2})\), w.h.p.
  \begin{itemize}
  \item [(i)] \(\hat{W}(k) \in \mathcal{W}(\tau)\) for all \(k \in [K]\).
  \item [(ii)] There exists \(k \in \{0, \ldots, K - 1\}\) with \(E_S(W(k)) \leq C \cdot m^{-2} \cdot (K_\eta)^2 (\log \frac{2}{\delta})^4 \gamma^{-2}\).
  \end{itemize}
- Bound for Rademacher complexity in \(\mathcal{W}(\tau)\): for \(m \geq \tilde{O}(\tau^{-d} \delta^{-1})\) and \(f_{\mathcal{W}(\tau)} := \{f_W : W \in \mathcal{W}(\tau)\}\),
  \[
  \mathbb{R}_m(f_{\mathcal{W}(\tau)}) \leq C_2 \left(\tau \sqrt{m \log m + \frac{\tau \gamma^2}{\sqrt{n}}} \right).
  \]
- Bound for test error: for \(m \geq m_{\text{res}}^*\), \(K^* \leq K_{\text{res}}, \text{s.t.} \mathbb{E}[\mathbb{I}(y \neq \text{sign} f_{\mathcal{W}(\tau)}(x))] \leq 2 \mathbb{E}[\mathbb{I}(y \neq f_{\mathcal{W}(\tau)}(x))] \leq \epsilon \).

Comparison with Non-Residual Results

- Above results at most log dependence on \(L\).
- Width and sample requirement is reduced: \(m_m^* > \text{poly}(L)m_m^*, n_m > \text{poly}(L)n_m\).
- Step size and iterations required are better:
  \[
  \eta_{\text{res}} < \text{poly}(L^{-1})\eta_{\text{nonres}}, K_{\text{res}} > \text{poly}(L)K_{\text{nonres}}.
  \]
- Distance from initialization is key to above bounds:
  \[
  \tau_{\text{res}} > \text{poly}(L)\tau_{\text{nonres}}, \tau_{\text{res}} = \tilde{O}(\gamma^{-d} \epsilon^{-d - 1/2}).
  \]

Key Ingredients for Proof

- Backpropagation and forward propagations are bounded independent of depth: if \(x_1\) represents layers from input \(x\) to \(l\)-th layer, and \(b_l(x)\) represents layers \(l\) to final layer, then \(\|x_1\|_2 \leq C, \|b_l(x)\|_2 \leq C\).
- Network output is almost linear in \(\mathcal{W}(\tau)\): for \(m\) large and \(W, \tilde{W} \in \mathcal{W}(\tau)\),
  \[
  f_{\tilde{W}}(x) \approx f_W(x) + \left(\tilde{W} - W, \nabla_W f_W(x)\right).
  \]
- Loss is Lipschitz and almost convex in \(\mathcal{W}(\tau)\): for \(m\) large and \(W, \tilde{W} \in \mathcal{W}(\tau)\),
  \[
  \|\nabla_W L_S(W)\|_F \leq C\tau^{1/2} \sqrt{m} \leq \sqrt{m},
  L_S(W) - L_S(\tilde{W}) \geq (\tilde{W} - W, \nabla_W L_S(\tilde{W})).
  \]
- Width at most logarthmic in \(L\) is required for above approximations, rather than usual poly/exponential.

How Does Residual Architecture Help?

- Skip connections and scaling factor \(\theta\) prevents blowup by forcing Lipschitz constant of network output to be bounded independent of depth:
  \[
  ||x_i||_2 = ||I + \Theta_\Sigma W_i^T x_{i-1}||_2 \leq (1 + C\theta) ||x_{i-1}||_2 \leq \cdots \leq (1 + C\theta)^L ||x_0||_2.
  \]
- Representations from earlier layers are not lost in forward propagation, allowing separability of R.F. model in first layer to persist through all layers. If \(\alpha\) can separate at margin \(\gamma\) in first layer, then \(y \cdot \langle \alpha, x_i \rangle = \langle \alpha, x_i \rangle + \theta \sum_{l=2}^L \langle \alpha, \sigma(W_l^T x_{l-1}) \rangle \geq \gamma\).