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

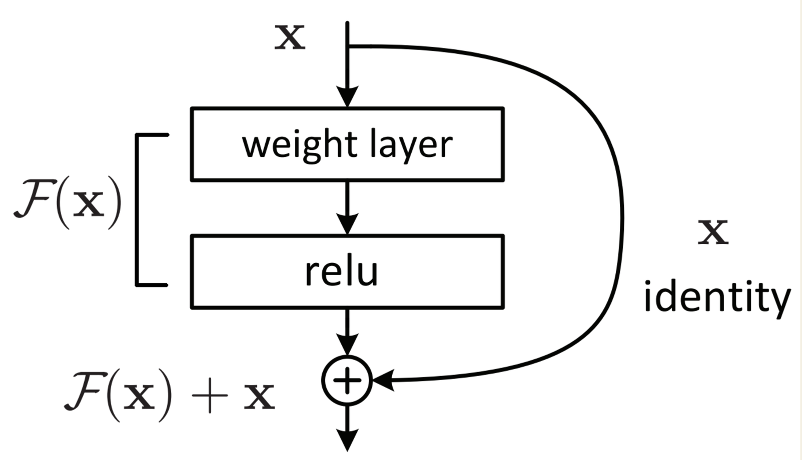
Background

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

Problem Description

- Input $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$, binary classification under cross-entropy loss $\ell(z) := \log(1 + \exp(-z))$.
- $f_W(x)$ = output of $L+1$ hidden layer residual network,

$$x_1 = W_1^\top x,$$
$$x_l = x_{l-1} + \theta\sigma(W_l^\top x_{l-1}),$$
$$l = 2, \ldots, L,$$
$$x_{L+1} = W_{L+1}^\top x_L,$$
$$f_W(x) = v^\top x_{L+1},$$
$$\sigma = \mathrm{ReLU}.$$



- Layer weights $W_l \in \mathbb{R}^{m_{l-1} \times m_l}$ trained by G.D.:

$$W_l^{(t+1)} = W_l^{(t)} - \eta \nabla_{W_l} L_S(W_1, \ldots, W_{L+1}),$$
$$L_S(W) := L_S(W_1, \ldots, W_{L+1}) = \frac{1}{n}\sum_{i=1}^n \ell(y_i \cdot f_W(x_i)),$$
$$\mathcal{E}_S(W) := \frac{1}{n}\sum_{i=1}^n -\ell'(y_i \cdot f_W(x_i)) = \text{surrogate error}.$$

Assumptions

- **Gaussian initialization**: $\left[W_l^{(0)}\right]_{i,j} \overset{\text{i.i.d.}}{\sim} N(0, 2/m_l)$.
- **Separability by random feature model**: there exists
$$f(x) = \mathbb{E}_{u\sim N(0,1)}[c(u)\sigma(u^\top x)], \quad \|c(\cdot)\|_\infty \le 1,$$
such that $y \cdot f(x) \ge \gamma > 0$ for all $(x, y) \in \operatorname{supp}\mathcal{D}$.
- **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**: $\theta = 1/\Omega(L)$.

Main Theorems

- **Weights stay close to init.** and **bound for surrogate error**: denote τ-neighborhood of init. by
$$\mathcal{W}(\tau) := \left\{(W_1, \ldots, W_{L+1}) : \left\|W_l - W_l^{(0)}\right\|_F \le \tau \ \forall l\right\}.$$
Let $\widetilde{O}$ hide logarithmic terms in L, n, δ^{-1}.
If $\tau = \widetilde{O}(\gamma^{12})$, $\eta = \widetilde{O}(\tau m^{-\frac{1}{2}} \wedge \gamma^4 m^{-1})$, and $K\eta = \widetilde{O}(\tau^2\gamma^4)$, then provided $m \ge \widetilde{O}(\tau^{-\frac{4}{3}} d\gamma^{-2})$, w.h.p.

> (i) $W^{(k)} \in \mathcal{W}(\tau)$ for all $k \in [K]$.
> (ii) There exists $k \in \{0, \ldots, K-1\}$ with
> $\mathcal{E}_S(W^{(k)}) \le C \cdot m^{-\frac{1}{2}} \cdot (K\eta)^{-\frac{1}{2}} \left(\log \frac{n}{\delta}\right)^{\frac{1}{4}} \cdot \gamma^{-2}$.

- **Bound for Rademacher complexity in $\mathcal{W}(\tau)$:** for $m \ge \widetilde{O}(\tau^{-\frac{4}{3}} d)$ and $f_{\mathcal{W}(\tau)} := \{f_W : W \in \mathcal{W}(\tau)\}$,
$$\mathfrak{R}_n\left(f_{\mathcal{W}(\tau)}\right) \le C_2\left(\tau^{\frac{4}{3}}\sqrt{m\log m} + \frac{\tau\sqrt{m}}{\sqrt{n}}\right).$$

- **Bound for test error:** for $m \ge m^*_{\mathrm{res}}$,
$$m^*_{\mathrm{res}} = \widetilde{O}(\mathrm{poly}(\gamma^{-1})) \cdot \max(d, \varepsilon^{-14}),$$
$$n_{\mathrm{res}} = \widetilde{O}(\mathrm{poly}(\gamma^{-1})) \cdot \varepsilon^{-4},$$
$$\eta_{\mathrm{res}} = O(\gamma^4 \cdot m^{-1}), \quad K_{\mathrm{res}} = \widetilde{O}(\mathrm{poly}(\gamma^{-1})) \cdot \varepsilon^{-2},$$
G.D. with step size η_{res} finds $W^{(k^*)}$, $k^* \le K_{\mathrm{res}}$, s.t.
$$\mathbb{E}[\mathbb{1}(y \ne \operatorname{sign}(f_{W^{(k^*)}}(x)))] \le 2\mathbb{E}[-\ell'(y f_{W^{(k^*)}}(x)] \le \varepsilon.$$

Comparison with Non-Residual Results

- **Above results at most log dependence on L.**
- **Width and sample requirement is reduced:**
$$m^*_{\mathrm{nonres}} > \mathrm{poly}(L) m^*_{\mathrm{res}}, \quad n_{\mathrm{nonres}} > \mathrm{poly}(L) n_{\mathrm{res}}.$$
- **Step size and iterations required are better**:
$$\eta_{\mathrm{nonres}} < \mathrm{poly}(L^{-1}))\eta_{\mathrm{res}}, \quad K_{\mathrm{nonres}} > \mathrm{poly}(L) K_{\mathrm{res}}.$$
- **Distance from initialization** is key to above bounds.
$$\tau_{\mathrm{nonres}} > \mathrm{poly}(L)\tau_{\mathrm{res}}, \quad \tau_{\mathrm{res}} = \widetilde{O}(\gamma^{-4}\varepsilon^{-1} m^{-1/2}).$$

Key Ingredients for Proof

- **Backpropagation and forward propagations are bounded independent of depth**: if x_l represents layers from input x to l-th layer, and $b_l(x)$ represents layers l to final layer, then
$$\|x_l\|_2 \le C, \quad \|b_l(x)\|_2 \le C.$$
- **Network output is almost linear in $\mathcal{W}(\tau)$**: for m large and $\widehat{W}, \widetilde{W} \in \mathcal{W}(\tau)$,
$$f_{\widehat{W}}(x) \approx f_{\widetilde{W}}(x) + \left\langle \widehat{W} - \widetilde{W}, \nabla_W f_{\widetilde{W}}(x) \right\rangle.$$
- **Loss is Lipschitz and almost convex in $\mathcal{W}(\tau)$**: for m large and $\widehat{W}, \widetilde{W} \in \mathcal{W}(\tau)$,
$$\left\|\nabla_{W_l} L_S(\widehat{W})\right\|_F \le C\theta^{\mathbb{1}(2\le l\le L)}\sqrt{m},$$
$$L_S(\widehat{W}) - L_S(\widetilde{W}) \gtrsim \left\langle \widehat{W} - \widetilde{W}, \nabla_W L_S(\widetilde{W}) \right\rangle.$$
- **Width at most logarithmic in L** is required for above approximations, rather than usual poly/exponential.

How Does Residual Architecture Help?

- **Skip connections and scaling factor θ prevents blowup** by forcing Lipschitz constant of network output to be bounded independent of depth:
$$\|x_l\|_2 = \left\|(I + \theta\Sigma_l W_l^\top x_{l-1})\right\|_2 \le (1 + C\theta)\|x_{l-1}\|_2$$
$$\le \cdots \le (1 + C\theta)^L \|x\|_2.$$
- **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 α can separate at margin γ in first layer, then
$$y \cdot \langle \alpha, x_l \rangle = \langle \alpha, x_1 \rangle + \theta\sum_{l'=2}^{l} \langle \alpha, \sigma(W_{l'}^\top x_{l'-1}) \rangle \gtrsim \gamma.$$