Q. What are the optimization/generalization properties of  single ResNet(consts of residual blocks) vs than fully-connected networks [1], but theory is elusive.

Q. Can we extend this result to multi-block ResNets?

- Adding parallel shortcut networks can remove bad local min [3, 4].
- Adding skip-connections from hidden nodes to output removes bad local valleys [5].
- However, these results consider direct skip-connections to output.

Q. Can we also show that a chain of skip-connections improves the loss landscape?

- Near-identity regions of linear ResNets have good optimization landscape and expressive power [6].
- Extension to nonlinear function space is possible [7].
- Initialization near-identity regions leads to stable training and good generalization [8].

Q. What are the optimization/generalization properties of near-identity regions?

Benign Landscape of Deep ResNets

Given input $x \in \mathbb{R}^d$, consider the following ResNet:

$h_1(x) = x + V_1 \phi_1(x)$
$h_l(x) = h_{l-1}(x) + V_l \phi_l(U_l h_{l-1}(x)), \ l = 2, \ldots, L$
$\theta_l(x) = w^\top h_l(x)$

- $V_l \in \mathbb{R}^{d \times d}, U_l \in \mathbb{R}^{m \times d}, w \in \mathbb{R}^d$ are parameters
- $\phi_l : \mathbb{R}^m \rightarrow \mathbb{R}^d$ is any feed-forward network parametrized by $z$
- $\theta$ is the collection of all $U_l$, $V_l$, $z$, $w$

For loss $\ell(p; y)$ twice differentiable and convex in $p$, and data distribution $\mathcal{P}$,

$$\mathcal{R}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{P}}[\ell(h_l(x); y)], \ \mathcal{R}_{\text{lin}} = \inf_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{P}}[\ell(t^\top x; y)].$$

Theorem. Let $\theta^*$ be any twice-differentiable critical point of $\mathcal{R}(\cdot)$. If

- $\mathbb{E}_{(x,y) \sim \mathcal{P}}[\ell''(h_{l-1}(x); y)h_l(x) h_l(x)]$ is full rank; and
- $\text{coll}((U_1^T, \ldots, U_L^T)) \neq \mathbb{R}^d$,

Then, at least one of the following holds:

$$\mathcal{R}(\theta^*) \leq \mathcal{R}_{\text{lin}}, \text{or } \lambda_{\text{min}}(\nabla^2 \mathcal{R}(\theta^*)) < 0.$$

Near-identity Regions of ResNets

- Consider a ResNet with residual blocks:

$$h_1(x) = h_{l-1}(x) + \phi_l(h_{l-1}(x)), \ l = 1, \ldots, L.$$
- $\phi_l$ is any $O(1/L)$-Lipschitz function and $\phi^*_l(0) = 0$.

Theorem (informal). Assume the loss $\ell(p; y)$ is Lipschitz, convex, and differentiable in $p$. For any critical point $\theta^*$ of $\mathcal{R}(\cdot)$,

$$\mathcal{R}(\theta^*) \leq \mathcal{R}_{\text{lin}} + C.$$

- Consider a ResNet with residual blocks:

$$h(x) = h_{l-1}(x) + V_l \text{ReLU}(U_l h_{l-1}(x)), \ l = 1, \ldots, L.$$

Theorem (informal). Given any dataset $\mathcal{S} = \{x_i\}_{i=1}^n$, define a class

$$\mathcal{F}_L = \{f : \mathbb{R}^d \rightarrow \mathbb{R} \mid \|w\| \leq 1, \|V_l\| \leq 1/\sqrt{L}\}.$$

Then, the empirical Radamacher complexity satisfies

$$\mathcal{R}(\mathcal{F}_L) \leq e^2 \max_{\|w\| \leq 1} \|w\| \sqrt{n},$$

Both bounds are independent of depth $L$, which is difficult to achieve (e.g., $\mathcal{R}$ of fully-connected nets typically grows with $L$)

Conclusion

- Under geometric conditions, any critical point of the risk function of a deep ResNet is either 1) better than linear predictors or 2) the Hessian at the critical point is a strictly negative eigenvalue.