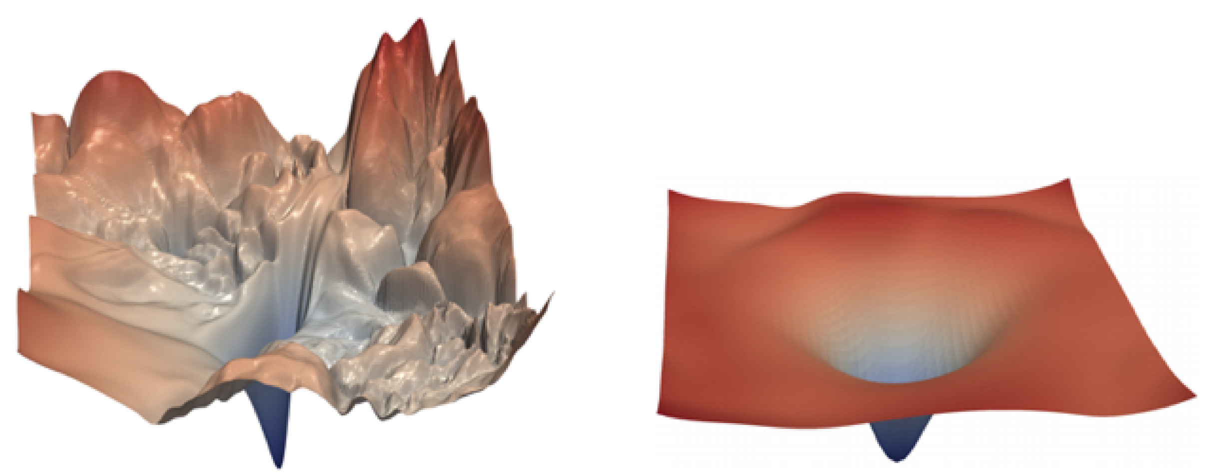
Are deep ResNets provably better than linear predictors?

Chulhee Yun (chulheey@mit.edu), Suvrit Sra, and Ali Jadbabaie

Laboratory for Information and Decision Systems, Massachusetts Institute of Technology

Introduction & Questions

- Residual nets (ResNets) consist of **residual blocks** $x \mapsto x + \Phi(x)$.
- Risks of deep ResNets are known to have more benign landscapes than fully-connected networks [1], but theory remains elusive.



(a) without skip connections (b) with skip connections

- Any local minimum of a **single**-block ResNet $x \mapsto \boldsymbol{w}^T(x + \boldsymbol{V}\phi(x))$ has risk value at least as good as linear predictors [2].

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 at 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 $x \in \mathbb{R}^{d_x}$, consider the following ResNet:

$$\begin{aligned} h_1(x) &= x + \boldsymbol{V}_1 \phi^1_{\boldsymbol{z}}(x) \\ h_l(x) &= h_{l-1}(x) + \boldsymbol{V}_l \phi^l_{\boldsymbol{z}}(\boldsymbol{U}_l h_{l-1}(x)),\ l = 2, \dots, L, \\ f_{\boldsymbol{\theta}}(x) &= \boldsymbol{w}^T h_L(x). \end{aligned}$$

- $\boldsymbol{V}_l \in \mathbb{R}^{d_x \times n_l}$, $\boldsymbol{U}_l \in \mathbb{R}^{m_l \times d_x}$, $\boldsymbol{w} \in \mathbb{R}^{d_x}$ are parameters
- $\phi^l_{\boldsymbol{z}} : \mathbb{R}^{m_l} \to \mathbb{R}^{n_l}$ is any feed-forward network parametrized by $\boldsymbol{z}$
- $\boldsymbol{\theta}$ is the collection of all $\boldsymbol{U}_l$, $\boldsymbol{V}_l$, $\boldsymbol{z}$, $\boldsymbol{w}$
- For loss $\ell(p; y)$ twice differentiable and convex in p, and data distribution $\mathcal{P}$,

$$\mathfrak{R}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f_{\boldsymbol{\theta}}(x); y)],\ \mathfrak{R}_{\text{lin}} = \inf_t \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(t^T x; y)].$$

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

- $\mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell''(f_{\boldsymbol{\theta}^*}(x); y) h_L(x) h_L(x)^T]$ is full rank; and
- $\operatorname{col}\left(\left[(\boldsymbol{U}_2^*)^T \cdots (\boldsymbol{U}_L^*)^T\right]\right) \neq \mathbb{R}^{d_x}$,

Then, at least one of the following holds:

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

- Under geometric conditions, a critical point of multi-block ResNet is **better than linear predictors** or is **a strict saddle point**.
- If $L = 1$, any critical point with $\boldsymbol{w}^* \neq 0$ satisfies $\mathfrak{R}(\boldsymbol{\theta}^*) \le \mathfrak{R}_{\text{lin}}$, recovering [2] in the same setting.
- A **chain of multiple skip-connections** (as opposed to direct) can improve the loss landscape.
- 1st condition requires representation of h_L to cover the full space.
- 2nd condition requires row space of $\boldsymbol{U}_l$'s not to cover the full space, giving room for improvement. Always satisfied if $\sum_{l=2}^L m_l < d_x$.
- Removal of these condition is left for future work.

Near-identity Regions of ResNets

- Consider a ResNet with residual blocks:

$$h_l(x) = h_{l-1}(x) + \phi^l_{\boldsymbol{z}}(h_{l-1}(x)),\ l = 1, \dots, L.$$

- $\phi^l_{\boldsymbol{z}}$ is any $O(1/L)$-Lipschitz function and $\phi^l_{\boldsymbol{z}}(0) = 0$.

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

$$\mathfrak{R}(\boldsymbol{\theta}^*) \le \mathfrak{R}_{\text{lin}} + C.$$

- Consider a ResNet with residual blocks:

$$h_l(x) = h_{l-1}(x) + \boldsymbol{V}_l \mathrm{ReLU}(\boldsymbol{U}_l h_{l-1}(x)),\ l = 1, \dots, L.$$

Theorem (informal). Given any dataset $S = \{x_i\}_{i=1}^n$, define a class $\mathcal{F}_L = \{f_{\boldsymbol{\theta}} : \mathbb{R}^{d_x} \to \mathbb{R} \mid \|\boldsymbol{w}\| \le 1, \|\boldsymbol{V}_l\|_{\mathrm{F}}, \|\boldsymbol{U}_l\|_{\mathrm{F}} \le 1/\sqrt{L}\}$. Then, the empirical Radamacher complexity satisfies

$$\mathcal{R}(\mathcal{F}_L|_S) \le \frac{e^2 \max_i \|x_i\|}{\sqrt{n}}.$$

- 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 has a strictly negative eigenvalue.
- Near-identity regions of ResNets enjoy size-indep. upper bounds on the risk value of critical points & Rademacher complexity.

[1] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Advances in Neural Information Processing Systems*, pp. 6389–6399, 2018.

[2] O. Shamir, "Are ResNets provably better than linear predictors?," *arXiv preprint arXiv:1804.06739*, 2018.

[3] S. Liang, R. Sun, Y. Li, and R. Srikant, "Understanding the loss surface of neural networks for binary classification," in *International Conference on Machine Learning*, pp. 2840–2849, 2018.

[4] S. Liang, R. Sun, J. D. Lee, and R. Srikant, "Adding one neuron can eliminate all bad local minima," in *Advances in Neural Information Processing Systems*, pp. 4355–4365, 2018.

[5] Q. Nguyen, M. C. Mukkamala, and M. Hein, "On the loss landscape of a class of deep neural networks with no bad local valleys," *arXiv preprint arXiv:1809.10749*, 2018.

[6] M. Hardt and T. Ma, "Identity matters in deep learning," in *International Conference on Learning Representations*, 2017.

[7] P. L. Bartlett, S. N. Evans, and P. M. Long, "Representing smooth functions as compositions of near-identity functions with implications for deep network optimization," *arXiv preprint arXiv:1804.05012*, 2018.

[8] H. Zhang, Y. N. Dauphin, and T. Ma, "Fixup initialization: Residual learning without normalization," in *International Conference on Learning Representations (ICLR)*, 2019.