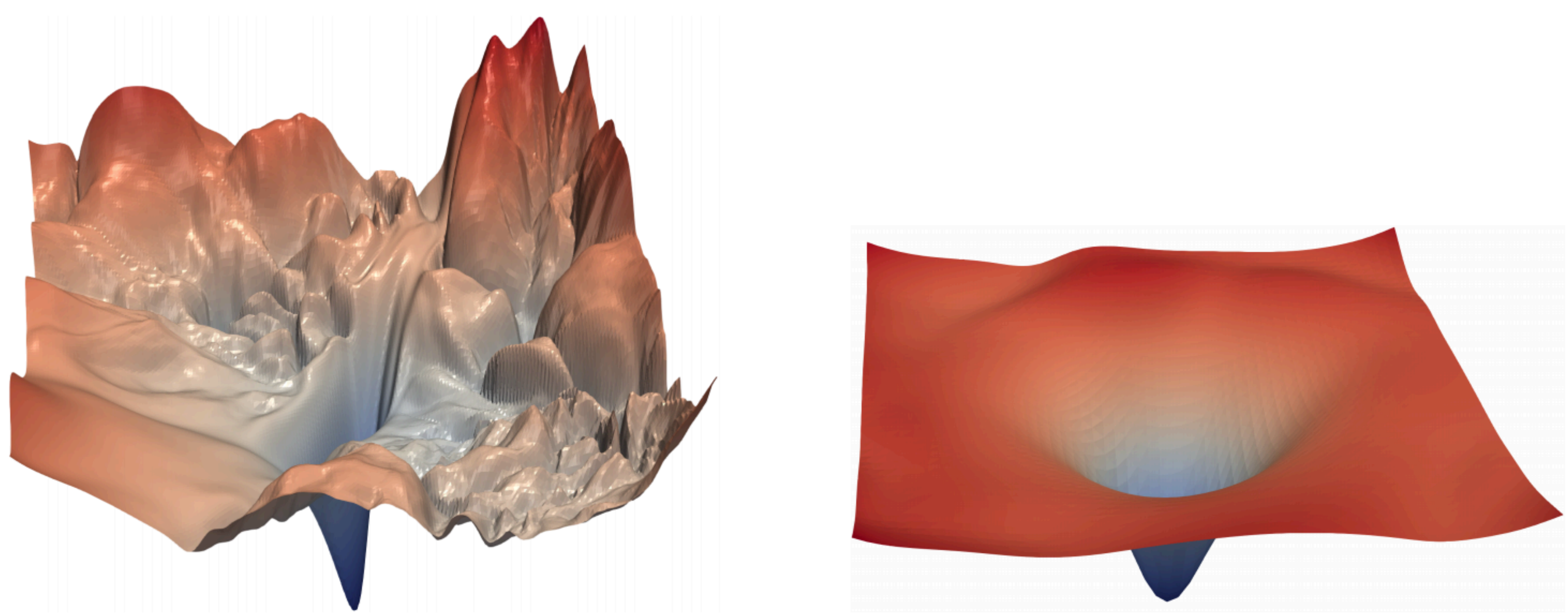
Are deep ResNets provably better than linear predictors?

Chulhee Yun, Suvrit Sra, Ali Jadbabaie

Laboratory for Information and Decision Systems, MIT

Introduction

- Residual networks have been observed to have benign loss landscapes than fully-connected networks
  [Li et al, NeurIPS 2018]



(a) without skip connections

(b) with skip connections

[Li et al, NeurIPS 2018]

Introduction

- Any local minimum of a single-block residual network $x \mapsto \boldsymbol{w}^T(x + \boldsymbol{V}\phi(x))$ has risk value *at least as good as* linear predictors [Shamir, NeurIPS 2018]

Can we extend this result to <u>multi-block</u> ResNets?

Introduction

- Adding parallel shortcut networks can eliminate bad local minima [Liang et al., ICML 2018, NeurIPS 2018]

- Adding many skip-connections from hidden nodes to output removes bad local valleys [Nguyen et al., ICLR 2019]

- However, they only consider direct skip-connection to output.

Can a chain of skip-connections improve the loss landscape?

Introduction

- Near-identity regions of *linear* ResNets have good optimization landscape and expressive power
  [Hardt & Ma, ICLR 2017]

- Nonlinear function space extension is possible
  [Bartlett et al., arXiv 2018]

- Initialization at near-identity regions leads to stable training and good generalization performance
  [Zhang et al., ICLR 2019]

What are the optimization/generalization properties of <u>near-identity regions</u>?

Multi-block ResNets

- ResNet operation for $x \in \mathbb{R}^{d_x}$

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

where $\boldsymbol{U}_l \in \mathbb{R}^{m_l \times d_x}$, $\phi_{\boldsymbol{z}}^l : \mathbb{R}^{m_l} \to \mathbb{R}^{n_l}$, $\boldsymbol{V}_l \in \mathbb{R}^{d_x \times n_l}$, $\boldsymbol{w} \in \mathbb{R}^{d_x}$,
$\boldsymbol{\theta}$ is the collection of all parameters

Multi-block ResNets

- For loss $\ell(z; y)$ and data distribution $\mathcal{P}$,

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

Multi-block ResNets

Theorem. Suppose the loss function $\ell(z; y)$ is convex and twice-differentiable in z. Let $\boldsymbol{\theta}^*$ be any twice-differentiable critical point of $\mathfrak{R}(\,\cdot\,)$. If

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

then at least one of the following holds:

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

Multi-block ResNets

- Under geometric conditions, a critical point of multi-block ResNet is better than linear predictors or is a strict saddle.

- If $L = 1$, any critical point with $\boldsymbol{w}^* \neq 0$ satisfies $\mathfrak{R}(\boldsymbol{\theta}^*) \leq \mathfrak{R}_{\text{lin}}$ (recovers [Shamir, NeurIPS 2018])

- Shows that a chain of multiple skip-connections, as opposed to direct connections to output, can improve optimization landscapes

- 2nd condition is satisfied whenever $\sum_{l=2}^{L} m_l < d_x$

Near-identity regions

- Consider ResNet with residual blocks:

$$h_l(x) = h_{l-1}(x) + \phi_z^l(h_{l-1}(x)), \quad l = 1, \ldots, L$$

Theorem. Suppose ϕ_z^l is $O(1/L)$-Lipschitz, and loss function $\ell(z; y)$ is a convex, differentiable, and Lipschitz function of z. Then, for any critical point $\boldsymbol{\theta}^*$ of $\mathfrak{R}(\,\cdot\,)$,

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

where the constant C doesn't depend on L.

Near-identity regions

- Consider ResNet with residual blocks:

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

Theorem. Given a dataset $S = (x_1, \ldots, x_n)$, define the function class $\mathcal{F}_L = \left\{ f_{\boldsymbol{\theta}} : \mathbb{R}^{d_x} \to \mathbb{R} \mid \|\boldsymbol{w}\| \le 1, \|\boldsymbol{V}_l\|_{\mathrm{F}}, \|\boldsymbol{U}_l\|_{\mathrm{F}} \le 1/\sqrt{L} \right\}$.

Then, the empirical Radamacher complexity satisfies

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

Near-identity regions

- When the residual part is $O(1/L)$-Lipschitz, each residual block is near-identity.
- Risk value $\mathfrak{R}(\boldsymbol{\theta}^*)$ attained is not too far off from $\mathfrak{R}_{\text{lin}}$
- Bounds on $\mathfrak{R}(\boldsymbol{\theta}^*)$ and Radamacher complexity are **independent of depth** of the network, which is difficult to achieve in general [Golowich et al., COLT 2018]

Thank you for your attention!

Reference:

Are deep ResNets provably better than linear predictors?
https://arxiv.org/abs/1907.03922