An approximation theory of deep residual networks

Instructor: Weinan E

Mathematical Introduction to Machine Learning
MAT 490/APC 490

Princeton University, Spring 2021

Residual networks

- Consider the **scaled** residual network (ResNet):

$$\begin{aligned} \boldsymbol{z}_0(\boldsymbol{x}) &= V\tilde{\boldsymbol{x}} \\ \boldsymbol{z}_{l+1}(\boldsymbol{x}) &= \boldsymbol{z}_l(\boldsymbol{x}) + \frac{1}{L}\frac{1}{m}U_l\sigma(W_l\boldsymbol{z}_l(\boldsymbol{x})), \quad l = 0, \ldots, L-1 \\ f_L(\boldsymbol{x};\theta) &= \boldsymbol{\alpha}^T\boldsymbol{z}_L(\boldsymbol{x}) \end{aligned} \tag{1}$$

  where $\tilde{\boldsymbol{x}} = (\boldsymbol{x}^T, 1)^T \in \mathbb{R}^{d+1}, W_l \in \mathbb{R}^{m\times D}, U_l \in \mathbb{R}^{D\times m}, \boldsymbol{\alpha} \in \mathbb{R}^D$ and

$$V = \begin{pmatrix} I_{d+1} \\ 0 \end{pmatrix} \in \mathbb{R}^{D\times(d+1)}.$$

  We use $\theta = \{W_1, U_1, \ldots, W_L, U_L, \boldsymbol{\alpha}\}$ to denote all the parameters to be learned.
- We assume that $\sigma(t) = \max(0, t)$ and $\boldsymbol{x} \in X := [0, 1]^d$.

The continuum limit

- Taking $m \to \infty$, the update of hidden state becomes

$$\boldsymbol{z}_{l+1}(\boldsymbol{x}) = \boldsymbol{z}_l(\boldsymbol{x}) + \frac{1}{L}\mathbb{E}_{(\boldsymbol{u},\boldsymbol{w})\sim\rho_l}[\boldsymbol{u}\sigma(\boldsymbol{w}^T\boldsymbol{z}_l(\boldsymbol{x}))]. \tag{2}$$

- The above iteration can be viewed as the forward Euler disretization of the ODE:

$$\frac{d\boldsymbol{z}(\boldsymbol{x},t)}{dt} = \mathbb{E}_{(\boldsymbol{u},\boldsymbol{w})\sim\rho_t}[\boldsymbol{u}\sigma(\boldsymbol{w}^T\boldsymbol{z}(\boldsymbol{x},t))]. \tag{3}$$

  The scaling factor $1/L$ corresponds to the step size of disretization.
- In this continuous level, the parameters are $\{\boldsymbol{\alpha}, (\rho_t)\}$.

The compositional law of large numbers

Theorem 1 (LNN-type approximation)

Let $(\rho_t)_{t\in[0,1]}$ be a sequence of probability distributions on $\mathbb{R}^D \times \mathbb{R}^D$ with the property that there exist constants c_1 and c_2 such that

$$\begin{aligned} &\mathbb{E}_{\rho_t}\||\boldsymbol{u}||\boldsymbol{w}^T|\|_F^2 < c_1 \\ &\left|\mathbb{E}_{\rho_t}[\boldsymbol{u}\sigma(\boldsymbol{w}^T\boldsymbol{z})] - \mathbb{E}_{\rho_s}[\boldsymbol{u}\sigma(\boldsymbol{w}^T\boldsymbol{z})]\right| \le c_2|t-s||\boldsymbol{z}|,\ \forall\, s,t \in [0,1]. \end{aligned} \tag{4}$$

Let $\boldsymbol{z}$ be the solution of the following ODE,

$$\begin{aligned} \boldsymbol{z}(\boldsymbol{x},0) &= V\boldsymbol{x}, \\ \frac{d}{dt}\boldsymbol{z}(\boldsymbol{x},t) &= \mathbb{E}_{(\boldsymbol{u},\boldsymbol{w})\sim\rho_t}[\boldsymbol{u}\sigma(\boldsymbol{w}^T\boldsymbol{z}(\boldsymbol{x},t))]. \end{aligned} \tag{5}$$

Then, for any fixed $\boldsymbol{x} \in X$, we have

$$\boldsymbol{z}_L(\boldsymbol{x}) \to \boldsymbol{z}(\boldsymbol{x},1)$$

in probability as $L \to +\infty$. Moreover, the convergence is uniform in $\boldsymbol{x}$.

The compositional law of large numbers

Remarks:

- The moment boundedness of (ρ_t) is required to ensure the convergence of Monte-Carlo discretization.
- The continuity *wrt* t of (ρ_t) is required to ensure the convergence of the forward Euler discretization.
- In this theorem, we view the ResNet (1) as a forward Euler discretization of ODE (5) with a **stochastic approximation** of the expectation in RHS. As a result, the width m **can be fixed**.
- This approximation does not provide any rate. The CLT-type approximation require stronger regularity.

Intuition of stochastic approximation

Consider the case of $m = 1$. Let $L = L'M$ with $L', M \gg 1$, and $dt = \frac{1}{L}, \Delta t = \frac{M}{L} \ll 1$. Let $t = l\,dt$ and $\hat{\boldsymbol{z}}(\boldsymbol{x}; t) = \boldsymbol{z}_l(\boldsymbol{x})$.

$$\begin{aligned}
\hat{\boldsymbol{z}}(\boldsymbol{x}; t + \Delta t) &= \boldsymbol{z}_{l+M-1}(\boldsymbol{x}) + \frac{1}{L}\boldsymbol{u}_{l+M}\sigma(\boldsymbol{w}_{l+M}^T\sigma(\boldsymbol{z}_{l+M-1}(\boldsymbol{x}))) \\
&= \boldsymbol{z}_l(\boldsymbol{x}) + \frac{1}{L}\sum_{j=l+1}^{j=l+M}\boldsymbol{u}_j\sigma(\boldsymbol{w}_j^T\sigma(\boldsymbol{z}_j(\boldsymbol{x}))) \\
&= \boldsymbol{z}(\boldsymbol{x}; t) + \frac{M}{L}\frac{1}{M}\sum_{j=l+1}^{j=l+M}\boldsymbol{u}_j\sigma(\boldsymbol{w}_j^T\sigma(\boldsymbol{z}_j(\boldsymbol{x}))) \qquad (\boldsymbol{u}_j, \boldsymbol{w}_j) \sim \rho_{t+(j-l)dt}.
\end{aligned} \tag{6}$$

Note that $(j-l)dt \leq \Delta t \ll 1$, ρ_t and $\boldsymbol{z}(\boldsymbol{x}; t)$ are Lipschitz continuous in t. Therefore,

$$\frac{1}{M}\sum_{j=l+1}^{j=l+M}\boldsymbol{u}_j\sigma(\boldsymbol{w}_j^T\sigma(\boldsymbol{z}_j(\boldsymbol{x}))) = \mathbb{E}_{(\boldsymbol{u},\boldsymbol{w})\sim\rho_t}[\boldsymbol{u}\sigma(\boldsymbol{w}^T\hat{\boldsymbol{z}}(\boldsymbol{x}; t))] + o(\Delta t).$$

Hence, the ResNet can be viewed as a coarse discretization of the ODE:

$$\hat{\boldsymbol{z}}(\boldsymbol{x}; t + \Delta t) \approx \hat{\boldsymbol{z}}(\boldsymbol{x}; t) + \Delta t\,\mathbb{E}_{(\boldsymbol{u},\boldsymbol{w})\sim\rho_t}[\boldsymbol{u}\sigma(\boldsymbol{w}^T\hat{\boldsymbol{z}}(\boldsymbol{x}; t))], \tag{7}$$

Flow-induced functions

- Motivated by previous results, consider the set of functions $f_{\boldsymbol{\alpha},\{\rho_t\}}$ defined by:

$$\begin{aligned} \boldsymbol{z}(\boldsymbol{x},0) &= V\boldsymbol{x}, \\ \frac{d\boldsymbol{z}(\boldsymbol{x},t)}{dt} &= \mathbb{E}_{(\boldsymbol{u},\boldsymbol{w})\sim\rho_t}\boldsymbol{u}\sigma(\boldsymbol{w}^T\boldsymbol{z}(\boldsymbol{x},t)) \\ f_{\boldsymbol{\alpha},(\rho_t)}(\boldsymbol{x}) &= \boldsymbol{\alpha}^T\boldsymbol{z}(\boldsymbol{x},1), \end{aligned} \tag{8}$$

- Let $\boldsymbol{e}$ be the all-one vector. Define the following linear ODE:

$$\begin{aligned} N_p(0) &= \boldsymbol{e}, \\ \dot{N}_p(t) &= 3\left(\mathbb{E}_{\rho_t}(|\boldsymbol{u}||\boldsymbol{w}|^T)^p\right)^{1/p} N_p(t), \end{aligned} \tag{9}$$

  where $|\boldsymbol{v}|$ and $|\boldsymbol{v}|^q$ are defined element-wise for any vector or matrix $\boldsymbol{v}$.
- We will use this linear ODE to control the complexity of the original nonlinear ODE (8).
- The factor 3 is only required for the control of Rademacher complexity. For controlling the approximation error, we can replace 3 by 1. But for simplicity, we use 3 for both scenarios.

Flow-induced function spaces

- Let $\|(\rho_t)\|_{Lip}$ be the smallest constant C such that for any $t, s \in [0,1]$, we have

$$
\begin{aligned}
|\mathbb{E}_{\rho_t} U\sigma(W\boldsymbol{z}) - \mathbb{E}_{\rho_s} U\sigma(W\boldsymbol{z})| &\le C|t-s||\boldsymbol{z}|, \\
\Big| \|\mathbb{E}_{\rho_t}|U||W|\|_{1,1} - \|\mathbb{E}_{\rho_s}|U||W|\|_{1,1} \Big| &\le C|t-s|,
\end{aligned} \tag{10}
$$

where $\|\cdot\|_{1,1}$ is the sum of the absolute values of all the entries in a matrix.

Definition 2

Let f be a function that satisfies $f = f_{\boldsymbol{\alpha},(\rho_t)}$ for a pair of $\{\boldsymbol{\alpha},(\rho_t)\}$. We define

$$
\begin{aligned}
\|f\|_{\mathcal{D}_p} &= \inf_{f = f_{\boldsymbol{\alpha},(\rho_t)}} |\boldsymbol{\alpha}|^T N_p(1) \\
\|f\|_{\tilde{\mathcal{D}}_p} &= \inf_{f = f_{\boldsymbol{\alpha},(\rho_t)}} |\boldsymbol{\alpha}|^T N_p(1) + \|N_p(1)\|_1 - D + \|(\rho_t)\|_{Lip},
\end{aligned}
$$

The space $\mathcal{D}_p$ and $\tilde{\mathcal{D}}_p$ are defined as the set all continuous functions that admit the ODE representation with finite $\mathcal{D}_p$ and $\tilde{\mathcal{D}}_p$ norm, respectively.

Flow-induced function spaces

- $\mathcal{D}_p$ norm does no control the regularity of representation (ρ_t), while $\tilde{\mathcal{D}}_p$ does.
- We add a "$-D$" term in the definition of $\tilde{\mathcal{D}}_p$ norm because $\|N_p(1)\|_1 \geq D$ and we want the norm of the zero function to be 0.
- We use the terminology "norm" loosely, and we do not care whether these are really norms. Strictly speaking, they are just some quantities that can be used to bound approximation/estimation errors.

The embedding result

Proposition 1

Assume that $D \geq d+2$ and $m \geq 1$. For any function $f \in \mathcal{B}$, we have $f \in \tilde{\mathcal{D}}_1$, and

$$\|f\|_{\tilde{\mathcal{D}}_1} \leq 2\|f\|_{\mathcal{B}} + 1.$$

Moreover, $f = f_{\boldsymbol{\alpha},(\rho_t)}$ with $\rho_t = \rho$ for any $t \in [0,1]$.

Proof:

- Since $f \in \mathcal{B}$, there exit a distribution ρ such that

$$f(\boldsymbol{x}) = \mathbb{E}_{(a,\boldsymbol{b},c)\sim\rho}[a\sigma(\boldsymbol{b}^T\boldsymbol{x} + c)]$$
$$\|f\|_{\mathcal{B}} = \mathbb{E}_{(a,\boldsymbol{b},c)\sim\rho}[|a|(\|\boldsymbol{b}\| + |c|)].$$

The embedding result

Proof:

- It is easy to verify that f can be represented by an ODE as follows

$$z(\boldsymbol{x},0) = \begin{bmatrix} \boldsymbol{x} \\ 1 \\ 0 \end{bmatrix}$$
$$\frac{d}{dt} z(\boldsymbol{x},t) = \mathbb{E}_{(a,\boldsymbol{b},c)\sim\rho} \begin{bmatrix} 0 \\ 0 \\ a \end{bmatrix} \sigma([\boldsymbol{b}^T, c, 0] z(\boldsymbol{x},t)) \tag{11}$$
$$f(\boldsymbol{x}) = \boldsymbol{e}_{d+2}^T z(\boldsymbol{x},1),$$

  where $e_{d+2} = (0,0,\ldots,0,1)^T \in \mathbb{R}^{d+2}$.
- It is obviously that $\rho_t = \tilde{\rho}$ for some $\tilde{\rho}$ and any $t \in [0,1]$. Hence, $\|(\rho_t)\|_{lip} = 0$. An explicit calculation gives us that

$$|\boldsymbol{\alpha}|^T N_1(1) + N_1(1) - D = 2\|f\|_{\mathcal{B}} + 1.$$

- Using the definitions of $\tilde{\mathcal{D}}_1$ norm, we complete the proof.

Weighted path norms for ResNets

When L is finite, the complexity is controlled by the quantity defined below.

- Given a ResNet $f_L(\cdot;\theta)$ define the **weighted path norm** as

$$\|\theta\|_{\mathcal{P}} := |\boldsymbol{\alpha}|^T \left(I + \frac{3}{Lm}|U_L||W_L|\right) \cdots \left(I + \frac{3}{Lm}|U_1||W_1|\right) \boldsymbol{e}. \tag{12}$$

  It is a discrete analog of the $\mathcal{D}_1$ norm.

- This weighted path norm is a weighted sum over all paths from the input to the output, and gives larger weight to the paths that go through more nonlinearities. Given a path P, let $w_1^P, u_1^P, \ldots, w_L^P, u_L^P$ be the weights, and $a(P)$ be number of nonlinearities that P goes through. Then,

$$\|\theta\|_{\mathcal{P}} = \sum_{P\,:\,\text{all paths}} \left(\frac{3}{mL}\right)^{a(P)} \prod_{l=1}^{L} |w_l^P||u_l^P|. \tag{13}$$

Direct approximation

Theorem 3

Let $f \in \tilde{\mathcal{D}}_2$, $\delta \in (0,1)$. Then, there exists an absolute constant C, such that for any

$$L \geq C \left(m^4 D^6 \|f\|_{\tilde{\mathcal{D}}_2}^5 (\|f\|_{\tilde{\mathcal{D}}_2} + D)^2 \right)^{\frac{3}{\delta}},$$

there is an L-layer residual network $f_L(\cdot;\Theta)$ that satisfies

$$\|f - f_L(\cdot;\Theta)\|^2 \leq \frac{\|f\|_{\tilde{\mathcal{D}}_2}^2}{L^{1-\delta}},$$

and

$$\|\Theta\|_{\mathcal{P}} \leq 9\|f\|_{\tilde{\mathcal{D}}_1}.$$

Inverse approximation

Theorem 4

Let f be a function defined on X. Assume that there is a sequence of residual networks $\{f_L(\cdot;\theta_L)\}_{L=1}^{\infty}$ such that $f_L(\boldsymbol{x};\theta) \to f(\boldsymbol{x})$ for every $\boldsymbol{x} \in X$ as $L \to \infty$. Assume further that the parameters in $\{f_L(\cdot;\theta)\}_{L=1}^{\infty}$ are (entry-wise) bounded by c_0. Then, we have $f \in \mathcal{D}_\infty$, and

$$\|f\|_{\mathcal{D}_\infty} \le \frac{2e^{m(c_0^2+1)} D^2 c_0}{m}$$

Moreover, if for some constant c_1, $\|f_L\|_{\mathcal{D}_1} \le c_1$ holds for all $L > 0$, then we have

$$\|f\|_{\mathcal{D}_1} \le c_1$$

Rademacher complexity

Theorem 5

Let $\tilde{\mathcal{D}}_2^Q = \{f \in \tilde{\mathcal{D}}_2 : \|f\|_{\tilde{\mathcal{D}}_2} \leq Q\}$*, then we have*

$$\widehat{Rad}_n(\tilde{\mathcal{D}}_2^Q) \lesssim Q\sqrt{\frac{2\log(2d)}{n}}.$$

The proof of the above theorem is a simple combination of the direct approximation theorem with the following proposition.

Proposition 2

Let $\mathcal{F}^Q = \{f_L(\cdot;\theta) : \|\theta\|_{\mathcal{P}} \leq Q\}$ *where* $f_L(\cdot;\theta)$ *is the* L*-layer ResNet. We have*

$$\widehat{Rad}_n(\mathcal{F}^Q) \leq 3Q\sqrt{\frac{2\log(2d)}{n}}$$

Rademacher complexity

Proof: By the direct approximation theorem, for any $\varepsilon \in (0,1)$ and $f \in \tilde{\mathcal{D}}_2^Q$, there exist a L (sufficiently large), a constant $c > 0$, and θ^f such that

$$\frac{1}{n}\sum_{i=1}^{n} |f(x) - f_L(x;\theta^f)|^2 \le \varepsilon^2 \qquad \|\theta^f\|_{\mathcal{P}} \le cQ.$$

Proof: By the direct approximation theorem, for any $\varepsilon \in (0,1)$ and $f \in \tilde{\mathcal{D}}_2^Q$, there exist a L (sufficiently large), a constant $c > 0$, and θ^f such that

$$\frac{1}{n}\sum_{i=1}^{n} |f(x) - f_L(x;\theta^f)|^2 \leq \varepsilon^2 \qquad \|\theta^f\|_{\mathcal{P}} \leq cQ.$$

Therefore,

$$\begin{aligned}
\widehat{\mathrm{Rad}}_n(\mathcal{D}_2^Q) &= \frac{1}{n}\mathbb{E}_\xi[\sup_{f\in\tilde{\mathcal{D}}_2^Q}\sum_{i=1}^{n}\xi_i f(x_i)] \\
&\leq \frac{1}{n}\mathbb{E}_\xi[\sup_{f\in\tilde{\mathcal{D}}_2^Q}\left(\sum_{i=1}^{n}\xi_i(f(x_i) - f_L(x_i;\theta)) + \sum_{i=1}^{n}\xi_i f_L(x_i;\theta^f)\right)] \\
&\leq \frac{1}{n}\mathbb{E}_\xi[\sup_{f_L(\cdot;\theta)\in\mathcal{F}_L^{cQ}}\sum_{i=1}^{n}\xi_i f_L(x_i;\theta)] + \varepsilon \\
&\leq \widehat{\mathrm{Rad}}_n(\mathcal{F}_L^{cQ}) + \varepsilon \leq 3cQ\sqrt{\frac{2\log(2d)}{n}} + \varepsilon.
\end{aligned} \tag{14}$$

Where the last inequality follows from Prop. 2. Taking $\varepsilon \to 0$, we complete the proof.

- **Proof of the upper bound for the Rademacher complexity of ResNets.**

Define the intermediate quantities

- let $\boldsymbol{g}_l(\boldsymbol{x}) = \sigma(W_l \boldsymbol{z}_{l-1})$, and g_l^i be the i-th element of $\boldsymbol{g}_l$. Then, we have the following recurrence relation:

$$g_l^i = \sigma(W_l^{i,:}(\gamma U_{l-1}\boldsymbol{g}_{l-1} + \gamma U_{l-2}\boldsymbol{g}_{l-2} + \cdots + \gamma U_1 \boldsymbol{g}_1 + \boldsymbol{z}_0),$$

  where $W_l^{i,:}$ is the i-th row of W_l, $\gamma = \frac{1}{Lm}$ is the scaling factor, and $\boldsymbol{z}_0 = V\boldsymbol{x}$.
- g_l^i is l-layer ResNet. We define its *weighted path norm* by

$$\|g_l^i\|_{\mathcal{P}} = 3|W_l^{i,:}|(I + 3\gamma|U_{l-1}||W_{l-1}|)\cdots(I + 3\gamma|U_1||W_1|)|V|\boldsymbol{e}, \tag{15}$$

Recurrence relation of path norms

With an abuse of notation, let $\|f_L\|_{\mathcal{P}}$ and $\|g_l^i\|_{\mathcal{P}}$ denote the path norm of the parameters. We have

$$\|f_L\|_{\mathcal{P}} = \gamma \sum_{l=1}^{L} \sum_{j=1}^{m} \left(|\boldsymbol{\alpha}|^T |U_l^{:,j}| \right) \|g_l^j\|_{\mathcal{P}} + |\boldsymbol{\alpha}|^T |V| \boldsymbol{e}$$

$$\|g_{l+1}^i\|_{\mathcal{P}} = \sum_{k=1}^{l} \sum_{j=1}^{m} 3\gamma \left(|W_{l+1}^{i,:}| |U_k^{:,j}| \right) \|g_k^j\|_{\mathcal{P}} + 3|W_{l+1}^{i,:}| |V| \boldsymbol{e},$$

where $U_l^{:,j}$ is the j-th column of U_l.

Recurrence relation of path norms

With an abuse of notation, let $\|f_L\|_{\mathcal{P}}$ and $\|g_l^i\|_{\mathcal{P}}$ denote the path norm of the parameters. We have

$$\|f_L\|_{\mathcal{P}} = \gamma \sum_{l=1}^{L} \sum_{j=1}^{m} \left(|\boldsymbol{\alpha}|^T |U_l^{:,j}| \right) \|g_l^j\|_{\mathcal{P}} + |\boldsymbol{\alpha}|^T |V| \boldsymbol{e}$$

$$\|g_{l+1}^i\|_{\mathcal{P}} = \sum_{k=1}^{l} \sum_{j=1}^{m} 3\gamma \left(|W_{l+1}^{i,:}| |U_k^{:,j}| \right) \|g_k^j\|_{\mathcal{P}} + 3|W_{l+1}^{i,:}| |V| \boldsymbol{e},$$

where $U_l^{:,j}$ is the j-th column of U_l.
Proof: Recall the definition of $\|f_L\|_{\mathcal{P}}$, we have

$$\begin{aligned}
\|f_L\|_{\mathcal{P}} &= |\boldsymbol{\alpha}|^{\mathsf{T}} (I + 3\gamma |U_L||W_L|) \cdots (I + 3\gamma |U_1||W_1|) |V| \boldsymbol{e} \\
&= \sum_{l=1}^{L} |\boldsymbol{\alpha}|^{\mathsf{T}} |U_l| \cdot 3\gamma |W_l| \prod_{j=1}^{l-1} (I + 3\gamma |U_{l-j}||W_{l-j}|) |V| + |\boldsymbol{\alpha}|^{\mathsf{T}} |V| \boldsymbol{e} \\
&= \gamma \sum_{l=1}^{L} \sum_{j=1}^{m} \left(|\boldsymbol{\alpha}|^{\mathsf{T}} |U_l^{:,j}| \right) \|g_l^j\|_{\mathcal{P}} + |\boldsymbol{\alpha}|^{\mathsf{T}} |V| \boldsymbol{e},
\end{aligned}$$

The proof for the recurrence relation of g_l^i is similar.

Recursion of hypothesis space

Lemma 6

Let $\mathcal{G}_l^Q = \{g_l^i : \|g_l^i\|_{\mathcal{P}} \leq Q\}$*, then*

(1) $\mathcal{G}_k^Q \subseteq \mathcal{G}_l^Q$ *for* $k \leq l$*;*

(2) $\mathcal{G}_l^q \subseteq \mathcal{G}_l^Q$ *and* $\mathcal{G}_l^q = \frac{q}{Q}\mathcal{G}_l^Q$ *for* $q \leq Q$*.*

Proof:

- $\mathcal{G}_k^Q \subseteq \mathcal{G}_l^Q$ and $\mathcal{G}_l^q \subseteq \mathcal{G}_l^Q$ are obvious.
- For any $g_l \in \mathcal{G}_l^q$, define $\tilde{g}_l$ by replacing the output parameters $\boldsymbol{w}$ by $\frac{Q}{q}\boldsymbol{w}$, then we have $\|\tilde{g}_l\|_{\mathcal{P}} = \frac{Q}{q}\|g_l\|_{\mathcal{P}} \leq Q$, and hence $\tilde{g}_l \in \mathcal{G}_l^Q$. Therefore, we have $\frac{Q}{q}\mathcal{G}_l^q \subseteq \mathcal{G}^Q$. Similarly we can obtain $\frac{q}{Q}\mathcal{G}_l^Q \subseteq \mathcal{G}^q$. Consequently, we have $\mathcal{G}_l^q = \frac{q}{Q}\mathcal{G}_l^Q$.

Proof of Prop. 2

- To prove Prop. 2, we only need to prove that for any $l = 0, 1, \ldots, L$
$$\widehat{\mathsf{Rad}}_n(\mathcal{G}_l^Q) \le Q\sqrt{\frac{2\log(2d)}{n}}. \quad (16)$$
This will be done by induction.
- When $l = 1$, $g_1^i(\boldsymbol{x}) = \sigma(W_1^{i,:}V\boldsymbol{x})$. By the contraction lemma and the bound of Rademacher complexity of linear class, (16) holds.
- Now assume that the result holds for $1, 2, \ldots, l$. For $l+1$, we have

$$\begin{aligned}
n\widehat{\mathsf{Rad}}_n(\mathcal{G}_{l+1}^Q) &= \mathbb{E}_\xi \sup_{g_{l+1}\in\mathcal{G}_{l+1}^Q} \sum_{i=1}^n \xi_i g_{l+1}(\boldsymbol{x}_i) \\
&= \mathbb{E}_\xi \sup_{(1)} \sum_{i=1}^n \xi_i \sigma(\boldsymbol{w}_{l+1}^T(\gamma U_l \boldsymbol{g}_l + \gamma U_{l-1}\boldsymbol{g}_{l-1} + \cdots + \gamma U_1 \boldsymbol{g}_1 + \boldsymbol{z}_0)) \\
&\le \mathbb{E}_\xi \sup_{(1)} \sum_{i=1}^n \xi_i (\boldsymbol{w}_{l+1}^T(\gamma U_l \boldsymbol{g}_l + \gamma U_{l-1}\boldsymbol{g}_{l-1} + \cdots + \gamma U_1 \boldsymbol{g}_1 + \boldsymbol{z}_0)), \quad \text{(contraction lemma)}
\end{aligned}$$

where the condition (1) is $\sum_{k=1}^{l}\sum_{j=1}^{m} 3\gamma \left(|\boldsymbol{w}_{l+1}|^T|U_k^{:,j}|\right)\|g_k^j\|_{\mathcal{P}} + 3|\boldsymbol{w}_{l+1}|^T|V|\boldsymbol{e} \le Q$

Proof of Prop. 2

- Let $a_k = \gamma \sum_{j=1}^{m} \left(|\boldsymbol{w}_{l+1}|^T |U_k^{:,j}| \right) \|g_k^j\|_{\mathcal{P}}$ and $b = |\boldsymbol{w}_{l+1}|^T |V| \boldsymbol{e}$. Then, the constraint becomes

$$3 \sum_{k=1}^{l} a_k + 3b \leq Q. \tag{17}$$

- Therefore, we have

$$\begin{aligned} n\widehat{\mathsf{Rad}}_n(\mathcal{G}_{l+1}^Q) &\underset{(2)}{\overset{(i)}{\leq}} \mathbb{E}_\xi \sup \left\{ \sum_{k=1}^{l} a_k \sup_{g \in \mathcal{G}_k^1} \left| \sum_{i=1}^{n} \xi_i g(\boldsymbol{x}_i) \right| + b \sup_{\|\boldsymbol{u}\|_1 \leq 1} \left| \sum_{i=1}^{n} \xi_i \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{x}_i \right| \right\} \\ &\overset{(ii)}{\leq} \mathbb{E}_\xi \sup_{\substack{a+b \leq \frac{Q}{3} \\ a,b \geq 0}} \left\{ a \sup_{g \in \mathcal{G}_l^1} \left| \sum_{i=1}^{n} \xi_i g(\boldsymbol{x}_i) \right| + b \sup_{\|\boldsymbol{\alpha}\|_1 \leq 1} \left| \sum_{i=1}^{n} \xi_i \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{x}_i \right| \right\} \\ &\leq \frac{Q}{3} \left[\mathbb{E}_\xi \sup_{g \in \mathcal{G}_l^1} \left| \sum_{i=1}^{n} \xi_i g(\boldsymbol{x}_i) \right| + \mathbb{E}_\xi \sup_{\|\boldsymbol{u}\|_1 \leq 1} \left| \sum_{i=1}^{n} \xi_i \boldsymbol{\alpha}^\mathsf{T} \boldsymbol{x}_i \right| \right], \end{aligned} \tag{18}$$

where (i) is due to the scaling invariance, and (ii) follows from Lemma 6.

Proof of Prop. 2

- By symmetry,

$$\begin{aligned}\mathbb{E}_\xi \sup_{g\in\mathcal{G}_l^1}\left|\sum_{i=1}^n \xi_i g(\boldsymbol{x}_i)\right| &\le \mathbb{E}_\xi \sup_{g\in\mathcal{G}_l^1}\sum_{i=1}^n \xi_i g(\boldsymbol{x}_i) + \mathbb{E}_\xi \sup_{g\in\mathcal{G}_l^1} -\sum_{i=1}^n \xi_i g(\boldsymbol{x}_i)\\ &= 2\mathbb{E}_\xi \sup_{g\in\mathcal{G}_l^1}\sum_{i=1}^n \xi_i g(\boldsymbol{x}_i) = 2n\widehat{\mathsf{Rad}}_n(\mathcal{G}_l^1) \le 2n\sqrt{\frac{2\log(2d)}{n}}. \qquad (19)\end{aligned}$$

  And

$$\mathbb{E}_\xi \sup_{\|\boldsymbol{u}\|_1\le 1}\left|\sum_{i=1}^n \xi_i \boldsymbol{u}^\intercal \boldsymbol{x}_i\right| = \mathbb{E}_\xi \sup_{\|\boldsymbol{u}\|_1\le 1}\sum_{i=1}^n \xi_i \boldsymbol{u}^\intercal \boldsymbol{x}_i \le n\sqrt{\frac{2\log(2d)}{n}}, \qquad (20)$$

  where the supremum is reached at $\boldsymbol{u} = \sum_{i=1}^n \xi_i \boldsymbol{x}_i$.
- Plugging the above bounds into (18) gives us

$$\widehat{\mathsf{Rad}}_n(\mathcal{G}_{l+1}^Q) \le \frac{Q}{3}\left[2\sqrt{\frac{2\log(2d)}{n}} + \sqrt{\frac{2\log(2d)}{n}}\right] \le Q\sqrt{\frac{2\log(2d)}{n}}.$$

Summary

- The continuum limit of deep ResNet is an ODE: $\dot{\boldsymbol{z}}(\boldsymbol{x},t) = \mathbb{E}_{(\boldsymbol{u},\boldsymbol{w})\sim\rho_t}[\boldsymbol{u}\sigma(\boldsymbol{w}^T\boldsymbol{z}(\boldsymbol{x};t))]$.
- The ResNet can be viewed as the forward Euler discretization of this ODE with stochastic approximation for the RHS.
- To control the complexity of the flow map of the nonlinear ODE, we define the linear ODE: $\dot{N}_1(t) = \mathbb{E}_{\rho_t}[|\boldsymbol{u}||\boldsymbol{w}|^T]N_1(t)$.
- Bound the Rademacher complexity via the weighted path norm.

All the missing proofs can be found in the following papers.

- https://arxiv.org/abs/1903.02154.
- https://arxiv.org/abs/1906.08039.