Generalization of Scaled Deep ResNets in the Mean-Field Regime

Yihang Chen[1] Fanghui Liu[2] Yiping Lu[3] Grigorios G Chrysos[4] Volkan Cevher[1]

[1]EPFL [2]University of Warwick

[3]New York University [4]University of Wisconsin - Madison

Overview

Question:

- Can we build a generalization analysis of trained Deep ResNets in the mean-field setting?

Contributions:

- The paper provides the first **minimum eigenvalue estimation** (lower bound) of the Gram matrix of the gradients for deep ResNet parameterized by the ResNet encoder's parameters and MLP predictor's parameters in the mean-field regime.
- The paper proves that the **KL divergence** of feature encoder ν and output layer ν can be bounded by a constant (depending only on network architecture parameters) during the training, which facilitates our generalization analysis.
- This paper builds the connection between the Rademacher complexity result and KL divergence, and then derive the **convergence rate** $\mathcal{O}(1/\sqrt{n})$ **for generalization**.

Gradient Evolution

- The evolution of the ResNet layers $\nu(\boldsymbol{\theta}, s)$ can be characterized as

$$\frac{\partial \nu}{\partial t}(\boldsymbol{\theta}, s, t) = \nabla_{\boldsymbol{\theta}} \cdot \left(\nu(\boldsymbol{\theta}, s, t) \nabla_{\boldsymbol{\theta}} \frac{\delta \widehat{L}(\tau, \nu)}{\delta \nu}(\boldsymbol{\theta}, s, t) \right), \quad t \geq 0, \tag{2}$$

- The evolution of the final layer distribution $\tau(\boldsymbol{\omega})$ can be characterized as

$$\frac{\partial \tau}{\partial t}(\boldsymbol{\omega}, t) = \nabla_{\boldsymbol{\omega}} \cdot \left(\tau(\boldsymbol{\omega}, t) \nabla_{\boldsymbol{\omega}} \frac{\delta \widehat{L}(\tau, \nu)}{\delta \tau}(\boldsymbol{\omega}, t) \right), \quad t \geq 0, \tag{3}$$

where the functional derivative

$$\frac{\delta \widehat{L}(\tau, \nu)}{\delta \tau}(\boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_n} [\beta \cdot (f_{\tau,\nu}(\boldsymbol{x}) - y(\boldsymbol{x})) \cdot h(\boldsymbol{Z}_\nu(\boldsymbol{x}, 1), \boldsymbol{\omega})] .$$

- We define one Gram matrix for the ResNet layers, $\boldsymbol{G}_1(\tau, \nu)$ by $\boldsymbol{G}_1(\tau, \nu) = \int_0^1 \boldsymbol{G}_1(\tau, \nu, s) \mathrm{d}s, \quad \boldsymbol{G}_1(\tau, \nu, s) = \mathbb{E}_{\boldsymbol{\theta} \sim \nu(\cdot, s)} \boldsymbol{J}_1(\tau, \nu, \boldsymbol{\theta}, s) \boldsymbol{J}_1(\tau, \nu, \boldsymbol{\theta}, s)^\top$.
- We define the Gram matrix for the MLP parameter distribution τ, $\boldsymbol{G}_2(\tau, \nu)$ by $\boldsymbol{G}_2(\tau, \nu) = \mathbb{E}_{\boldsymbol{\omega} \sim \tau(\cdot)} \boldsymbol{J}_2(\nu, \boldsymbol{\omega}) \boldsymbol{J}_2(\nu, \boldsymbol{\omega})^\top$, where the row vector of $\boldsymbol{J}_2$ is defined as

$$(\boldsymbol{J}_2(\nu, \boldsymbol{\omega}))_{i,\cdot} = \nabla_{\boldsymbol{\omega}} h(\boldsymbol{Z}_\nu(\boldsymbol{x}_i, 1), \boldsymbol{\omega}), \quad 1 \leq i \leq n .$$

- The Gram matrix for the whole network is $\boldsymbol{G} = \alpha^2 \boldsymbol{G}_1 + \boldsymbol{G}_2$.

Problem Setting

Basic Settings:

- The training set $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is drawn from an unknown distribution μ on $\mathcal{X} \times \mathcal{Y}$, and μ_X is the marginal distribution of μ over $\mathcal{X}$.
- We consider a binary classification task, denoted by minimizing the expected risk, let $\ell_{0-1}(f, y) := \mathbb{1}\{yf < 0\}$.
- We employ the squared loss in ERM in training, i.e, $\ell(f, y) := \frac{1}{2}(y - f)^2$.
- The hypothesis f is parameterized by ResNet feature encoder and a non-linear predictor, $f_{\tau,\nu}$. The empirical loss $\widehat{L}(\tau, \nu) := \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}_n} \ell(f_{\tau,\nu}(\boldsymbol{x}), y(\boldsymbol{x}))$.

Network Structure: (α, β will be determined later)

- The following ODE models the infinite death infinite width ResNet.

$$\frac{\mathrm{d}\boldsymbol{z}(\boldsymbol{x}, s)}{\mathrm{d}s} = \alpha \cdot \int_{\mathbb{R}^{k_\nu}} \boldsymbol{\sigma}(\boldsymbol{z}(\boldsymbol{x}, s), \boldsymbol{\theta}) \mathrm{d}\nu(\boldsymbol{\theta}, s), \; s \in [0, 1], \; \boldsymbol{z}(\boldsymbol{x}, 0) = \boldsymbol{x} . \tag{1}$$

We denote the solution of Equation (1) as $\boldsymbol{Z}_\nu(\boldsymbol{x}, s)$.

- The whole network can be written as

$$f_{\tau,\nu}(\boldsymbol{x}) := \beta \cdot \int_{\mathbb{R}^{k_\tau}} h(\boldsymbol{Z}_\nu(\boldsymbol{x}, 1), \boldsymbol{\omega}) \mathrm{d}\tau(\boldsymbol{\omega}) ,$$

Assumptions:

Assumption 1 (Assumptions on data). *We assume that for $\boldsymbol{x}_i \neq \boldsymbol{x}_j \sim \mu_X$, the following holds with probability 1,*

$$\|\boldsymbol{x}_i\|_2 = 1, |y(\boldsymbol{x}_i)| \leq 1, \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \leq C_{\max} < 1 , \forall i, j \in [n] .$$

Assumption 2 (Assumption on initialization). *The initial distribution τ_0, ν_0 is standard Gaussian: $(\tau_0, \nu_0)(\boldsymbol{\omega}, \boldsymbol{\theta}, s) \propto \exp\left(-\frac{\|\boldsymbol{\omega}\|_2^2 + \|\boldsymbol{\theta}\|_2^2}{2}\right), \forall s \in [0, 1]$.*

Assumption 3 (Assumptions on activation $\boldsymbol{\sigma}, h$). *Let $\boldsymbol{\theta} := (\boldsymbol{u}, \boldsymbol{w}, b) \in \mathbb{R}^{k_\nu}$, where $\boldsymbol{u}, \boldsymbol{w} \in \mathbb{R}^{k_\nu}, b \in \mathbb{R}$, i.e. $k_\nu = 2d + 1$; $\boldsymbol{\omega} := (a, \boldsymbol{w}, b) \in \mathbb{R}^{k_\tau}$, where $\boldsymbol{w} \in \mathbb{R}^{k_\nu}, a, b \in \mathbb{R}$, i.e. $k_\tau = d + 2$. For any $\boldsymbol{z} \in \mathbb{R}^{k_\nu}$, we assume*

$$\boldsymbol{\sigma}(\boldsymbol{z}, \boldsymbol{\theta}) = \boldsymbol{u}\sigma_0(\boldsymbol{w}^\top \boldsymbol{z} + b), \quad h(\boldsymbol{z}, \boldsymbol{\omega}) = a\sigma_0(\boldsymbol{w}^\top \boldsymbol{z} + b), \quad \sigma_0 : \mathbb{R} \to \mathbb{R}.$$

In addition, we have the following assumption on σ_0. $|\sigma_0(x)| \leq C_1 \max(|x|, 1), |\sigma_0'(x)| \leq C_1, |\sigma_0''(x)| \leq C_1$, and let $\mu_i(\sigma_0)$ be the i-th Hermite coefficient of σ_0.

Main Results

KL divergence:

Lemma 4. *Under Assumption 1, 2, 3, there exist a constant $\Lambda := \Lambda(d)$, only depending on the dimension d, such that $\lambda_{\min}[\boldsymbol{G}(\tau_0, \nu_0)]$ is lower bounded by*

$$\lambda_0 := \lambda_{\min}(\boldsymbol{G}(\tau_0, \nu_0)) \geq \lambda_{\min}(\boldsymbol{G}_2(\tau_0, \nu_0)) \geq \Lambda(d) .$$

Theorem 5. *Assume the PDE (3) has solution $\tau_t \in \mathcal{P}^2$, and the PDE (2) has solution $\nu_t \in \mathcal{C}(\mathcal{P}^2; [0, 1])$. Under Assumption 1, 2, 3, for some constant C_{KL} dependent on d, α, taking $\bar{\beta} := \frac{\beta}{n} > \frac{4\sqrt{C_{\mathrm{KL}}(d,\alpha)}}{\Lambda r_{\max}}$, the following results hold for all $t \in [0, \infty)$:*

$$\mathrm{KL}(\tau_t \| \tau_0) \leq \frac{C_{\mathrm{KL}}(d, \alpha)}{\Lambda^2 \bar{\beta}^2}, \quad \mathrm{KL}(\nu_t \| \nu_0) \leq \frac{C_{\mathrm{KL}}(d, \alpha)}{\Lambda^2 \bar{\beta}^2} .$$

where the radius $r_{\max}$ is defined such that if $\nu \in \mathcal{C}(\mathcal{P}^2; [0, 1]), \tau \in \mathcal{P}^2$, $\max\{\mathcal{W}_2(\nu, \nu_0), \mathcal{W}_2(\tau, \tau_0)\} \leq r_{\max}$, we have $\lambda_{\min}(\boldsymbol{G}_2(\tau, \nu)) \geq \frac{\lambda_0}{2}$.

Generalization:

Theorem 6 (Generalization). *Assume $\tau_y \in \mathcal{C}(\mathcal{P}^2; [0, 1])$ and $\nu_y \in \mathcal{P}^2$ be the ground truth distributions, such that, $y(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\omega} \sim \tau_y} h(\boldsymbol{Z}_{\nu_y}(\boldsymbol{x}, 1), \boldsymbol{\omega})$. Under the Assumption 1, 2 and 3, we set $\beta > \Omega(\sqrt{n})$. For any $\delta > 0$, with probability at least $1 - \delta$, the following bound holds:*

$$\mathbb{E}_{\boldsymbol{x} \sim \mu_X} \ell_{0-1}(f_{\tau_\star, \nu_\star}(\boldsymbol{x}), y(\boldsymbol{x})) \lesssim 1/\sqrt{n} + 6\sqrt{\log(2/\delta)/2n},$$

where $\lesssim$ hides the constant dependence on d, α.

Related Works:

- Lu, Yiping, et al. "Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations." International Conference on Machine Learning. PMLR, 2018.
- Ding, Zhiyan, et al. "On the global convergence of gradient descent for multi-layer resnets in the mean-field regime." arXiv preprint arXiv:2110.02926 (2021).