Introduction of Analyzing Over-parameterized Neural Networks

Nov. 29, 2021

ILLINOIS

Outline

- Background
- Motivations
- **Work #1:** A Convergence Theory for Deep Learning via Over-Parameterization
- **Work #2:** Gradient Descent Finds Global Minima of Deep Neural Networks
- Extension: Convergence of Linearized GNN
- Summary

Natural Questions

1. What are over-parameterized neural networks (NNs)?

2. Why we want to analyze them?

Background (what?)

- Over-parameterized NNs:
  - Neural Networks with **(much) more parameters** than the total number of training samples $\boldsymbol{n}$.
- **Fact**: Lazy training for over-parameterized NNs:
  - Over-parameterized NNs trained with gradient-based methods could converge linearly to **zero training loss**, with their parameters **hardly varying**.

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.
[2] Chizat, Lenaic, and Francis Bach. "A note on lazy training in supervised differentiable programming." , NeurIPS 2018.

Motivations (why?)

- Deep neural networks (DNNs) with non-linear activations have demonstrated a great success. But theoretically analyzing them is non-trivial.

- Many NN architectures are highly over-parameterized in practice (e.g., Wide Residual Networks with 100x parameters).

- After the GD / SGD, bounding the changes of
  - Intermediate variables / outputs
  - Network outputs
  - Gradient flows

  helps to analyze algorithms (e.g., Neural Contextual Bandits).

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.
[2] Du, Simon, et al. "Gradient descent finds global minima of deep neural networks." , ICML 2019.

Comparison of these two concurrent works

- **Comparison**:
  - Different assumption of non-linear activations
    - ReLU vs. c-Lipschitz Smoothness (Sigmoid & Softplus)
  - Different focuses on NN types
  - Different assumption of data points

Will mainly focus on the **first paper (Zeyuan Allen-Zhu et al.)** in this presentation

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.
[2] Du, Simon, et al. "Gradient descent finds global minima of deep neural networks." , ICML 2019.

- Background
- Motivations
- **Work #1:** A Convergence Theory for Deep Learning via Over-Parameterization
- **Work #2:** Gradient Descent Finds Global Minima of Deep Neural Networks
- Extension: Convergence of Linearized GNN
- Summary

Work #1: Initialization and Assumptions

- Data points :
  - With total n data points $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$:

Assumption 2.1. *For every pair $i, j \in [n]$, we have $\|x_i - x_j\| \geq \delta$.*

- Ensure the separateness of the data points
  - Help to bound the network gradients & keep GD efficient

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Initialization and Assumptions

- **FC Network** (L hidden layers + input & output layers):

$$
\begin{aligned}
g_{i,0} &= \mathbf{A}x_i & h_{i,0} &= \phi(\mathbf{A}x_i) & &\text{for } i \in [n] \\
g_{i,\ell} &= \mathbf{W}_\ell h_{i,\ell-1} & h_{i,\ell} &= \phi(\mathbf{W}_\ell h_{i,\ell-1}) & &\text{for } i \in [n], \ell \in [L] \\
y_i &= \mathbf{B}h_{i,L} & & & &\text{for } i \in [n]
\end{aligned}
$$

  - $\boldsymbol{A}, \boldsymbol{W_l}, \boldsymbol{B}$ are trainable parameters. $\phi(\cdot)$ **as the ReLU activation.**

- **Objective:**

$$
F(\overrightarrow{\mathbf{W}}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} F_i(\overrightarrow{\mathbf{W}}) \quad \text{where} \quad F_i(\overrightarrow{\mathbf{W}}) \stackrel{\text{def}}{=} \frac{1}{2}\|\mathbf{B}h_{i,L} - y_i^*\|^2 \quad \text{for each } i \in [n]
$$

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Initialization and Assumptions

- Network Width:

Assumption 2.4. *Throughout this paper we assume* $m \geq \Omega\big(\mathrm{poly}(n, L, \delta^{-1}) \cdot d\big)$ *for some sufficiently large polynomial. To present the simplest proof, we did not try to improve such polynomial factors.*

d := Dim. of output

- Parameter initialization:

Definition 2.3. *We say that* $\overrightarrow{\mathbf{W}} = (\mathbf{W}_1, \ldots, \mathbf{W}_L)$, $\mathbf{A}$ *and* $\mathbf{B}$ *are at random initialization if*

- $[\mathbf{W}_\ell]_{i,j} \sim \mathcal{N}(0, \frac{2}{m})$ *for every* $i, j \in [m]$ *and* $\ell \in [L]$;
- $\mathbf{A}_{i,j} \sim \mathcal{N}(0, \frac{2}{m})$ *for every* $(i, j) \in [m] \times [\mathfrak{d}]$; *and*
- $\mathbf{B}_{i,j} \sim \mathcal{N}(0, \frac{1}{d})$ *for every* $(i, j) \in [d] \times [m]$.

Dim. of data points

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Initialization and Assumptions

- Diagonal sign matrix $\boldsymbol{D_{i,l}}$:
  - **Matrix substitution** for ReLU activation $\phi(\cdot)$
  - For $i \in [n]$ and $l \in [L]$, elements of diagonal matrix $\boldsymbol{D_{i,l}}$:
    - $\left(\boldsymbol{D}_{i,l}\right)_{k,k} = \mathbb{I}[\left(\boldsymbol{W}_l \cdot \boldsymbol{h}_{i,l-1}\right)_k \geq \boldsymbol{0}]$
  - Thus,
    - $\boldsymbol{h}_{i,l} = \phi(\boldsymbol{W}_l \cdot \boldsymbol{h}_{i,l-1}) = \boldsymbol{D_{i,l}} \cdot \boldsymbol{W}_l \boldsymbol{h}_{i,l-1}$

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Initialization and Assumptions

- Bounds after random initialization:

Lemma 7.1 (forward propagation). *If* $\varepsilon \in (0,1]$, *with probability at least* $1 - O(nL) \cdot e^{-\Omega(m\varepsilon^2/L)}$ *over the randomness of* $\mathbf{A} \in \mathbb{R}^{m \times \mathfrak{d}}$ *and* $\overrightarrow{\mathbf{W}} \in (\mathbb{R}^{m \times m})^L$, *we have*

$$\forall i \in [n], \ell \in \{0, 1, \ldots, L\} \quad : \quad \|h_{i,\ell}\| \in [1-\varepsilon, 1+\varepsilon] \ .$$

- Remark:
  - Random matrix initialization + ReLU will give stable intermediate variables.
  - $\|h_{i,\ell}\| \approx 1$ for all $i \in [n]$ and $\ell \in [L]$.

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Initialization and Assumptions

- Bounds after random initialization:

Lemma 7.3 (intermediate layers). *Suppose* $m \geq \Omega(nL\log(nL))$. *With probability at least* $\geq 1 - e^{-\Omega(m/L)}$ *over the randomness of* $\overrightarrow{\mathbf{W}} \in (\mathbb{R}^{m\times m})^L$, *for all* $i \in [n], 1 \leq a \leq b \leq L$,

(a) $\|\mathbf{W}_b\mathbf{D}_{i,b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_{i,a}\mathbf{W}_a\|_2 \leq O(\sqrt{L})$.

(b) $\|\mathbf{W}_b\mathbf{D}_{i,b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_{i,a}\mathbf{W}_a v\| \leq 2\|v\|$ *for all vectors* v *with* $\|v\|_0 \leq O\big(\frac{m}{L\log m}\big)$.

(c) $\|u^\top\mathbf{W}_b\mathbf{D}_{i,b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_{i,a}\mathbf{W}_a\| \leq O(1)\|u\|$ *for all vectors* u *with* $\|u\|_0 \leq O\big(\frac{m}{L\log m}\big)$.

For any integer s *with* $1 \leq s \leq O\big(\frac{m}{L\log m}\big)$, *with probability at least* $1 - e^{-\Omega(s\log m)}$ *over the randomness of* $\overrightarrow{\mathbf{W}} \in (\mathbb{R}^{m\times m})^L$:

(d) $|u^\top\mathbf{W}_b\mathbf{D}_{i,b-1}\mathbf{W}_{b-1}\cdots\mathbf{D}_{i,a}\mathbf{W}_a v| \leq \|u\|\|v\|\cdot O\big(\frac{\sqrt{s\log m}}{\sqrt{m}}\big)$ *for all vectors* u, v *with* $\|u\|_0, \|v\|_0 \leq s$.

a-th layer to b-th layer

- Intermediate layers / outputs are bounded with random initialization.

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Convergence with Gradient Descent

Theorem 1 (gradient descent). *Suppose* $m \geq \widetilde{\Omega}\big(\mathsf{poly}(n, L, \delta^{-1}) \cdot d\big)$. *Starting from random initialization, with probability at least* $1 - e^{-\Omega(\log^2 m)}$, *gradient descent with learning rate* $\eta = \Theta\big(\frac{d\delta}{\mathsf{poly}(n,L)\cdot m}\big)$ *finds a point* $F(\overrightarrow{\mathbf{W}}) \leq \varepsilon$ *in* $T = \Theta\big(\frac{\mathsf{poly}(n,L)}{\delta^2} \cdot \log \varepsilon^{-1}\big)$ *iterations.*

- Remark:
  - **Linear convergence rate** (ϵ drops exponentially fast in T)
  - Result is **independent** from the **dimensionality of input data**
  - Similar result for SGD is also given in the paper
  - Proved with technical **Theorems 3 & 4** (introduced next).

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Technical Theorems

Theorem 3 (no critical point). *With probability* $\geq 1-e^{-\Omega(m/\mathsf{poly}(n,L,\delta^{-1}))}$ *over randomness* $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}, \mathbf{B}$, *it satisfies for every* $\ell \in [L]$, *every* $i \in [n]$, *and every* $\overrightarrow{\mathbf{W}}$ *with* $\|\overrightarrow{\mathbf{W}} - \overrightarrow{\mathbf{W}}^{(0)}\|_2 \leq \frac{1}{\mathsf{poly}(n,L,\delta^{-1})}$,

$$\|\nabla F(\overrightarrow{\mathbf{W}})\|_F^2 \leq O\Big(F(\overrightarrow{\mathbf{W}}) \times \frac{Lnm}{d}\Big) \quad \textit{and} \quad \|\nabla F(\overrightarrow{\mathbf{W}})\|_F^2 \geq \Omega\Big(F(\overrightarrow{\mathbf{W}}) \times \frac{\delta m}{dn^2}\Big) \ .$$

Upper bound

Lower bound

- **Remark:**
  - If the objective is large, the gradient norm is also large.
  - If we are sufficiently close to the random initialization, there is no saddle point or critical point.
  - The hope of finding global minima of the objective $F(\overrightarrow{\boldsymbol{W}})$.

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Technical Theorems

Theorem 4 (semi-smoothness). *With probability at least* $1 - e^{-\Omega(m/\mathrm{poly}(L,\log m))}$ *over the randomness of* $\overrightarrow{\mathbf{W}}^{(0)}, \mathbf{A}, \mathbf{B}$*, we have :*

for every $\breve{\overrightarrow{\mathbf{W}}} \in (\mathbb{R}^{m\times m})^L$ *with*

$$\|\breve{\overrightarrow{\mathbf{W}}} - \overrightarrow{\mathbf{W}}^{(0)}\|_2 \le \frac{1}{\mathrm{poly}(L,\log m)},$$

Parameter close to initialization **after GD**

and for every $\overrightarrow{\mathbf{W}}' \in (\mathbb{R}^{m\times m})^L$ *with*

$$\|\overrightarrow{\mathbf{W}}'\|_2 \le \frac{1}{\mathrm{poly}(L,\log m)},$$

Small Perturbation

the following inequality holds

$$F(\breve{\overrightarrow{\mathbf{W}}} + \overrightarrow{\mathbf{W}}') \le F(\breve{\overrightarrow{\mathbf{W}}}) + \langle \nabla F(\breve{\overrightarrow{\mathbf{W}}}), \overrightarrow{\mathbf{W}}' \rangle + O\Big(\frac{nL^2 m}{d}\Big)\|\overrightarrow{\mathbf{W}}'\|_2^2 + \frac{\mathrm{poly}(L)\sqrt{nm\log m}}{\sqrt{d}} \cdot \|\overrightarrow{\mathbf{W}}'\|_2 \big(F(\breve{\overrightarrow{\mathbf{W}}})\big)^{1/2}$$

Use this semi-smoothness in stead of Lipschitz smoothness for the optimization

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

[2] Yurii Nesterov. “Introductory Lectures on Convex Programming Volume: A Basic course”, volume I. Kluwer Academic Publishers, 2004.

Work #1: Technical Theorems

- Supplement for Theorem 3 & Theorem 4:
  - Authors also proved that: GD / SGD can converge fast enough s.t. the **trained parameter $\vec{W}$** would **stay close** to the randomly initialized $\vec{W}^{(0)}$ by

$$\| \vec{W} - \vec{W}^{(0)} \|_2 \leq \frac{1}{Poly(n, L, \delta^{-1})}$$

  - And this ensures both Theorem 3 & Theorem 4 would apply.

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Technical Theorems

- Proof flow sketch of technical Theorems 3 & 4:
  - Step 1: properties at random initialization.
  - Step 2: stability after adversarial perturbation.
  - Step 3: gradient bound.
  - Step 4: smoothness.

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Step 1: Properties at Random Initialization

- Given arbitrary $a, b \in [L]$, $i \in [n]$:

Back-prop. : $\|\mathbf{B}\mathbf{D}_{i,L}\mathbf{W}_L \cdots \mathbf{D}_{i,a}\mathbf{W}_a\|_2 \leq O(\sqrt{m/d})$

Bounding the backward matrix

Forward-prop. : $\|\mathbf{D}_{i,a}\mathbf{W}_a \cdots \mathbf{D}_{i,b}\mathbf{W}_b\|_2 \leq O(\sqrt{L})$

Bounding the intermediate matrix

- With $\| x_i - x_j \| \geq \delta$:

$$\|h_{i,\ell} - h_{j,\ell}\| \geq \Omega(\delta) \text{ for each layer } \ell \in [L].$$

Separateness of h_l

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Step 2: Stability after Adversarial Perturbation

- "Forward Stability":
  - For $\overrightarrow{\boldsymbol{W}} = [\boldsymbol{W_l}]_{l\in[L]}$, s.t., $\forall l \in [L], \| \boldsymbol{W_l} - \boldsymbol{W_l^{(0)}} \|_2 \leq \omega \leq \frac{1}{poly(L)}$

(a) the number of sign changes $\|\mathbf{D}_{i,\ell} - \mathbf{D}_{i,\ell}^{(0)}\|_0$ is at most $O(m\omega^{2/3}L) \ll m$, and

(b) the perturbation amount $\|h_{i,\ell} - h_{i,\ell}^{(0)}\| \leq O(\omega L^{5/2}) \ll 1$.

- Remark:
  - Proof cannot be derived by induction --- constants will blow up exponentially in the number of layers L

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Step 3: Gradient Bound

- **Upper bound:**
  - With random initialization: $\|h_{i,\ell-1}^{(0)}\| \approx 1$
  - After GD, the spectral norm shift of $\|\mathbf{B}\mathbf{D}_{i,L}\mathbf{W}_L \cdots \mathbf{D}_{i,a}\mathbf{W}_a\|_2$ is at most $O(\omega^{1/3}L^2\sqrt{m/d})$

A part of $\nabla_{W_a} F(\vec{W})$

- **Lower bound:**
  - Bound the contribution to the gradient matrix of each sample $i \in [n]$.
  - Apply the separateness of intermediate outputs (previous slide) to bound the whole gradient matrix.

- Step 1, 2 and 3 help to prove technical Theorem 3.

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Step 4: Smoothness

- Proof sketch:
  - Suppose we are currently at $\vec{\boldsymbol{W}}$ (after GD) from initial $\vec{\boldsymbol{W}}^{(\mathbf{0})}$.
  - Considering the perturbation $\vec{\boldsymbol{W}}'$, we have

$$\| \boldsymbol{h}_{i,l} - \boldsymbol{h}'_{i,l} \| \leq O(L^{1.5}) \| \vec{\boldsymbol{W}}' \|_2$$

$\vec{\boldsymbol{W}}$ $\qquad$ $\vec{\boldsymbol{W}} + \vec{\boldsymbol{W}}'$

    - The change to each hidden layer output is proportional to the amount of perturbation.
  - Along with other properties, we ensure the semi-smoothness.

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

Work #1: Relationship with Neural Kernels

- With any

$$x, \widetilde{x} \in \mathbb{R}^{\mathfrak{d}} \qquad \|\overrightarrow{\mathbf{W}}'\|_2 \le \omega \le \frac{1}{\mathsf{poly}(L, \log m)}$$

- We have:

$$(a)\ \|\nabla y(\overrightarrow{\mathbf{W}}^{(0)} + \overrightarrow{\mathbf{W}}'; x) - \nabla y^{\mathsf{ntk}}(\overrightarrow{\mathbf{W}}'; x)\|_F \le \widetilde{O}\big(\omega^{1/3} L^3\big) \cdot \|\nabla y^{\mathsf{ntk}}(\overrightarrow{\mathbf{W}}'; x)\|_F;$$

$$(b)\ y(\overrightarrow{\mathbf{W}}^{(0)} + \overrightarrow{\mathbf{W}}'; x) = y(\overrightarrow{\mathbf{W}}^{(0)}; x) + y^{\mathsf{ntk}}(\overrightarrow{\mathbf{W}}'; x) \pm \widetilde{O}\big(L^3 \omega^{4/3} \sqrt{m}\big);\ \textit{and}$$

$$(c)\ \big\langle \nabla y(\overrightarrow{\mathbf{W}}^{(0)} + \overrightarrow{\mathbf{W}}'; x), \nabla y(\overrightarrow{\mathbf{W}}^{(0)} + \overrightarrow{\mathbf{W}}'; \widetilde{x}) \big\rangle = K^{\mathsf{ntk}}(x, \widetilde{x}) \pm \widetilde{O}\big(\omega^{1/3} L^3\big) \cdot \sqrt{K^{\mathsf{ntk}}(x, x) K^{\mathsf{ntk}}(\widetilde{x}, \widetilde{x})}\ .$$

- with

$$y(\overrightarrow{\mathbf{W}}; x) \stackrel{\text{def}}{=} y = \mathbf{B} h_L \in \mathbb{R}^d$$

Network Output

$$K^{\mathsf{ntk}}(x, \widetilde{x}) \stackrel{\text{def}}{=} \langle \nabla y(\overrightarrow{\mathbf{W}}^{(0)}; x), \nabla y(\overrightarrow{\mathbf{W}}^{(0)}; \widetilde{x}) \rangle$$

NTK Function

$$y^{\mathsf{ntk}}(\overrightarrow{\mathbf{W}}'; x) \stackrel{\text{def}}{=} \langle \nabla y(\overrightarrow{\mathbf{W}}^{(0)}; x), \overrightarrow{\mathbf{W}}' \rangle = \textstyle\sum_{\ell=1}^{L} \langle \nabla_{\mathbf{W}_\ell} y(\overrightarrow{\mathbf{W}}^{(0)}; x), \mathbf{W}'_\ell \rangle$$

NTK Objective

[1] Allen-Zhu, Zeyuan, Yuanzhi Li, and Zhao Song. "A convergence theory for deep learning via over-parameterization." , ICML 2019.

- Background
- Motivations
- **Work #1:** A Convergence Theory for Deep Learning via Over-Parameterization
- **Work #2:** Gradient Descent Finds Global Minima of Deep Neural Networks
- Extension: Convergence of Linearized GNN
- Summary

Work #2: Assumptions

- c-Lipschitz Smoothness (e.g., Sigmoid & Softplus)

Condition 3.1 (Lipschitz and Smooth). *There exists a constant $c > 0$ such that $|\sigma(0)| \leq c$ and for any $z, z' \in \mathbb{R}$,*

$$|\sigma(z) - \sigma(z')| \leq c|z - z'|,$$
$$\text{and } |\sigma'(z) - \sigma'(z)| \leq c|z - z'|.$$

- $\sigma(\cdot)$ is analytic and is not a polynomial function.

[1] Du, Simon, et al. "Gradient descent finds global minima of deep neural networks." , ICML 2019.

Work #2: Convergence Rate of GD

- Definition of recursively defined Gram matrix (FC Net.):

Definition 5.1. *The Gram matrix* $\mathbf{K}^{(H)}$ *is recursively defined as follows, for* $(i,j) \in [n] \times [n]$, *and* $h = 1, \ldots, H-1$

$$
\begin{aligned}
\mathbf{K}_{ij}^{(0)} =& \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \\
\mathbf{A}_{ij}^{(h)} =& \begin{pmatrix} \mathbf{K}_{ii}^{(h-1)} & \mathbf{K}_{ij}^{(h-1)} \\ \mathbf{K}_{ji}^{(h-1)} & \mathbf{K}_{jj}^{(h-1)} \end{pmatrix}, \qquad (7) \\
\mathbf{K}_{ij}^{(h)} =& c_\sigma \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \mathbf{A}_{ij}^{(h)}\right)} \left[\sigma(u)\sigma(v)\right], \\
\mathbf{K}_{ij}^{(H)} =& c_\sigma \mathbf{K}_{ij}^{(H-1)} \mathbb{E}_{(u,v)^\top \sim N\left(\mathbf{0}, \mathbf{A}_{ij}^{(H-1)}\right)} \left[\sigma'(u)\sigma'(v)\right].
\end{aligned}
$$

- Analogous to the Neural Tangent Kernel.

Comparable definitions (for Gram matrix) are also given for CNN / ResNet

[1] Du, Simon, et al. "Gradient descent finds global minima of deep neural networks." , ICML 2019.

Work #2: Convergence Rate of GD

Theorem 5.1 (Convergence Rate of Gradient Descent for Deep Fully-connected Neural Networks). *Assume for all* $i \in [n]$, $\|\mathbf{x}_i\|_2 = 1$, $|y_i| = O(1)$ *and the number of hidden nodes per layer*

$$m = \Omega\left(2^{O(H)} \max\left\{\frac{n^4}{\lambda_{\min}^4\left(\mathbf{K}^{(H)}\right)}, \frac{n}{\delta}, \frac{n^2 \log(\frac{Hn}{\delta})}{\lambda_{\min}^2\left(\mathbf{K}^{(H)}\right)}\right\}\right)$$

Width requirement

where $\mathbf{K}^{(H)}$ *is defined in Equation (7). If we set the step size*

$$\eta = O\left(\frac{\lambda_{\min}\left(\mathbf{K}^{(H)}\right)}{n^2 2^{O(H)}}\right),$$

LR requirement

then with probability at least $1 - \delta$ *over the random initialization the loss, for* $k = 1, 2, \ldots$, *the loss at each iteration satisfies*

$$L(\theta(k)) \leq \left(1 - \frac{\eta \lambda_{\min}\left(\mathbf{K}^{(H)}\right)}{2}\right)^k L(\theta(0)).$$

Linear convergence rate (similar to work #1)

Comparable Theorems are also given for CNN / ResNet

[1] Du, Simon, et al. "Gradient descent finds global minima of deep neural networks." , ICML 2019.

Extension: Convergence of Linearized GNN

- Definition of linearized GNN:

Definition 1. (Linear GNN). Given data matrix $X \in \mathbb{R}^{m_x \times n}$, aggregation matrix $S \in \mathbb{R}^{n \times n}$, weight matrices $W \in \mathbb{R}^{m_y \times m_H}$, $B_{(l)} \in \mathbb{R}^{m_l \times m_{l-1}}$, and their collection $B = (B_{(1)}, \ldots, B_{(H)})$, a linear GNN with H layers $f(X, W, B) \in \mathbb{R}^{m_y \times n}$ is defined as

$$f(X, W, B) = W X_{(H)}, \quad X_{(l)} = B_{(l)} X_{(l-1)} S. \qquad (2)$$

- Result brief:

  - Similar result is also given for GNN with skip-connections (JK-Net).
  - Theorem of linear convergence rate is given.
  - Training of GNNs are accelerated with:
    - Skip-connections / more depth / a good label distribution (labels that are more correlated to features).

[1] Xu, Keyulu, et al. "Optimization of graph neural networks: Implicit acceleration by skip connections and more depth.", ICML 2021.

Summary

- Over-parameterized NNs:
  - NNs with (much) more parameters than training samples.
- Analyzing different types of over-parameterized NNs:
  - FC-Network / CNN / ResNet
- Components could be bounded after GD / SGD:
  - Intermediate variables
  - Network outputs
  - Gradient flows
- Still a developing topic of DL theory.
  - E.g., GNNs with non-linear activations.

Thanks for listening!
Questions?