9. Approximation: Near-initial approximation

This slide

- A brief excursion to the behaviors of neural nets near its random initialization

- **Motivation.** Overparametrized nets stay near its initialization after training
    - Little movement = better generalization guarantee

This slide

- We want to show that:
    - if a neural net is
        - overparameterized
        - near its initialization

    then it is can be approximated by its linearization at initialization (thus generalize well?)

- See MJT for
    - Full extension to NTK
    - Universal approximation with NTK

Setup

- We consider a bias-free two-layer net

$$f(\mathbf{x};\mathbf{W}) = \sum_{i=1}^{m} a_i \cdot \sigma(\mathbf{w}_i^\top \mathbf{x})$$

  - $\mathbf{x} \in \mathbb{R}^d$
  - $\mathbf{w}_i \in \mathbb{R}^d$
  - $a_i \in \mathbb{R}$

  - $\mathbf{W}^\top = [\mathbf{w}_1 \,|\, \mathbf{w}_2 \,|\, \cdots \,|\, \mathbf{w}_m] \in \mathbb{R}^{d \times m}$

- We study this, under the regime where $m \to \infty$
- **Assumption.** The 2nd layer weights are frozen; we only update $\mathbf{w}_i$

Initialization

- **2nd layer.** Random binary initialization

$$a_i \sim \mathrm{Unif}(\{-1, +1\})$$

- **1st layer.** Random Gaussian initialization

$$\mathbf{w}_i \sim \mathcal{N}(0, I_d)$$

- <u>Note.</u> Should be scaled by the factors $1/\sqrt{m}$ and $1/\sqrt{d}$
    - But we skip for now, for simple notations

Taylor approximation

- We are interested in the following approximation

$$f_0(\mathbf{x}; \mathbf{W}) := f(\mathbf{x}; \mathbf{W}_0) \; + \; \langle \partial_{\mathbf{W}} f(\mathbf{x}; \mathbf{W}_0), \mathbf{W} - \mathbf{W}_0 \rangle$$

- This is a classic 1st order Taylor approximation
    - The differential $\partial_{\mathbf{W}}$ is called the Clarke subdifferential
        - Roughly, the set of all gradient candidates for non-differentiable functions
        - By default, we select the minimum-norm gradient

Taylor approximation

- More tediously, we can write the approximation as:

$$
\begin{aligned}
f_0(\mathbf{x}; \mathbf{W}) &= \sum_{i=1}^{m} a_i \sigma(\mathbf{w}_{0,i}^\top \mathbf{x}) \; + \; \sum_{i=1}^{m} a_i \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x}) \mathbf{x}^\top (\mathbf{w}_i - \mathbf{w}_{0,i}) \\
&= \sum_{i=1}^{m} a_i \cdot \left(\sigma(\mathbf{w}_{0,i}^\top \mathbf{x}) - \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x}) \mathbf{w}_{0,i}^\top \mathbf{x} + \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x}) \mathbf{w}_i^\top \mathbf{x} \right)
\end{aligned}
$$

- This is an affine approximation of $f(\mathbf{x}; \mathbf{W})$
  - Affine with respect to $\mathbf{W}$
  - Nonlinear with respect to $\mathbf{x}$

Nets near init are almost linear

Claim

- **Roughly.** Whenever $\mathbf{W} \approx \mathbf{W}_0$, then we have

$$f(\,\cdot\,;\mathbf{W}) \approx f_0(\,\cdot\,;\mathbf{W})$$

  - Smooth activation: easy
  - ReLU: difficult

Claim

- Slightly more concretely, we want results like:

Claim (informal)

With a high probability, we have

$$f(\mathbf{x}; \mathbf{W}) - f_0(\mathbf{x}; \mathbf{W}) \leq \frac{C \cdot \|\mathbf{W} - \mathbf{W}_0\|^{(\text{pow.})}}{m^{(\text{pow.})}}$$

- Tricky part is that $\|\mathbf{W} - \mathbf{W}_0\|$ may have some dependencies on m
    - If it is a Frobenius norm…

Nets near initialization

Proposition 4.1.

Suppose that $\|\mathbf{x}\|_2 \leq 1$, and let $\sigma : \mathbb{R} \to \mathbb{R}$ be a β-smooth function.   (i.e., gradient is β-Lipschitz)

Then, for any parameters $\mathbf{W}, \mathbf{W}_0$, we have

$$f(\mathbf{x}; \mathbf{W}) - f_0(\mathbf{x}; \mathbf{W}) \leq \frac{\beta}{2} \|\mathbf{W} - \mathbf{W}_0\|_F^2$$

- If we revive the 2nd layer's scaling factors $1/\sqrt{m}$, we get the desired property.

Proof idea

$$f(\mathbf{x}; \mathbf{W}) - f_0(\mathbf{x}; \mathbf{W}) \leq \frac{\beta}{2} \|\mathbf{W} - \mathbf{W}_0\|_F^2$$

- Proceed in two steps:
  - **Step 1.** Show that, for β-smooth function, we have:

$$|\sigma(x) - \sigma(x_0) - \sigma'(x_0)(x - x_0)| \leq \frac{\beta(x - x_0)^2}{2}$$

    - Any volunteer? 🙋‍♂️

Proof idea

$$f(\mathbf{x}; \mathbf{W}) - f_0(\mathbf{x}; \mathbf{W}) \leq \frac{\beta}{2} \|\mathbf{W} - \mathbf{W}_0\|_F^2$$

- Proceed in two steps:
  - **Step 1.** Show that, for β-smooth function, we have:

$$|\sigma(x) - \sigma(x_0) - \sigma'(x_0)(x - x_0)| \leq \frac{\beta(x - x_0)^2}{2}$$

    - Any volunteer? 🙋

    - Taylor's theorem.

$$f(x) = f(a) + f'(a)(x - a) + \int_a^x f''(t) \frac{(x - t)^2}{2} \mathrm{d}t$$

Proof idea

$$f(\mathbf{x};\mathbf{W}) - f_0(\mathbf{x};\mathbf{W}) \leq \frac{\beta}{2}\|\mathbf{W}-\mathbf{W}_0\|_F^2$$

$$|\sigma(x) - \sigma(x_0) - \sigma'(x_0)(x-x_0)| \leq \frac{\beta(x-x_0)^2}{2}$$

- **Step 2.** Use the step 1 result, to examine the LHS
  - Recall that we had:

$$f_0(\mathbf{x};\mathbf{W}) = \sum_{i=1}^{m} a_i \cdot \left(\sigma(\mathbf{w}_{0,i}^\top \mathbf{x}) - \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x})\mathbf{w}_{0,i}^\top \mathbf{x} + \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x})\mathbf{w}_i^\top \mathbf{x}\right)$$

  - Also recall that we had:

$$\|\mathbf{x}\|_2 \leq 1$$

Extension to ReLU

- For ReLU, things are not that easy…

- **Tool.** Thankfully, we know that, for ReLU:

$$\sigma(x) = x \cdot \sigma'(x)$$

  - Thus, we also have:

$$\begin{aligned} f_0(\mathbf{x}; \mathbf{W}) &= \sum_{i=1}^{m} a_i \cdot \left(\sigma(\mathbf{w}_{0,i}^\top \mathbf{x}) - \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x}) \mathbf{w}_{0,i}^\top \mathbf{x} + \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x}) \mathbf{w}_i^\top \mathbf{x} \right) \\ &= \sum_{i=1}^{m} a_i \cdot \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x}) \mathbf{w}_i^\top \mathbf{x} \end{aligned}$$

Extension to ReLU

- Thus, we also have:

$$
\begin{aligned}
f(\mathbf{x};\mathbf{W}) - f_0(\mathbf{x};\mathbf{W}) &= \sum_{i=1}^{m} a_i \cdot \left(\sigma(\mathbf{w}_i^\top \mathbf{x}) - \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x})\mathbf{w}_i^\top \mathbf{x}\right) \\
&= \sum_{i=1}^{m} a_i \cdot \mathbf{w}_i^\top \mathbf{x}\left(\sigma'(\mathbf{w}_i^\top \mathbf{x}) - \sigma'(\mathbf{w}_{0,i}^\top \mathbf{x})\right) \\
&= \sum_{i=1}^{m} a_i \cdot \mathbf{w}_i^\top \mathbf{x}\left(\mathbf{1}\{\mathbf{w}_i^\top \mathbf{x} \geq 0\} - \mathbf{1}\{\mathbf{w}_{0,i}^\top \mathbf{x} \geq 0\}\right) \qquad \cdots (⭐)
\end{aligned}
$$

- Question. How do we bound this ⭐?

Extension to ReLU

$$\sum_{i=1}^{m} a_i \cdot \mathbf{w}_i^\top \mathbf{x} \Big(\mathbf{1}\{\mathbf{w}_i^\top \mathbf{x} \geq 0\} - \mathbf{1}\{\mathbf{w}_{0,i}^\top \mathbf{x} \geq 0\} \Big) \qquad \cdots (⭐)$$

- Naïve. Maybe use something like Cauchy-Schwarz
  - Will get something like

$$\leq \sqrt{m} \|\mathbf{W}\|_F$$

  - Non-diminishing as $m \to \infty$, even after multiplying $1/\sqrt{m}$

- Intuition. Exploit the randomness of the matrix $\mathbf{W}_0$

Concentration inequality

- The key intuition is formalized in the following lemma.

Lemma 4.2.

Let $\mathbf{u}_i \sim \mathcal{N}(0, I_d)$. Then, for any $\tau > 0$ and $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\| > 0$, we have:

$$\sum_{i=1}^{m} \mathbf{1}\left\{ |\mathbf{u}_i^\top \mathbf{x}| \le \tau \|\mathbf{x}\| \right\} \le m\tau + \sqrt{m \log(1/\delta)}, \qquad \text{with probability at least } 1 - \delta$$

- Any useful intuitions?

Proof sketch

$$\sum_{i=1}^{m} \mathbf{1}\left\{ |\mathbf{u}_i^\top \mathbf{x}| \le \tau \|\mathbf{x}\| \right\} \le m\tau + \sqrt{m \log(1/\delta)}, \qquad \text{with probability at least } 1-\delta$$

- Define $P_i = \mathbf{1}\left\{ |\mathbf{u}_i^\top \mathbf{x}| \le \tau \|\mathbf{x}\| \right\}$.
- Then, proceed in three steps:
  - **Step 1.** By rotational invariance, we have

$$P_i = \mathbf{1}\{ |\mathbf{u}_{i,1}| \le \tau \}$$

Proof sketch

$$\sum_{i=1}^{m} \mathbf{1}\left\{ |\mathbf{u}_i^\top \mathbf{x}| \le \tau \|\mathbf{x}\| \right\} \le m\tau + \sqrt{m \log(1/\delta)}, \qquad \text{with probability at least } 1-\delta$$

- Define $P_i = \mathbf{1}\left\{ |\mathbf{u}_i^\top \mathbf{x}| \le \tau \|\mathbf{x}\| \right\}$.
- Then, proceed in three steps:
  - **Step 1.** By rotational invariance, we have

$$P_i = \mathbf{1}\{ |\mathbf{u}_{i,1}| \le \tau \}$$

  - **Step 2.** Inspecting the Gaussian density, we have:

$$\mathbf{Pr}[P_i = 1] = \int_{-\tau}^{+\tau} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} \, \mathrm{d}z \le \frac{2\tau}{\sqrt{2\pi}} \le \tau$$

Proof sketch

$$\sum_{i=1}^{m} \mathbf{1}\left\{ |\mathbf{u}_i^\top \mathbf{x}| \leq \tau \|\mathbf{x}\| \right\} \leq m\tau + \sqrt{m \log(1/\delta)}, \qquad \text{with probability at least } 1 - \delta$$

- Define $P_i = \mathbf{1}\left\{ |\mathbf{u}_i^\top \mathbf{x}| \leq \tau \|\mathbf{x}\| \right\}$.
- Then, proceed in three steps:
  - **Step 1.** By rotational invariance, we have

$$P_i = \mathbf{1}\{ |\mathbf{u}_{i,1}| \leq \tau \}$$

  - **Step 2.** Inspecting the Gaussian density, we have:

$$\mathbf{Pr}[P_i = 1] = \int_{-\tau}^{+\tau} \frac{\exp(-z^2/2)}{\sqrt{2\pi}} \, \mathrm{d}z \leq \frac{2\tau}{\sqrt{2\pi}} \leq \tau$$

  - **Step 3.** Apply Hoeffding's inequality to get the claim

The result

- Given the previous lemma, we are ready to prove today's main result

Lemma 4.1.

For any radius $B \geq 0$, any fixed $\mathbf{x} \in \mathbb{R}^d$ with $\|\mathbf{x}\| \leq 1$, for any $\mathbf{W} \in \mathbb{R}^{m \times d}$ with $\|\mathbf{W} - \mathbf{W}_0\|_F \leq B$, we have:

$$\left| f(\mathbf{x}; \mathbf{W}) - f_0(\mathbf{x}; \mathbf{W}) \right| \leq m^{\frac{1}{3}} \left(\sqrt{2} B^{\frac{4}{3}} + B \left(\log(1/\delta) \right)^{1/4} \right), \qquad \text{with probability at least } 1 - \delta$$

- Rough intuitions: Combine two claims
  - With high probability, $\|\mathbf{w}_{0,i}^\top \mathbf{x}\|$ won't be small
    - Reason: Gaussian initialization $\mathbf{W}_0$ concentrates around its "shell
  - If $\|\mathbf{W} - \mathbf{W}_0\|_F$ is small, then $\|\mathbf{w} - \mathbf{w}_{0,i}\|$ will be small for many i

  - Putting these together, we know that $\mathbf{w}_i^\top \mathbf{x}$ and $\mathbf{w}_{0,i}^\top \mathbf{x}$ have same signs quite often!

Proof sketch

- Concretely, for each index $i \in [m]$, define the subset of indices:

$$S_1 = \left\{ i \in [m] \;\middle|\; |\mathbf{w}_{0,i}^\top \mathbf{x}| \leq \tau \|\mathbf{x}\| \right\}$$

$$S_2 = \left\{ i \in [m] \;\middle|\; \|\mathbf{w}_i - \mathbf{w}_{0,i}\| \geq \tau \right\}$$

- **Claim.** These are the only bad cases — i.e., $\mathbf{w}_i^\top \mathbf{x}$ and $\mathbf{w}_{0,i}^\top \mathbf{x}$ have different signs

Proof sketch

- Concretely, for each index $i \in [m]$, define the subset of indices:

$$S_1 = \left\{ i \in [m] \;\middle|\; |\mathbf{w}_{0,i}^\top \mathbf{x}| \le \tau \|\mathbf{x}\| \right\}$$

$$S_2 = \left\{ i \in [m] \;\middle|\; \|\mathbf{w}_i - \mathbf{w}_{0,i}\| \ge \tau \right\}$$

- **Claim.** These are the only bad cases — i.e., $\mathbf{w}_i^\top \mathbf{x}$ and $\mathbf{w}_{0,i}^\top \mathbf{x}$ have different signs
- Suppose that we have $i \notin S_1 \cup S_2$.
- Suppose that we have $\mathbf{w}_i^\top \mathbf{x} > 0$.
    - As $i \notin S_1$, we know that $\mathbf{w}_{0,i}^\top \mathbf{x}$ is either $> \tau\|\mathbf{x}\|$ or $< -\tau\|\mathbf{x}\|$
    - However, we cannot have $< -\tau\|\mathbf{x}\|$, as

$$\mathbf{w}_{0,i}^\top \mathbf{x} = \mathbf{w}_i^\top \mathbf{x} - (\mathbf{w}_i^\top - \mathbf{w}_{0,i}^\top)\mathbf{x} > 0 - \tau\|\mathbf{x}\|$$

    - Thus, in this case, we have $\mathbf{w}_{0,i}^\top \mathbf{x} > \tau\|\mathbf{x}\|$, meaning that they have a same sign

Proof sketch

$$S_1 = \left\{ i \in [m] \,\middle|\, |\mathbf{w}_{0,i}^\top \mathbf{x}| \le \tau \|\mathbf{x}\| \right\} \qquad S_2 = \left\{ i \in [m] \,\middle|\, \|\mathbf{w}_i - \mathbf{w}_{0,i}\| \ge \tau \right\}$$

- Now, let's control the size of $S_1 \cup S_2$
  - By the union bound, we have

$$|S| := |S_1 \cup S_2| \le |S_1| + |S_2|$$

    - $|S_1|$: By Lemma 4.2, we know that

$$|S_1| \le m\tau + \sqrt{m \log(1/\delta)}, \qquad \text{w.p. at least } 1 - \delta$$

    - $|S_2|$: Notice that

$$B^2 \ge \|\mathbf{W} - \mathbf{W}_0\|_F^2 \ge \sum \mathbf{1}\{i \in S_2\} \cdot \|\mathbf{w}_i - \mathbf{w}_{0,i}\|^2 \ge |S_2| \cdot \tau^2$$

      - Thus, we have $|S_2| \le B^2/\tau^2$

Proof sketch

- Combine these two bounds and optimize the sum w.r.t. τ, to get:

$$|S| \leq 2m^{2/3}B^{2/3} + \sqrt{m\log(1/\delta)} \leq m^{2/3}\left(2B^{2/3} + \sqrt{\log(1/\delta)}\right) \quad \text{w.p. } 1-\delta$$

- Plus this into ⭐ and finish the proof

Wrapping up

- **Takeaway.** Wide width = More linearizable
    - If we take an infinite-width limit, perhaps NNs behave just like f_0?
        - Motivates NTK

        - NTK will be covered, if we have some time…