Neural Tangent Kernel (NTK) Analysis

We will mainly consider feedforward neural neural networks.

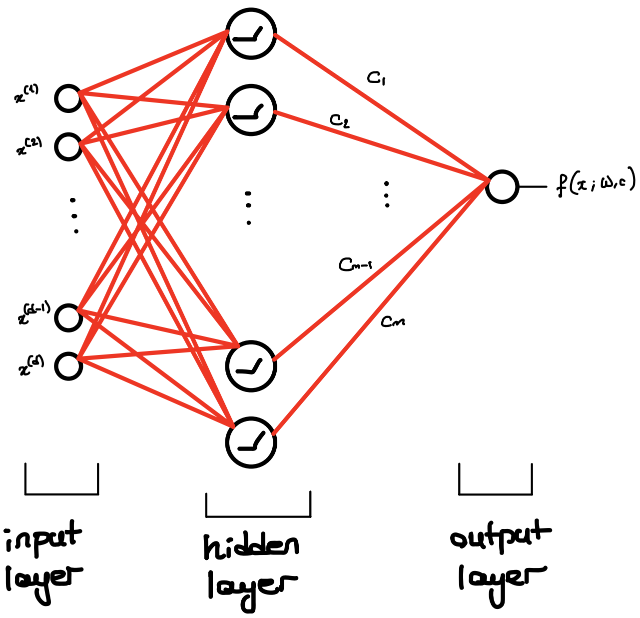
$$f(x; \omega, c) = \sum_{i=1}^{m} c_i \, \sigma(\omega_i^T x) \quad \text{where} \quad \omega_i \in \mathbb{R}^d,\ c_i \in \mathbb{R} \quad \text{for all } i \in \{1, 2, \ldots, m\}.$$

neuron or hidden unit

m is the <u>width</u> of the neural network.

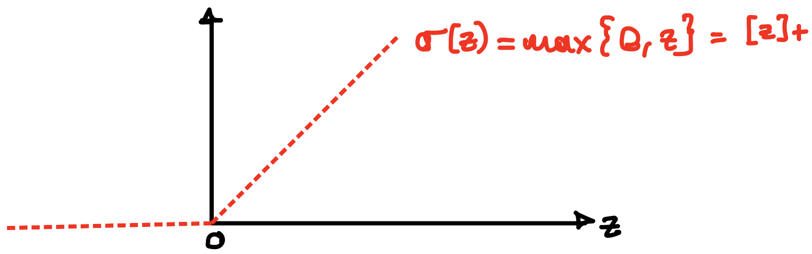
$$\omega^T = [\omega_1^T \ \ldots \ \omega_m^T] \in \mathbb{R}^{md}$$

$$c = [c_1 \ \ldots \ c_m] \in \mathbb{R}^m$$



$\sigma : \mathbb{R} \to \mathbb{R}$ is the activation function.

Mainly used: $\sigma(z) = \max\{0, z\}$ (<u>Rectified</u> <u>Linear</u> <u>Unit</u>, ReLU)



Recall $\mathbb{1}\{z \geq 0\} := \begin{cases} 1, & \text{if } z \geq 0 \\ 0, & \text{otherwise.} \end{cases}$

Then we also have $\sigma(z) = z \, \mathbb{1}\{z \geq 0\}$.

<u>Lemma 1</u> (Some elementary properties of ReLU networks)

1. Given $x \in \mathbb{R}^d$, $x \, \mathbb{1}\{\theta^T x \geq 0\} \in \partial_\theta \sigma(\theta^T x)$ where $\sigma(z) = [z]_+$. (∂_θ is subdifferential)

2. Define $\nabla_{\omega_i} f(x; \omega, c) := c_i \, x \, \mathbb{1}\{\omega_i^T x \geq 0\}$ for $i = 1, 2, \ldots, m$,

$$\nabla_\omega^T f(x; \omega, c) := [\nabla_{\omega_1}^T f(x; \omega, c) \ \cdots \ \nabla_{\omega_m}^T f(x; \omega, c)].$$

Then

$$\nabla_\omega^T f(x; \omega, c)\, \omega = f(x; \omega, c), \quad \forall x \in \mathbb{R}^d,\ (\omega, c) \in \mathbb{R}^{md} \times \mathbb{R}$$

Pf: ① $\sigma(\theta^T x) + x^T \mathbb{1}\{\theta^T x \geq 0\}[\theta' - \theta]$

$$= \sigma(\theta^T x) - \sigma(\theta^T x) + x^T\theta' \mathbb{1}\{\theta^T x \geq 0\}$$

$$= x^T \theta' \mathbb{1}\{\theta^T x \geq 0\}$$

$$\leq x^T\theta' \mathbb{1}\{x^T\theta' \geq 0\} = \sigma(x^T\theta').$$

② $\nabla_\omega^T f(x;\omega,c)\,\omega = \sum_{i=1}^m \nabla_{\omega_i}^T f(x;\omega,c)\,\omega_i$

$$= \sum_{i=1}^m c_i\, \omega_i^T x\, \mathbb{1}\{\omega_i^T x \geq 0\} = \sum_{i=1}^m c_i\, \sigma(\omega_i^T x).$$

Note: If $c_i \geq 0, \forall i$, then f is convex and

$$\nabla_{\omega_i} f(x;\omega,c) \in \partial_{\omega_i} f(x;\omega,c).$$

However, c can take on any value. Thus, we define $\nabla_{\omega_i} f(x;\omega,c)$ as it may <u>not</u> be a subgradient due to negative c_i.

PART I: LINEARIZATION OF OVERPARAMETERIZED NETWORKS NEAR INITIALIZATION

<u>DEF1</u> (Symmetric Xavier initialization)

Suppose that $m \in \mathbb{Z}_+$ is even. $\mathbb{R}^{md} \times \mathbb{R}^m$-valued random variable $(\omega(0), c)$ is called a (symmetric) Xavier initialization if

$$\omega_i(0) = \omega_{i+\frac{m}{2}}(0) \overset{iid}{\sim} \mathcal{N}(0, I_d)$$

$$c_i = -c_{i+\frac{m}{2}} \overset{iid}{\sim} \text{Rademacher},$$

for $i = 1, 2, \dots, m$.

We consider $F(x;\omega,c) := \frac{1}{\sqrt{m}} \sum_{i=1}^m c_i\, \sigma(\omega_i^T x).$

Note that the scale $\frac{1}{\sqrt{m}}$ will be important.

<u>Lemma 2</u> Let $(\omega(0), c)$ be a symmetric Xavier initialization. Then, show that $F(x;\omega(0),c) = 0$ w.p. 1 for any $x \in \mathbb{R}^d$.

LEMMA 3 (Taylor-like expansion of ReLU networks)

Given $x \in \mathbb{R}^d$, let $(w(0), c)$ be a symmetric Xavier initialization and $w \in \mathbb{R}^{md}$ be an arbitrary weight vector. Then, w.p. 1,

$$F(x; w, c) = \overbrace{\nabla_w^T F(x; w(0), c)\Big(w - w(0)\Big)}^{\text{linear in } w} + \overbrace{\Big(\nabla_w F(x; w, c) - \nabla_w F(x; w(0))\Big)^T w}^{\text{nonlinearity}}$$

Pf: By Lemma 1.(2), $\nabla_w^T F(x; w(0), c)\, w(0) = F(x; w(0), c) = 0.$

Then, $\nabla_w^T F(x; w(0), c)\, w + \Big(\nabla_w F(x; w, c) - \nabla_w F(x; w(0), c)\Big)^T w$

$$= \nabla_w^T F(x; w, c)\, w = F(x; w, c). \quad \blacksquare$$

Remark: The idea will be to control the nonlinearity by large width m (i.e., overparameterization) when $\|w - w(0)\|_2$ is small, i.e., lazy training or near-initialization or kernel regime.

We will see that $x \mapsto \nabla_w^T F(x; w(0), c)\,[w - w(0)]$ will be a powerful linear model (linear in w, nonlinear x).

The main result of this discussion: if $\|w_i - w_i(0)\|_2 \le \frac{\alpha}{\sqrt{m}}$ for all i, then $\left|F(x; w, c) - \nabla^T F(x; w(0), c)[w - w(0)]\right| = O\left(\frac{1}{\sqrt{m}}\right)$ with high prob.

THEOREM 1 (ReLU networks are almost linear near initialization)

Let $(w(0), c)$ be a symmetric random initialization, and $w \in \mathbb{R}^{md}$ be any vector s.t.

$$\max_{1 \le i \le m} \|w_i - w_i(0)\|_2 \le \frac{\alpha}{\sqrt{m}},$$

for some $\alpha > 0$. Then, for any $x \in \mathbb{R}^d$, $\delta \in (0,1)$,

$$\left|F(x; w, c) - \nabla_w^T F(x; w(0), c)\,\big[w - w(0)\big]\right| \le \frac{\alpha}{\sqrt{m}}\big(1 + \|x\|_2\big)\left[\alpha\|x\|_2 + \sqrt{\log(1/\delta)}\right],$$

with probability at least $1-\delta$ over the random initialization.

Proof: By using Lemma 3, we can write:

$$F(x;\omega,c) = F(x;\omega(0),c) + \nabla_\omega^T F(x;\omega(0),c)\left[\omega-\omega(0)\right] + \left[\nabla_\omega F(x;\omega,c) - \nabla F(x;\omega(0),c)\right]^T \omega$$

By Lemma 2, we have $F(x;\omega(0),c) = 0$. Then, we can decompose $F(x;\omega,c)$ into linear and nonlinear parts (in ω):

$$F(x;\omega,c) = \nabla_\omega^T F(x;\omega(0),c)\left[\omega-\omega(0)\right] + \Delta(\omega,m)$$

where

$$\Delta(\omega,c) := \left[\nabla_\omega F(x;\omega,c) - \nabla_\omega F(x;\omega(0),c)\right]^T \omega,$$

is the nonlinear term.

Lazy training or kernel regime:

$$\omega_i = \omega_i(0) + u_i \quad , \quad \|u_i\|_2 \le \frac{\alpha}{\sqrt{m}} \quad \text{for all } i = 1,2,\dots,m.$$

$$\Delta(\omega,m) \triangleq \left[\nabla_\omega F(x;\omega,c) - \nabla_\omega F(x;\omega(0),c)\right]^T \omega$$

$$= \Delta_1(\omega,m) + \Delta_2(\omega,m)$$

where

$$\Delta_1(\omega,m) := \left[\nabla_\omega F(x;\omega,c) - \nabla_\omega F(x;\omega(0),c)\right]^T \omega(0)$$

$$\Delta_2(\omega,m) := \left[\nabla_\omega F(x;\omega,c) - \nabla_\omega F(x;\omega(0),c)\right]^T u$$

Note that, by definition,

$$\left[\nabla_\omega F(x;\omega,c) - \nabla_\omega F(x;\omega(0),c)\right]^T V = \frac{1}{\sqrt{m}} \sum_i c_i \left[\mathbb{1}\{\omega_i^T x \ge 0\} - \mathbb{1}\{\omega_i^T(0) x \ge 0\}\right] V_i^T x$$

for any $V = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_m \end{bmatrix} \in \mathbb{R}^{md}$.

We will need Lemma 4 to conclude the proof of Theorem 1.

Lemma 4 Let $w \in \mathbb{R}^{md}$ be s.t.

$$\max_{1 \le i \le m} \|w_i - w_i(0)\|_2 \le \frac{\alpha}{\sqrt{m}} \quad \text{for some } \alpha > 0. \text{ Then,}$$

for any $x \in \mathbb{R}^d$ and $\delta \in (0,1)$, $\exists A_\delta(x) \in \sigma(w(0), c)$ s.t.

$$\mathbb{P}(A_\delta(x)) \ge 1 - \delta, \text{ and}$$

$$|\Delta_1(w,m)| \le \frac{\alpha}{\sqrt{m}} \cdot \left[\alpha \cdot \|x\|_2 + \sqrt{\log(1/\delta)}\right] \quad \text{and}$$

$$|\Delta_2(w,m)| \le \frac{\alpha \|x\|_2}{\sqrt{m}} \left[\alpha \cdot \|x\|_2 + \sqrt{\log(1/\delta)}\right] \quad \text{in } A_\delta(x).$$

Proof of Lemma 4

Then,

$$|\Delta_1(w,m)| = \left| \frac{1}{m} \sum_{i=1}^m c_i \left[\mathbb{1}\{w_i^T x \ge 0\} - \mathbb{1}\{w_i^T(0) x \ge 0\}\right] w_i^T(0) x \right|$$

$$\le \frac{1}{\sqrt{m}} \sum_{i=1}^m \left|\mathbb{1}\{w_i^T x \ge 0\} - \mathbb{1}\{w_i^T(0) x \ge 0\}\right| \cdot |w_i^T(0) x|$$

Now, we start counting sign changes from $w_i^T(0)x$ to $w_i^T x$.

Let
$$S(x) := \{ i \in [m] : \mathbb{1}\{w_i^T(0) x \ge 0\} \ne \mathbb{1}\{w_i^T x \ge 0\} \}.$$

$$\left|\mathbb{1}\{w_i^T x \ge 0\} - \mathbb{1}\{w_i^T(0) x \ge 0\}\right| = \mathbb{1}\{i \in S(x)\}.$$

Also,

$$i \in S(x) \Rightarrow |w_i^T(0) x| \le |w_i^T(0) x - w_i^T x|$$
$$\le \|w_i(0) - w_i\|_2 \cdot \|x\|_2$$
$$= \|u_i\|_2 \cdot \|x\|_2 \le \frac{\alpha \|x\|_2}{\sqrt{m}}.$$

Then,

$$|\Delta_1(w,m)| \le \frac{1}{\sqrt{m}} \sum_i \mathbb{1}\{i \in S(x)\} \cdot |w_i^T(0) x|$$

$$\le \frac{\alpha}{m} \sum_{i=1}^m \mathbb{1}\{i \in S(x)\} \le \alpha \cdot \frac{1}{m} \sum_{i=1}^m \mathbb{1}\left\{ |w_i^T(0) x| \le \frac{\alpha \|x\|_2}{\sqrt{m}} \right\}.$$

Lemma 5 (Gaussian Anti-Concentration)

Let $u \sim \mathcal{N}(0, I_d)$, $x \neq 0$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left(|u^T x| \leq \|x\| \cdot \varepsilon \right) \leq \sqrt{\frac{2}{\pi}} \cdot \varepsilon.$$

Proof: For any $x \neq 0$, $\frac{u^T x}{\|x\|_2} \sim \mathcal{N}(0,1)$ since

$$\mathbb{E}\left[\frac{u^T x}{\|x\|_2} \right] = \left(\mathbb{E}[u]\right)^T \frac{x}{\|x\|_2} = 0 \text{ , and}$$

$$\mathrm{Var}\left(u^T \frac{x}{\|x\|_2}\right) = \mathbb{E}\left[\frac{x^T u}{\|x\|_2} \cdot \frac{u^T x}{\|x\|_2} \right] = \frac{x^T \mathbb{E}[u u^T] x}{\|x\|_2^2} = 1.$$

Then,

$$\mathbb{P}\left(|u^T x| < \|x\| . \varepsilon \right) = \mathbb{P}\left(-\varepsilon \leq \frac{u^T x}{\|x\|_2} \leq \varepsilon\right)$$

$$= \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}} \underbrace{e^{-\frac{y^2}{2}}}_{\leq 1} dy \leq \int_{-\varepsilon}^{\varepsilon} \frac{1}{\sqrt{2\pi}} dy$$

$$\leq \frac{\sqrt{2}\varepsilon}{\sqrt{\pi}}. \quad \blacksquare$$

Also,

$$\frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left\{ |\omega_i^T(0) x| \leq \frac{\alpha}{\sqrt{m}} \cdot \|x\|_2 \right\} - \mathbb{P}\left(|u^T x| \leq \frac{\alpha}{\sqrt{m}} \|x\|_2 \right) \leq \sqrt{\frac{\log(1/\delta)}{m}}$$

with probability at least $1-\delta$.

Thus, let

$$A_1(x) := \left\{ \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left\{ |\omega_i^T(0) x| \leq \frac{\alpha \|x\|_2}{\sqrt{m}} \right\} \leq \frac{1}{\sqrt{m}} \left[\alpha \|x\|_2 + \sqrt{\log(1/\delta)} \right] \right\}.$$

Then, $\mathbb{P}(A_1(x)) \geq 1-\delta$, and

$$|\Delta_1(\omega, x)| \leq \frac{\alpha}{\sqrt{m}} \left[\alpha . \|x\|_2 + \sqrt{\log(1/\delta)} \right] \text{ in } A_1(x).$$

$$\Delta_2(\omega, x) = \left[\nabla F(x;\omega,c) - \nabla F(x;\omega(0),c)\right]^T u$$

$$= \frac{1}{\sqrt{m}} \sum_{i=1}^{m} c_i \left[\mathbb{1}\{\omega_i^T x \geq 0\} - \mathbb{1}\{\omega_i^T(0)x \geq 0\}\right] u_i^T x$$

Recall that $\max_i \|u_i\|_2 \leq \frac{\alpha}{\sqrt{m}}$. Then,

$$|\Delta_2(\omega, x)| \leq \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \left|\mathbb{1}\{\omega_i^T x \geq 0\} - \mathbb{1}\{\omega_i^T(0)x \geq 0\}\right| \cdot |\underbrace{u_i^T x}|$$

Cauchy-Schwarz $\leq \|u_i\|_2 \cdot \|x\|_2 \leq \frac{\alpha}{\sqrt{m}} \cdot \|x\|_2$

$$\leq \frac{\alpha\|x\|_2}{m} \sum_i \mathbb{1}\{i \in S(x)\}$$

$$\leq \frac{\beta}{\sqrt{m}} \left[\alpha \cdot \|x\|_2 + \sqrt{\log(1/\delta)}\right] \quad \text{in } A_1(x).$$

Thus, we conclude the proof of Lemma 4. ■

Coming back to the proof of Theorem 1,

$$\left|F(x;\omega,c) - \nabla^T F(x;\omega(0),c)[\omega - \omega(0)]\right| = |\Delta(\omega, m)|$$

$$= |\Delta_1(\omega, m) + \Delta_2(\omega, m)|$$

$$\leq |\Delta_1(\omega, m)| + |\Delta_2(\omega, m)|.$$

Substituting the bounds on $|\Delta_1|$ and $|\Delta_2|$ that we found in Lemma 4, we conclude the proof of Theorem 1. ■