CS 839: Theoretical Foundations of Deep Learning		Spring 2022
	Lecture 12 Neural Tangent Kernel II	
Instructor: Yingyu Liang	*Date: Mar 3^{rd}, 2022*	*Scriber: Jitian Zhao*

1 NTK on Two-layer Neural Networks with ReLU

Consider regression setting with dataset $(x_i, y_i)_{i=1}^n, x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$ and $\|x_i\| = 1, |y_i| \leq 1$. The squared loss is defined to be:

$$L(w) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i; w))^2 \tag{1}$$

Define prediction vector $u = [f(x_1; w), ..., f(x_n; w)]^\top \in \mathbb{R}^n$ and for gradient flow, we assume chain rule holds here:

$$\frac{dw(t)}{dt} = -\nabla L(w) \tag{2}$$

Consider two-layer neural networks with ReLU activation

$$f(x; w) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m} a_i \sigma(\langle w_i, x\rangle),$$

where $\sigma(z) = \max\{0, z\}$. Initialize the weights by $a_i(0) \sim \text{uniform}\{-1, 1\}$ and $w_i(0) \sim N(0, I_d)$, and the training updates only w_i's.

For weights w, let

$$H_{ij} = \left\langle \frac{\partial f(x_i; w)}{\partial w}, \frac{\partial f(x_j; w)}{\partial w} \right\rangle.$$

Let $H(t)$ be a shorthand for $H(w(t))$ and let $H^* = \mathbb{E}_{w(0)}[H(0)]$.

Theorem 1. Assume $\lambda_0 = \lambda_{\min}(H^*) > 0$. If $m = \Omega(\frac{n^6}{\lambda^4\delta^3})$, then with probability $\geq 1 - \delta$,

$$\|u(t) - y\|_2^2 \leq \exp\left(-\lambda_0 t\right)\|u(0) - y\|_2^2.$$

The proof of the theorem is based on the following lemma on the dynamics of u:

Lemma 2.

$$\frac{du(t)}{dt} = -H(t)[u(t) - y] \tag{3}$$

Proof. This lemma was proved in the previous lecture. □

To apply the above lemma, we need to lower bound $H(t)$. We first show that $H(0) \approx H^*$.

Lemma 3. Assume $\|x_i\| \leq 1$ and $\sigma(z) = \max\{0, z\}$. If the number of hidden neuron $m \geq \Omega(\epsilon^{-2} n^2 \log(\frac{n}{\delta}))$, then with probability at least $1 - \delta$ over the random initialization,

$$\|H(0) - H^*\|_2 \leq \epsilon.$$

Proof. This lemma was proved in the previous lecture. □

We then show that if the weight $w(t)$ is near $w(0)$, then $H(t) \approx H(0)$.

Lemma 4. With probability $\geq 1-\delta$ over $w(0)$, for any $\{w_k\}_{k=1}^m$ satisfying

$$\|w_k - w_k(0)\|^2 \leq \frac{\sqrt{2\pi}\delta\lambda_0}{16n^2} := R, \forall k \in [m],$$

we have $\|H - H(0)\|_2 \leq \frac{\lambda_0}{4}$ and thus $\lambda_{\min}(H) \geq \frac{\lambda_0}{2}$.

Proof. Define event $A_{ik} = \{\exists w_k, \|w_k - w_k(0)\| \leq R, \mathbf{1}[x_i^\top w_k(0) \geq 0] \neq \mathbf{1}[x_i^\top w_k \geq 0]\}$.

We first bound the probability of A_{ik}:

$$\Pr(A_{ik}) \leq \Pr(|x_i^\top w_k(0)| \leq R) \leq \frac{2R}{\sqrt{2\pi}}, \tag{4}$$

where the first inequality comes from the fact that $|x_i^\top w_k - x_i^\top w_k(0)| \leq \|x_i\|\|w_k - w_k(0)\| \leq R$, and the second from the anti-concentration of Gaussians.

Applying the above inequality we can bound individual entry as following:

$$\mathbb{E}[|H_{ij}(0) - H_{ij}|] \leq \mathbb{E}\left[\left|\frac{1}{m}x_i^\top x_j \sum_{k=1}^m \Big(\mathbf{1}[x_i^\top w_k(0) \geq 0]\mathbf{1}[x_j^\top w_k(0) \geq 0]\right.\right. \tag{5}$$

$$\left.\left. - \mathbf{1}[x_i^\top w_k \geq 0]\mathbf{1}[x_j^\top w_k \geq 0]\Big)\right|\right] \tag{6}$$

$$\leq \frac{1}{m}\sum_{i=1}^m \mathbb{E}[\mathbf{1}[A_{ik} \cup A_{jk}]] \tag{7}$$

$$\leq \frac{1}{m}\sum_{k=1}^m [\Pr(A_{ik}) + \Pr(A_{jk})] \tag{8}$$

$$\leq \frac{4R}{\sqrt{2\pi}}. \tag{9}$$

With the bound for the individual entry, we can further bound the difference between two matrices as following:

$$\mathbb{E}[\|H - H(0)\|_2] \leq \mathbb{E}[\|H - H(0)\|_F] \tag{10}$$

$$\leq \mathbb{E}\left[\sum_{ij} |H_{ij}(0) - H_{ij}|\right] \tag{11}$$

$$\leq \frac{4n^2 R}{\sqrt{2\pi}} \tag{12}$$

$$\leq \frac{4n^2}{\sqrt{2\pi}} \frac{\sqrt{2\pi}\delta\lambda_0}{16n^2} \tag{13}$$

$$= \frac{\delta\lambda_0}{4}. \tag{14}$$

Thus, according to Markov's inequality, we know that $\Pr[\|H - H(0)\|_2 \geq \frac{\lambda_0}{4}] \leq \frac{\delta\lambda_0/4}{\lambda_0/4} = \delta$ and the proof is done. □

Lemma 5. Suppose for $0 \leq s \leq t$, $\lambda_{\min}(H(s)) \geq \frac{\lambda_0}{2}$, then we have following result:

1. $\|u(t) - y\|_2^2 \leq \exp(-\lambda_0 y)\|u(0) - y\|_2^2$.
2. $\|w_k(t) - w_k(0)\|_2 \leq s\sqrt{n}\|u(0) - y\|_2/(\lambda_0\sqrt{m}) := R'$.

Proof. For the first result:

$$
\begin{aligned}
\frac{d\|u(t) - y\|_2^2}{dt} &= 2(u(t) - y)^\top \frac{du(t)}{dt} && (15)\\
&= -2(u(t) - y)^\top H(t)(u(t) - y) && (16)\\
&\leq -2\|u(t) - y\|_2^2 \frac{\lambda_0}{2} && (17)\\
&\leq -\lambda_0\|u(t) - y\|_2^2. && (18)
\end{aligned}
$$

This means we can further obtain the result from Grönwall's inequality (see e.g., wiki link):

$$\|u(t) - y\|_2^2 \leq \exp(\lambda_0 t)\|u(0) - y\|_2^2.$$

For the second result, define $\dot{w}(s) := -\nabla L(w(s))$:

$$
\begin{aligned}
\|w_k(t) - w_k(0)\|_2 &= \|\int_0^t \dot{w}_k(s) ds\|_2 && (19)\\
&\leq \int_0^t \|\dot{w}_k(s)\|_2 ds. && (20)
\end{aligned}
$$

$$
\begin{aligned}
\|\dot{w}_k(s)\| &= \|\sum_{i=1}^{n} (f(x_i; w(s)) - y_i)\frac{1}{\sqrt{m}} a_k \mathbf{1}[w_k(s)^\top x_i \geq 0] x_i\|_2 && (21)\\
&\leq \frac{1}{\sqrt{m}} \sum_{i=1}^{n} |f(x_i; w(s)) - y_i| && (22)\\
&\leq \frac{1}{\sqrt{m}} \sqrt{n} \sqrt{\sum_{i=1}^{n} (u_i(s) - y_i)^2} && (23)\\
&\leq \sqrt{\frac{n}{m}} \exp(-\lambda_0 s/2)\|u(0) - y\|_2. && (24)
\end{aligned}
$$

Plug (24) into (20) we have:

$$
\begin{aligned}
\|w_k(t) - w_k(0)\|_2 &\leq \sqrt{\frac{n}{m}} \|u(0) - y\|_2 \int_0^t \exp(-\lambda_0 s/2) ds && (25)\\
&= \|w_k(t) - w_k(0)\|_2 && (26)\\
&\leq \sqrt{\frac{n}{m}} \|u(0) - y\|_2 \frac{2}{\lambda_0} := R'. && (27)
\end{aligned}
$$

□

With all the lemmas, to prove Theorem 1, it is sufficient to ensure that $R' \leq R$, which requires

$$m = \Omega\left(\frac{n^5\|u(0) - y\|_2^2}{\lambda_0^4\delta^2}\right).$$

One can show that $\mathbb{E}\|u(0)-y\|_2^2 = O(n)$, and then by Markov's inequality, $\|u(0)-y\|_2^2 \leq O(\frac{n}{\delta})$ with probability $\geq 1-\delta$. The proof of Theorem 1 is then completed.