Mathematical Theory of Neural Network Models

Lecture 7: The analysis for two-layer neural networks and kernel method

July 16

Lecturer: Chao Ma *Scribe: Jiawei Wang,Yiming Qiao*

1 Neural Tangent Kernel

We begin from the two-layers neural networks:

$$f(x;a,w)=\frac{\sum_{k=1}^{m}a_k\sigma(w_k^Tx)}{\sqrt{m}}(w_k\sim\Omega(0,I)) \tag{1}$$

The initial value of $a_k and w_k(0)$ is as follows:

$$\begin{cases} a_k=\pm 1 \\ w_k(0)\sim\pi_0 \end{cases} \tag{2}$$

The loss function is:

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

Using the gradient descent:

$$\frac{\mathrm{d}w_k}{\mathrm{d}t}=-\frac{\partial L(w)}{\partial W_k}=-\sum_{i=1}^{n}(f(x_i;a,w)-y_i)a_k\sigma'(w_k^Tx_i)x_i \tag{3}$$

We need to optimize the $w_k(t)$:

$$u_i(t)=f(x_i;a,w(t))=\frac{\sum_{k=1}^{m}a_k\sigma(w_k^T(t)x_i)}{\sqrt{m}} \tag{4}$$

$$\frac{\mathrm{d}u_j(t)}{\mathrm{d}t}=\frac{\sum_{k=1}^{m}a_k\sigma'(w_k^T(t)x_j)\frac{\mathrm{d}w_k^T(t)}{\mathrm{d}t}x_j}{\sqrt{m}} \tag{5}$$

Then put the Equation 3 into Equation 5:

$$\frac{\mathrm{d}u_j(t)}{\mathrm{d}t}=\sum_{i=1}^{n}(y_i-u_i(t))\frac{\sum_{k=1}^{m}a_k^2\sigma'(w_k^T(t)x_i)\sigma'(w_k^T(t)x_j)x_i^Tx_j}{m} \tag{6}$$

We difine the $H(t)$:

$$H_{ij}(t) = \frac{\sum_{k=1}^{m} a_k^2 \sigma^{'}(w_k^T(t)x_i)\sigma^{'}(w_k^T(t)x_j)x_i^T x_j}{m} \tag{7}$$

So we put 7 into 5:

$$\frac{\mathrm{d}}{\mathrm{d}t}u(t) = H(t)(y-u) \tag{8}$$

And the loss function can also be showed that:

$$L = \frac{\|y-u\|^2}{2} \tag{9}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}L(t) = -(y-u)^T H(t)(y-u) \tag{10}$$

When $t=0$:

$$H_{ij}(0) = \frac{\sum_{k=1}^{m} a_k^2 \sigma^{'}(w_k^T(0)x_i)\sigma^{'}(w_k^T(0)x_j)x_i^T x_j}{m} \tag{11}$$

Let $m \longrightarrow \infty$:

$$H_{ij}^{\infty} = E_{w\sim\pi_0}\sigma^{'}(w^T x_i)\sigma^{'}(w^T(0)x_j)x_i^T x_j \tag{12}$$

Then we define $K(x,x^{'}) = E_w \sigma^{'}(w^T x)\sigma^{'}(w^T(0)x^{'})x^T x^{'}$ which is named **Neural Tangent Kernel**. Now we consider the property of $H^m(t)$.

2 Some lemma about $H^m(t)$

Lemma 2.1. $if x_i \nparallel x_j, i \neq j$,*then* $\lambda_{min}(H^{\alpha}) = \lambda_0 > 0$

Corollary 2.2. $\lambda_{min}(H(0)) \geq \frac{3}{4}\lambda_0$

Lemma 2.3. $w_1(0), w_2(0), \cdots, w_m(0) \sim N(0,I), w.p. \geq 1-\varepsilon$,*for* $w_1, w_2, \cdots, w_m$ *satisfying:*$\|w_k - w_k(0)\|_2 \leq \frac{c\varepsilon\lambda_0}{n^2}$, *then* $\|H(w) - H(w(0))\|_2 < \frac{\lambda_0}{4}$

Corollary 2.4. *if for* $s \in [0,t], \lambda_{min}(H(w(s))) \geq \frac{\lambda_0}{2}$,*then:* $\|y-u(t)\|_2^2 \leq e^{-\lambda_0 t}\|y-u(0)\|_2^2$

3 Mean field method

Now we introduce a **mean field method** when $m \to \infty$.In this method, we think of the parameters as following a distribution.

$$f(x;a,w) = \frac{1}{m}\sum_{k=1}^{m} a_k \sigma(w_k^T x)$$

Because $a \in \mathbf{R}$,$w \in \mathbf{R^d}$,

$$(a,w) \in \mathbf{R^{d+1}}$$

We define ρ:Distribution on $\mathbf{R^{d+1}}$,so:

$$f(x;a,w) = \int a\sigma(w^T x)\rho(da,dw) = E_{(a,w)\sim\rho} a\sigma(w^T x)$$

4 Compare the NN and kernel method

We have two conclusion:

- *1.* NN is close to kernel when $m \geq \mathcal{O}(n^2)$.
- *2.* When there is no $m \geq \mathcal{O}(n^2)$, at some time point before in the beginning NN is close to kernel.

For the conclusion **1**, it is obvious in previous class. Now we consider the conclusion **2**.
We define the **Target function** f^* and the norm $\gamma(f^*)$:

$$f^* = \int_{\pi_0 \sim s^{d-1}} a^*(b)\sigma(b^T x) d\pi_0(b) \tag{13}$$

$$\gamma(f^*) = max1, sup|a^*(b)| < \infty \tag{14}$$

There are three properties of target function:

- *1.*random feature model can learn f^* well.
- *2.*In some time point t_0,when $t \in [0, t_0]$,NN $\approx$ RF.
- *3.*NN with early stopping can learn f_* well.

Define $R(a, B)$ and $f(x, a, B)$:

$$R(a, B) = \|f(x, a, B) - f^*\|^2$$

$$f(x, a, B) = \sum_{k=1}^{m} a_k \sigma(b_k^T x)$$

Theorem 4.1. *For RF,*$\forall \varepsilon > 0, w.p. \geq 1 - \varepsilon, for b_k \in B_0$,$\exists a^*$ *s.t.*

$$R(a^*, B) \leq \frac{\gamma^2(f^*)}{m}(1 + \sqrt[2]{2\log\frac{1}{\varepsilon}})^2 \tag{15}$$

$$\|a^*\|_1 \leq \frac{\gamma(f^*)}{\sqrt{m}} \tag{16}$$

Theorem 4.2. *For NN,* $f(x, a, B) = \sum_{k=1}^{m} a_k \sigma(b_k^T x)$ *with constraints:*

- $D_k(0) \sim \pi_0$ *same as RF*
- $a_k(0) = \pm\beta, \beta = \frac{c}{m}$
- $\|f^*\|_\infty \leq 1$

Then, $\forall \varepsilon > 0, w.p. \geq 1 - 4\varepsilon$*, we have:*

$$\hat{R} \leq C(\frac{1}{m} + \frac{1}{mt} + \frac{1}{\sqrt{n}}) \tag{17}$$

In order to proof 4.2, we need to introduce some lemmas as follows.

Lemma 4.3. $\beta = \frac{c}{m}$*,T is constant,* $\exists C_T$ *s.t. for* $0 \leq t \leq T$*:*

$$\|a_t\| \leq C_T(\frac{c}{\sqrt{m}} + \sqrt{m}t) \tag{18}$$

$$\|B_t\| \leq C_T(\frac{ct}{\sqrt{m}} + \sqrt{m}) \tag{19}$$

$$\|B_t - B_0\| \leq C_T(c+1)(\frac{ct}{\sqrt{m}} + \frac{\sqrt{m}}{2}t^2) \tag{20}$$

Lemma 4.4. *for* $m \geq r^2, \forall \varepsilon > 0, w.p. \geq 1 - 4\varepsilon, \forall t \in [0, T], \exists \tilde{C_T} s.t.$

$$\|\tilde{a_t}\| \leq \tilde{C_T}(\frac{1}{\sqrt{m}} + \frac{\sqrt{t}}{\sqrt{m}} + \frac{\sqrt{t}}{n^{\frac{1}{4}}}) \tag{21}$$

Lemma 4.5. *for* $t \in [0, T]$*,*

$$\|a_t - \tilde{a_t}\| \leq C_T\frac{t^2}{m}(1 + mt)(t + m)(\frac{1}{\sqrt{m}} + \frac{\sqrt{t}}{\sqrt{m}} + \frac{\sqrt{t}}{n^{\frac{1}{4}}}) \tag{22}$$

Theorem 4.6. *under the same assumption of 4.2,*$\forall T, \varepsilon > 0, w.p. \geq 1 - \varepsilon$*,for* $\forall t \leq T$*, have:*

$$R(a_t, B_t) \leq C(\frac{1}{m} + \frac{1}{mt} + \frac{1}{\sqrt{n}}) \tag{23}$$