Mathematical Theory of Neural Network Models

# Lecture 7: Property of Two Layer Neural Network

July 16

*Lecturer: Chao Ma* *Scribe: Yuncong Peng, Yilei Han,Jingsong Zhang*

## 1 Estimation of NN vs. RF at a limited time scale

Consider NN vs. RF
(1)$m >= O(n^2)$ ,NN is always close to RF
(2)$m < O(n^2)$ ,NN is only close to RF at the beginning

Assume target function as $f^*$

$$f^* = \int_{S^{d-1}} a^* \sigma(b^t x) d(\pi_0(b))$$
$$\gamma(f^*) = max\{1, sup\,|a^*(b)|\} < \infty$$

- RF can learn $f^*$ well,
- $NN \approx RF, t \in [0, t_0]$
- NN with early stopping can also learn well.

**Theorem 1.1** (RF). *$\forall\ \delta$,with $prob >= 1 - \delta$ over the chance of $\{b_0\}$ there exist $a^*$,s.t.*

$$R(a^*, B) \leq \frac{\gamma^2(f^*)}{m}(1 + \sqrt{2\log(\frac{1}{\delta})})^2$$
$$\|a^*\|_1 \leq \frac{\gamma(f^*)}{\sqrt{m}}$$

Notations:

$$f(x; \boldsymbol{a}, B) = \sum_{k=1}^{m} a_k \sigma(b_k^T x)$$
$$\mathcal{R}\left(\boldsymbol{a}, B\right) = \|f(x; \boldsymbol{a}, B) - f^*\|^2$$
$$\hat{\mathcal{R}}\left(\boldsymbol{a}, B\right) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i; \boldsymbol{a}, B) - f^*(x_i))^2$$

**Theorem 1.2.** *For a two-layer Neural Network, which is defined as $f(x; a, B) = \sum_{k=1}^{m} a_k \sigma(b_k^T x)$,with the assumption below:*

- $b_k(0) \sim \pi_0$ *(same as Random Feature)*
- $a_k(0) = \pm\beta, \beta = \frac{C}{m}$
- $\|f^*\|_\infty \leq 1$

*Then* $\forall \delta > 0$*, with* $p \geq 1 - 4\delta$ *, we have*

$$\hat{\mathcal{R}}_n\left(\boldsymbol{a}_t, B_t\right) \leq C\left(\frac{1}{m} + \frac{1}{mt} + \frac{1}{\sqrt{n}}\right)$$

**proof:**
Let $\hat{\rho} = \frac{1}{n}\sum_{i=1}^{n}\delta_{\boldsymbol{x}_i}$,then we have

$$\begin{aligned}\hat{\mathcal{R}}_n\left(\boldsymbol{a}_t, B_t\right) &= \left\|f\left(\cdot; \boldsymbol{a}_t, B_t\right) - f^*(\cdot)\right\|_{\hat{\rho}}^2 \\ &\leq 3\left(\left\|f\left(\cdot; \boldsymbol{a}_t, B_t\right) - f\left(\cdot; \tilde{\boldsymbol{a}}_t, B_t\right)\right\|_{\hat{\rho}}^2 + \left\|f\left(\cdot; \tilde{\boldsymbol{a}}_t, B_t\right) - f\left(\cdot; \tilde{\boldsymbol{a}}_t, B_0\right)\right\|_{\hat{\rho}}^2 + \hat{\mathcal{R}}_n\left(\tilde{\boldsymbol{a}}_t, B_0\right)\right)\end{aligned}$$

By Cauchy-Schwartz, we have

$$\left\|f\left(\boldsymbol{x}; \boldsymbol{a}_t, B_t\right) - f\left(\boldsymbol{x}; \tilde{\boldsymbol{a}}_t, B_t\right)\right\|_{\hat{\rho}}^2 \leq \left\|\boldsymbol{a}_t - \tilde{\boldsymbol{a}}_t\right\|^2 \left\|B_t\right\|^2$$

$$\left\|f\left(\boldsymbol{x}; \tilde{\boldsymbol{a}}_t, B_t\right) - f\left(\boldsymbol{x}; \tilde{\boldsymbol{a}}_t, B_0\right)\right\|_{\hat{\rho}}^2 \leq \left\|\tilde{\boldsymbol{a}}_t\right\|^2 \left\|B_t - B_0\right\|^2$$

For $\hat{\mathcal{R}}_n\left(\tilde{\boldsymbol{a}}_t, B_0\right)$, with probability $1 - \delta$, there exists $a^*$ as a good estimator. Thus we have

$$\begin{aligned}\hat{\mathcal{R}}_n\left(\tilde{\boldsymbol{a}}_t, B_0\right) &= \left(\hat{\mathcal{R}}_n\left(\tilde{\boldsymbol{a}}_t, B_0\right) - \hat{\mathcal{R}}_n\left(\boldsymbol{a}^*, B_0\right)\right) + \left(\hat{\mathcal{R}}_n\left(\boldsymbol{a}^*, B_0\right) - \mathcal{R}\left(\boldsymbol{a}^*, B_0\right)\right) + \mathcal{R}\left(\boldsymbol{a}^*, B_0\right) \\ &=: I_1 + I_2 + I_3\end{aligned}$$

As the *generalization error*, We can bound $I_2$ as follows:

$$I_2 \leq \frac{2\left(2\sqrt{m}\left\|a^*\right\| + 1\right)^2}{\sqrt{n}}\left(1 + \sqrt{2\ln\left(\frac{2}{\delta}\left(\left\|a^*\right\| + \frac{1}{\left\|a^*\right\|}\right)\right)}\right)$$

For $I_1$, consider the Lyapunov function

$$J(t) = t\left(\hat{\mathcal{R}}_n\left(\tilde{\boldsymbol{a}}_t, B_0\right) - \hat{\mathcal{R}}_n\left(\boldsymbol{a}^*, B_0\right)\right) + \frac{1}{2}\left\|\tilde{\boldsymbol{a}}_t - \boldsymbol{a}^*\right\|^2$$

Since $\hat{\mathcal{R}}_n\left(\tilde{\boldsymbol{a}}_t, B_0\right)$ is convex with respect to $\tilde{\boldsymbol{a}}_t$, we have $\frac{d}{dt}J(t) \leq 0$, which implies $J(t) \leq J(0)$. Hence we have

$$\hat{\mathcal{R}}_n\left(\tilde{\boldsymbol{a}}_t, B_0\right) \leq \hat{\mathcal{R}}_n\left(\boldsymbol{a}^*, B_0\right) + \frac{\left\|\boldsymbol{a}_0 - \boldsymbol{a}^*\right\|^2}{2t}$$

Estimate the norm of parameters.

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left(a_t-\tilde{a}_t\right)\\
=&-\frac{1}{n}\sum_{i=1}^{n}\left[\left(f\left(x_i,a_t,B_t\right)-y_i\right)\tau\left(B_H,x_i\right)-\left(f\left(x_i;\tilde{a}_t,B_0\right)-y_i\right)\tau\left(B(0)x_i\right)\right]\\
=&-\frac{1}{n}\sum_{i=1}^{n}\left(a_k^T\tau\left(B_tx_i\right)\tau\left(B_tx_i\right)-\tilde{a}_t^T\tau\left(B_0x_i\right)\tau\left(B_0x_i\right)\right)+\frac{1}{n}\sum_{i=1}^{n}y^T\left(\tau\left(B_tx_i\right)-\tau\left(B_0x_i\right)\right)\\
=&-\frac{1}{n}\sum_{i=1}^{n}\tau\left(B_tx_i\right)\tau\left(B_tx_i\right)^T\left(a_t-\tilde{a}_t\right)+\frac{1}{n}\sum_{i=1}^{n}\left(\tau\left(B_tx_i\right)\tau\left(B_tx_i\right)^T-\tau\left(B_0x_i\right)\tau\left(B_0x_i\right)^T\right)\tilde{a}_t\\
&+\frac{1}{n}\sum_{i=1}^{n}y^T\left(\tau\left(B_tx_i\right)-\tau\left(B_0x_i\right)\right)
\end{aligned}
$$

Then

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left\|a_t-\tilde{a}_t\right\|^2\\
=&2\left(a_t-\tilde{a}_t\right)^T\frac{\mathrm{d}}{\mathrm{d}t}\left(a_t-\tilde{a}_t\right)\\
\leq&-\left(a_t-\tilde{a}_t\right)^T\frac{1}{n}\sum_{i=1}^{n}\tau\left(B_tx_i\right)\tau\left(B_tx_i\right)^T\left(a_t-\tilde{a}_t\right)\\
&+\left(a_t-\tilde{a}_t\right)^T\left[\frac{2}{n}\sum_{i=1}^{n}\left(\tau\left(B_tx_i\right)\tau\left(B_tx_i\right)^T-\tau\left(B_0x_i\right)\tau\left(B_0x_i\right)^T\right)\right]\tilde{a}_t\\
&+\left(a_t-\tilde{a}_t\right)^T\left[\frac{2}{n}\sum_{i=1}^{n}y^T\left(\tau\left(B_tx_i\right)-\tau\left(B_0x_i\right)\right)\right]\\
\leq&\left(a_t-\tilde{a}_t\right)^T\left[\frac{2}{n}\sum_{i=1}^{n}\left(\tau\left(B_tx_i\right)\tau\left(B_tx_i\right)^T-\tau\left(B_0x_i\right)\tau\left(B_0x_i\right)^T\right)\right]\tilde{a}_t\\
&+\left(a_t-\tilde{a}_t\right)^T\left[\frac{2}{n}\sum_{i=1}^{n}y^T\left(\tau\left(B_tx_i\right)-\tau\left(B_0x_i\right)\right)\right]\\
\leq&2\left\|B_t-B_0\right\|\left(\left\|B_t\right\|\left\|\tilde{a}_t\right\|+\left\|B_0\right\|\left\|\tilde{a}_t\right\|+1\right)\left\|a_t-\tilde{a}_t\right\|
\end{aligned}
$$

Then estimate the bound of $\left\|a_t\right\|,\left\|B_t-B_0\right\|,\left\|\tilde{a}_t\right\|,\left\|a_t-\tilde{a}_t\right\|$

**Lemma 1.3.** *Let $T$ is Constant, $\exists C_T$ s.t. $\forall 0 \le t \le T$,*

$$\|a_t\| \le C_T\left(\frac{C}{\sqrt{m}} + \sqrt{m}t\right)$$
$$\|B_t\| \le C_T\left(\frac{ct}{\sqrt{m}} + \sqrt{m}\right)$$
$$\|B_t - B_0\| \le C_T(c+1)\left(\frac{ct}{\sqrt{m}} + \frac{\sqrt{m}}{2}t^2\right)$$

**Lemma 1.4.** *Assume that $m \ge r^2$,$\forall \epsilon > 0$, with probability $\ge 1-4\epsilon$, $\forall t \in [0,T]$*

$$\|\tilde{a}_t\| \le \tilde{C}_T\left(\frac{1}{\sqrt{m}} + \frac{t}{\sqrt{m}} + \frac{\sqrt{t}}{n^{\frac{1}{4}}}\right)$$

**Lemma 1.5.** *With assumptions above, $\forall t \in [0,T]$*

$$\|a_t - \tilde{a}_t\| \le C_T\frac{t^2}{m}(1+mt)(t+m)\left(\frac{1}{\sqrt{m}} + \frac{\sqrt{t}}{\sqrt{m}} + \frac{\sqrt{t}}{n^{\frac{1}{4}}}\right)$$

With all above, we have

$$\hat{\mathcal{R}}_n(\boldsymbol{a}_t, B_t) \le C\left(\frac{1}{m} + \frac{1}{mt} + \frac{1}{\sqrt{n}}\left(1+\sqrt{t}+\frac{\sqrt{mt}}{n^{1/4}}\right)^2 \right.$$
$$\left. + \frac{t^2}{m^2}(1+mt)^2\left(1+\frac{t^2}{m^2}(t+m)^4\right)\left(1+\sqrt{t}+\frac{\sqrt{mt}}{n^{1/4}}\right)^2\right)$$

for $t \in [0,T]$, and some constant $C$.
If we assume $m \ge n$, and take $t \in \left[0, \frac{\sqrt{n}}{m}\right]$,then we can take $T=1$ and obtain

$$\hat{\mathcal{R}}(a_t, B_t) \le C\left(\frac{1}{m} + \frac{1}{mt} + \frac{1}{\sqrt{n}}\right) \quad \forall 0 \le t \le 1$$

One can refer to *A Comparative Analysis of the Optimization and Generalization Property of Two-layer Neural Network and Random Feature Models Under Gradient Descent Dynamics* (arXiv:1904.04326v1) for more details.