Mathematical Theory of Neural Network Models

Lecture 9.3: Theory of the Frequency Principle

July 19

Lecturer: Tao Luo *Scribe: Yunzhen Feng, Haocheng Ju, Dingwen Kong*

1 Supervised Learning

In the previous section, we only have partial theoratical guarantee for only one hidden layer with activation tanh. Besides, it only reveals little about dynamics.

In the following section, we hope to prove the emergence of F-Principle phenomenon with universality(different structure and multiple activation function) and dynamics(time dependence and long time behavior). For more details, please see [1]

1.1 Parameters

We consider a DNN with (H-1)-hidden layers with the following parameters:

$$\begin{aligned} W^{(l)} &= \left(W_i^{(l)}\right)_{i=1}^{n_l}, \quad W_i^{(l)} \in \mathbb{R}^{n_{l-1}} \\ b^{(l)} &= \left(b_i^{(l)}\right)_{i=1}^{n_l}, \quad b^{(l)} \in \mathbb{R} \\ \theta &= \left(W^{(l)}, b^{(l)}\right)_{l=1}^{H} \\ n_l &= \#\text{ neurals in layer } l, \quad (n_0 = d, n_H = 1) \\ N &= \dim \theta. \end{aligned}$$

1.2 Target function

We are only interested in f_{target} in a compact domain Θ:

$$\begin{aligned} h(x,\theta) &= h^{(H)}(x,\theta)\chi(x) \\ f(x) &= f_{\text{target}}(x)\chi(x), \end{aligned}$$

where $\chi(x)$ is a bump function.

For loss function, we consider MSE loss and general loss with a dynamic continuous parameter updates.

1.3 Assumptions

1.3.1 Assumption 1(regularity)

The bump function χ satisfies $\chi(x) = 1, x \in \Omega$ and $\chi(x) = 0, x \in \mathbb{R}^d \backslash \Omega'$ for domains Ω and Ω' with $\Omega \subset\subset \Omega' \subset\subset \mathbb{R}^d$. There is a positive integer k such that $f_{\text{target}} \in W^{k,\infty}_{\text{loc}}\left(\mathbb{R}^d; \mathbb{R}\right)$ and $\chi \in W^{k,\infty}_{\text{loc}}\left(\mathbb{R}^d; [0, +\infty)\right)$, and$\sigma_i^{(l)} \in W^{k,\infty}_{\text{loc}}(\mathbb{R}; \mathbb{R})$ for $l = 1, \cdots, H-1, i = 1, \cdots, n_l$.

1.3.2 Assumption 2(bounded population density

There exists a function $\rho \in L^\infty\left(\mathbb{R}^d; [0, +\infty)\right)$ satisfying $\mathrm{d}\mu = \rho \mathrm{d}x$.

1.3.3 Assumption 3(bounded trajectory)

The training dynamics is nontrivial,i.e.$\theta(t) \neq$ const. There exists a constant $R > 0$ such that $\sup_{t\geq 0} |\theta(t)| \leq R$ where the parameter vector $\theta(t)$ is the solution to the dynamic continuous parameter updates functions.

For common activation function ReLU, tanh, and sigmoid, they are all in this category.
In the following part, we have two different assumption for the two kind of loss.

1.3.4 Assumption 4(used for L^2 loss)

The density ρ satisfies $\sqrt{\rho} \in W^{k,\infty}_{\text{loc}}\left(\mathbb{R}^d; [0, +\infty)\right)$.

1.3.5 Assumption 5(used for general loss)

The function ℓ in the general loss function $\tilde{L}_\rho(\theta)$ satisfies $\ell \in C^2(\mathbb{R}; [0, +\infty))$ and there exist positive constants C and r_0 such that $C^{-1}\left[\ell'(z)\right]^2 \leq \ell(z) \leq C|z|^2$ for $|z| \leq r_0$

Note that, L^p functions all satisfy Assumption 5 and the L^p function here in definition does not go through root; it is actually $(L^p)^p$.

2 Loss in frequency domain

Here, is where the assumption of $\sqrt{\rho}$ comes from. Then we consider the two ratio of the high and low frequency part.

$$\frac{\left|\mathrm{d}L^-_{\rho,\eta}/\mathrm{d}t\right|}{|\mathrm{d}L_\rho/\mathrm{d}t|} \text{ and } \frac{\left|\mathrm{d}L^+_{\rho,\eta}/\mathrm{d}t\right|}{|\mathrm{d}L_\rho/\mathrm{d}t|}$$

where B_η and $B_\eta^c = \mathbb{R}^d \backslash B_\eta$ are a ball centered at the origin with radius $\eta > 0$ and its complements,and

$$L^-_{\rho,\eta}(\theta) = \int_{B_\eta} \left|\hat{h}_\rho(\xi,\theta) - \hat{f}_\rho(\xi)\right|^2 \mathrm{d}\xi, \quad L^+_{\rho,\eta}(\theta) = \int_{B_c^c} \left|\hat{h}_\rho(\xi,\theta) - \hat{f}_\rho(\xi)\right|^2 \mathrm{d}\xi$$

Note $L_\rho = L^-_{\rho,\eta} + L^+_{\rho,\eta}$ for any $\eta > 0$ We wants to prove the dominance of the first term. For the L_2 loss, we have the following theorem:

Theorem 1 (F-principle in the initial stage)(L^2 loss function) Suppose that Assumption 1,2,3,4 hold.Then for any $1 \le m \le 2k-1$ and any $T > 0$ satifying $|\nabla_\theta L_\rho(\theta(T))| > 0$(if k=1, we futher require that $\inf_{t\in(0,T]} |\nabla_\theta L_\rho(\theta(t))| > 0$),there ia a constant $C > 0$ such that

$$\frac{\left|\mathrm{d}L^+_{\rho,\eta}/\mathrm{d}t\right|}{|\mathrm{d}L_\rho/\mathrm{d}t|} \le C\eta^{-m} \quad \text{and} \quad \frac{\left|\mathrm{d}L^-_{\rho,\eta}/\mathrm{d}t\right|}{|\mathrm{d}L_\rho/\mathrm{d}t|} \ge 1 - C\eta^{-m}, \quad t \in (0,T]$$

As for the general loss, we define two similar observation ratio. From different techniques, we can have similar results:

Theorem 2 (F-principle in the initial stage)(general loss function) Suppose that Assumption 1,2,3,5 hold.Then for any $1 \le m \le 2k-1$ and any $T > 0$ satifying $\left|\nabla_\theta \tilde{L}_\rho(\theta(T))\right| > 0$,there is a constant $C > 0$ such that

$$\frac{\|\mathrm{d}\hat{h}/\mathrm{d}t\|_{L^2(B_{\bar{\eta}})}}{\|\mathrm{d}\hat{h}/\mathrm{d}t\|_{L^2(\mathbb{R}^d)}} \le C\eta^{-m} \quad and \quad \frac{\|\mathrm{d}\hat{h}/\mathrm{d}t\|_{L^2(B_\eta)}}{\|\mathrm{d}\hat{h}/\mathrm{d}t\|_{L^2(\mathbb{R}^d)}} \ge 1 - C\eta^{-m}, \quad t \in (0,T]$$

However, the C here rely on the t. Can we have a uniform estimation on C? What's the dependence $C(T)$? Is there an upper bound of C when $T \to \infty$?
If we have another assumption of a existence of a non-degenerate global minimizer $\theta*$, there dooes exist a time-independent upper bound on C.
Then what's the dependence $C(T)$? We analyze the following item and do Fourier transform for the nonlinear term. But the dynamics here, is NOT a gradient flow. Thus, the deliminator may become 0 at some time. To avoid it, we do integration and find a appropriate time scale in this weak form:

$$\frac{\int_{T_1}^{T_2} \left|\frac{dL^-_\eta}{dt}\right| dt}{\int_{T_1}^{T_2} \left|\frac{dL}{dt}\right| dt} \text{ and } \frac{\int_{T_1}^{T_2} \left|\frac{dL^+_\eta}{dt}\right| dt}{\int_{T_1}^{T_2} \left|\frac{dL}{dt}\right| dt}.$$

For the time scale, there is a half life, $\frac{1}{2}L\left(\theta\left(T_1\right)\right) = L\left(\theta\left(T_2\right)\right)$, which results in a ultimate theorem for all the time period.
The following corollary and the theorem all have $\sqrt{T}$ relationship.

Theorem 3 (F-principle in the intermediate stage)(general loss function) Suppose that assumption 1,2,3,5 hold.Then for any $1 \leq m \leq k-1$, there is a constant $C > 0$ such that for any $0 < T_1 < T_2$ satisfying $\frac{1}{2} L\left(\theta\left(T_1\right)\right) \geq L\left(\theta\left(T_2\right)\right)$ we have

$$\frac{\int_{T_1}^{T_2} \left| \frac{dL_\eta^+}{dt} \right| \mathrm{d}t}{\int_{T_1}^{T_2} \left| \frac{\mathrm{d}L}{\mathrm{d}t} \right| \mathrm{d}t} \leq C \sqrt{T_2 - T_1} \eta^{-m}$$

Corollary Under the same assumptions of theorem 3, for any $1 \leq m \leq k-1$, there is a constant $C > 0$ such that for any $0 < T_1 < T_2$ satisfying $\frac{1}{2} L\left(\theta\left(T_1\right)\right) \geq L\left(\theta\left(T_2\right)\right)$ and $L\left(\theta\left(T_1\right)\right) \geq L(\theta(t))$ for all $t \in [T_1, T_2]$, we have

$$\frac{\left| L_\eta^+\left(\theta\left(T_1\right)\right) - L_\eta^+\left(\theta\left(T_2\right)\right) \right|}{\left| L\left(\theta\left(T_1\right)\right) - L\left(\theta\left(T_2\right)\right) \right|} \leq C \sqrt{T_2 - T_1} \eta^{-m}$$

References

[1] Luo, Ma, Xu, and Zhang. Theory on frequency principle in gerenal deep neural networks. *arXiv preprint arXiv:1906.09235*, 2019.