Mathematical Theory of Neural Network Models

Lecture 9.2: Quantitative understanding of F-principle

July 19

Lecturer: Yaoyu Zhang *Scribe: Yunzhen Feng, Haocheng Ju, Dingwen Kong*

Why overparameterized NNs often generalize well?

1 Notation

In this lecture, we consider the two-layer NN case.

$$h(x,\theta)=\sum_{i=1}^{N} w_i\sigma(r_i x-|r_i|l_i),$$

where $r_i, x\in\mathbb{R}^d, \theta=(w,R,l), w,l\in\mathbb{R}^N$ and $R\in\mathbb{R}^{N\times d}$, and by default $\sigma(x)=max(x,0)$ is the activation function of ReLU. The target function is denoted by $f(x)$. The following notation will be used in studying the training dynamics: $u(x,t)=h(x,\theta(t))-f(x), u_p(x,t)=h_p(x,\theta(t))-f_p(x)$, where $h_p(x,\theta(t))=h(x,\theta(t))p(x), f_p(x)=f(x)p(x)$ and $p(x)$ is the population probability density given by $p(x)=\frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i)$. For u and u_p, their Fourier transforms are written as $\hat{u}(\xi,t)$ and $\hat{u}_p(\xi,t)$, respectively. $\theta(t)=(w(t),R(t),l(t))$ are the parameters at training time t.

The following notes consists three aspects:

- Dynamics,
- Optimization,
- Generalization.

2 Linear F-Principle dynamics

Analyze the stationary stage:

$$h(x,\theta(t))=h\left(x,\theta_0\right)+\nabla_\theta h\left(x,\theta_0\right)\left(\theta(t)-\theta_0\right)\ \text{for any } t>0,$$

Using linear dynamic analysis:

$$\frac{d\theta(t)}{dt}=-\nabla_\theta h\left(X,\theta_0\right)^T\left(h(X,\theta(t))-Y\right).$$

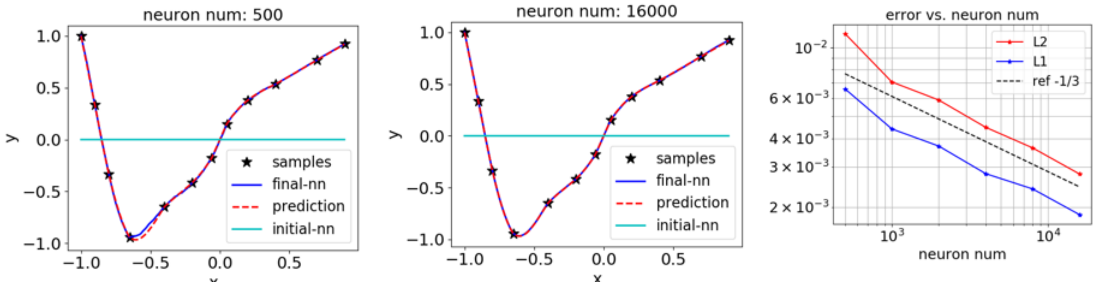

Figure 1: First part dominates

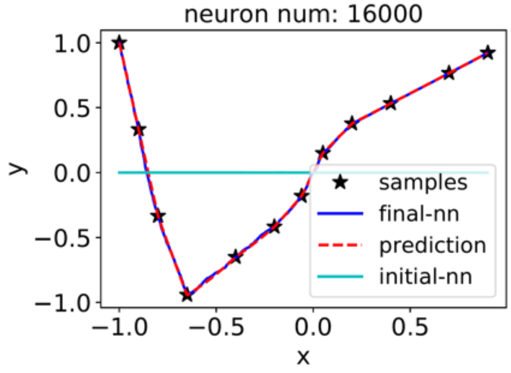

Figure 2: Second part dominates

How to understand it? For a network with one single hidden layer, we can get an effective dynamic model of the output

$$\partial_t \hat{h}(\xi, t) = \left[\frac{\frac{1}{N}\sum_{i=1}^N \left(r_i(0)^2 + w_i(0)^2\right)}{\xi^4} + \frac{4\pi^2 \frac{1}{N}\sum_{i=1}^N \left(r_i(0)^2 w_i(0)^2\right)}{\xi^2} \right] \left(\widehat{f_p}(\xi, t) - \widehat{h_p}(\xi, t) \right).$$

Here, we consider the process of sampling. The f is the target function $(\cdot)_p = (\cdot)p$,where$p(x) = \frac{1}{M}\sum_{i=1}^M \delta\left(x - x_i\right)$. Here, we use $\hat{\cdot}$ to denote the Fourier transfer and ξ for the frequency.For simplicity, it can be rewritten as

$$\partial_t \hat{u}(\xi, t) = -\left[\frac{\langle r^2 \rangle + \langle w^2 \rangle}{\xi^4} + \frac{4\pi^2 \langle r^2 w^2 \rangle}{\xi^2} \right] \widehat{u_p}(\xi, t).$$

Thus, the different part of frequency should evolve clearly from the following equation. But, the $\hat{u_p}$ brings interactions between high frequency and low frequency parts. Here, In experiment, we can set different variance to analyze each part. As you can see, when the first part dominate the whole equation(Figure 1), the performance becomes better along with the increase in neuron number. On the contrary, when the second term dominates(Figure 2), it is close to a piece-wise linear one. For high dimension case, r_i is also a high dimension vector, and the gradient flow is slightly different,

$$\partial_t \widehat{u}(\xi, t) = -\left[\frac{\langle |r|^2\rangle + \langle w^2\rangle}{\xi^{d+3}} + \frac{4\pi^2 \langle |r|^2 w^2\rangle}{\xi^{d+1}}\right] \widehat{u_p}(\xi, t).$$

Notice that the decay rate is also related to the dimension.

3 Optimization framework

On a physical viewpoint, we still not know much about how the final solution was selected, i.e. the implicit bias. To answer this question, we set up another optimization framework:

$$\min_{h-h_{ini}\in F_Y} \int \left[\frac{\langle |r|^2\rangle + \langle w^2\rangle}{\xi^{d+3}} + \frac{4\pi^2 \langle |r|^2 w^2\rangle}{\xi^{d+1}}\right]^{-1} \left|\hat{h}(\xi) - \hat{h}_{ini}(\xi)\right|^2 \mathrm{d}\xi$$

$$s.t. \qquad h(X) = Y$$

where the scalar above is the weight here. This penalty is small when the frequency is low, thus revealing the preference. First, we present a general theorem that the long-time limit solution of a gradient flow dynamics is equivalent to the solution of a constrained optimization problem. Let H_1 and H_2 be two separable Hilbert space and $P : H_1 \to H_2$ is a surjective linear operator, P^* is the adjoint operator of P. Given $g \in H_2$, we consider the e following two problems.
(i) The initial value problem

$$\frac{du}{dt} = P^*(g - Pu),\ u(0) = u_{ini}$$

Since this equation is linear and with nonpositive eigenvalues on the right hand side, there exists a unique global-in-time solution $u(t)$ for all $t \in [0, +\infty)$ satisfying the initial condition. Moreover, the long-time limit $lim_{t->+\infty} u(t)$ exists and will be denoted as u_∞
(ii) The minimization problem

$$\min_{u-u_{ini}\in H_1} \|u - u_{ini}\|_{H_1},\ s.t. Pu = g$$

Theorem 1. *Suppose that PP^* is surjective. The above problems (i) and (ii) are equivalent in the sense that $u_\infty = u_{min}$. More precisely, we have*

$$u_\infty = u_{min} = P^*(PP^*)^{-1}(g - Pu_i) + u_{ini}$$

The following corollaries is obtained directly from Theorem 1.

Corollary 1. *Let u be the parameter vector θ in $H_1 = \mathbb{R}^{N_p}$, g be the outputs of the training data Y, and P be a full rank matrix in the linear DNN model. Then the following two problems are*

equivalent in the sense that $\theta_\infty = \theta_{min}$.
(i) The initial value problem

$$\frac{d\theta}{dt} = P^*(Y - P\theta),\ \theta(0) = \theta_{ini}$$

(ii) The minimization problem

$$\min_{\theta-\theta_{ini}\in H_1} \|\theta - \theta_{ini}\|_{\mathbb{R}^{N_p}},\ s.t. P\theta = g$$

Corollary 2. *Let $\gamma : \mathbb{R}^d \to \mathbb{R}^+$ be a positive function and h be a function in $L^2(\mathbb{R})$. The operator $\Gamma : L^2((R)^d) \to L^2((R)^d)$ is defined by $[\Gamma\hat{h}](\xi) = \gamma(\xi)\hat{h}(\xi), \xi \in \mathbb{R}^d$. Define the Hilbert space $H_\Gamma := Im(\Gamma)$. Let $X = (x_i)_{i=1}^M \in \mathbb{R}^{d\times M}, Y = (y_i)_{i=1}^M \in \mathbb{R}^M$ and $P : H_\Gamma \to \mathbb{R}^M$ be a surjective operator*

$$P : \hat{h} \to \left(\int_{\mathbb{R}^d} \hat{h}(\xi) e^{2\pi i x_i \xi} d\xi\right)_{i=1}^M = (h(x_i))_{i=1}^M$$

Then the following two problems are equivalent in the sense that $\hat{h}_\infty = \hat{h}_{min}$.
(i) The initial value problem

$$\frac{d\hat{h}(\xi)}{dt} = (\gamma(\xi))^2 \sum_{i=1}^M (y_i e^{-2\pi i x_i \xi} - \hat{h}(\xi) * e^{-2\pi i x_i \xi}),\ \hat{h}(0) = \hat{h}_{ini}$$

(ii) The minimization problem

$$\min_{\hat{h}-\hat{h}_{ini}\in H_\Gamma} \int_{\mathbb{R}^d} (\gamma(\xi))^{-2} |\hat{h}(\xi) - \hat{h}_{ini}(\xi)|^2 d\xi,\ s.t. h(x_i) = y_i, i = 1, \ldots, M$$

Then we consider the optimization part using the similar viewpoint. We can have the following theorem.

Theorem 2. *Let $\theta(t)$ be the solution of gradient flow dynamics*

$$\frac{d\theta(t)}{t} = -\nabla_\theta h(X, \theta_0)^T_{h(X,\theta(t))} D(h(X, \theta(t)), Y) \tag{1}$$

with initial value $\theta(0) = \theta_0$, where $\nabla_\theta h(X, \theta_0)^T$ T is a full rank (rank M) matrix of size $N_P \times M$ with $N_P > M$. Then $\theta(\infty) = lim_{t\to\infty}\theta(t)$ exists and uniquely solves the constrained optimization problem

$$\min_\theta \|\theta - \theta_0\|_2,\ s.t.\ h(X, \theta) = Y$$

Remark 1. *Compared with the nonlinear gradient flow of DNN, the linearization in Eq.(1) is only performed on the hypothesis function h but not on the loss function or the gradient flow.*
In all these three step, from dynamics to optimization, we show that how the deep neural network favor the low frequency part.

4 An apriori generalization error bound

We begin with the definition of an FP-norm, which naturally induces a FP-space containing all possible solutions of a target NN, whose Rademacher complexity can be controlled by the FP-norm of the target function. Thus we obtain an *a priori* estimate of the generalization error of NN by the theory of Rademacher complexity. The a priori estimates follows the Monte Carlo error rates with respect to the sample size. Importantly, the estimate unravels how frequency components of the target function affect the generalization performance of DNN.

4.1 FP-norm and FP-space

We denote $Z^{d*} := \mathbb{Z}^d \backslash \{0\}$. Given a frequency weight function $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$ or $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ satisfying

$$\|\gamma\|_{l^2} = (\sum_{k \in \mathbb{Z}^d} (\gamma(k))^2)^{\frac{1}{2}} < +\infty or \, \|\gamma\|_{l^2} = (\sum_{k \in \mathbb{Z}^{d*}} (\gamma(k))^2)^{\frac{1}{2}} < +\infty$$

we define the FP-norm for all function $h \in L^2(\Omega)$:

$$\|h\|_\gamma := \left\|\hat{h}\right\|_{H_\Gamma} = (\sum_{k \in \mathbb{Z}^d} (\gamma(k))^{-2} |\hat{h}(k)|^2)^{\frac{1}{2}}$$

If $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ is not defined at 0, we set $(\gamma(0))^{-1} := 0$ in the above definition and $\|\cdot\|_\gamma$ is only a semi-norm of h. This norm comes directly from the regularization in the training process. Next we define the FP-space

$$F_\gamma(\Omega) = \left\{ h \in L^2(\Omega) : \|h\|_\gamma < \infty \right\}$$

4.2 *a priori* generalization error bound

The following lemma shows that the FP-norm closely relates to the Rademacher complexity, which is defined as

$$\hat{R}(\mathcal{H}) = \frac{1}{M} \mathbb{E}_\epsilon \left[\underset{h \in \mathcal{H}}{sup} \sum_{i=1}^{M} \epsilon_i h(x_i) \right]$$

for the function space H

Lemma 1. *(i)For $\mathcal{H} = \{h : \|h\|_\gamma < Q\}$ with $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, we have*

$$\hat{R}(\mathcal{H}) \leq \frac{1}{\sqrt{M}} Q \|\gamma\|_{l^2}$$

(ii)For $\mathcal{H}' = \{h : \|h\|_\gamma < Q, |\hat{h}(0)| \leq c_0\}$ with $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ and $(\gamma(0))^{-1} := 0$, we have

$$\hat{R}(\mathcal{H}') \leq \frac{c_0}{\sqrt{M}} + \frac{1}{\sqrt{M}} Q \|\gamma\|_{l^2}$$

Notice that the complexity is bounded by the Q, which directly comes from the property of target function. Combining the above lemma and Lemma 14 in [2], we obtain the following estimate of the generalization error bound.

Theorem 3. *Suppose that the real-valued target function $f \in F_\gamma(\Omega)$, the training dataset $\{x_i; y_i\}_{i=1}^M$ satisfies $y_i = f(x_i), i = 1, \ldots, M$, and h_M is the solution of the regularized model*

$$\min_{h - h_{ini} \in F_\gamma(\Omega)} \|h - h_{ini}\|_\gamma, \ s.t. \ h(x_i) = y_i, \ i = 1, \ldots, M$$

Then we have
(i) given $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, for any $\delta \in (0,1)$, with probability at least $1 - \delta$ over the random training samples, the population risk has the bound

$$L(h_M) \leq \|f - h_{ini}\|_\gamma \|\gamma\|_{l^2} \left(\frac{2}{\sqrt{M}} + 4\sqrt{\frac{2log(4/\delta)}{M}} \right)$$

(ii)given $\gamma : \mathbb{Z}^{d*} \to \mathbb{R}^+$ with $\gamma(0)^{-1} := 0$, for any $\delta \in (0,1)$, with probability at least $1 - \delta$ over the random training samples, the population risk has the bound

$$L(h_M) \leq (\|f - h_{ini}\|_\infty + 2\|f - h_{ini}\|_\gamma \|\gamma\|_{l^2}) \left(\frac{2}{\sqrt{M}} + 4\sqrt{\frac{2log(4/\delta)}{M}} \right)$$

When the data is sampled from low frequency function, this theorem provide a good guarantee of the generalization performance. After shuffled, the data can be seen as generated from a high frequency function, thus the generalization should be worse naturally. This comparison can partly answer the experiments in [1].

Remark 2. *By the assumption in the theorem, the target function f belongs to $F_\gamma(\Omega)$ which is a subspace of $L^2(\Omega)$. In most applications, f is also a continuous function. In any case, f can be well-approximated by a large neural network due to universal approximation theory.*

To sum up, when the h_{ini} is large, the output function has much vacillation. To decrease the h_{ini} or even set to zero, we propose a method that can achieve this goal, keep the good property of initialization, and also speed up the optimization. For more details, please read [2]

References

[1] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[2] Y. Zhang, Z.-Q. J. Xu, T. Luo, and Z. Ma. Explicitizing an Implicit Bias of the Frequency Principle in Two-layer Neural Networks. *arXiv e-prints*, May 2019.