Mathematical Theory of Neural Network Models

Lecture 9.1: Deep learning Theory: F-principle

July 19

Lecturer: Zhiqin Xu *Scribe: Yunzhen Feng, Haocheng Ju, Dingwen Kong*

1 Supervised Learning

Most important problems :

- Error analysis
- Multiple layers:what is the advantage of it?
- High-dimensional problems
- Huge number of parameters: why algorithms can find good solutions in large para space?

Philosophy : started from simple structure to take a glimpse: function with single variable.
Three error analysis:

- Approximation error
- Generalization error
- Training error

We know that single hiden layer can fit any data. But fitting is definitely not enough. Overfitting? Deep neural network says NO! DNN has large complexity but good generalization. To understand it, there are several methods.

Current approaches:

- Complexity measure and regularizer.
- Local properties of minima.
- Other technique including noise compression and etc.

We started from a toy case: the function with only one variable x. The results show somehow flatness. How to define this flatness? Fourier analysis may be useful.
The activation is smooth or with low Lipschitz, thus the learned function is flat. From the experiment in [2], the network catch the low frequency first then move to higher part. When capturing low-frequency components, it keep high components low.

The reason behind:

- Amplitude?
- Frequency?
- Does activation function matter?
- For any network structure?
- Real dataset?

Numerical simulation on one hidden layer proves the results in three frequency components.
But there are more problems. How to do Fourier transform in high dimension space?
For $\{x_i, y_i\}_{i=1}^N$ with $x_i \in \mathcal{R}^d, d = 784$ and $y \in \{0, 1, \cdots, 9\}$. The frequency comes from two aspects: its own picture and the reaction frequency of y w.r.t. x. We define

$$y_i^\delta = \sum_j y_j \exp\left(-\frac{|x_j - x_i|^2}{2\delta}\right).$$

The Gaussian filter keeps the low frequency of y_i. Then we use this y_i^δ to do the test. Since the higher frequency part is not preserved, the loses data should decrease first and then peak up. The results support it.
Now consider a network

$$\gamma = \sum_{k=1}^N a_k \sigma(w_k x + b_k),$$

with loss function

$$L = \frac{1}{2^N} \sum \left(\gamma(x_i) - f(x_i)\right)^2,$$

and tanh activation function. View it in a Fourier transfer standpoint

$$\hat{\gamma}(k) = \sum_{i=1}^N \frac{2\pi a_i \mathbf{i}}{|w_i|} \exp\left(\frac{\mathrm{i} b_i k}{w_i}\right) \frac{1}{\exp\left(-\frac{\pi k}{2w_i}\right) - \exp\left(\frac{\pi k}{2w_i}\right)}$$

The amplitude deviation between DNN outout and target function can be defined as

$$\begin{aligned} D(k) &\triangleq \gamma(k) - \hat{f}(k) \\ D(k) &= A(k) e^{j\theta(k)} \\ L(k) &= \tfrac{1}{2}|D(k)|^2. \end{aligned}$$

Theorem 1. *Consider a DNN with one hidden layer using tanh function $\sigma(x)$ as the activation function. For any frequecies k_1 and k_2 such that $|\hat{f}(k_1)| > 0$, $|\hat{f}(k_2)| > 0$, and $|k_1| > |k_2| > 0$, there exists constant c and C such that for sufficiently small δ,*

$$\frac{\mu\left(\left\{W : \left|\frac{\partial L(k_1)}{\partial \Theta_{ij}}\right| > \left|\frac{\partial L(k_2)}{\partial \theta_{ij}}\right| \text{ for all } i, j\right\} \cap B_\delta\right)}{\mu\left(B_\delta\right)} \geq 1 - C\exp(-c/\delta) \tag{1}$$

where B_δ is a ball with radius δ centered at the origin and $\mu(\cdot)$ is the Lebesgue measure.

2 For higher dimension

2.1 Qualitative understanding

- Generalization difference in different network
  The difference between performance of MNIST, CIFAR10 and parity: in MNIST and CIFAR10, the low-frequency part is the dominant part. On the contrary, parity has different frequency map.
- Early stopping is effective
  Observation: test error gets worse after some training step. Training and test only overlap at LOW frequency part.
- Compression phase

2.2 Application

Solving PDE :

- Jacobi: high frequency faster
- DNN: low frequency faster
- DNN+Jacobi: all frequency could be faster

Phase DNN Shift high frequency to low frequency
$\Rightarrow$ DNN converge low frequency
$\Rightarrow$ Shift low frequency back to high frequency

Example: Poisson Equation

$$-\Delta u(x) = g(x), \quad x \in \Omega = [-1, 1] \tag{2}$$

$$u(x) = 0, \quad x = -1, 1 \tag{3}$$

$$-\Delta u_i = -\frac{u_{i+1} - 2u_i + u_{i-1}}{(\triangle x)^2} = g\left(x_i\right) \tag{4}$$

Jacobi method:

$$\begin{aligned} &A = D - L - U \\ &u^{l+1} = D^{-1}(L + U)u^l + D^{-1}g \\ &R_J = D^{-1}(L + U) \end{aligned} \tag{5}$$

Drawback: lower frequency converges slower
DNN:

Loss: $I(u) = \int_\Omega \left(\frac{1}{2} |\nabla_x u(x)|^2 - g(x)u(x) \right) dx + \beta \int_{\partial\Omega} u(x)^2 ds$
Numerical test did show that DNN has a better performance than Jacobi method.

3 More results

3.1 One-hidden Relu DNN

Details will be presented in Lecture 9.2

3.2 An apriori generalization error bound

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

$$\min_{h - h_{\text{ini}} \in F_\gamma(\Omega)} \|h - h_{\text{ini}}\|_\gamma, \quad \text{s.t.} \quad h(x_i) = y_i, \quad i = 1, \cdots, M. \tag{6}$$

Then we have
(1)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 riskhas the bound

$$L(h_M) \leq \|f - h_{\text{ini}}\|_\gamma \|\gamma\|_{\ell^2} \left(\frac{2}{\sqrt{M}} + 4\sqrt{\frac{2\log(4/\delta)}{M}} \right) \tag{7}$$

for more details, please read [1]

References

[1] Xu, Zhang, Luo, Xiao, and Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*, 2019.

[2] Xu, Zhang, and Xiao. Training behavior of deep neural network in frequency domain. *arXiv preprint arXiv:1807.01251*, 2018.