Deeper or Wider: A Perspective from Optimal Generalization Error with Sobolev Loss

Yahong Yang

(joint work with Juncai He (KAUST)
The Pennsylvania State University

yxy5498@psu.edu

May 29, 2024

Deep Neural Networks

Define $L, N \in \mathbb{N}_+$, $N_0 = d$ and $N_{L+1} = 1$, $N_i \in \mathbb{N}_+$ for $i = 1, 2, \ldots, L$, then a σ-NN ϕ with the width N and depth L can be described as follows:

$$\boldsymbol{x} = \tilde{\boldsymbol{h}}_0 \overset{W_1, b_1}{\longrightarrow} \boldsymbol{h}_1 \overset{\sigma}{\longrightarrow} \tilde{\boldsymbol{h}}_1 \ldots \overset{W_L, b_L}{\longrightarrow} \boldsymbol{h}_L \overset{\sigma}{\longrightarrow} \tilde{\boldsymbol{h}}_L \overset{W_{L+1}, b_{L+1}}{\longrightarrow} \phi(\boldsymbol{x}) = \boldsymbol{h}_{L+1},$$

where $\boldsymbol{h}_i := \boldsymbol{W}_i \tilde{\boldsymbol{h}}_{i-1} + \boldsymbol{b}_i$, and $\tilde{\boldsymbol{h}}_i = \sigma_i(\boldsymbol{h}_i)$. Here, σ denotes the activation function, such as $ReLU(x) := \max\{x, 0\}$, $ReLU^2$, or $tanh(x)$, among others. ($L = 1$ will reduce to the shallow neural networks)

Error Analysis for Sobolev Training

Denote

$$\boldsymbol{\theta}_D := \arg\inf_{\boldsymbol{\theta}} \mathcal{R}_D(\boldsymbol{\theta}) := \arg\inf_{\boldsymbol{\theta}} \int_{(0,1)^d} |\nabla(f - \phi(\boldsymbol{x};\boldsymbol{\theta}))|^2 + |f - \phi(\boldsymbol{x};\boldsymbol{\theta})|^2 \,\mathrm{d}\boldsymbol{x}, \tag{1}$$

$$\boldsymbol{\theta}_S := \arg\inf_{\boldsymbol{\theta}} \mathcal{R}_S(\boldsymbol{\theta}) := \arg\inf_{\boldsymbol{\theta}} \sum_{i=1}^{M} \frac{|\nabla(f_i - \phi(\boldsymbol{x}_i;\boldsymbol{\theta}))|^2 + |f_i - \phi(\boldsymbol{x}_i;\boldsymbol{\theta})|^2}{M}, \tag{2}$$

where $f_i = f(\boldsymbol{x}_i)$.

Error Analysis for Sobolev Training

The overall inference error (generalization error) is $\mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S)$, which can be divided into two parts:

$$\mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S) \leq \underbrace{\mathcal{R}_D(\boldsymbol{\theta}_D)}_{\text{approximation error}} + \underbrace{\mathbf{E}\mathcal{R}_D(\boldsymbol{\theta}_S) - \mathbf{E}\mathcal{R}_S(\boldsymbol{\theta}_S).}_{\text{sample error}} \tag{3}$$

Discussion about deep neural networks

• Advantages: The approximation rate $O(W^{-\frac{2(n-m)}{d}})$ of deep neural networks is much better than traditional methods and shallow or not very deep neural networks $O(W^{-\frac{(n-m)}{d}})$.

• Disadvantages: The structure of deep neural networks is too complex, and the absolute value of parameters in deep neural networks can be very large, which can cause a large sample error, making it very challenging to train the neural network effectively.

Deep or shallow neural network, how to choose between them?[1]

[1] Y. Yang and J. He. Deeper or wider: A perspective from optimal generalization error with Sobolev Loss. ICML, 2024.

Generalization error about deep neural networks

Theorem

Let $d, L, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. *For any* $f \in W^{n,\infty}([0,1]^d)$ *with* $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$ *for* $n > k$ *and* $k = 0, 1, 2$, *we have*

$$\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k}) \leq C\left[\left(\frac{W}{(\log W)^2}\right)^{-\frac{4(n-k)}{d}} + \frac{W^2}{M}\log M\right]$$

where $W = \mathcal{O}(L(\log L)^3)$ *is the number of parameters in DeNNs,* $\mathbf{E}$ *is expected responding to* X, $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ *is an independent random variables set uniformly distributed on* $[0,1]^d$, *and* C *is independent with* M, L.

- k represents the regularity of loss functions, M denotes the number of sample points, W signifies the number of parameters of neural networks.

Optimality

Corollary

Let $d, M \in \mathbb{N}_+$, $B, C_1, C_2 \in \mathbb{R}_+$. *For any* $f \in W^{n,\infty}([0,1]^d)$ *with* $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$ *for* $n > k$ *and* $k = 0, 1, 2$, *we have*

$$\mathbf{E}\mathcal{R}_{D,k}(\boldsymbol{\theta}_{S,k}) \leq CM^{-\frac{2(n-k)}{2(n-k)+d}}$$

where the result is up to the logarithmic term, $\mathbf{E}$ *is expected responding to* X, *and* $X := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ *is an independent set of random variables uniformly distributed on* $[0,1]^d$. C *is a constant independent of* M.

Generalization error about shallow neural networks

Proposition ((Schmidt-Hieber(2020), Suzuki(2018)))

Let $\mathcal{F}$ be a set of functions defined as follows:

$$\mathcal{F} := \{\phi \text{ is a } \sigma_2\text{-NN with number of parameters } \mathcal{O}(W) \text{ and depth } \log(W) \text{ and parameters bounded by } F\},$$

where F is a universal constant larger than 1. Assume that $\|f\|_{W^{n,\infty}([0,1]^d)} \leq 1$ for $n > k$ and $k = 0, 1, 2$. If $\varepsilon > 0$ satisfies $\mathcal{N}(\varepsilon, \mathcal{F}, n) \geq 3$, then it holds that

$$\mathbf{E}\mathcal{R}_{D,k}(\theta_{S,k}) \leq C\Big[W^{-2(n-k)/d} + \varepsilon(1+\sigma) + \left(1+\sigma^2\right) \cdot \frac{\sum_{i=1}^{d} \log \mathcal{N}(\varepsilon, D_i^k \mathcal{F}, \|\cdot\|_\infty) + \mathcal{N}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty)}{M}\Big]$$

for $k = 0, 1, 2$, where C, σ are universal constants, $\mathcal{N}(\varepsilon, D_i\mathcal{F}, \|\cdot\|_\infty)$ is covering number.

Deep or shallow neural network

- Generalization error of deeper neural networks: $\mathcal{O}\left(W^{-\frac{4(n-k)}{d}} + \frac{W^2}{M}\right)$.
- Generalization error of wider neural networks: $\mathcal{O}\left(W^{-\frac{2(n-k)}{d}} + \frac{W}{M}\right)$.

When $M \geq W^{\frac{2n+2d-2k}{d}}$, the order of the generalization error in DNNs with an arbitrary number of hidden layers surpasses that of shallow neural networks.

Deep or shallow neural network

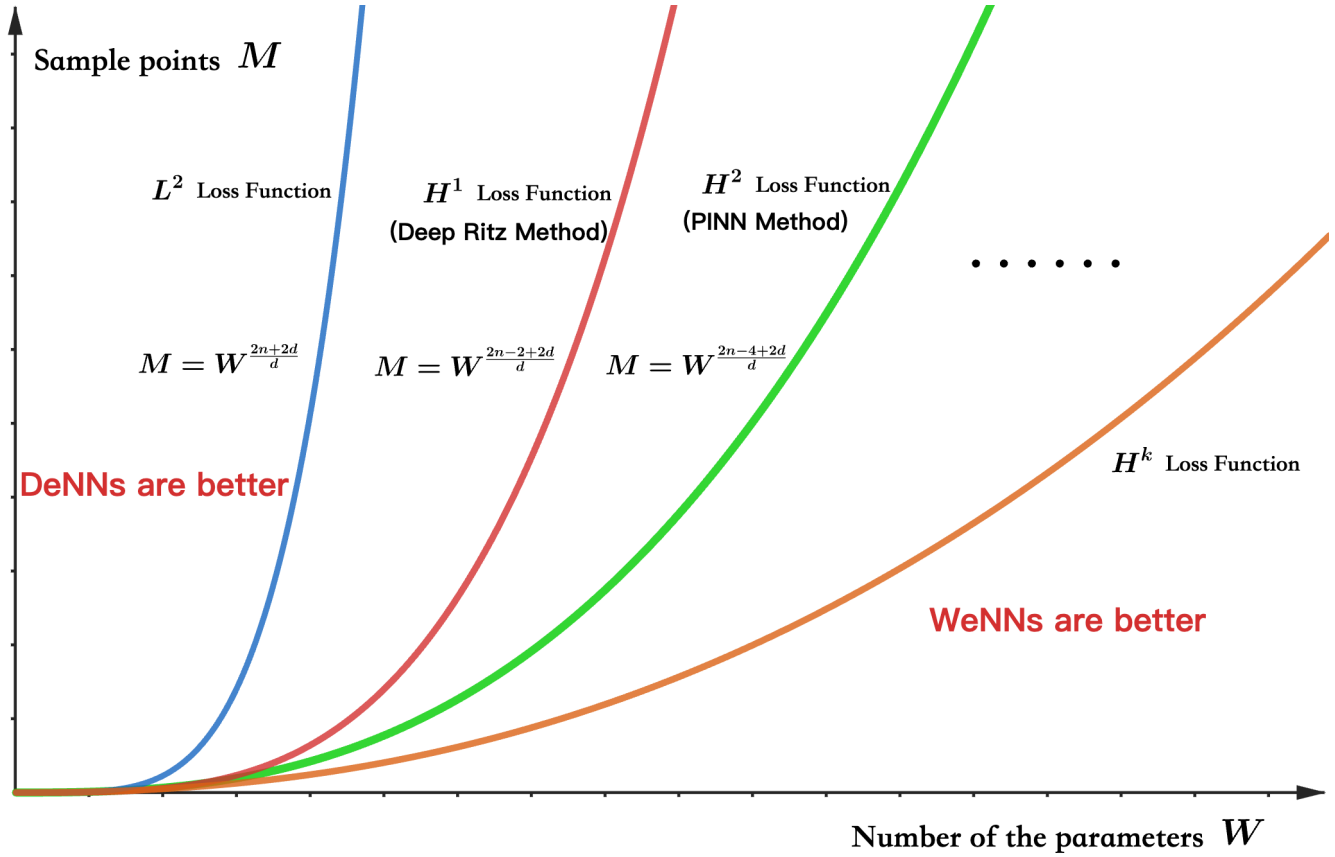

Figure: Type 1 represents NNs with shallow or moderately deep hidden layers, where the number of hidden layers is deliberately confined to be $\mathcal{O}(1)$ or $\mathcal{O}(\log(1/\varepsilon))$. Type 2 denotes DNNs with an arbitrary number of hidden layers.

Deep or shallow neural network

Neural Network	Large Data Regime	Small Data Regime
Shallow (Dep.1, Wid.20)	Mean: 0.000618, Std: 9.86e-05	Mean: 0.001512, Std: 0.000252
Deep (Dep.4, Wid.10)	Mean: 0.000369, Std: 2.57e-05	Mean: 0.004956, Std: 0.004462

Table: This table compares the performance of shallow and deep neural networks in terms of mean test performance and standard deviation across different data regimes. It illustrates how network depth and data availability impact learning outcomes, with shallow networks performing better in small data scenarios, while deep networks excel with larger datasets.

Deep or shallow neural network

• If the number of parameters in our neural networks is fixed, the choice between shallow or moderately deep hidden layers and DNNs depends on the availability of sample points.

• When the number of sample points is fixed, the decision between establishing a neural network with few parameters or one with a higher parameter count depends on the specific requirements.

• The space between the two curves signifies a transition region. Within this region, it is advisable to shift from using shallow neural networks to DNNs to effectively address the problem, especially when the pair (W, M) falls within such transitional areas.

Discussions and future works

•: Considering approximation and generalization errors for more precise functions.

•: We specifically compare DNNs, characterized by an abstract number of hidden layers, with shallow or not very DNNs in the underparameterized case. For the overparameterized case, we consider this as a topic for future research.

•: Add the training analysis like NTK, and combine three part errors.

References

1. Y. Yang, H. Yang, and Y. Xiang. Nearly optimal VC-dimension and pseudo-dimension bounds for deep neural network derivatives. Conference on Neural Information Processing Systems (NeurIPS), 2023.
2. Y. Yang, Y. Wu, H. Yang, and Y. Xiang. Nearly optimal approximation rates for deep super ReLU networks on Sobolev spaces. arXiv preprint arXiv:2310.10766, 2023.
3. Y. Yang, Y. Lu, Optimal Deep Neural Network Approximation for Korobov Functions with respect to Sobolev Norms. arXiv preprint arXiv:2311.04779, 2023.
4. Y. Yang and J. He. Deeper or wider: A perspective from optimal generalization error with Sobolev loss. arXiv preprint arXiv:2402.00152, 2024.

Thank you for your listening!