

On the Impacts of the Random Initialization in the Neural Tangent Kernel Theory

Guhan Chen, Yicheng Li, Qian Lin
Tsinghua University

November 11, 2024

Setup and Notations

Network Structure: We consider a fully-connected network with weights initialized using a standard random Gaussian distribution. The network structure is defined as follows:

$$\begin{aligned}\alpha^{(1)}(x) &= \sqrt{\frac{2}{m_1}} \left(W^{(0)}x + b^{(0)} \right); \\ \alpha^{(l)}(x) &= \sqrt{\frac{2}{m_l}} W^{(l-1)} \sigma(\alpha^{(l-1)}(x)), \quad l = 2, 3, \dots, L; \\ f(x; \theta) &= W^{(L)} \sigma(\alpha^{(L)}(x)),\end{aligned}\tag{1}$$

Network Width and Initialization: The network width m satisfies the following bounds:

$$cm \leq \min\{m_l : l = 0, 1, \dots, L\} \leq \max\{m_l : l = 0, 1, \dots, L\} \leq Cm$$

for some positive constants c and C . The elements of matrices W and vector b are all initialized as standard Gaussian random variables.

Setup and Notations

Distribution of data: For sample pairs $\{(x_i, y_i)\}_{i=1, \dots, n}$, we assume that they follows:

$$y_i = f^*(x_i) + \epsilon_i, \quad (2)$$

where f^* is the real function and $\{\epsilon_i\}$ are the noise terms. The assumption on f^* and $\{\epsilon_i\}$ will be stated later.

Training Procedure: Given training samples $\{(x_i, y_i)\}_{i=1, \dots, n}$, where $x \in \mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{X}$ is a domain with smooth boundary, the network is trained under a Mean Squared Error (MSE) loss function through gradient flow:

$$L(\theta) = \frac{1}{2n} \sum_{i=1}^n (f(x_i; \theta) - y_i)^2$$

Motivation

When the network is wide enough, we observe that the L-2 generalization error relationship between **Mirrored Initialization** and **Standard Initialization** is:

$$\|f_t^{\text{NN}} - f^*\|_{L_2}^2 \approx \|f_t^{\text{NN},(0)} - (f^* - f_0^{\text{NN}})\|_{L_2}^2$$

- ▶ f_t^{NN} : Network trained from f_0^{NN} (Standard fully-connected initialization) at time t .
- ▶ $f_t^{\text{NN},(0)}$: Network trained from initial output 0 (Mirrored fully-connected initialization) at time t .
- ▶ f^* : goal function of the regression problem.

It shows that the non-zero output works as introducing an **implicit bias** in the training process.

Motivation (Continued)

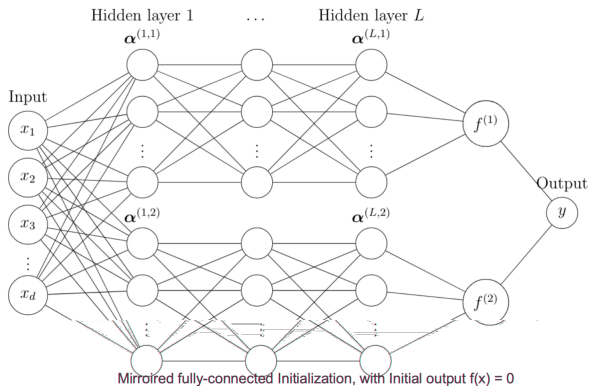


Figure: Mirrored fully-connected initialization, with initial output $f \equiv 0$.

Neural Tangent Kernel Theory

The **Gradient Flow (GF)** of the network is:

$$\frac{d}{dt}f_t^{\text{NN}}(x) = -\frac{1}{n} \sum_{i=1}^n \langle \nabla_{\theta} f_t^{\text{NN}}(x), \nabla_{\theta} f_t^{\text{NN}}(x_i) \rangle (f(x_i) - y_i).$$

When network is wide enough ($m \rightarrow \infty$), it falls into the **NTK regime** [1, 2]:

$$\lim_{m \rightarrow \infty} \langle \nabla_{\theta} f_t^{\text{NN}}(x), \nabla_{\theta} f_t^{\text{NN}}(x') \rangle \rightarrow K^{\text{NTK}}(x, x')$$

In this way, the GF of network can be approximated by KGF (Kernel Gradient Flow):

$$\frac{d}{dt}f_t^{\text{NTK}}(x) = \frac{1}{n} \sum_{i=1}^n K^{\text{NTK}}(x, x_i) (f_t^{\text{NTK}}(x_i) - y_i)$$

Methods

The generalization ability of **KGF predictor** depends on the smoothness of the regression function. Denote by $\mathcal{H}$ the RKHS of kernel $k(\cdot, \cdot)$. For $f^* \in [\mathcal{H}]^S$:

- ▶ The generalization error of KGF is about $\Theta(n^{-\frac{\beta}{d+1}})$, where β is the **EDR** (eigenvalue decay rate) of kernel $k(\cdot, \cdot)$:

$$\lambda_i(k) \asymp i^{-\beta}. \quad (3)$$

Especially, the **EDR** of NTK is $\frac{d+1}{d}$.

Key Intuition

Key point: Calculate the Smoothness of the Implicit Bias Caused by Initial Output Function

The revised goal function f^{**} converges to a GP (Gaussian Process):

$$f^{**} = f^* - f_0^{\text{NN}} \Rightarrow f^* - f^{\text{GP}} \sim \mathcal{GP}(f^*, K^{\text{RF}})$$

If f^* is smooth, the smoothness of f^{**} depends on the smoothness of f^{GP} . Denote by $\mathcal{H}^{\text{NTK}}$ the RKHS with respect to NTK. Our results shows that (Theorem 4.2)

$$\mathbb{P}(f^{\text{GP}} \in [\mathcal{H}^{\text{NTK}}]^s) = 0, \quad s \geq \frac{3}{d+1}$$

$$\mathbb{P}(f^{\text{GP}} \in [\mathcal{H}^{\text{NTK}}]^s) = 1, \quad s < \frac{3}{d+1}$$

In this way, we can directly derive the generalization error of the KGF predictor, as well as the network when width m is large enough.

Main Results

Generalization Ability of Network under Different Initialization

1. **Assumption 1:** Source condition (Smoothness of goal function)
 $f^* \in [\mathcal{H}^{\text{NTK}}]^s$, where $s \geq \frac{3}{d+1}$.
2. **Assumption 2:** Noise The training samples $\{(x_i, y_i)\}_{i=1}^n$ are generated by $y_i = f^*(x_i) + \epsilon_i$ where the noise term ϵ satisfies the following condition:

$$\mathbb{E}[|\epsilon|^m | x] \leq \frac{1}{2} m \sigma^2 L^{m-2}, \quad a.e. x \in \mathcal{X}$$

for some constant $\sigma, L, m, n \geq 2$.

Main Results (Continued)

Results on Generalization Ability:

► Mirrored Initialization (Existing Result)[3]:

$$\|f_t^{\text{NN}} - f^*\|_{L_2}^2 \leq \mathcal{O}\left(n^{-\frac{s(d+1)}{s(d+1)+d}}\right)$$

► Standard Initialization (Theorem 4.3, 4.4):

$$\|f_t^{\text{NN}} - f^*\|_{L_2}^2 \approx \Theta\left(n^{-\frac{3}{d+3}}\right)$$

When the smoothness s is close to 1 (a common assumption), the generalization error of mirrored initialization is approximately $n^{-\frac{1}{2}}$ and is shown to be minimax optimal. However, in contrast, the generalization error of **commonly used** standard initialization scales as $n^{-\frac{3}{d+3}}$, highlighting the so-called Curse of Dimensionality.

Comparison of Mirrored Initialization and Standard Initialization

We train wide networks under **mirrored initialization** and **standard initialization** with a smooth goal function and under different sample sizes n . The figure compares the MSE generalization error of the two initialization methods across varying n values.

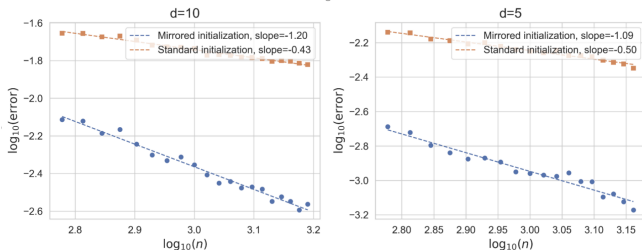


Figure: Comparison plot of generalization error.

Smoothness of Real Datasets

We evaluate the smoothness of different real-world datasets by calculating the smoothness of their goal functions.

With the input dimension $d = 784, 3072, 784$, the smoothness of initialization function is equal to $\frac{3}{d+1} \approx 0$. However, the smoothness of real datasets is far better than $\frac{3}{d+1}$, which implies that **standard initialization will indeed destroy the generalization performance**.

Dataset	Dimension	Smoothness
MNIST	$28 \times 28 \times 1$	0.40
CIFAR-10	$32 \times 32 \times 3$	0.09
Fashion-MNIST	$28 \times 28 \times 1$	0.22

Table: Smoothness of goal functions for popular datasets.

Conclusion

Summary of Findings

- ▶ This study highlights the importance of initialization techniques in neural networks and their effects on generalization abilities, especially the superiority of mirrored initialization over standard initialization.
- ▶ Under NTK theory, the learning rate $n^{-\frac{3}{d+3}}$ with standard initialization performs so poorly that we have reason to believe NTK theory cannot fully explain the superior performance of neural networks.

Thank you!

References I



Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song.

A convergence theory for deep learning via over-parameterization.

In International conference on machine learning, pages 242–252. PMLR, 2019.



Arthur Jacot, Franck Gabriel, and Clément Hongler.

Neural tangent kernel: Convergence and generalization in neural networks.

Advances in neural information processing systems, 31, 2018.



Yicheng Li, Zixiong Yu, Guhan Chen, and Qian Lin.

On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains.

Journal of Machine Learning Research, 25(82):1–47, 2024.