Commun. Comput. Phys.
doi: 10.4208/cicp.OA-2025-0185

Vol. **40**, No. 1, pp. 61-87
July 2026

Structured First-Layer Initialization Pre-Training Techniques to Accelerate Training Process Based on ε-Rank

Tao Tang[4,7], Jiang Yang[1,3,5,6,*], Yuxiang Zhao[1] and Quanhui Zhu[1,2]

[1] *Department of Mathematics, Southern University of Science and Technology, Shenzhen, China.*
[2] *Department of Mathematics, National University of Singapore, Singapore.*
[3] *SUSTech International Center for Mathematics, Shenzhen, China.*
[4] *School of Mathematics and Statistics, Guangzhou Nanfang College, Guangzhou, China.*
[5] *Guangdong Provincial Key Laboratory of Computational Science and Material Design, Southern University of Science and Technology, Shenzhen, China.*
[6] *National Center for Applied Mathematics Shenzhen (NCAMS), Shenzhen, 518055, P.R. China.*
[7] *Zhuhai SimArk Technology Co., LTD, Zhuhai, Guangdong, China.*

Received 17 July 2025; Accepted (in revised version) 18 November 2025

Abstract. Training deep neural networks for scientific computing remains computationally expensive due to the slow formation of diverse feature representations in early training stages. Recent studies [37] identify a staircase phenomenon in training dynamics, where loss decreases are closely correlated with increases in ε-rank, reflecting the effective number of linearly independent neuron functions. Motivated by this observation, this work proposes a structured first-layer initialization (SFLI) pre-training technique to enhance the diversity of neural features at initialization by constructing ε-linearly independent neurons in the input layer. We present systematic initialization schemes compatible with various activation functions and integrate the strategy into multiple neural architectures, including modified multi-layer perceptrons and physics-informed residual adaptive networks. Only needing to add one line of code to conventional stochastic gradient descent algorithms, extensive numerical experiments on function approximation and PDE benchmarks, demonstrate that SFLI significantly improves the initial ε-rank, accelerates convergence, mitigates spectral bias, and enhances prediction accuracy.

AMS subject classifications: 68T07, 65Z05, 35Q68

Key words: Staircase phenomenon, ε-rank, structured first-layer initialization, deep neural network, training acceleration.

*Corresponding author. *Email addresses:* ttang@nfu.edu.cn (T. Tang), yangj7@sustech.edu.cn (J. Yang), 12131241@mail.sustech.edu.cn (Y. Zhao), zhuqh@nus.edu.sg (Q. Zhu)

1 Introduction

Neural networks have become a cornerstone of modern machine learning, achieving remarkable accuracy in both classical machine learning tasks and emerging areas such as scientific computing and partial differential equation (PDE) modeling. Despite their success, training such models remains computationally intensive, often requiring large-scale resources and prolonged optimization. This has motivated a broad range of efforts to accelerate the training process, including architectural innovations, advanced optimization techniques, and improvements in loss function design.

Among these efforts, particular attention has been paid to the learning dynamics of neural networks [6, 10, 17, 18, 22, 40]. A key observation in training dynamics is the so-called frequency principle (F-Principle) or spectral bias [24, 34–36, 39], which states that neural networks tend to fit low-frequency components of the target function earlier in training. To address this, multiscale network designs like MscaleDNN [19, 20] have been proposed, which incorporate hierarchical frequency encodings. The performance of such architectures strongly depends on the behavior of the first hidden layer, particularly the diversity of its initial activations, a critical aspect that remains insufficiently studied.

A standard multi-layer perceptron (MLP) can be formulated as follows:

$$\begin{cases} y_0 = x, \\ y_l = H(y_{l-1};\theta_l), \quad l = 1,\cdots,L, \\ y = \beta \cdot y_L, \end{cases} \tag{1.1}$$

where $x \in \mathbb{R}^d$ is the input, $y_l \in \mathbb{R}^{n_l}$ is the neurons of the l-th hidden layer, and $\beta \in \mathbb{R}^{n_L}$ is the coefficients in the output layer. Each layer mapping H represents the structure of the hidden layer, and a fully connected layer can be explicitly given by $H(y_{l-1};\theta_l) = \sigma(W_l y_{l-1} + b_l)$, where $W_l \in \mathbb{R}^{n_{l+1}\times n_l}$, $b_l \in \mathbb{R}^{n_{l+1}}$, are trainable parameters. The activation function σ is applied element-wise. The neurons $y_L(x)$ in the final hidden layer can be viewed as a collection of scalar neuron functions defined on the input domain, and the final output y of the network is a linear combination of these functions. This perspective aligns with interpretations in the deep finite element method [32] and the finite neuron method [33], where $y_L(x)$ serves as a set of basis functions.

The training dynamics of deep neural networks have attracted significant interest, particularly in understanding how low-dimensional representations are formed during optimization [1, 23]. Recent work [37] has identified *staircase phenomenon*: **In training dynamics, the loss function often decreases rapidly along with a significant growth of linear independence of neuron functions.** A consistently low ε-rank of y_L during training indicates a lack of functional diversity among neurons, which in turn limits the expressive power of the network. Such limitations pose a critical challenge for problems in scientific computing, where resolving fine-scale or high-frequency structures is essential.

Motivated by these insights, this work introduces a novel pre-training strategy