Initialization Strategies for MLPs

Tingyu Gao
Yifang Huang
Applied Machine Learning in Python – LMU
Gao.Tingyu@campus.lmu.de
Yifang.Huang@campus.lmu.de

July 12, 2025

1 Task Overview

This report investigates the impact of weight initialization strategies on the trainability and signal stability of fully connected multilayer perceptrons (MLPs). Experiments are conducted on the MNIST dataset. We evaluate the performance of He, Orthogonal, and other initialization strategies across networks of varying depths, using ReLU, Tanh, Sigmoid, and GELU activation functions. The results demonstrate that He initialization consistently outperforms other methods in terms of both trainability and signal stability, particularly when applied to ReLU networks.

2 Methods

We implemented MLPs with varying depths (5, 10, 20, 30, 40, 50 layers), a fixed hidden layer width (128 units), and tested multiple activation functions (ReLU, Tanh, Sigmoid, and GELU). Our primary focus was on comparing five initialization strategies:

- **He:** Gaussian: $\mathcal{N}\left(0, \frac{2}{\text{fan\_in}}\right)$, Uniform: $\mathcal{U}\left(-\sqrt{\frac{6}{\text{fan\_in}}}, \sqrt{\frac{6}{\text{fan\_in}}}\right)$.
- **Xavier:** Gaussian: $\mathcal{N}\left(0, \sqrt{\frac{2}{\text{fan\_in}+\text{fan\_out}}}\right)$, Uniform: $\mathcal{U}\left(-\sqrt{\frac{1}{\text{fan\_in}}}, \sqrt{\frac{1}{\text{fan\_in}}}\right)$.
- **Orthogonal:** $W = \sqrt{\frac{2}{\text{fan\_in}}} \cdot Q$, where Q is an orthogonal matrix obtained via QR decomposition of a random matrix.
- **Normal:** $\mathcal{N}(0, 0.1)$.
- **Truncated normal:** $W_{ij} \sim \text{Trunc}\,\mathcal{N}\left(0, \frac{2}{\text{fan\_in}}; -2\sigma, +2\sigma\right)$, where values beyond two standard deviations are discarded and re-sampled.

The activation functions Tanh $\left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)$, Sigmoid $\left(\frac{1}{1+e^{-x}}\right)$, and GELU $\left(x \cdot \Phi(x) = x \cdot \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]\right)$ were chosen for comparison with ReLU.

We compared Adam, standard gradient descent (GD), and stochastic gradient descent (SGD). GD was significantly slower due to full-batch updates, while Adam's adaptive learning rates altered optimization dynamics, making comparisons across initializations less consistent. To ensure fair evaluation under fixed training conditions, we adopted SGD as the optimizer in all experiments, using standard cross-entropy loss for the multi-class classification task on MNIST.

The MLPs were trained with an input size of 784, output size of 10, hidden layer size of 128, and depths ranging from 5 to 50 layers, using a batch size of 128, learning rate of 0.01, and 50 epochs.

3 Experiments and Results

We began by investigating the impact of different variance scaling constants in Gaussian-based He initialization. Smaller scaling factors were found to weaken signal propagation across layers, occasionally leading to neuron inactivity due to diminished activations. In contrast, larger scaling factors help preserve activation magnitudes throughout training, but may introduce instability and increase the risk of gradient explosion. Our experimental results demonstrate that a scaling factor of 2 strikes the best balance, enabling stable signal propagation without causing exploding or vanishing gradients, and achieving the fastest convergence across all tested settings. (see Figure 1).

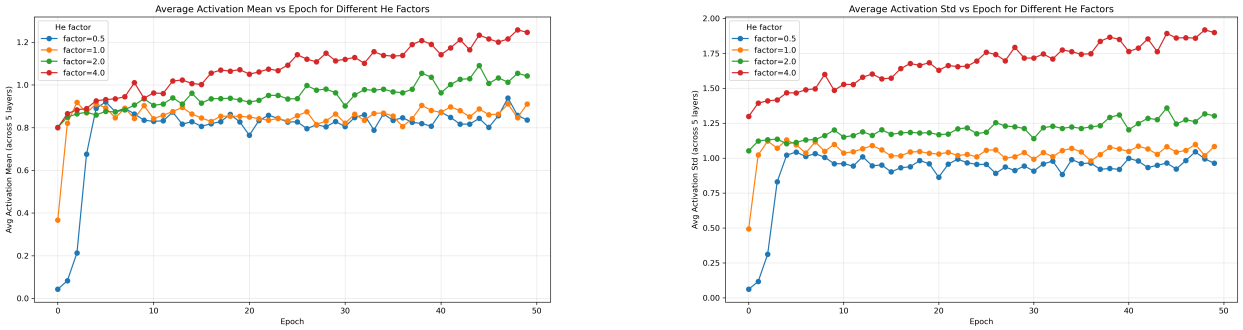

Figure 1: Activation mean and standard deviation across epochs for different He factors.

We compared combinations of different activation functions and initialization strategies. Across all activation functions tested, He initialization consistently achieved 95% accuracy in the fewest epochs. It also maintained high accuracy as network depth increased. For instance, with ReLU (see Figure 2).

In contrast, Orthogonal and other non-He initialization methods became unsuitable for very deep networks. With Orthogonal initialization, activation collapse occurred at around 30 layers, where the outputs of each layer converged to near-constant vectors, rendering the network ineffective. Signs of activation explosion were already observed as early as 20 layers, at which point the model's generalization ability began to degrade (see Figure 3).

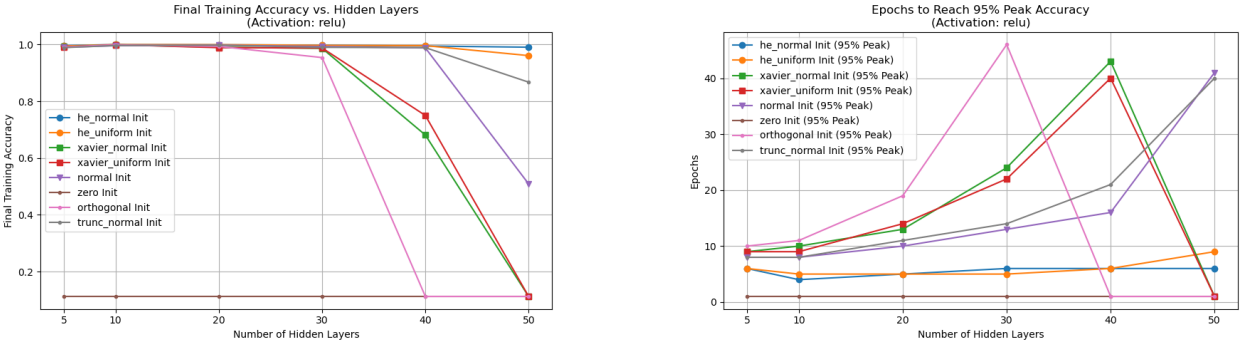

Figure 2: Accuracy and convergence speed across different He initialization scales.

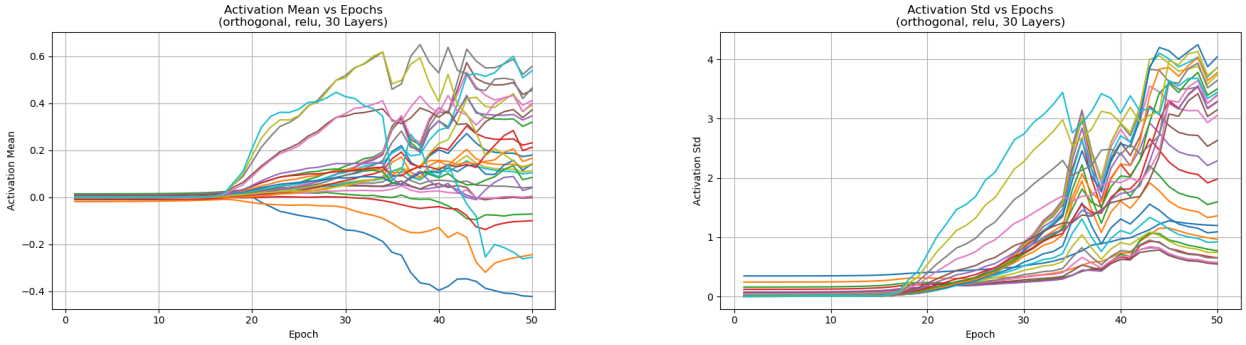

Figure 3: Activation mean and standard deviation for Orthogonal initialization at 30 layers.

4 Discussion

From a theoretical perspective, our empirical findings align well with predictions from mean-field theory. Specifically, in deep ReLU networks, mean-field analysis suggests that in order to preserve the variance of activations across layers, the weights should be initialized such that $\mathrm{Var}[x^\ell] \approx \mathrm{Var}[x^{\ell-1}]$.

To illustrate this, consider the ReLU activation $\phi(z) = \max(0, z)$, which effectively zeroes out half of its input. Assuming inputs z follow a zero-mean Gaussian distribution, we have $\mathbb{E}[\phi(z)^2] = \frac{1}{2}\mathrm{Var}[z]$. If the weights are initialized with variance $\sigma_w^2 = \frac{2}{\text{fan\_in}}$, then the output variance becomes:

$$\mathrm{Var}[x^\ell] \approx \sigma_w^2 \cdot \mathbb{E}[\phi(z)^2] = \frac{2}{\text{fan\_in}} \cdot \frac{1}{2} \cdot \mathrm{Var}[x^{\ell-1}] = \mathrm{Var}[x^{\ell-1}].$$

This derivation confirms that the variance of activations remains constant across layers under this initialization—precisely the behavior we observe empirically when using a scaling factor of 2.

While the observed behavior suggests that signal magnitudes are approximately preserved, our initialization does not fully ensure uniform propagation in all directions. In particular, although average gradient scales remain stable in He-initialized ReLU networks, the input-output Jacobian may still have highly variable singular values. Nevertheless, the empirical stability of activations and gradients indicates a partial realization of well-conditioned propagation, which likely contributes to the improved trainability observed.

Overall, our results highlight the strong interplay between activation functions and initialization strategies in deep networks. Tanh demonstrates broad compatibility, performing stably across various depths and initialization methods. In contrast, ReLU shows a strong dependence on He initialization, as only He consistently enables successful training as depth increases. Among all initialization strategies, He initialization stands out for its fast convergence, stable signal propagation, and robust performance across all tested activations and depths—even outperforming Xavier when paired with Tanh. On the other hand, our findings confirm that Sigmoid is fundamentally unsuitable for deep architectures, suffering from severe vanishing gradients beyond 5–6 layers, and exhibiting poor training stability and efficiency even at shallower depths.