A SMALL STUDY ON NETWORK DEPTH, GRADIENTS, AND LEARNING BEHAVIOR

Abstract

- This study investigates how increasing neural network depth influences learning behavior under controlled experimental conditions. Rather than optimizing performance, the objective is to observe how depth affects training dynamics, including loss behavior, gradient norms, and generalization patterns. Fully connected neural networks were implemented from scratch using NumPy, and depth was systematically varied while keeping datasets, initialization, training procedures, and random seeds fixed. Experiments were conducted on two synthetic nonlinear classification tasks of increasing complexity: a circle dataset and a nested rings dataset. On the simpler dataset, increasing depth did not improve performance, while gradient-related metrics and loss oscillations increased, suggesting growing optimization instability without generalization gains. On the more complex dataset, moderate increases in depth improved testing accuracy up to an intermediate level, after which both performance and gradient-related metrics plateaued. Overall, the findings suggest that network depth influences learning behavior within a limited effective range, and that additional depth beyond this range may increase optimization complexity without proportional performance benefits under the current experimental setting.

I/Introduction

Motivation:

This study was motivated by curiosity about how neural networks are structured, particularly how network depth influences learning behavior. Rather than assuming that deeper models are inherently superior, the goal is to observe how varying the number of layers affects training dynamics under controlled conditions. Specifically, the study explores whether there exists a range of depth within which learning becomes more effective or stable, and whether increasing depth beyond that range introduces diminishing returns or instability. This work does not aim to propose an optimal architecture or maximize performance. Instead, it serves as a small observational reference focused on understanding how depth shapes loss behavior, gradient dynamics, and generalization patterns.

Paper philosophy (observational-first):

This study follows an exploratory, observation-first approach. Instead of fixing conclusions in advance, depth-related patterns were examined progressively through experimentation. For each configuration, performance metrics, loss curves, and gradient statistics were recorded and compared across datasets. Several implementation-related issues emerged during experimentation—such as dead ReLU activations, incomplete metric tracking, and unsuitable initial hyperparameters. Rather than removing or hiding these inconsistencies, they were

documented and addressed systematically before continuing to the second dataset. All neural networks were implemented from scratch using NumPy, without relying on external deep learning frameworks. This design choice was made to ensure full visibility into forward and backward computations, allowing depth-related behaviors to be examined directly rather than through a black-box implementation.

II/Research Questions and Working Hypotheses

Research questions:

This study is motivated by a simple question I had when first learning about neural networks: does increasing depth actually make a model better at learning? Rather than interpreting "better" only in terms of accuracy, this study expands the question to consider multiple aspects of the learning process. Specifically, the research focuses on how changes in network depth affect not only final accuracy, but also gradient norms, loss behavior, and the shape of loss curves during training. By shifting the focus away from accuracy alone, the study aims to compare learning behavior across different depths from several perspectives, instead of relying on a single performance metric.

Working Hypotheses:

At the beginning of this study, I initially believed that increasing network depth would consistently lead to improved performance on nonlinear datasets, measured through loss values and accuracy. This belief came from the idea that adding depth also increases the number of neurons involved in processing the data, which could allow the model to represent more complex patterns and learn more effectively. However, as I reflected on this assumption before conducting the experiments, this initial belief was gradually replaced by a more cautious hypothesis. I began to consider that increasing depth might improve training dynamics only up to a certain point. Beyond that point, additional depth could introduce unnecessary complexity, leading to unstable loss behavior, increased gradient magnitude or variance, and diminishing improvements in testing accuracy. This study is therefore guided by the hypothesis that depth improves learning only within a limited range, rather than indefinitely.

III/Experimental Setup, Scope, Methodological Notes

1/From-scratch NN :

All modules used in this study were implemented entirely from scratch, with NumPy serving as the only supporting library for matrix operations. No deep learning frameworks such as PyTorch, TensorFlow, or Keras were used at any stage of the project. Several core modules were written and reused across experiments, including activation functions (ReLU and Sigmoid), linear layers, the loss function (Binary Cross-Entropy), the training loop, and a gradient diagnostics module used to measure gradient norms during training.

2/Controlled variables :

As stated in the project README, multiple experimental factors were kept fixed across all runs. These included the data generation logic, training procedure, loss function, activation functions, weight initialization strategy, and random seeds. By holding these variables constant and isolating network depth as the only changing factor, the study aims to make observations of loss behavior and gradient dynamics more direct and interpretable, in line with the exploratory nature of the research. During the course of experimentation, several issues were encountered, including dead ReLU activations and unstable training caused by unsuitable weight initialization. In particular, these issues required rebaselining at depths 2 and 4 in the circle dataset. To preserve experimental integrity and consistency, the affected experiments were restarted. Adjustments such as learning rate changes and the adoption of He initialization are documented in the Fairness in Depth Comparison section.

3/Dataset design (Circle → Nested Rings):

The study evaluates networks of varying depth (1, 2, 4, and 8 layers) across two synthetic datasets with different levels of complexity: the circle dataset and the nested rings dataset. The circle dataset serves as an initial sanity check, providing a simple nonlinear structure to verify the correctness of the implementation and observe basic loss behavior before moving to a more challenging setting. The nested rings dataset is then used as a stress test, where differences in loss curves, gradient norms, and training stability become more pronounced as depth increases.

4/Scope:

This study is limited to fully connected neural networks implemented from scratch, using fixed datasets and a fixed training procedure. The scope does not include convolutional architectures, large-scale real-world datasets, or comparisons with state-of-the-art models. These constraints are intentionally chosen to prioritize clarity and interpretability over performance, allowing the effects of network depth on learning behavior to be observed more directly.

5/Methodological Notes:

This study follows an exploratory and observational methodology rather than a performance-driven one. The experimental design is centered on isolating network depth as the primary variable, while keeping data generation, training procedure, loss function, activation functions, and initialization strategies consistent across experiments. Instead of optimizing architectures for maximum accuracy, the focus is placed on observing learning behavior through loss dynamics, gradient statistics, and stability patterns. Methodological adjustments, such as rebaselining or metric expansion, are treated as part of the learning process and are explicitly documented rather than hidden. This approach aims to preserve interpretability and honesty in observation, even when results are imperfect or unstable.

IV/Observed Failures, The Absence Of Crucial Metrics And Training Instabilities

1/Observed Failures :

Dead ReLU phenomenon: In Dataset 1 (circle), a dead ReLU phenomenon was observed following the baseline experiments. This issue arises when ReLU activations progressively truncate the variance of the pre-activation values, leading to gradients that are extremely small or effectively zero when propagated back to the initial layers. As a result, weight updates become largely ineffective, preventing meaningful training progress. This behavior was identified as a technical issue rather than a feature-related limitation. Leaving the issue unaddressed would distort the original purpose of the study by shifting the focus from analyzing learning metrics to merely evaluating model survivability. The problem was resolved by adopting He initialization, which was then consistently applied from the restarted experiments in Dataset 1 through all experiments in Dataset 2.

Initial hyperparameter configuration: Within this study, the primary hyperparameter adjustment involved the learning rate. The learning rate was reduced by a factor of 100 (from 0.1 to 0.001) to preserve the integrity of the research objective. This adjustment allowed deeper models to remain trainable and stable, while not significantly affecting models that were already able to converge under the original setting. Following this change, training behavior across all depths became noticeably more stable. The rebaselining of Dataset 1 included this learning rate modification, which was subsequently applied consistently to all models evaluated on both datasets.

Experimental limitations in Dataset 1: Several important metrics were initially absent during the early stages of Dataset 1 experiments. These included gradient-related metrics (mean gradient norm, coefficient of variation, and gradient norm standard deviation), training and testing accuracy, as well as repeated runs across multiple random seeds (10 runs). The absence of accuracy metrics and repeated trials reduced the objectivity of early observations and increased the risk of luck-based or cherry-picked interpretations. Gradient-related metrics were later introduced to provide a more comprehensive view of training behavior, beyond loss and accuracy alone. These limitations were addressed toward the later stage of Dataset 1 and fully adopted in Dataset 2. The updated experimental protocol and reflections are documented in dataset1_reflection_dataset2_protocol.md.

2/Training Instability :

1/Dataset 1 (circle): Increasing network depth does not lead to a noticeable change in performance, as both training and testing accuracy remain relatively similar across models. The overall shapes of the loss curves are also largely comparable, especially during the early stages of training. As network depth increases, loss curves remain convergent but exhibit progressively reduced smoothness. Shallower models, particularly Depth 2, achieve the most stable optimization, while deeper models show increasing oscillations, with Depth 8 standing out as a clear exception due to its noticeably larger fluctuations. Although these instabilities

do not prevent convergence, they indicate growing optimization difficulty as depth increases. Similarly, gradient-based metrics reveal a relatively identical pattern. The mean gradient norm, coefficient of variation, and gradient norm standard deviation all show a consistent increasing trend as network depth increases. Both the numerical metrics and gradient norm plots indicate an overall growth in gradient magnitude and variability with depth

2/Dataset 2 (nested rings) : In contrast to the circle dataset, the nested rings dataset reveals a clearer relationship between network depth and learning behavior. Testing accuracy increases as depth grows, but this improvement appears to saturate beyond a certain point. The most noticeable gains occur from the baseline model to depth 2 and depth 4, while performance plateaus at depth 6 and depth 8. Loss curve shapes also differ more clearly across depths compared to Dataset 1. The initial slopes are steeper, indicating more active early learning. For the baseline model, a clear plateau phase is not observed; instead, continuous improvement is recorded throughout training, although the rate of improvement slows during the late stage. Notably, training instability is most evident in the baseline model rather than in deeper architectures. An additional irregularity emerges when increasing depth from 2 to 4. The depth-4 model exhibits the largest degree of instability, as observed through loss curve fluctuations, regardless of comparison with the baseline model. At the same time, this model achieves the highest testing accuracy. Deeper models, such as depth 6 and depth 8, display comparatively smaller and more subtle instability than the depth-4 model. Gradient-based metrics further clarify this behavior. The mean gradient norm and coefficient of variation increase sharply from the baseline model up to depth 6, while depth 8 produces values close to depth 6, suggesting a possible saturation effect. In contrast, the gradient norm standard deviation follows a pattern similar to testing accuracy: it peaks at depth 4 and then decreases or stabilizes in deeper models.

V/Results Overview

1/Dataset 1 (Circle):

On a relatively simple nonlinear dataset such as the circle dataset, increasing network depth led to observations that differed from the initial expectation. While deeper models were expected to improve overall learning-related metrics, testing and training accuracy remained largely unchanged across depths, with all models achieving similarly high performance. In contrast, gradient-related metrics—including mean gradient norm, gradient norm standard deviation, and coefficient of variation—showed a clear upward trend as depth increased, rising noticeably up to depth 6 before saturating at depth 8. Across all depths, loss curves enter a similar plateau phase after approximately 100 epochs, indicating that increased depth does not extend effective learning beyond this point. At the same time, loss curves exhibited mildly increasing oscillations and occasional spikes as depth increased, although these instabilities were not severe enough to prevent convergence. Despite these changes in loss and gradient behavior, performance remained nearly constant. Overall, for this dataset, deeper and more complex models did not provide performance benefits, but instead introduced greater instability in gradient-related metrics. Taken together, these findings do not support the

hypothesis that increasing depth improves generalization on simple nonlinear datasets, as performance remains saturated while optimization instability increases.

2/Dataset 2 (Nested Rings):

In contrast to the circle dataset, the nested rings dataset revealed a clearer relationship between network depth and learning behavior. Increasing depth led to consistent improvements in testing accuracy from the baseline model up to depth 4, after which performance began to saturate at depths 6 and 8. Other metrics—including mean gradient norm, gradient norm standard deviation, coefficient of variation, and features of the loss curves—followed a similar trend, increasing alongside performance before reaching a plateau. One notable deviation from this pattern was observed in training accuracy. While testing accuracy peaked at depth 4, training accuracy reached its lowest point at this depth and showed little further improvement in deeper models. Overall, results from this dataset aligned more closely with the refined hypothesis that depth improves learning behavior only within a limited range, with the primary anomaly being the behavior of training accuracy at intermediate depths.

VI/Analysis / Interpretation

1/Dataset 1 (Circle)

Loss behavior and loss curve shape. As network depth increases, loss curves exhibit progressively stronger oscillations with increasing depth, with the most pronounced fluctuations observed in the deepest model, along with a wider range of loss variation. However, these changes do not prevent convergence across models. This suggests that increasing depth mainly introduces training instability rather than substantially altering the convergence outcome on this dataset.

Performance. Neither training nor testing accuracy shows a consistent upward trend across different depths. Instead, performance appears to saturate across all tested models. This indicates that increasing depth does not improve predictive performance on a dataset with relatively low complexity, where even shallow models already achieve high and stable accuracy.Dataset 1 suggests that for sufficiently simple nonlinear problems, even the limited beneficial range of depth may become negligible.These results do not support the hypothesis that increased depth improves generalization on the circle dataset, suggesting that depth alone is insufficient to yield performance gains under the current experimental setting.

Gradient norms. Gradient-related metrics, including mean gradient norm, standard deviation, and coefficient of variation, increase steadily with depth before saturating at the deepest models. When considered alongside the stable performance metrics, this pattern suggests that increasing depth primarily amplifies gradient activity without translating into meaningful performance gains.Compared to loss curves, gradient-based metrics make optimization instability more explicit, suggesting that gradient statistics provide a more sensitive indicator of depth-induced training difficulty.

2/Dataset 2 (Nested Rings)

Loss behavior and loss curve shape. In contrast to the circle dataset, increasing depth on the nested rings dataset produces clearer changes in loss behavior. Improvements in loss dynamics are observed up to a certain depth, after which further increases result in diminishing changes. This suggests that for more complex datasets, depth can influence convergence behavior, but only within a limited range before the effect saturates. The depth-4 model's case suggests moderate instability may reflect active representation reshaping rather than pathological training behavior.

Performance. Training and testing accuracy exhibit diverging trends as depth increases. Training accuracy improves substantially from the baseline to moderate depths and then saturates, with minor fluctuations rather than consistent further gains as depth increases, while testing accuracy improves up to an intermediate depth before saturating. This pattern suggests that moderately deep models can generalize better than shallow ones; however, further increases in depth do not yield additional generalization benefits and may even lead to mild performance degradation.

Gradient norms. Gradient-related metrics follow a similar pattern to testing accuracy: increasing with depth up to an intermediate model and then stabilizing at higher depths. The alignment between gradient norm saturation and performance saturation suggests that the influence of depth on internal gradient behavior becomes limited beyond a certain architectural scale.

VII/Limitations

- This study is intentionally limited in scope in order to maintain clarity, control, and interpretability in observing the effects of network depth. As a result, the findings and interpretations should be understood within the following constraints. First, the experiments are restricted to fully connected neural networks with relatively small depth and width. Architectural elements commonly used in modern deep learning systems—such as convolutional layers, residual connections, normalization techniques, or attention mechanisms—are not included. While this simplification allows depth to be isolated as the primary variable, it also limits the extent to which the observations can be generalized to more complex or modern architectures. Second, the training procedure is deliberately kept simple and fixed across all experiments. A single optimization setup, loss function, and activation configuration is used, without adaptive optimizers, learning rate schedules, or regularization techniques. Although this consistency is necessary for a controlled comparison, it may not reflect training dynamics observed under more advanced optimization strategies. Third, the study relies exclusively on synthetic two-dimensional datasets with controlled structure. While the circle and nested rings datasets are suitable for visualizing learning behavior and gradient dynamics, they do not capture the noise, dimensionality, or distributional complexity of real-world

data. Consequently, the role of depth observed in this study may differ when applied to higher-dimensional or more heterogeneous datasets. Fourth, the range of depth explored in this study is limited. The deepest models remain relatively shallow compared to large-scale deep learning systems. As such, conclusions about gradient behavior, instability, or saturation should not be extrapolated to very deep networks or large-capacity models without further empirical validation. Finally, although multiple metrics are used to characterize learning behavior, the study does not account for computational cost, memory usage, or training efficiency. The analysis therefore focuses on learning dynamics rather than practical deployment considerations. Together, these limitations define the scope within which the observations and interpretations of this study should be understood. Rather than weakening the conclusions, they clarify the specific context in which the effects of network depth are examined.

VIII/Conclusion

Under a controlled experimental setting where the core components were kept fixed—including the model modules and codebase, data generation logic, training procedure, loss function, activation functions, weight initialization strategy, and random seeds—increasing network depth does not produce a uniform or consistent effect across different datasets. In this study, Dataset 1 exhibited behavior that deviated from the initial hypothesis: training and test performance remained largely saturated as depth increased, while gradient norm mean, standard deviation, and coefficient of variation rose noticeably, indicating growing gradient-related training instability without corresponding performance gains. In contrast, Dataset 2 partially aligned with the hypothesis, showing initial benefits from increased depth; however, both test accuracy and gradient-related metrics plateaued beyond depth 4. This suggests that increasing depth alone may be insufficient to overcome limitations related to the fixed-width fully connected design, and that the benefits of deeper models diminish once the network becomes relatively deep under the current setup. Notably, Dataset 2 also revealed an inverse trend between training and test accuracy across depths, which may indicate that deeper models are not necessarily better at fitting the observed training data but can exhibit improved generalization to unseen data relative to shallower architectures.