
A rationale from frequency perspective for grokking in training neural network

Zhangchen Zhou^{1,2}, Yaoyu Zhang^{1,2}, Zhi-Qin John Xu^{1,2*}

¹ Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University

² School of Mathematical Sciences, Shanghai Jiao Tong University

Abstract

Grokking is the phenomenon where neural networks (NNs) initially fit the training data and later generalize to the test data during training. In this paper, we empirically provide a frequency perspective to explain the emergence of this phenomenon in NNs. The core insight is that the networks initially learn the less salient frequency components present in the test data. We observe this phenomenon across both synthetic and real datasets, offering a novel viewpoint for elucidating the grokking phenomenon by characterizing it through the lens of frequency dynamics during the training process. Our empirical frequency-based analysis sheds new light on understanding the grokking phenomenon and its underlying mechanisms.

1 Introduction

Neural networks (NNs) exhibit a remarkable phenomenon where they can effectively generalize the target function despite being over-parameterized (Breiman, 1995; Zhang et al., 2017). Conventionally, it has been believed that during the initial training process, the test and training losses remain relatively consistent. However, in recent years, the grokking phenomenon has been observed in Power et al. (2022), indicating that the training loss decreases significantly faster than the test loss (the test loss may even increase) initially, and then the test loss decreases rapidly after a certain number of training steps. Interpreting such training dynamics is crucial for understanding the generalization capabilities of NNs.

The grokking phenomenon has been empirically observed across a diverse array of problems, including algorithmic datasets (Power et al., 2022), MNIST, IMDb movie reviews, QM9 molecules (Liu et al., 2022), group operations (Chughtai et al., 2023), polynomial regression (Kumar et al., 2024), sparse parity functions (Barak et al., 2022; Bhattacharya et al., 2023), and XOR cluster data (Xu et al., 2023).

In this work, we provide an explanation for the grokking phenomenon from a frequency perspective, which does not need constraints on data dimensionality or explicit regularization in training dynamics. The key insight is that, in the initial stages of training, NNs learn the less salient frequency components present in the test data. Grokking arises due to a misalignment between the preferred frequency in the training dynamics and the dominant frequency in the test data, which is a consequence of insufficient sampling. We use three examples to illustrate this mechanism.

In the first two examples, we consider one-dimensional synthetic data and high-dimensional parity function, where the training data contains spurious low-frequency components due to aliasing effects caused by insufficient sampling. With common (small) initialization, NNs adhere to the frequency principle (F-Principle) that they often learn data from low to high frequencies. Consequently, during the initial stage of training, the networks learn the misaligned low-frequency components, resulting in

*Corresponding author: xuzhiqin@sjtu.edu.cn

an increase in test loss while the training loss decreases. In the third example, we consider the MNIST dataset trained by NNs with large initialization, where high frequencies are preferred in contrast to the F-Principle during the training. Since the MNIST dataset is dominated by low frequency, the test accuracy almost does not increase compared with the fast increase of training accuracy at the initial training stage.

Our work highlights that the distribution of the training data and the frequency preference in training dynamics are critical to the grokking phenomenon. The frequency perspective provides a rationale for the underlying mechanism of how these two factors work together to produce grokking phenomenon.

2 Related Works

Grokking The grokking phenomenon was first proposed by Power et al. (2022) on algorithmic datasets in Transformers (Vaswani et al., 2017), and Liu et al. (2022) attributed it to an effective theory of representation learning dynamics. Nanda et al. (2022) and Furuta et al. (2024) interpreted the grokking phenomenon of algorithmic datasets by investigating the Fourier transform of embedding matrices and logits. Liu et al. (2022) empirically discovered the grokking phenomenon on datasets beyond algorithmic ones, under the condition of large initialization scale with WD, which was theoretically proved by Lyu et al. (2023) for homogeneous NNs. Thilak et al. (2022) utilized cyclic phase transitions to explain the grokking phenomenon. Varma et al. (2023) explained grokking through circuit efficiency and discovered two novel phenomena called ungrokking and semi-grokking. Barak et al. (2022) and Bhattacharya et al. (2023) observed the grokking phenomenon in sparse parity functions. Merrill et al. (2023) investigated the training dynamics of two-layer neural networks on sparse parity functions and demonstrated that grokking results from the competition between dense and sparse subnetworks. Kumar et al. (2024) attributed grokking to the transition in training dynamics from the lazy regime to the rich regime without WD, while we do not require this transition. For example, in our first two examples, we only need the F-Principle without regime transition. Xu et al. (2023) discovered a grokking phenomenon on XOR cluster data and showed that NNs perform like high-dimensional linear classifiers in the initial training stage when the data dimension is larger than the number of training samples, a constraint that is not necessary for our examples.

Frequency Principle The implicit frequency bias of NNs that fit the target function from low to high frequency is named as frequency principle (F-Principle) (Xu et al., 2019, 2020, 2022) or spectral bias (Rahaman et al., 2019). The key insight of the F-principle is that the decay rate of a loss function in the frequency domain derives from the regularity of the activation functions (Xu et al., 2020). This phenomenon gives rise to a series of theoretical works in the Neural Tangent Kernel (Jacot et al., 2018) (NTK) regime (Luo et al., 2022; Zhang et al., 2019, 2021; Cao et al., 2021; Yang and Salman, 2019; Ronen et al., 2019; Bordelon et al., 2020) and in general settings (Luo et al., 2019). Ma et al. (2020) suggests that the gradient flow of NNs obeys the F-Principle from a continuous viewpoint. Also note that if the initialization of network parameters are too large, F-Principle may not hold in the training dynamics of NNs (Xu et al., 2020, 2022).

3 Preliminaries

3.1 Notations

We denote the dataset as $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$. A NN with trainable parameters $\boldsymbol{\theta}$ is denoted as $f(\mathbf{x}; \boldsymbol{\theta})$. In this article, we consider the Mean Square Error (MSE):

$$\ell(\mathcal{S}) = \frac{1}{2n} \sum_{i=1}^n (f(\mathbf{x}_i; \boldsymbol{\theta}) - y_i)^2. \quad (1)$$

3.2 Nonuniform discrete Fourier transform

We apply nonuniform discrete Fourier transform (NUDFT) on the dataset \mathcal{S} in a specific direction \mathbf{d} in the following way:

$$\mathcal{F}[\mathcal{S}](\mathbf{k}) = \frac{1}{n} \sum_{i=1}^n y_i \exp(-ik\mathbf{x}_i \cdot \mathbf{d}). \quad (2)$$

where $i = \sqrt{-1}$, frequency $\mathbf{k} = k\mathbf{d}$.

4 One-dimensional synthetic data

Dataset The nonuniform training dataset consisting of n points is constructed as follows:

- i) 10000 evenly-spaced points in the range $[-\pi, \pi]$ are sampled.
- ii) From this set, we select 20 points that minimize $|\sin(x) - \sin(6x)|$ and 10 points that minimize $|\sin(3x) - \sin(6x)|$.
- iii) These 30 points are then combined with $n - 30$ uniformly selected points to form the complete training dataset.

For the uniform dataset, the training data consists of n evenly-spaced points, sampled from the range $[-\pi, \pi]$. The test data, on the other hand, comprises 1000 evenly-spaced points, sampled from the range $[-\pi, \pi]$.

Experiment Settings The NN architecture employs a fully-connected structure with four hidden layers of widths 200-200-200-100. The network is in default initialization in Pytorch. The optimizer employed is Adam, with a learning rate of 2×10^{-6} . The target function is $\sin 6x$.

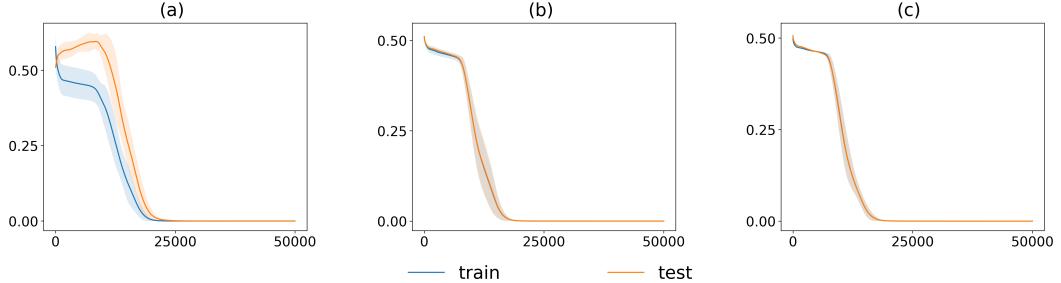


Figure 1: (a) (b) is the train and test loss of $n = 65$ and $n = 1000$ nonuniform experiment, respectively. (c) is the train and test loss of $n = 65$ uniform experiment. The activation function is $\sin x$. Each experiment is averaged over 10 trials and the shallow parts represent the standard deviation.

As illustrated in Fig. 1 (a), when training on a dataset of 65 nonuniform points, a distinct grokking phenomenon is observed during the initial stages of the training process, followed by a satisfactory generalization performance in the terminal stage. In contrast, when training on a larger dataset of 1000 points, as depicted in Fig. 1 (b), the grokking phenomenon is notably absent during the initial training phase. This observation aligns with the findings reported in Liu et al. (2022), which suggest that larger dataset sizes do not trigger the grokking phenomenon. Furthermore, as shown in Fig. 1 (c), we find that when maintaining the data size but uniformly sampling the data, the grokking phenomenon disappears, indicating that the data distribution influences the manifestation of grokking.

Upon examining the output of the NN, we observe that with the default initialization, the output is almost zero, as depicted in Fig. 2 (a). During the training process, the output of the NN resembles $\sin x$, as shown in Fig. 2 (b). When the loss approaches zero, the network finally fits the target function $\sin 6x$, as illustrated in Fig. 2 (c). In the subsequent subsection, we will explain this process from the perspective of the frequency domain.

4.1 The frequency spectrum of the synthetic data

We directly utilize Eq. (2) to compute the frequency spectrum of the synthetic data and examine the frequency domain. We evenly-spaced sample 1000 points in the range $[0, 10]$ as the frequency k . The

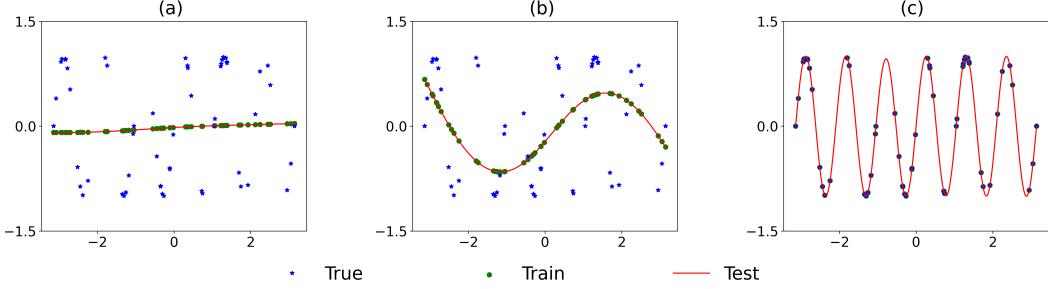


Figure 2: During training with a set of $n = 65$ non-uniformly sampled data points, the learned output function evolves across epochs as shown in (a)-(c) for 0, 2000, and 35000 epochs, respectively. The blue stars represent the exact training data points, the green dots are the network’s outputs on the training data, and the red curve shows the overall learned output function, drawing on 1000 evenly-spaced data points.

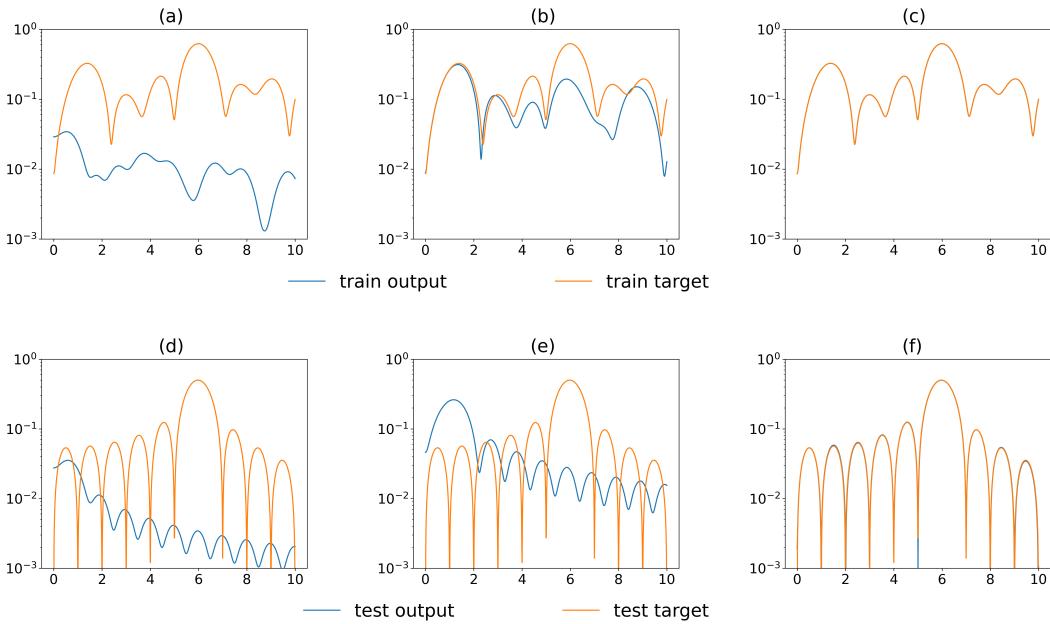


Figure 3: The evolution of the frequency spectrum during training for Fig. 2. The columns from left to right correspond to epochs 0, 2000, and 35000, respectively. The top row illustrates the frequency spectra of the target function (orange solid lines) and the network’s output (blue solid lines) on the training data. The bottom row shows the frequency spectra of the target function (orange solid lines) and the network’s output (blue solid lines) on the test data. The ordinate represents frequency and the abscissa represents the amplitude of the corresponding frequency components.

peak with the largest amplitude is the genuine frequency 6. The presence of numerous peaks in the frequency spectrum can be attributed to the fact that the target function multiplies a window function over the interval $[-\pi, \pi]$, which gives rise to spectral leakage.

A key observation is that insufficient and nonuniform sampling leads to a discrepancy between the frequency spectra of the training and test datasets, and the training dataset’s spectrum contains a spurious low-frequency component that is not dominant in the test dataset’s spectrum. In the case of $n = 65$ nonuniform data points, the NUDFT of the training data contrasts starkly with the NUDFT of the true underlying data, as illustrated in Fig. 3 (a) and (d). The training process adheres to the F-Principle, where the NNs initially fit the low-frequency components, as shown in Fig. 3 (b). At this time, as depicted in Fig. 3 (e), the low-frequency components of the output on the test dataset deviate substantially from the exact low-frequency components, inducing an initial ascent in the test loss.

As the training loss approaches zero, we note that the amplitude of the low-frequency components on the test dataset decreases, as illustrated in Fig. 3 (f), while it remains unchanged on the training dataset, as shown in Fig. 3 (c). This phenomenon arises due to two distinct mechanisms for learning the low-frequency components. Initially, following the F-Principle, the network genuinely learns the low-frequency components. Subsequently, as the network learns the high-frequency components, the insufficient sampling induces frequency aliasing onto the low-frequency components, thereby preserving their amplitude on the training dataset.

When training on data with sufficient or uniform sampling, as demonstrated in Fig. 9 and Fig. 10 in the appendix, the frequency spectrum of the training dataset aligns with that of the test dataset, thereby mitigating the occurrence of the grokking phenomenon.

Furthermore, this phenomenon is robust across different activation functions, such as ReLU and tanh, with the results shown in Fig. 11 and Fig. 14 in the appendix. The corresponding frequency spectra are depicted in Fig. 12, Fig. 13, Fig. 15, and Fig. 16.

5 Parity Function

Parity Function $\mathcal{I} = \{i_1, i_2, \dots, i_k\} \subseteq [m]$ is a randomly sampled index set. (m, \mathcal{I}) parity function $f_{\mathcal{I}}$ is defined as:

$$f_{\mathcal{I}} : \Omega = \{-1, 1\}^m \longrightarrow \{-1, 1\} \quad (3)$$

$$\mathbf{x} = \{x_1, x_2, \dots, x_m\} \longmapsto \prod_{i \in \mathcal{I}} x_i \quad (4)$$

The loss function refers to the MSE loss defined in (1), which directly computes the difference between the labels y_i and the outputs of the NNs $f(\mathbf{x}_i; \boldsymbol{\theta})$. When evaluating the error, we focus on the discrepancy between y_i and $\text{sign}(f(\mathbf{x}_i; \boldsymbol{\theta}))$, effectively treating the problem as a classification task.

Experiment Settings The experiment targets the $(10, [10])$ parity function, with a total of $2^{10} = 1024$ data points. The data is split into training and test datasets with different proportions. We employ width-1000 two-layer fully-connected NN with activation function ReLU. The training is performed using the Adam optimizer with a learning rate of 2×10^{-4} .

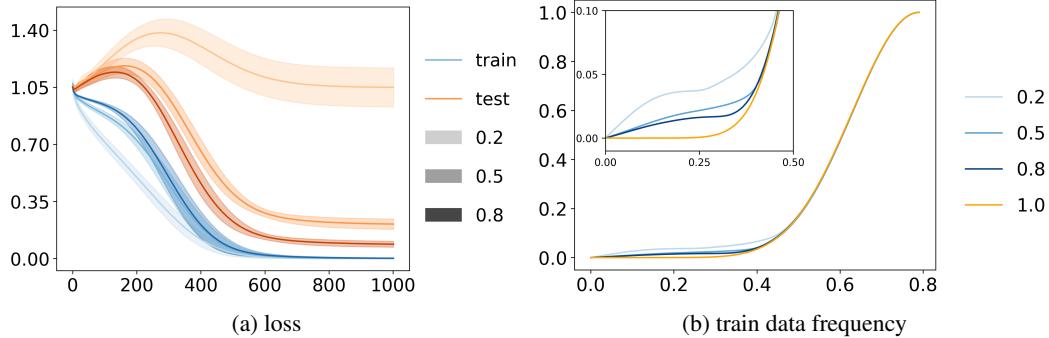


Figure 4: (a) Loss for $(10, [10])$ parity function with different train data proportion respectively. The blue solid lines are the training dataset, and the orange solid lines are the test dataset. (b) The frequency spectra of the training set over different proportion ratios. The blue solid lines are the frequency spectra of the training dataset, and the orange solid line is the frequency spectrum of all data. The proportion ratios of the training-test dataset are 0.2, 0.5, and 0.8, from shallow to deep. The ordinate represents frequency and the abscissa represents the amplitude of the corresponding frequency components. Each experiment is averaged over 10 trials and the shallow parts represent the standard deviation.

The grokking phenomenon is observed in the behavior of the loss across different proportions of the training and test set splits, as illustrated in Fig. 4 (a). When the training set constitutes a higher proportion of the data, the decline in test loss tends to occur earlier compared to scenarios where the training set has a lower proportion. This observation aligns with the findings reported in Barak et al. (2022); Bhattacharya et al. (2023), corroborating the described phenomenon. In the following subsection, we provide an explanation for this phenomenon from a frequency perspective.

5.1 The frequency spectrum of parity function

When we directly use Eq. (2) to compute the exact frequency spectrum of parity function on $\xi = (\xi_1, \xi_2, \dots, \xi_m)$, we obtain

$$\frac{1}{2^m} \sum_{\mathbf{x} \in \Omega} \prod_{j=1}^m x_j \exp(-i\xi \cdot \mathbf{x}) = (-i)^m \prod_{j=1}^m \sin \xi_j. \quad (5)$$

To mitigate the computation cost, we only compute on $\xi = k \times \mathbb{1}$, where $\mathbb{1} = (1, 1, \dots, 1)$. And k ranges from 0 to $\frac{\pi}{2}$.

As illustrated in Fig. 4 (b), in the parity function task, when the sampling is insufficient, applying the NUDFT to the training dataset introduces spurious low-frequency components. Moreover, the fewer the sampling points, the more pronounced these low-frequency components become.

To understand the evolution of the frequency domain during the NN’s attempt to solve this problem, we illustrate with a concrete example with the proportion 0.5. Fig. 5 (a) is the loss for this example and Fig. 5 (b) is the frequency spectra of the training dataset and all data. Although the test loss does not approach zero, the test error already goes zero and we regard it as a good generalization as a classification task, as is illustrated in Fig. 17 (b) in the appendix. During the training process, owing to the F-Principle that governs NNs, the low-frequency components are prioritized and learned first. Consequently, the NN initially fits the spurious low-frequency components arising from insufficient sampling, as is shown in Fig. 5 (c). This results in an initial increase in the test loss, as the low-frequency components on the training dataset are first overfitted. However, as the NN continues to learn the high-frequency components and discovers that utilizing high-frequency components enables a complete fitting of the parity function, we observe a subsequent decline in the low-frequency components, ultimately converging to the true frequency spectrum, which is depicted in Fig. 5 (d).

Based on these findings, we can explain why a lower proportion of training data leads to a later decrease in test loss. When the training data proportion is lower, neural networks spend more time initially fitting the spurious low-frequency components arising from insufficient sampling, before eventually learning the high-frequency components needed to generalize well to the test data.

6 MNIST with large network initialization

Experiment Settings We adopt similar experiment settings from Liu et al. (2022). The only difference is that we do not utilize explicit weight decay (WD). α is the scaling factor on the initial parameters. The training set consists of 1000 points, and the batch size is set to 200. The NN architecture comprises three fully-connected hidden layers, each with a width of 200 neurons, employing the ReLU activation function. For large initialization, the Adam optimizer is utilized for training, with a learning rate of 0.001. For default initialization, the SGD optimizer is employed, with a learning rate of 0.02.

As illustrated in Fig. 6 (a), a grokking phenomenon is observed, suggesting that the explanation provided in Liu et al. (2022) may not be applicable in this scenario. And when we use default initialization, the grokking phenomenon will not appear, as illustrated in Fig. 6 (b).

6.1 The frequency spectrum of MNIST

In this subsection, we explain the mechanism over the frequency domain.

To mitigate the computation cost, we adopt the projection method introduced in Xu et al. (2022). We first compute the principal direction d of the dataset (including training and test data) \mathcal{S} through

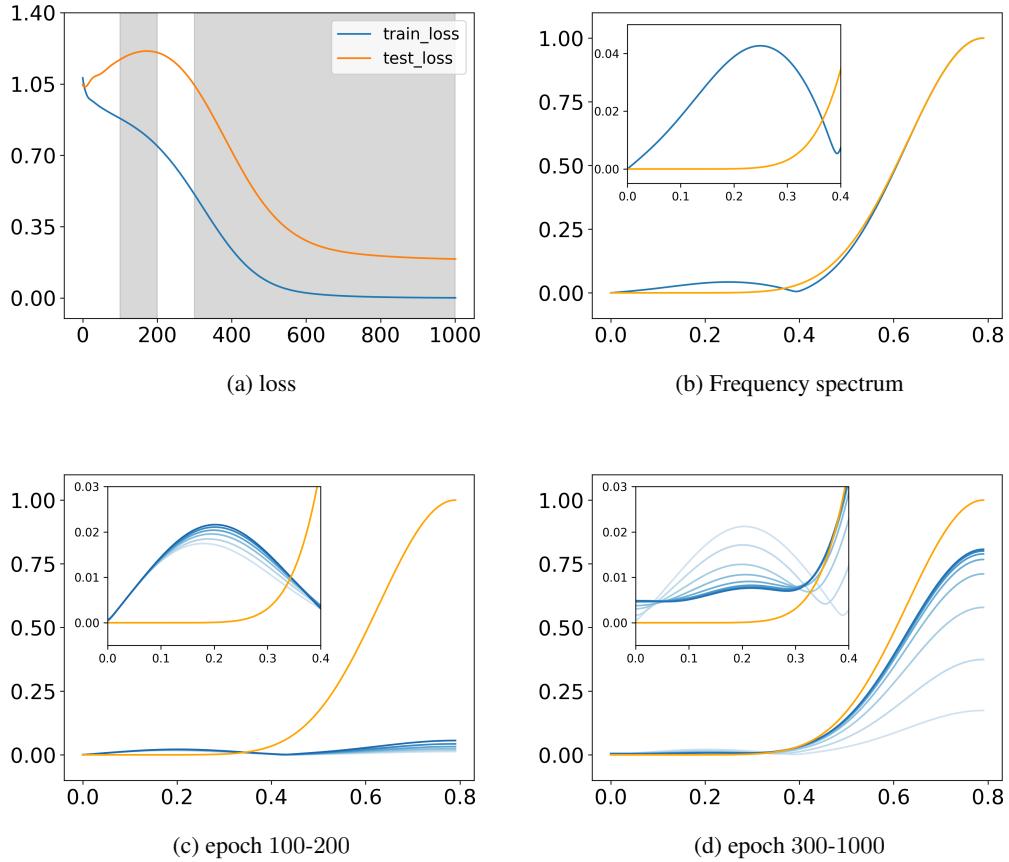


Figure 5: (a) The train and test loss for a specific experiment with proportion ratio 0.5. The ordinate represents epochs and the abscissa represents the loss. The grey background represents the selected intervals for (c)(d). (b) The frequency spectrum difference between the training dataset and the whole dataset. (c)(d) The frequency spectrum of the whole dataset. The orange solid line is the exact frequency spectrum of the $(10, [10])$ parity function. The blue solid lines are the frequency spectra during the training. From shallow to deep corresponds to an increase in epochs, with (c) recording every 20 epochs, and (d) recording every 100 epochs. The ordinate represents frequency and the abscissa represents the amplitude of the corresponding frequency components.

Principal Component Analysis (PCA) which is intended to preserve more frequency information. Then, we apply Eq. (2) to obtain the frequency spectrum of MNIST. The values of k are chosen by sampling 1000 evenly-spaced points in the interval $[0, 2\pi]$.

In this setting, due to the large initialization adopted, the frequency principle for NNs no longer holds (Xu et al., 2022), and the low-frequency components will not converge first. During the initial stages of training, the amplitude of the NN’s output in the frequency domain substantially exceeds the amplitude of the image itself after the NUDFT, as shown in Fig. 7 (a) and (d). After a period of training, the NN fits the frequencies present in the training set, which are primarily caused by high-frequency aliasing due to undersampling (Fig. 7 (b)), but it cannot fit the full dataset (Fig. 7 (e)). With further training, the NN learns to fit the true high frequencies, at which point the generalization performance of the NN improves, increasing the test set accuracy, as shown in Fig. 7 (c) and (f).

However, with default initialization, this phenomenon does not manifest because the NN’s fitting pattern adheres to the F-Principle, and image classification functions are dominated by low-frequency components (Xu et al., 2020). As shown in Fig. 8 (a) and (d), the low-frequency spectra of the training and test sets are well-aligned. Consequently, as the network learns the low-frequency components from the training data, it simultaneously captures the low-frequency characteristics of the test set,

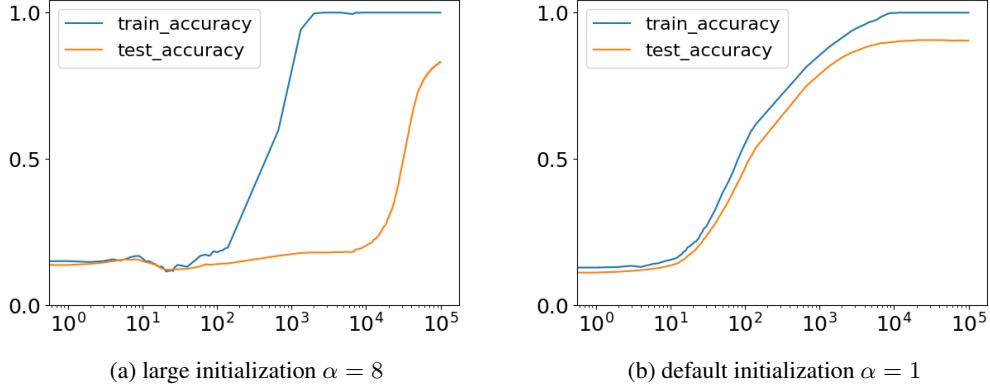


Figure 6: Accuracy on MNIST dataset for different initialization during the training process. The ordinate represents epochs and the abscissa represents the accuracy.

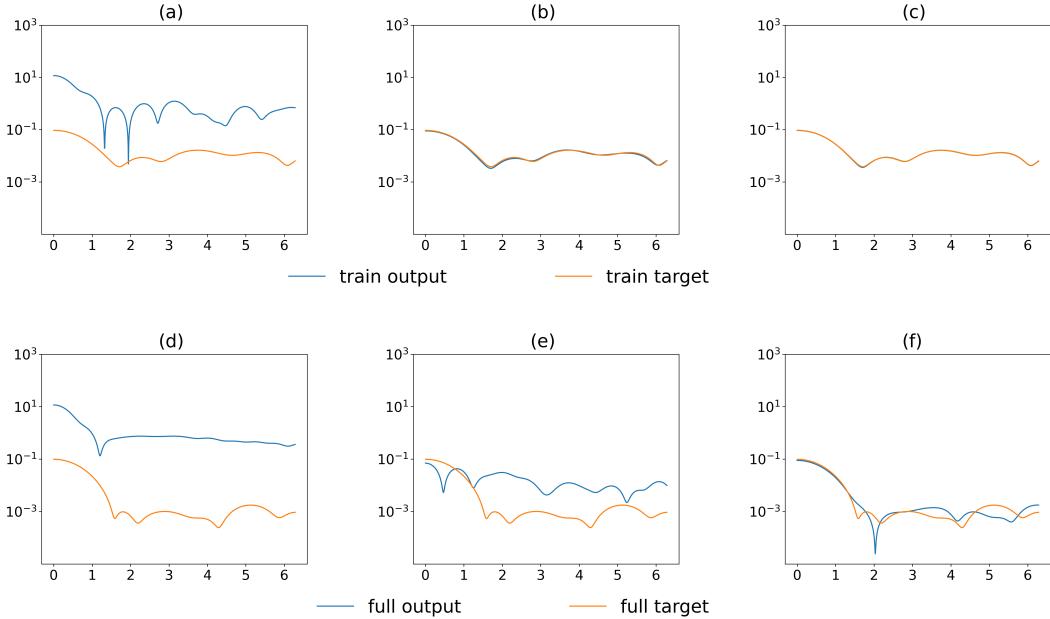


Figure 7: Frequency spectrum evolution during training on the MNIST dataset when $\alpha = 8$. The blue solid lines represent the frequency spectra of the network's outputs, while the orange solid lines depict the frequency spectra of the target data. The top row shows the frequency spectra on the training set, and the bottom row displays the spectra on the full dataset (train and test combined). The columns from left to right correspond to epochs 0, 2668, and 99383, respectively. The ordinate represents frequency and the abscissa represents the amplitude of the corresponding frequency components.

leading to a simultaneous increase in both training and test accuracy (Fig. 8 (b) and (e)). Ultimately, the network converges to accurately fit the frequency spectra of both the training set and the complete dataset, as illustrated in Fig. 8 (c) and (f).

7 Conclusion and limitations

In this work, we provided a frequency perspective to elucidate the grokking phenomenon observed in NNs. Our key insight is that during the initial training phase, networks prioritize learning the

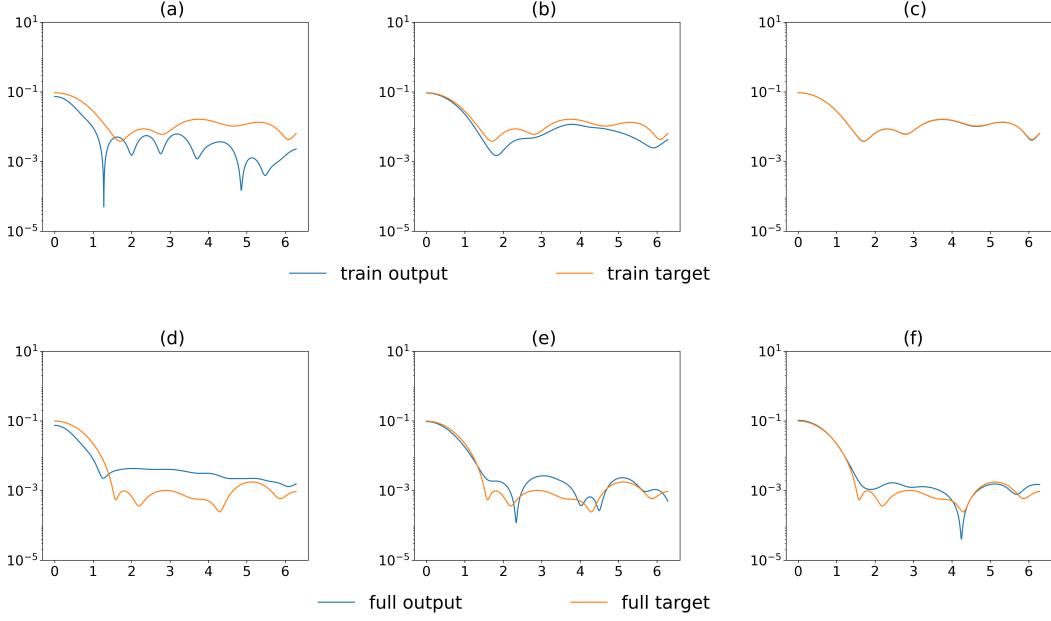


Figure 8: Frequency spectrum evolution during training on the MNIST dataset when $\alpha = 1$. The blue solid lines represent the frequency spectra of the network’s outputs, while the orange solid lines depict the frequency spectra of the target data. The top row shows the frequency spectra on the training set, and the bottom row displays the spectra on the full dataset (train and test combined). The columns from left to right correspond to epochs 0, 1334, and 99383, respectively. The ordinate represents frequency and the abscissa represents the amplitude of the corresponding frequency components.

salient frequency components present in the training data but not dominant in the test data, due to nonuniform and insufficient sampling.

With default initialization, NNs adhere to the F-Principle, fitting the low-frequency components first. Consequently, spurious low frequencies due to insufficient sampling in the training data lead to an initial increase in test loss, as demonstrated in our analyses on one-dimensional synthetic data and parity function learning. In contrast, with large initialization, NNs fit all frequency components at a similar pace, initially capturing spurious frequencies arising from insufficient sampling, as is shown in the MNIST examples.

This coarse-to-fine processing bears resemblance to the grokking phenomenon observed on XOR cluster data in Xu et al. (2023), where high-dimensional linear classifiers were found to first fit the training data.

However, we only demonstrate this mechanism empirically, leaving a comprehensive theoretical understanding of frequency dynamics during training for future work. Additionally, our mechanism may not apply to the grokking phenomenon observed in language-based data (like the algorithmic datasets).

Acknowledgments

Sponsored by the National Key R&D Program of China Grant No. 2022YFA1008200, the National Natural Science Foundation of China Grant No. 92270001, 12371511, Shanghai Municipal of Science and Technology Major Project No. 2021SHZDZX0102, and the HPC of School of Mathematical Sciences and the Student Innovation Center, and the Siyuan-1 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University.

References

- L. Breiman, Reflections after refereeing papers for nips, *The Mathematics of Generalization XX* (1995) 11–15.
- C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, in: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017.
- A. Power, Y. Burda, H. Edwards, I. Babuschkin, V. Misra, Grokking: Generalization beyond overfitting on small algorithmic datasets, arXiv preprint arXiv:2201.02177 (2022).
- Z. Liu, O. Kitouni, N. S. Nolte, E. Michaud, M. Tegmark, M. Williams, Towards understanding grokking: An effective theory of representation learning, *Advances in Neural Information Processing Systems* 35 (2022) 34651–34663.
- B. Chughtai, L. Chan, N. Nanda, A toy model of universality: Reverse engineering how networks learn group operations, in: International Conference on Machine Learning, PMLR, 2023, pp. 6243–6267.
- T. Kumar, B. Bordelon, S. J. Gershman, C. Pehlevan, Grokking as the transition from lazy to rich training dynamics, in: The Twelfth International Conference on Learning Representations, 2024. URL: <https://openreview.net/forum?id=vt5mnLVIVo>.
- B. Barak, B. Edelman, S. Goel, S. Kakade, E. Malach, C. Zhang, Hidden progress in deep learning: Sgd learns parities near the computational limit, *Advances in Neural Information Processing Systems* 35 (2022) 21750–21764.
- S. Bhattacharya, A. Patel, V. Kanade, P. Blunsom, Simplicity bias in transformers and their ability to learn sparse boolean functions, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 5767–5791.
- Z. Xu, Y. Wang, S. Frei, G. Vardi, W. Hu, Benign overfitting and grokking in relu networks for xor cluster data, in: The Twelfth International Conference on Learning Representations, 2023.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- N. Nanda, L. Chan, T. Lieberum, J. Smith, J. Steinhardt, Progress measures for grokking via mechanistic interpretability, in: The Eleventh International Conference on Learning Representations, 2022.
- H. Furuta, M. Gouki, Y. Iwasawa, Y. Matsuo, Interpreting grokked transformers in complex modular arithmetic, arXiv preprint arXiv:2402.16726 (2024).
- Z. Liu, E. J. Michaud, M. Tegmark, Omnidrok: Grokking beyond algorithmic data, in: The Eleventh International Conference on Learning Representations, 2022.
- K. Lyu, J. Jin, Z. Li, S. S. Du, J. D. Lee, W. Hu, Dichotomy of early and late phase implicit biases can provably induce grokking, in: The Twelfth International Conference on Learning Representations, 2023.
- V. Thilak, E. Littwin, S. Zhai, O. Saremi, R. Paiss, J. Susskind, The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon, arXiv preprint arXiv:2206.04817 (2022).
- V. Varma, R. Shah, Z. Kenton, J. Kramár, R. Kumar, Explaining grokking through circuit efficiency, arXiv preprint arXiv:2309.02390 (2023).
- W. Merrill, N. Tsilivis, A. Shukla, A tale of two circuits: Grokking as competition of sparse and dense subnetworks, in: ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models, 2023.

- Z.-Q. J. Xu, Y. Zhang, Y. Xiao, Training behavior of deep neural network in frequency domain, in: Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part I 26, Springer, 2019, pp. 264–274.
- Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, Z. Ma, Frequency principle: Fourier analysis sheds light on deep neural networks, *Communications in Computational Physics* 28 (2020) 1746–1767.
- Z.-Q. J. Xu, Y. Zhang, T. Luo, Overview frequency principle/spectral bias in deep learning, arXiv preprint arXiv:2201.07395 (2022).
- N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, A. Courville, On the spectral bias of neural networks, in: International conference on machine learning, PMLR, 2019, pp. 5301–5310.
- A. Jacot, F. Gabriel, C. Hongler, Neural tangent kernel: Convergence and generalization in neural networks, *Advances in neural information processing systems* 31 (2018).
- T. Luo, Z. Ma, Z.-Q. J. Xu, Y. Zhang, On the exact computation of linear frequency principle dynamics and its generalization, *SIAM Journal on Mathematics of Data Science* 4 (2022) 1272–1292.
- Y. Zhang, Z.-Q. J. Xu, T. Luo, Z. Ma, Explicitizing an implicit bias of the frequency principle in two-layer neural networks, arXiv preprint arXiv:1905.10264 (2019).
- Y. Zhang, T. Luo, Z. Ma, Z.-Q. J. Xu, A linear frequency principle model to understand the absence of overfitting in neural networks, *Chinese Physics Letters* 38 (2021) 038701.
- Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, Q. Gu, Towards understanding the spectral bias of deep learning, in: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, International Joint Conferences on Artificial Intelligence Organization, 2021.
- G. Yang, H. Salman, A fine-grained spectral perspective on neural networks, arXiv preprint arXiv:1907.10599 (2019).
- B. Ronen, D. Jacobs, Y. Kasten, S. Krithman, The convergence rate of neural networks for learned functions of different frequencies, *Advances in Neural Information Processing Systems* 32 (2019).
- B. Bordelon, A. Canatar, C. Pehlevan, Spectrum dependent learning curves in kernel regression and wide neural networks, in: International Conference on Machine Learning, PMLR, 2020, pp. 1024–1034.
- T. Luo, Z. Ma, Z.-Q. J. Xu, Y. Zhang, Theory of the frequency principle for general deep neural networks, arXiv preprint arXiv:1906.09235 (2019).
- C. Ma, L. Wu, et al., Machine learning from a continuous viewpoint, i, *Science China Mathematics* 63 (2020) 2233–2266.

A Hardware for the experiments

All experiments were conducted on a computer equipped with an Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz and an NVIDIA GeForce RTX 3090 GPU with 24GB of VRAM. Each experiment can be implemented within several minutes.

B One-dimensional synthetic data experiments

Fig. 9 shows the frequency spectrum of $n = 1000$ nonuniform experiment and Fig. 10 shows the frequency spectrum of $n = 65$ uniform experiment with the activation function $\sin x$.

ReLU activation The data is generated as in the experiment in the main text with activation function ReLU. The NN architecture employs a fully-connected structure with four layers of widths 200-200-200-100. The network is in default initialization in Pytorch. The optimizer employed is Adam, with a learning rate of 2×10^{-5} .

Tanh activation The data is generated as in the experiment in the main text with activation function Tanh. The NN architecture employs a fully-connected structure with four layers of widths 200-200-200-100. The network is in default initialization in Pytorch. The optimizer employed is Adam, with a learning rate of 2×10^{-5} .

Fig. 11 is the loss for $n = 65$ and $n = 1000$ nonuniform experiments when the activation function is ReLU. Fig.12 is the frequency spectrum of $n = 65$ nonuniform experiment, and Fig.13 is the frequency spectrum of $n = 1000$ nonuniform experiment.

Fig. 14 is the loss for $n = 65$ and $n = 1000$ nonuniform experiments when the activation function is Tanh. Fig.15 is the frequency spectrum of $n = 65$ nonuniform experiment, and Fig.16 is the frequency spectrum of $n = 1000$ nonuniform experiment.

We show that the mechanism of grokking in the frequency domain is the same varying in different activations.

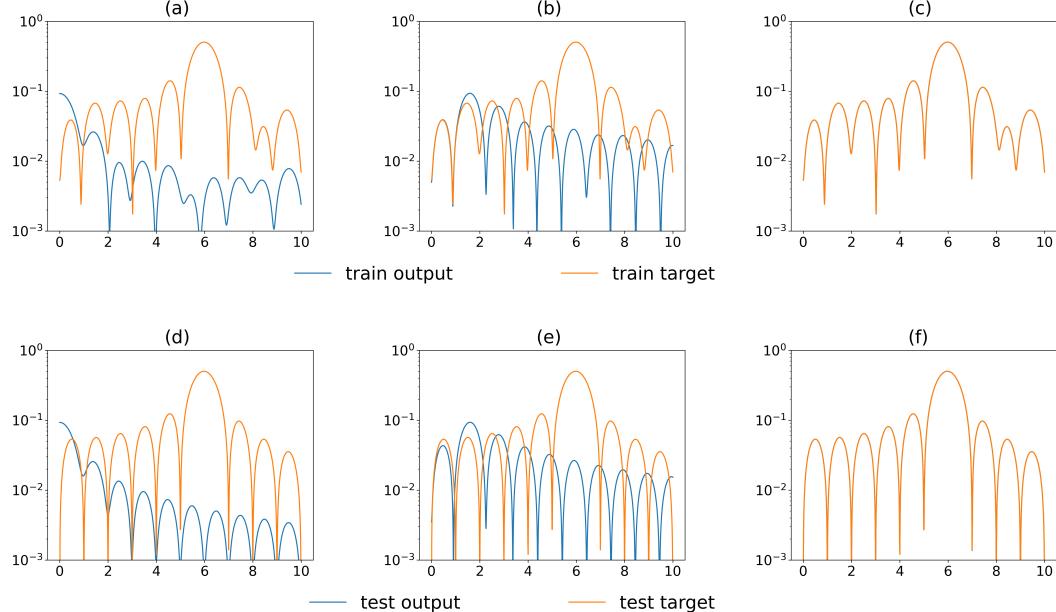


Figure 9: The frequency spectrum of target and output in different epoch during $n = 1000$ nonuniform experiment. The columns from left to right correspond to epochs 0, 2000, and 35000, respectively. In the first row, the orange solid line and the blue solid line is about the training target and training output respectively. In the second row, the orange solid line and the blue solid line is about the test target and test output respectively.

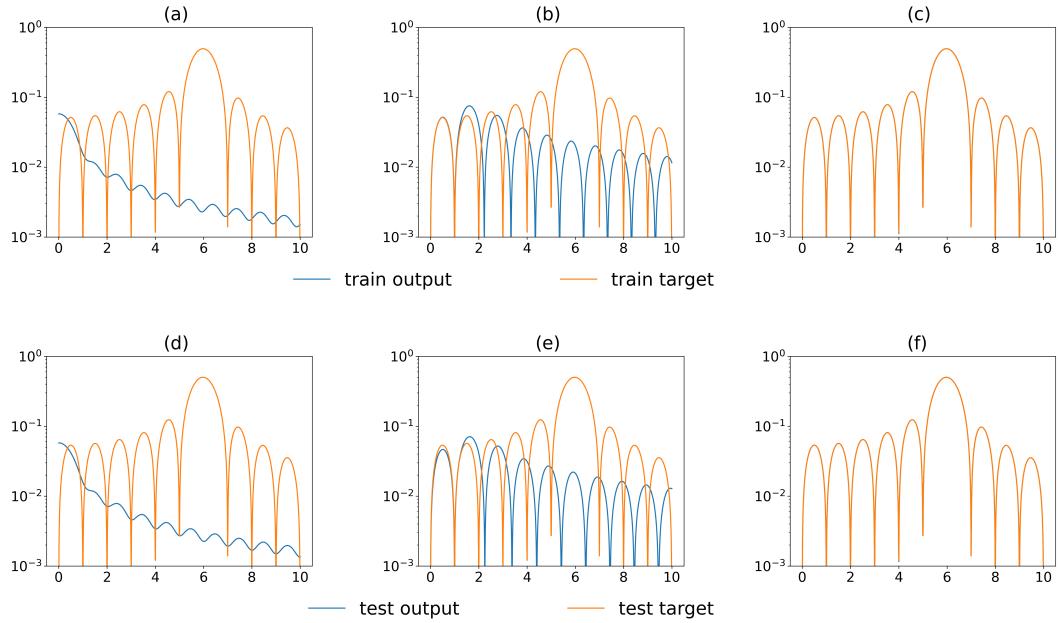


Figure 10: The frequency spectrum of target and output in different epoch during $n = 65$ uniform experiment. The columns from left to right correspond to epochs 0, 2000, and 35000, respectively. In the first row, the orange solid line and the blue solid line is about the training target and training output respectively. In the second row, the orange solid line and the blue solid line is about the test target and test output respectively.

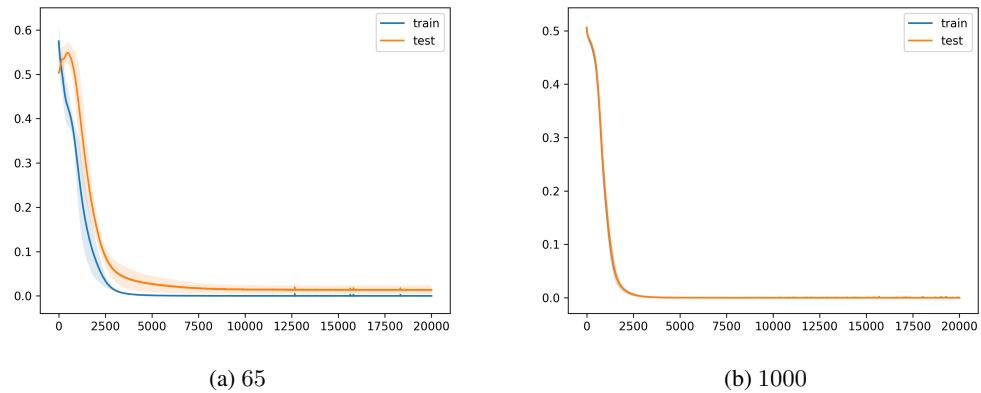


Figure 11: (a) (b) is the train and test loss of $n = 65$ and $n = 1000$ nonuniform experiment, respectively. The activation function is ReLU.

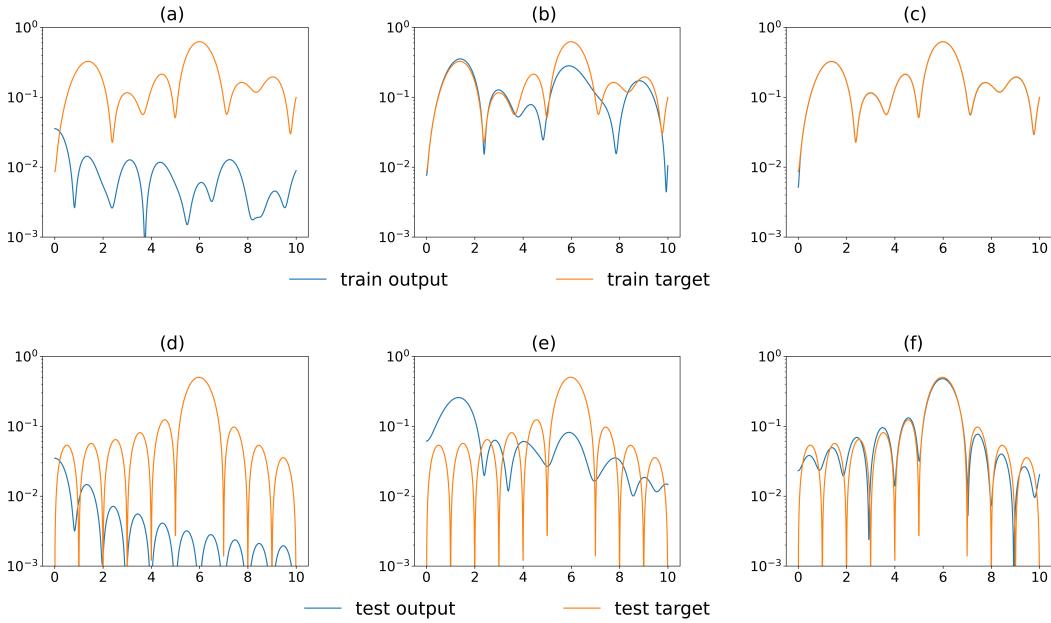


Figure 12: The frequency spectrum during training of Fig. 11 (a). The columns from left to right correspond to epochs 0, 1000, and 19900, respectively. The first row illustrates the frequency spectrum of the target function (orange solid line) and the network’s output (blue solid line) on the training data. The second row shows the frequency spectrum of the target function (orange solid line) and the network’s output (blue solid line) on the test data.

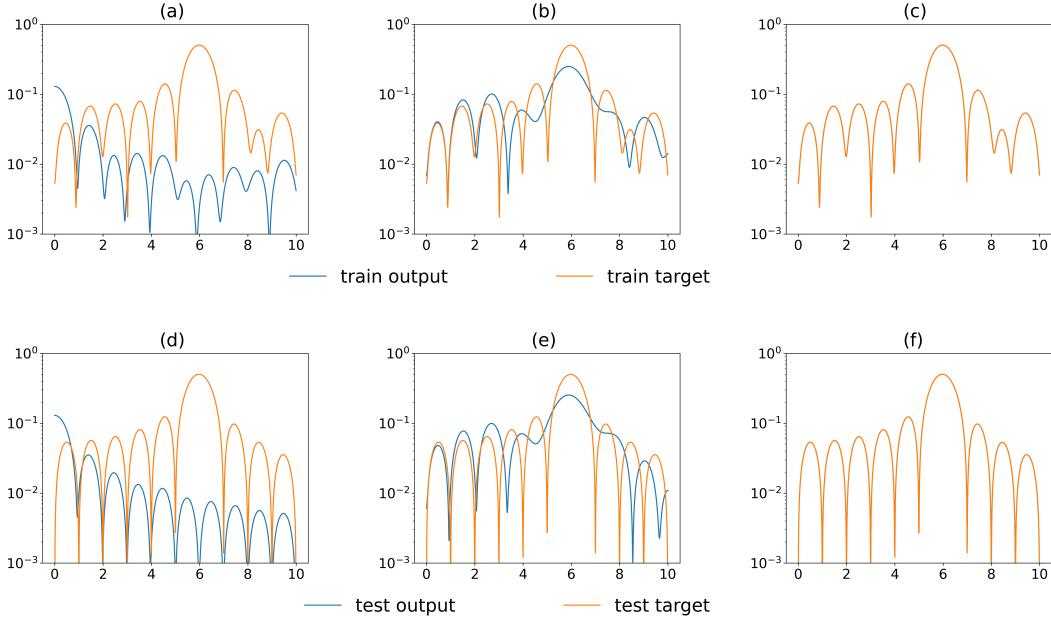


Figure 13: The frequency spectrum during training of Fig. 11 (b). The columns from left to right correspond to epochs 0, 1000, and 19900, respectively. The first row illustrates the frequency spectrum of the target function (orange solid line) and the network’s output (blue solid line) on the training data. The second row shows the frequency spectrum of the target function (orange solid line) and the network’s output (blue solid line) on the test data.

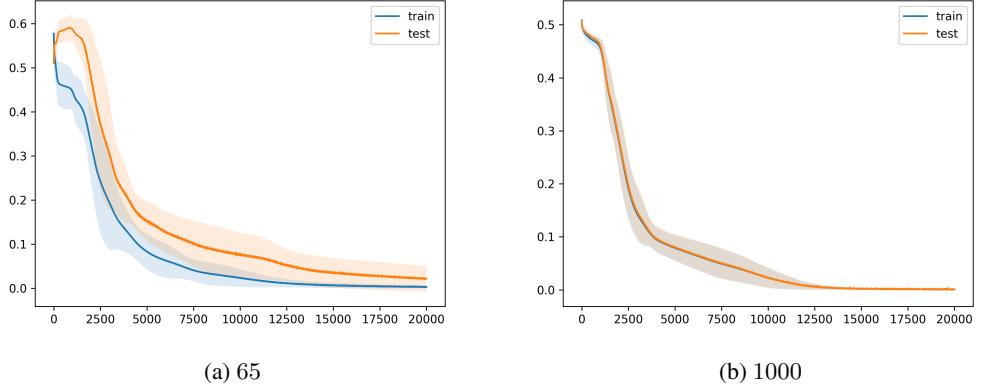


Figure 14: (a) (b) is the train and test loss of $n = 65$ and $n = 1000$ nonuniform experiment, respectively. The activation function is $\text{Tanh}x$.

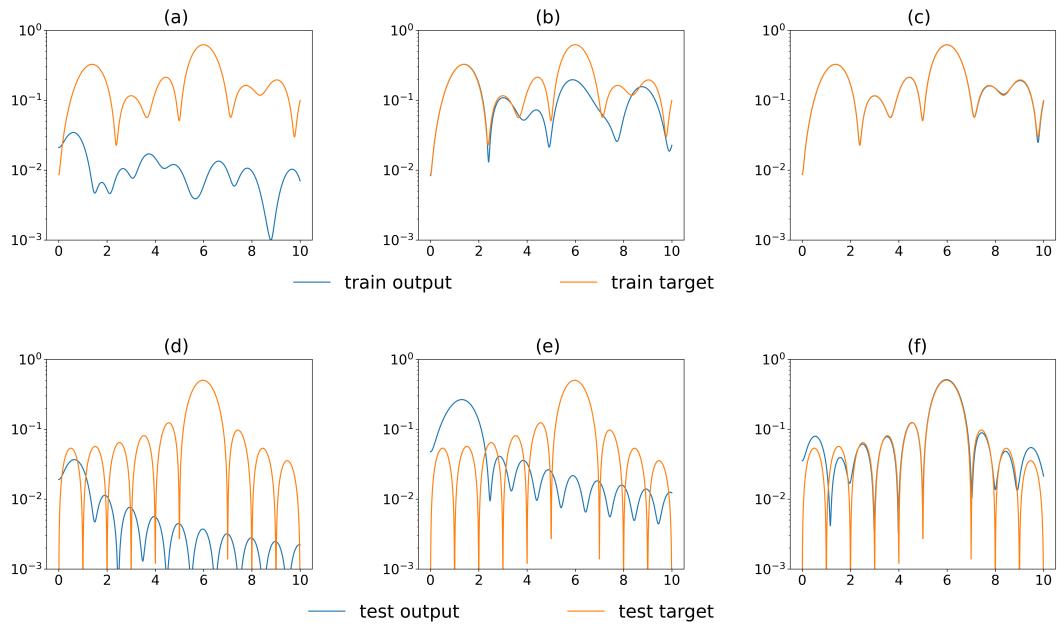


Figure 15: The frequency spectrum during training of Fig. 14 (a). The columns from left to right correspond to epochs 0, 1000, and 18000, respectively. The first row illustrates the frequency spectrum of the target function (orange solid line) and the network's output (blue solid line) on the training data. The second row shows the frequency spectrum of the target function (orange solid line) and the network's output (blue solid line) on the test data.

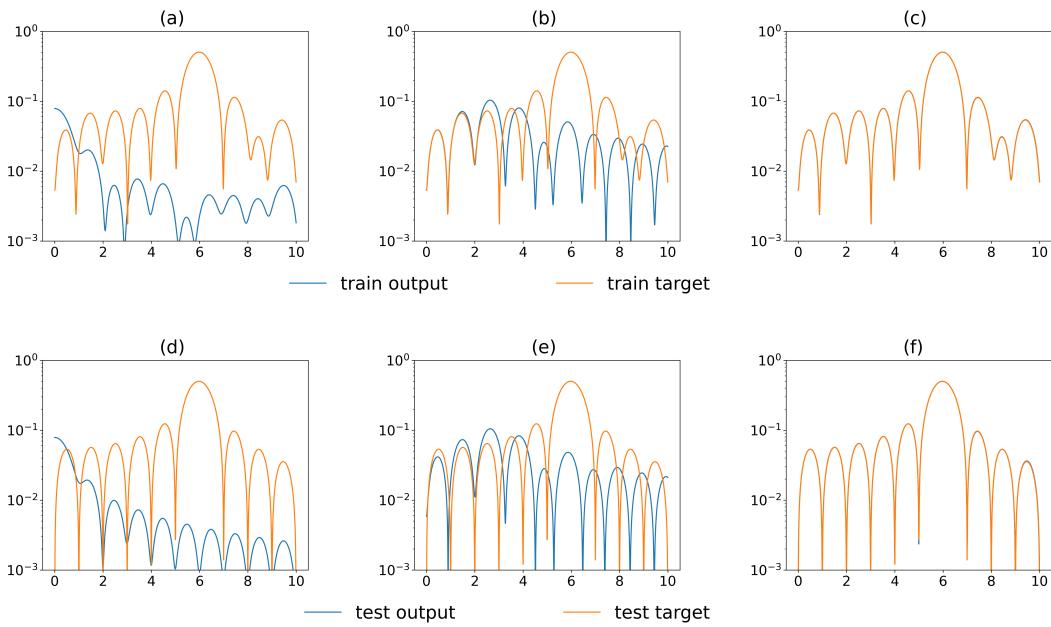


Figure 16: The frequency spectrum during training of Fig. 14 (b). The columns from left to right correspond to epochs 0, 1000, and 18000, respectively. The first row illustrates the frequency spectrum of the target function (orange solid line) and the network's output (blue solid line) on the training data. The second row shows the frequency spectrum of the target function (orange solid line) and the network's output (blue solid line) on the test data.

C Errors in the parity function task

The error function computes the difference between the label y_i and $\text{sign}(f(\mathbf{x}_i; \theta))$ under MSE loss, indicating whether the neural networks have learned the parity function problem.

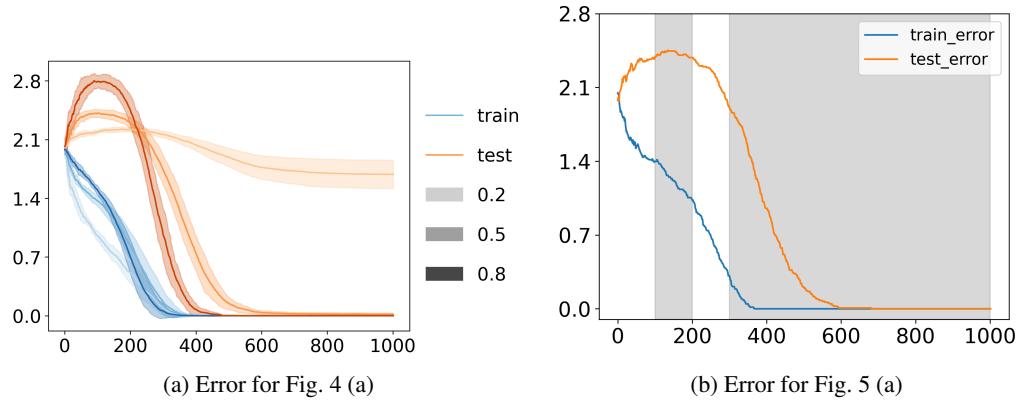


Figure 17: The error for the experiments