Addressing Spectral Bias of Deep Neural Networks by Multi-Grade Deep Learning

Ronglong Fang[1] and Yuesheng Xu[1]

Department of Mathematics and Statistics, Old Dominion University[1]

NeurIPS 2024

Supported by NSF and NIH National Institutes of Health Turning Discovery Into Health

Background

- High-frequency components imbedded in data are essential for many physical problems, such as medical image reconstruction, seismic wavefield modeling. Thus, it is imperative to extract them for their solutions.

Background

- High-frequency components imbedded in data are essential for many physical problems, such as medical image reconstruction, seismic wavefield modeling. Thus, it is imperative to extract them for their solutions.
- Standard deep neural networks, which will be called single grade deep learning (SGDL), suffer from the *spectral bias* [1] (N. Rahaman, et al., On the spectral bias of neural networks, PMLR, 2019, p. 5301–5310) :
  - SGDLs prioritize learning lower-frequency components of a function but struggle to capture its high-frequency features.

Background

- High-frequency components imbedded in data are essential for many physical problems, such as medical image reconstruction, seismic wavefield modeling. Thus, it is imperative to extract them for their solutions.
- Standard deep neural networks, which will be called single grade deep learning (SGDL), suffer from the *spectral bias* [1] (N. Rahaman, et al., On the spectral bias of neural networks, PMLR, 2019, p. 5301–5310) :
  - SGDLs prioritize learning lower-frequency components of a function but struggle to capture its high-frequency features.
- The multi-grade deep learning (MGDL) model, a model recently introduced in [2] (Y. Xu, Multi-grade deep learning, arXiv preprint arXiv:2302.00150, Feb. 1, 2023), trains a DNN **incrementally**, grade by grade, with each grade learning only a shallow neural network (SNN).

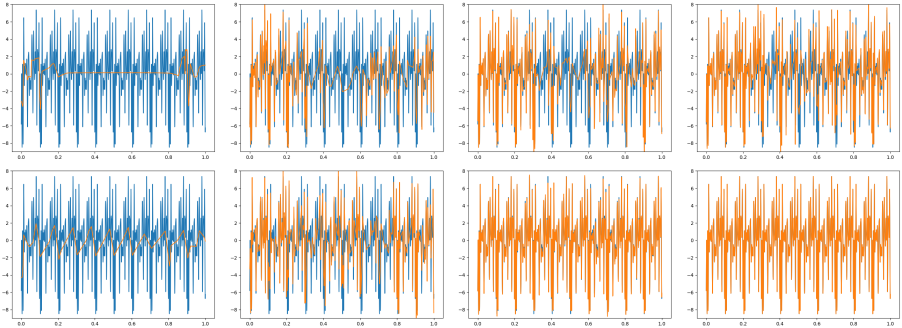


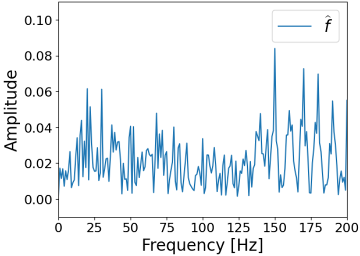
Figure: Comparison of functions learned by SGDL and MGDL models (yellow) vs. target function (blue). Top row: SGDL-learned function at training steps 1,000, 10,000, 20,000, and 30,000. Bottom row: MGDL-learned function at grades 1, 2, 3, and 4. Total training times: 32,402s (SGDL), 27,817s (MGDL).

Motivation

Consider f, with Fourier transform shown in Fig. 2 (Left) and represent f as

$$f = f_1 + f_2 \circ f_1 + f_3 \circ f_2 \circ f_1 + f_4 \circ f_3 \circ f_2 \circ f_1, \quad \text{(Sum – Composition Form)} \quad (1)$$

where the Fourier transforms $\hat{f}_j$, $j = 1, 2, 3, 4$, are displayed in Fig. 2 (Right).



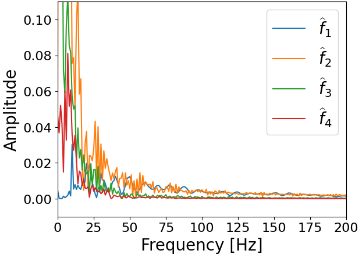


Figure: Spectrum comparison of f and f_j: Amplitude versus one-side frequency plots for f (Left) and f_j for $j \in \mathbb{N}_4$ (Right).

Motivation

Consider f, with Fourier transform shown in Fig. 2 (Left) and represent f as

$$f = f_1 + f_2 \circ f_1 + f_3 \circ f_2 \circ f_1 + f_4 \circ f_3 \circ f_2 \circ f_1, \quad (\text{Sum} - \text{Composition Form}) \quad (1)$$

where the Fourier transforms $\hat{f}_j$, $j = 1, 2, 3, 4$, are displayed in Fig. 2 (Right).

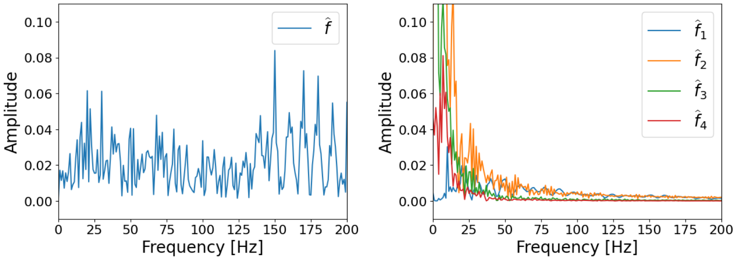


Figure: Spectrum comparison of f and f_j: Amplitude versus one-side frequency plots for f (Left) and f_j for $j \in \mathbb{N}_4$ (Right).

A high-frequency function can be decomposed as a sum-composition form of lower-frequency functions.

- The real Jacobi–Anger identity, named after the 19th-century, gives

$$\cos(a\sin(b\mathbf{x})) = \sum_{n=-\infty}^{\infty} J_n(a)\cos(nb\mathbf{x}). \tag{2}$$

  where $J_n(a)$ denotes the n-th Bessel function of the first kind.

- The left-hand side of (2) is a composition of two low-frequency functions $\cos(a\mathbf{x})$ and $\sin(b\mathbf{x})$, having frequencies $a/(2\pi)$ and $b/(2\pi)$, respectively, while the right-hand side is a linear combination of $\cos(nb\mathbf{x})$ with n taking all integers.

- The real Jacobi–Anger identity, named after the 19th-century, gives

$$\cos(a \sin(b\mathbf{x})) = \sum_{n=-\infty}^{\infty} J_n(a) \cos(nb\mathbf{x}). \tag{2}$$

where $J_n(a)$ denotes the n-th Bessel function of the first kind.

- The left-hand side of (2) is a composition of two low-frequency functions $\cos(a\mathbf{x})$ and $\sin(b\mathbf{x})$, having frequencies $a/(2\pi)$ and $b/(2\pi)$, respectively, while the right-hand side is a linear combination of $\cos(nb\mathbf{x})$ with n taking all integers.

Both **Sum-Composition Form** and the **Jacobi–Anger identity** suggest:

- A high-frequency function can be well approximated by a composition of several lower-frequency functions.

Multi-Grade Deep Learning

- Human education is arranged in grades. In such a system, students learn a complex subject in grades, by decomposing it into sequential, simpler topics. Inspired by human learning, the multi-grade deep learning (MGDL) model was introduced in [2].

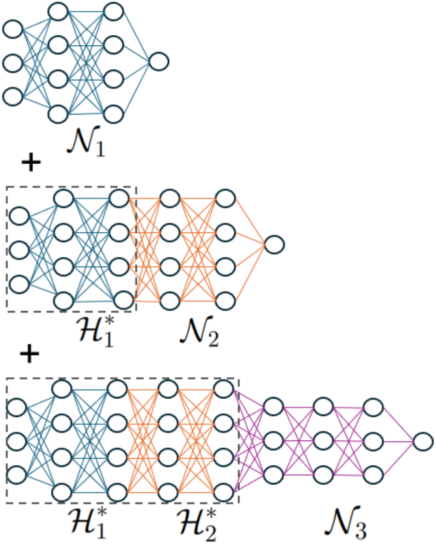
Multi-Grade Deep Learning

- Human education is arranged in grades. In such a system, students learn a complex subject in grades, by decomposing it into sequential, simpler topics. Inspired by human learning, the multi-grade deep learning (MGDL) model was introduced in [2].

- MGDL trains a DNN incrementally, grade by grade, each grade training only a shallow neural network (SNN) using the SNNs trained in the previous grades as features ("bases"), from the residue of its previous grade.

Multi-Grade Deep Learning

- Human education is arranged in grades. In such a system, students learn a complex subject in grades, by decomposing it into sequential, simpler topics. Inspired by human learning, the multi-grade deep learning (MGDL) model was introduced in [2].
- MGDL trains a DNN incrementally, grade by grade, each grade training only a shallow neural network (SNN) using the SNNs trained in the previous grades as features ("bases"), from the residue of its previous grade.
- After all grades are learned, MGDL sums the functions learned in each grade into a "Sum-Composition Form".



$$\mathcal{N}_1 \Leftarrow f$$

$$\mathcal{N}_2 \circ \mathcal{H}_1^* \Leftarrow f - \mathcal{N}_1^*$$

$$\mathcal{N}_3 \circ \mathcal{H}_2^* \circ \mathcal{H}_1^* \Leftarrow f - (\mathcal{N}_1^* + \mathcal{N}_2^* \circ \mathcal{H}_1^*)$$

Figure: Multi-grade network with 3 grades.

$$\mathcal{N}_1^* + \mathcal{N}_2^* \circ \mathcal{H}_1^* + \mathcal{N}_3^* \circ \mathcal{H}_2^* \circ \mathcal{H}_1^*$$

Theorem (Xu, 2023) Let $\mathbb{D}$ be a compact subset of $\mathbb{R}^s$ and $L_2(\mathbb{D}, \mathbb{R}^t)$ denote the space of t-dimensional vector-valued square integral functions on $\mathbb{D}$. If $\mathbf{f} \in L_2(\mathbb{D}, \mathbb{R}^t)$, then for all $i = 1, 2, \ldots$,

$$\mathbf{f} = \sum_{l=1}^{i} \mathbf{f}_l + \mathbf{e}_i, \quad \mathbf{f}_l := \mathcal{N}_l \circ \mathcal{N}_{l-1} \circ \cdots \circ \mathcal{N}_1.$$

where $\mathcal{N}_l$ is the SNN learned in grade l, and for $i = 1, 2, \ldots$, either $\mathbf{f}_{i+1} = \mathbf{0}$ or

$$\|\mathbf{e}_{i+1}\| < \|\mathbf{e}_i\|.$$

This theorem shows that the error strictly decreases as a new grade is added.

Numerical Experiments

Regression on the synthetic data

- Approximating the function $\lambda : [0,1] \to \mathbb{R}$ defined by

$$\lambda(\mathbf{x}) := \sum_{j=1}^{M} \alpha_j \sin\left(2\pi\kappa_j \mathbf{x} + \varphi_j\right), \quad \mathbf{x} \in [0,1] \tag{3}$$

where κ ranges from 0 to 200, $\varphi_j \sim \mathcal{U}(0, 2\pi)$, and amplitudes α are considered in four cases: constant, decreasing, varying as a function, and increasing.

Table: Comparison relative mean square error on testing data between SGDL and MGDL

	constant	decreasing	varying	increasing
SGDL	1.2×10^{-1}	5.7×10^{-3}	1.1×10^{-1}	7.7×10^{-1}
MGDL	$\mathbf{1.7 \times 10^{-5}}$	$\mathbf{6.5 \times 10^{-6}}$	$\mathbf{2.1 \times 10^{-5}}$	$\mathbf{1.3 \times 10^{-3}}$

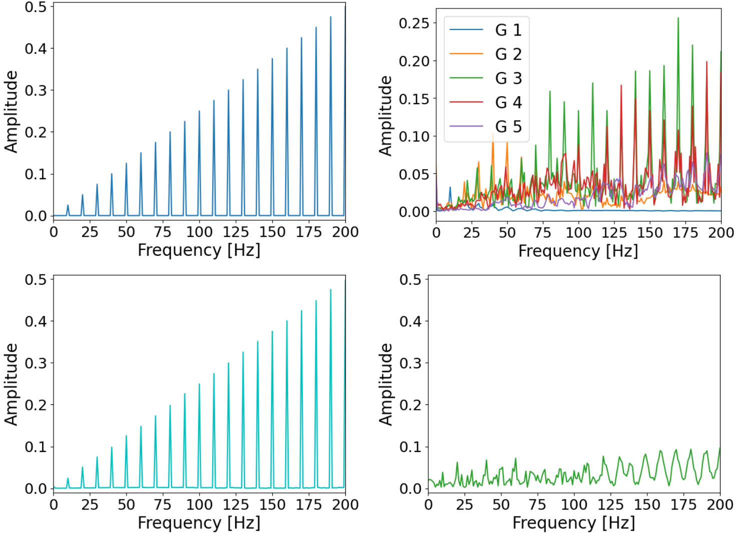


Figure: Amplitude versus one-side frequency. Top left: target function frequency. Top right: MGDL function frequency for grades 1 to 5. Bottom left: overall MGDL function frequency. Bottom right: SGDL function frequency.

Regression on the manifold data

- Given an injective mapping γ from $[0,1] \to \mathbb{R}^2$, a function λ from $[0,1] \to \mathbb{R}$, we wish to learn a network function $\tau : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\lambda(\mathbf{x}) = (\tau \circ \gamma)(\mathbf{x}). \tag{4}$$

The function τ is not defined on the entire $\mathbb{R}^2$ but on the manifold $\gamma([0,1])$.

- We choose λ as (3) with an increase amplitude α, and for $q = 4, 0$, we choose γ as

$$\gamma_q(\mathbf{x}) := [1 + \sin(2\pi q\mathbf{x})/2]\,(\cos(2\pi\mathbf{x}), \sin(2\pi\mathbf{x}))\,, \quad \mathbf{x} \in [0,1]. \tag{5}$$

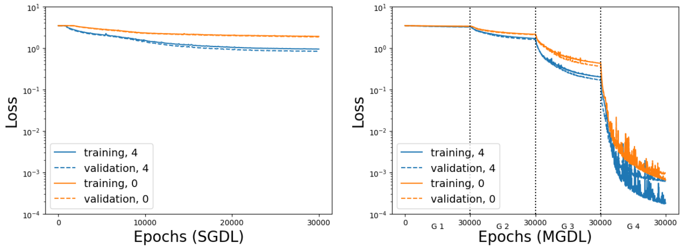
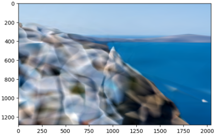


Figure: Comparison of the loss for learning τ with SGDL and MGDL vs epochs.

Regression on two-dimensional colored Images

- The models take pixel coordinates as input and output corresponding RGB values.
- We train the models on a grid of 1/4 of the image pixels and test on the full image.

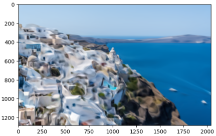


(a) G 1: 18.62

(b) G 2: 21.50

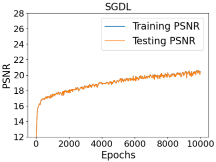
(c) G 3: 23.42

(d) G 4: 24.32

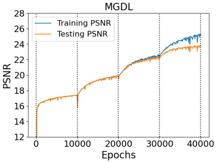


(e) SGDL: 20.39

(f) Ground Truth

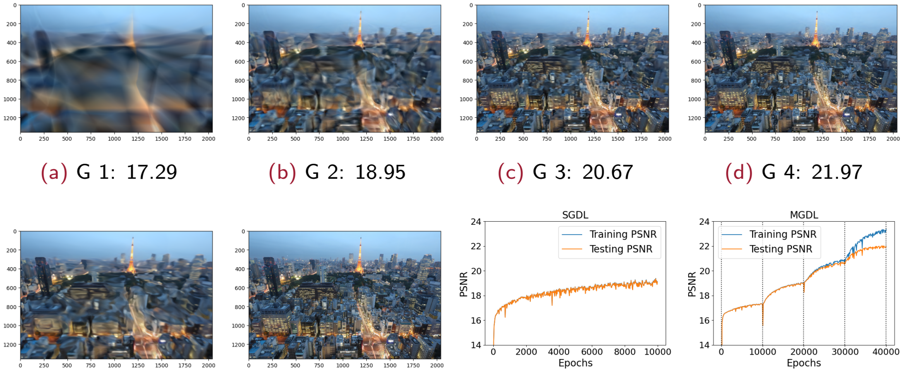


(g) SGDL: PSNR



(h) MGDL: PSNR

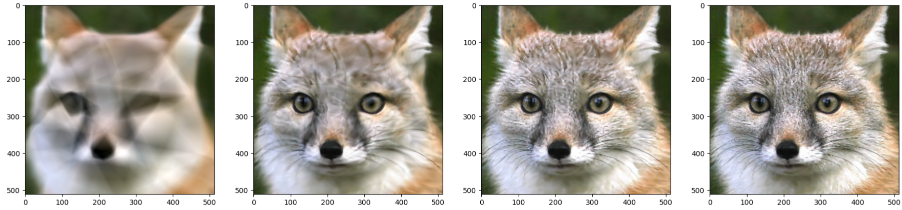
Figure: Comparison of MGDL and SGDL for image sea. (a)-(d): Predictions of MGDL for grades 1-4 with testing PSNR. (e): Prediction of SGDL with testing PSNR. (f): Ground truth image. (g)-(h): PSNR for SGDL and MGDL during training process. Training times: MGDL - 689 seconds, SGDL - 685 seconds.



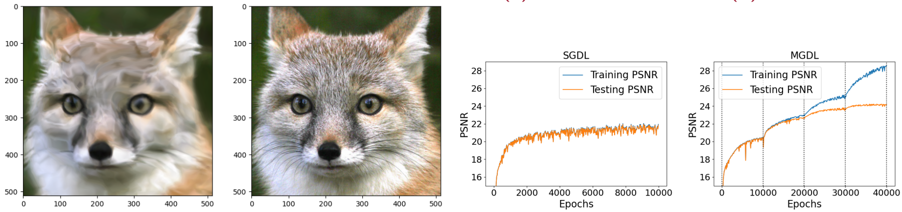
(a) G 1: 17.29 (b) G 2: 18.95 (c) G 3: 20.67 (d) G 4: 21.97

(e) SGDL: 19.09 (f) Ground Truth (g) SGDL: PSNR (h) MGDL: PSNR

Figure: Comparison of MGDL and SGDL for image building. (a)-(d): Predictions of MGDL for grades 1-4 with testing PSNR. (e): Prediction of SGDL with testing PSNR. (f): Ground truth image. (g)-(h): PSNR for MGDL and MGDL during training process. Training times: MGDL - 716 seconds, SGDL - 742 seconds.



(a) G 1: 20.41 (b) G 2: 22.67 (c) G 3: 23.71 (d) G 4: 24.18



(e) SGDL: 21.83 (f) Ground truth (g) SGDL (h) MGDL

Figure: Comparison of MGDL and SGDL for image cat. (a)-(d): Predictions of MGDL for grades 1-4 with testing PSNR. (e): Prediction of SGDL with testing PSNR. (f): Ground truth image. (g)-(h): PSNR for SGDL and MGDL during training process. Training times: MGDL - 138 seconds, SGDL - 77 seconds.

Conclusion

- We propose a novel approach to address the spectral bias issue by decomposing a function in the sum-composition form, in which the high-frequency functions are represented as compositions of low-frequency functions.

Conclusion

- We propose a novel approach to address the spectral bias issue by decomposing a function in the sum-composition form, in which the high-frequency functions are represented as compositions of low-frequency functions.

- We investigate the efficacy of MGDL in decomposing a function of high-frequency into its “sum-composition” form of SNNs.

Conclusion

- We propose a novel approach to address the spectral bias issue by decomposing a function in the sum-composition form, in which the high-frequency functions are represented as compositions of low-frequency functions.
- We investigate the efficacy of MGDL in decomposing a function of high-frequency into its "sum-composition" form of SNNs.
- We successfully apply the proposed approach to three datasets, showing that it can effectively address the spectral bias issue.

Conclusion

- We propose a novel approach to address the spectral bias issue by decomposing a function in the sum-composition form, in which the high-frequency functions are represented as compositions of low-frequency functions.
- We investigate the efficacy of MGDL in decomposing a function of high-frequency into its "sum-composition" form of SNNs.
- We successfully apply the proposed approach to three datasets, showing that it can effectively address the spectral bias issue.
- In future work, we will apply MGDL to real-world problems like medical image reconstruction, and further investigate the mathematical foundations behind its ability to address spectral bias.

[1] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, *On the spectral bias of neural networks*, in International conference on machine learning, PMLR, 2019, pp. 5301–5310.

[2] Y. Xu, *Multi-grade deep learning*, arXiv preprint arXiv:2302.00150, (Feb. 1, 2023).