On the Spectral Bias of Neural Networks

Nasim Rahaman Aristide Baratin Devansh Arpit Felix Draxler Min Lin Fred A. Hamprecht
Yoshua Bengio Aaron Courville

Proceedings of the 36 th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Introduction

- Observe characteristics of neural networks which show bias against frequencies in various operations.
- Transform input manifold to facilitate performance against different frequencies

Experiment 1: LF Learned First

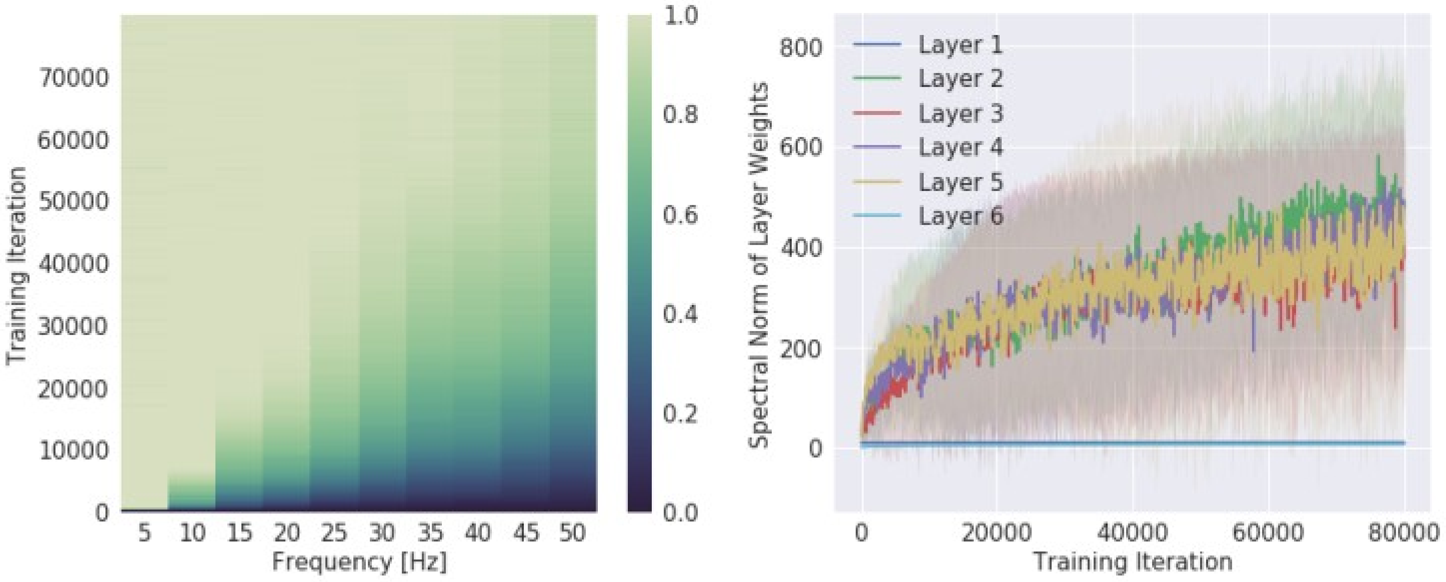
- We want to approximate mappin $\lambda : [0,1] \to \mathbb{R}$ given by:

$$\lambda(z) = \sum_i A_i \sin(2\pi k_i z + \varphi_i).$$

via a neural network (6-layer deep 256-unit wide ReLU network).

where $\kappa = (5, 10, ..., 45, 50)$ with 200 equally-spaced samples over [0, 1].

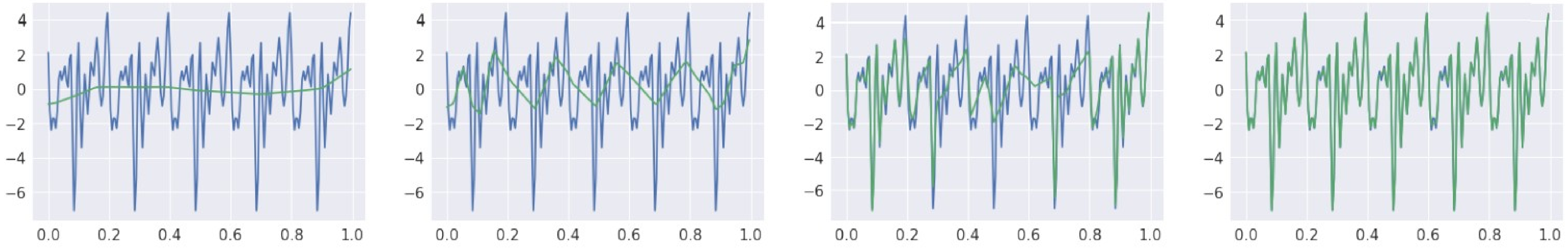
Experiment 1: LF Learned First



(a) Equal Amplitudes

- Left: color represents normalized magnitu $|\tilde{f}_\theta(k_i)|/A_i$ (f: DFT)
- Observe that lower frequencies are learned first
- Different amplitudes have minimal effect

Experiment 1: LF Learned First



(a) Iteration 100 (b) Iteration 1000 (c) Iteration 10000 (d) Iteration 80000

- Green: learnt function
- Blue: target functior $\lambda(z) = \sum_i A_i \sin(2\pi k_i z + \varphi_i).$)

Experiment 2: LF Robust to Parameter Perturbations

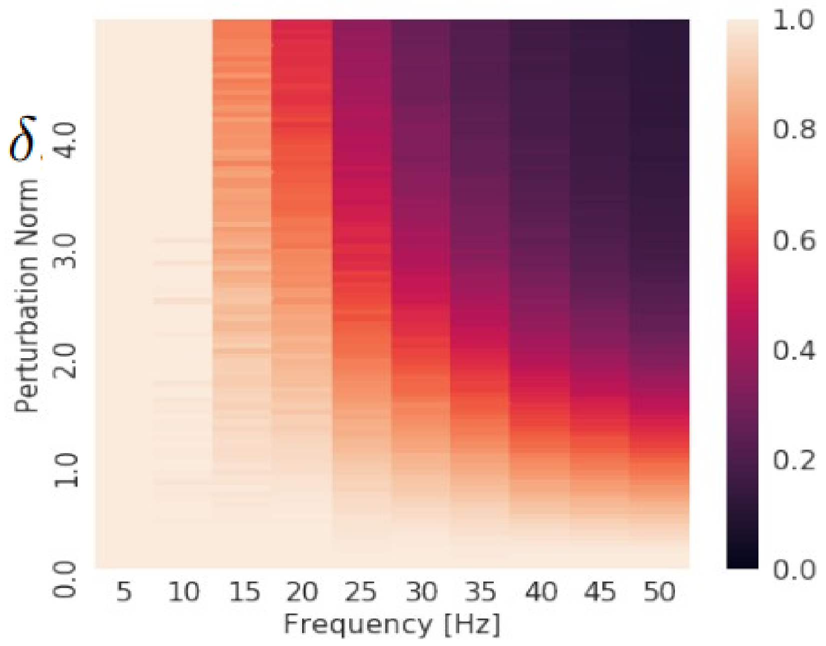
- Same target function $\lambda : [0,1] \to \mathbb{R}$, $\lambda(z) = \sum_i A_i \sin(2\pi k_i z + \varphi_i)$. $\kappa = (5, 10, ..., 45, 50)$ $A_i = 1 \, \forall i$.

- After convergence to θ^* we consider random isotropic perturbations:

$$\theta = \theta^* + \delta\hat{\theta}$$

- We average over 100 samples of $\hat{\theta}$ to obtain $|\tilde{f}_{\mathbb{E}\theta}(k_i)|$ (DFT), normalize by $|\tilde{f}_{\theta*}(k_i)|$ then average over phases ϕ.

Experiment 2: LF Robust to Parameter Perturbations



- Values close to 0: HF components not observed in average?
- *Higher frequencies are significantly less robust than the lower ones, guiding the intuition that expressing higher frequencies requires the parameters to be finely-tuned to work together, i.e. parameters that contribute towards expressing high-frequency components occupy a small volume in the parameter space.*

Experiment 3 & 4: LF Robust to Parameter Perturbations (Real Data)

- Test on MNIST dataset; target binary functi $\tau_0 : X \to \{0,1\}$

defined on input space $X = [0,1]^{784}$

- Can only evaluate by loss on validation set with LF/HF noise and their amplitudes (Experiment 3) because of large input dimension: 28x28=784.

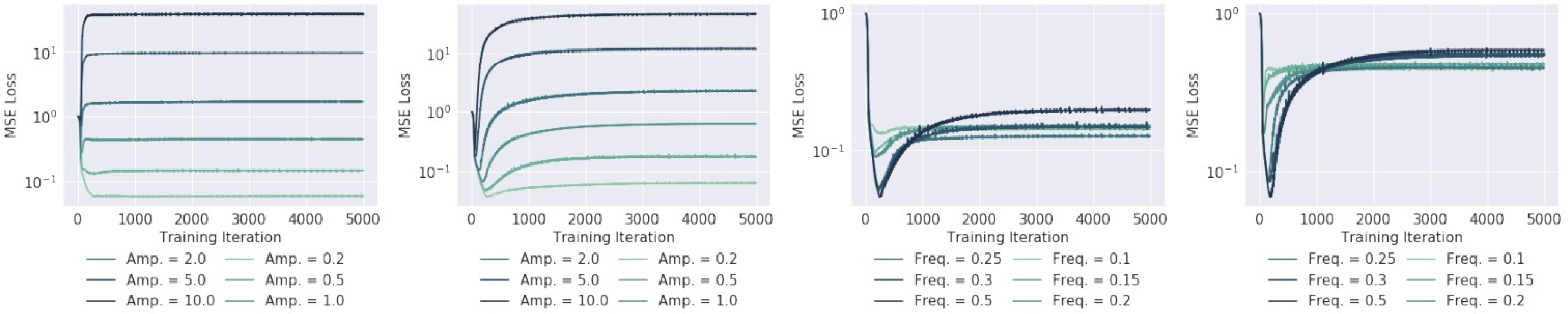
Experiment 3 & 4: LF Robust to Parameter Perturbations (Real Data)

- Test on MNIST dataset; target binary functi $\tau_0 : X \to \{0, 1\}$

defined on input space $X = [0, 1]^{784}$

- Can only evaluate by loss on validation set with LF/HF noise and their amplitudes (Experiment 3) because of large input dimension: 28x28=784.

Experiment 3 & 4: LF Robust to Parameter Perturbations
(Real Data)



(a) $k = 0.1$ (b) $k = 1$ (c) $\beta = 0.5$ (d) $\beta = 1.$

- *Validation performance is adversely affected by the amplitude of the LF noise, whereas Figure 4b shows that the amplitude of HF noise does not significantly affect the best validation score.*
- This is because HF signal is only fit later in the training.

Experiment 3 & 4: LF Robust to Parameter Perturbations (Real Data)

- We project the network function to the space spanned by orthonormal eigenfunctions φ_n of Gaussian RBF kernel (Braun et al., 2006),

sorted by decreasing eigenvalues which resemble sinusoids, index n being thought of as "frequency".

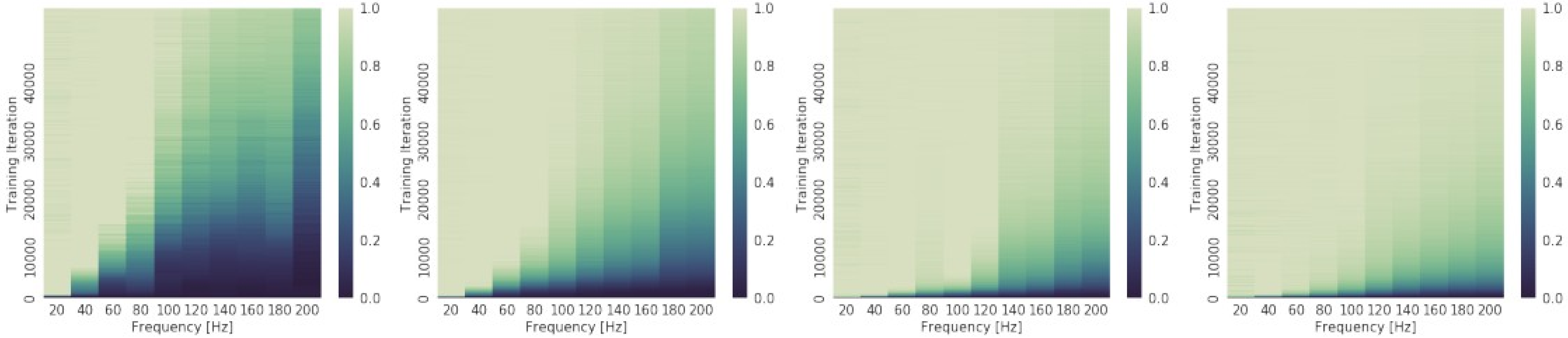
Experiment 5: Manifold hypothesis

- Identify functio $\lambda = \tau \circ \gamma$ where $\gamma : [0,1]^m \to \mathbb{R}^d$ and $\tau : \mathbb{R}^d \to \mathbb{R}$
- We approximate this wit $f : \mathbb{R}^d \to \mathbb{R}$ via a neural network.
- Specifically, we desi $\gamma_L : [0,1] \to \mathbb{R}^2$:

$$\gamma_L(z) = R_L(z)(\cos(2\pi z), \sin(2\pi z))$$
$$\text{where } R_L(z) = 1 + \frac{1}{2}\sin(2\pi L z)$$

- We use same function as in the previous experiments ($\kappa = (20, 40, ..., 180, 200)$

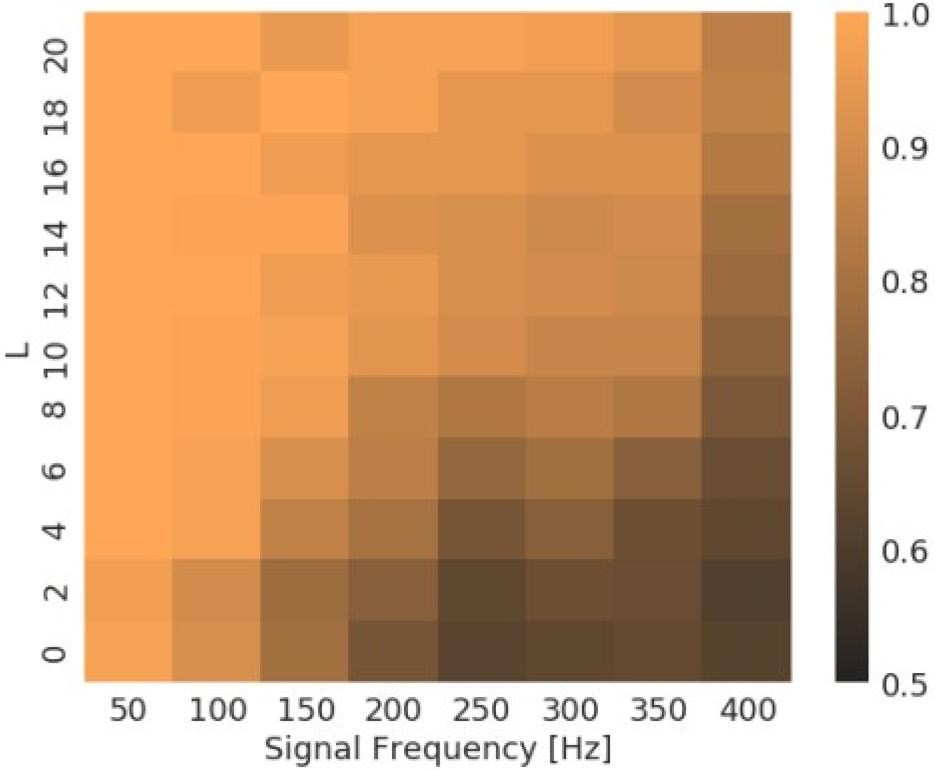
Experiment 5: Manifold hypothesis



(a) $L = 0$ (b) $L = 4$ (c) $L = 10$ (d) $L = 16$

- *Some manifolds (here with larger L) make it easier for the network to learn higher frequencies than others.*

Experiment 6: Manifold hypothesis II



- adapt the setting of Experiment 5 to binary classification by simply thresholding the function λ at 0.5 to obtain a binary

$$\lambda(z) = \sin(2\pi k z + \varphi)$$

$k \in \{50, 100, ..., 350, 400\}$ and $L \in \{0, 2, ..., 18, 20\}$.

Discussion

- Mathematical relationships can be derived btw. k and L
- E6 is also tested on MNIST dataset (See Experiment 8 in supp. mat.)

- To apply to our own training, identify what HF/LF components mean for our data.
- Can run frequency analyses on our own model.