Improving Neural Network Training Using Sobolev Loss functions

Aparna Gupte [1] Liu Zhang [2] Yunan Yang [3] Alex Townsend [4]

Main Contributions

- We propose a Sobolev norm-based loss function to modify the frequency bias property and accelerate training for functions supported on higher frequencies.
- We study how the frequency bias property depends on the choice of activation function σ.

Introduction

Neural networks (NNs) are known to learn lower Fourier frequency components first before higher components when trained with gradient descent [1]. This frequency bias property means that NNs take a long time to learn target functions that are supported on higher-frequency components (Figure 1).

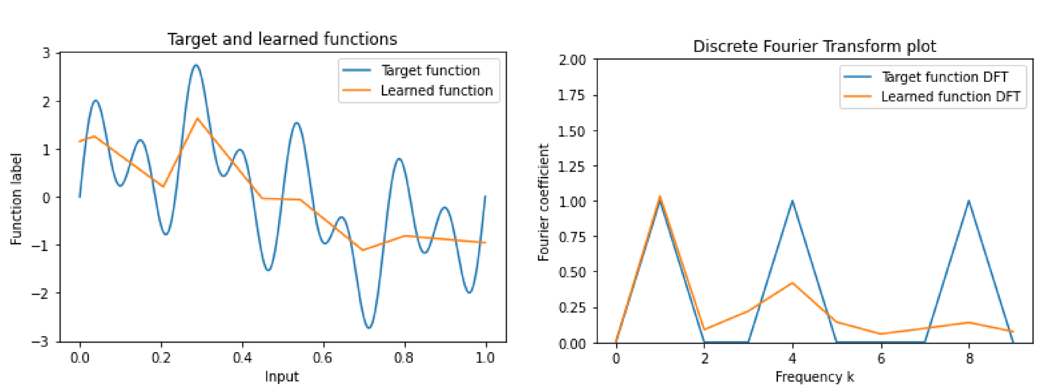

Figure 1: Training a NN with the target function $y(x) = \sin(2\pi x) + \sin(2\pi \cdot 4x) + \sin(2\pi \cdot 8x)$. After 4000 iterations, we plot the target and learned functions (left) and their Discrete Fourier Transforms (DFTs) (right). The NN has learned the low-frequency component $\sin(2\pi x)$ but not the high-frequency component $\sin(2\pi \cdot 8x)$.

Preliminaries

We study an over-parameterized NN with one-hidden layer:

$$f(\mathbf{x}; \mathbf{W}, \mathbf{a}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \, \sigma(\mathbf{w}_r^T \mathbf{x})$$

where $\mathbf{W}$, $\mathbf{a}$ are the weights and σ is the activation function. The NN takes in inputs that are uniformly distributed on the unit sphere, $\mathbf{x} \in \mathbb{S}^{d-1} \subset \mathbb{R}^d$. Given samples $\{\mathbf{x}_i, y_i\}_{i=1}^n$, we train the NN using gradient descent, with a learning rate η, to learn the inner weights $\mathbf{W}$ while keeping the outer weights $\mathbf{a}$ fixed. We seek to minimize the L^2 loss function

$$\Phi(\mathbf{W}) = \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i; \mathbf{W}, \mathbf{a}))^2.$$

Sobolev norm-based Loss Function

For a function $g : \mathbb{R}^d \to \mathbb{R}$, the H^s norm is a special case of the Sobolev norm, defined as

$$\|g\|_{H^s}^2 = \int_{\mathbb{R}^d} \left(1 + |\xi|^2\right)^{s/2} \hat{g}(\xi) \, d\xi,$$

where $\hat{g} : \mathbb{R}^d \to \mathbb{R}$ is the Fourier transform of g. We discretize this suitably to obtain a loss function (parameterized by $s \in \mathbb{R}$) that weighs different frequencies differently,

$$\Phi(\mathbf{W}) = \|\mathbf{r}\|_{H^s}^2 = \mathbf{r}^\top \mathbf{P} \mathbf{r},$$

where $\mathbf{r} = (f(\mathbf{x}_1) - y_1, \ldots, f(\mathbf{x}_n) - y_n)^\top$ is the residue vector. For $s = 0, -1, 1$, we observe the resulting frequency bias behaviour of the NN:

- $s = 0$: the NN has inherent low frequency bias, since this case is equivalent to L^2 loss,
- $s > 0$: larger weights are given to higher frequencies,
- $s < 0$: larger weights are given to lower frequencies.

Neural Tangent Kernel (NTK)

Based on the NTK framework, we analyze the training dynamics in the infinite-width limit (as $m \to \infty$). Arora et al. [2] showed that

$$\|\mathbf{r}(t)\|_2^2 \approx \sum_{i=1}^{n} (1 - \eta\lambda_i)^{2t} (\mathbf{v}_i^\top \mathbf{y})^2,$$

where λ_i, $\mathbf{v}_i$ are the eigenvalues and eigenvectors of the NTK matrix $\mathbf{K}$. As a result, components of the target function $\mathbf{y}$ along eigenvectors with larger eigenvalues are learned first. Further, by giving an explicit expression of $\mathbf{K}$, Basri et al. [3] showed that for the L^2 loss function and ReLU activation, eigenvectors of $\mathbf{K}$ are the spherical harmonics, and the eigenvalues $\lambda_k = \Theta(1/k^d)$. This confirms the low-frequency bias of neural networks.

With a Sobolov norm-based loss function, the NTK matrix becomes $\mathbf{KP}$. Since $\mathbf{P}$ has the same eigenvectors as $\mathbf{K}$ and its eigenvalues $\mu_k = \Theta(k^{2s})$, we can choose a H^s norm where $s = d/2$ to counterbalance the inherent low frequency bias.

Our Results

- Sobolev norm-based loss function with $s > 0$ reinforces inherent low frequency bias of NN. When $s < 0$, it counterbalances low frequency bias and accelerates training on target functions with higher frequency components. See Figure 2.
- The eigenvalues of the NTK matrix $\mathbf{K}$ decay polynomially for ReLU and Leaky ReLU activation functions (weaker low frequency bias) and exponentially for Sigmoid and Tanh functions (stronger low frequency bias). See Figure 3.

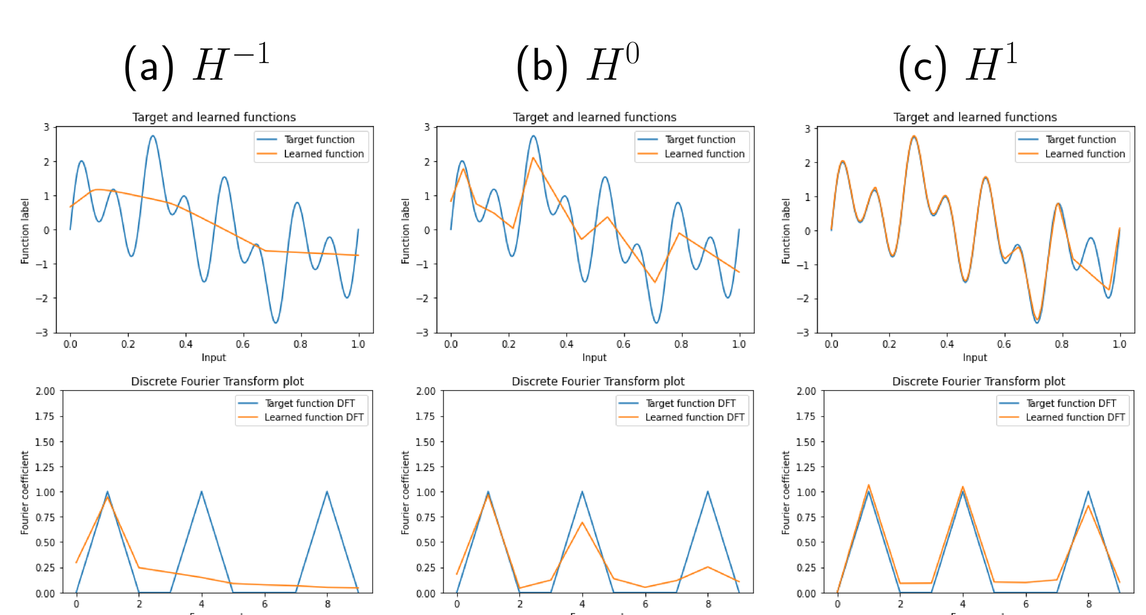

Figure 2: Frequency bias properties with Sobolev norm-based loss functions for different values of $s \in \{-1, 0, 1\}$. For $s = -1$, underfitting is observed: the intermediate and high frequency components are not learned by the NN. For $s = 1$, the NN has learned all three frequency components of the target function.

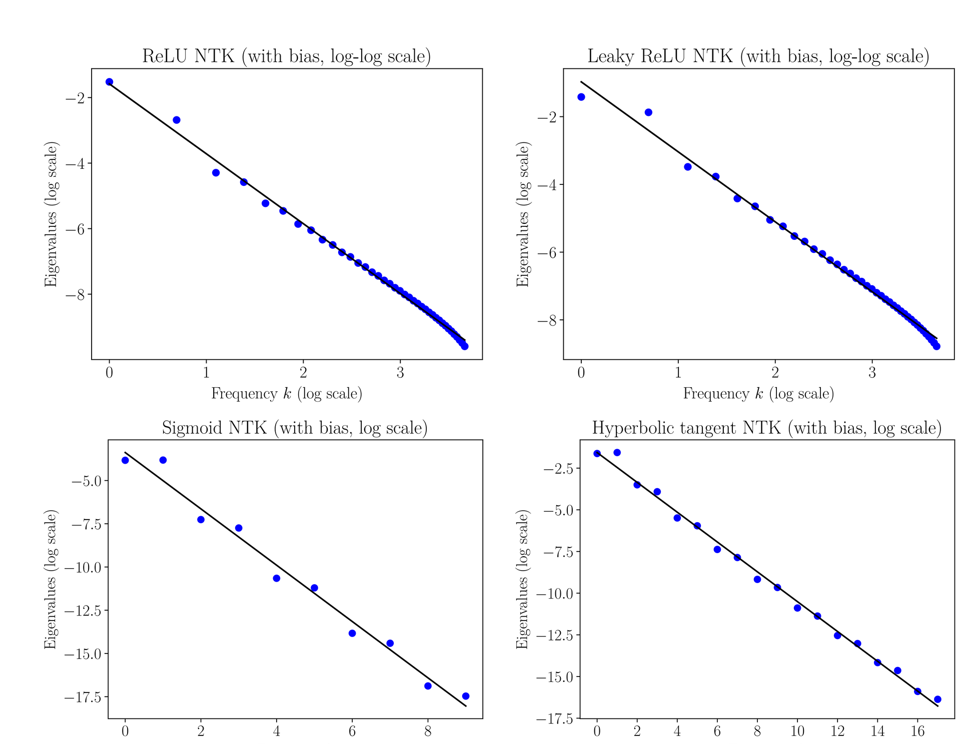

Figure 3: Numerically estimated eigenvalues of the NTK matrix $\mathbf{K}$ for different activation functions. The eigenvalues of $\mathbf{K}$ decay polynomially for ReLU and Leaky ReLU and exponentially for Sigmoid and Tanh.

Effect of Activation Function

We study how the activation function affects the frequency bias property by numerically estimating the eigenvalues of the NTK matrix $\mathbf{K}$ for different activation functions, including ReLU, Leaky ReLU, Sigmoid and Tanh (Figure 3).

Our numerical experiments are based on the Funk-Hecke theorem: For kernel K (which depends on the activation function), the eigenvalue corresponding to the k-th degree zonal harmonic, is given by

$$\lambda_k^d = \mathrm{Vol}(\mathbb{S}^d) \int_{-1}^{1} K(t) P_{k,d}(t) (1 - t^2)^{\frac{d-2}{2}} dt, \qquad (1)$$

where $P_{k,d}(t)$ denotes the Gegenbauer polynomial.

Applications

While it is currently difficult to apply the Sobolev norm-based loss function to real world data sets because of the lack of uniform data distributions, one promising application is to accelerate PINNs (Physics-Informed Neural Networks) in solving ordinary and partial differential equations.

Future Work

A future direction to explore would be to extend the theory and experiments to non-uniform data distributions and higher dimensions. This will potentially allow us to apply the Sobolev norm-based loss function to real world datasets.

Acknowledgements

We thank our mentors Prof. Yunan Yang and Prof. Alex Townsend for their guidance during the Cornell Maths REU, which has been a wonderful learning experience.

References

[1] Nasim Rahaman et al.
On the spectral bias of neural networks.
In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[2] Sanjeev Arora et al.
Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks.
May 2019.

[3] Ronen Basri et al.
The convergence rate of neural networks for learned functions of different frequencies.
2019.

[1] MIT, agupte@mit.edu.
[2] Yale-NUS, zhangliu@u.yale-nus.edu.sg.
[3] ETH Zurich.
[4] Cornell University.