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 $\sigma$.

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 $g(x) = \sin(2\pi x) + \sin(2\pi 4x) + \sin(2\pi 8x)$. 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 $\sin(2\pi 8x)$.

Sobolev norm-based Loss Function

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

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

where $\hat{g} : \mathbb{R}^d \rightarrow \mathbb{R}$ is the Fourier transform of $g$, and $s \in \mathbb{R}$.

We propose a Sobolev norm-based loss function $L^s$, defined as

$$L^s = \frac{1}{m} \sum_{i=1}^m (g(x_i) - y_i)^2 + \frac{\sigma(w_i^T x_i)}{\|w_i\|_{H^s}^2},$$

where $m$ is the number of training samples, $\sigma$ is the activation function, and $w_i$ are the learned weights.

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 $K$ 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.

Neural Tangent Kernel (NTK)

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

$$\|r(t)\|_{\mathcal{H}^s}^2 \approx \sum_{i=1}^n (1 - \eta \lambda_i)^{2s} (v_i^T y_i)^2,$$

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

Effect of Activation Function

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

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 data sets.

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

On the spectral bias of neural networks.

[2] Saikarth Aeron et al.
Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks.
May 2023.

The convergence rate of neural networks for learned functions of different frequencies.
2019.

1 MIT, aapunte@mit.edu
2 Yale-NUS, zliang@yale-nus.edu.sg
3 ETH Zurich
4 Cornell University.