Tuning Frequency Bias in Neural Network Training with Nonuniform Data

Annan Yu[†], Yunan Yang[‡], and Alex Townsend[†] [†]Cornell University [‡]ETH Zürich

ICLR International Conference On Learning Representations

What is Frequency Bias?

In the early epochs of neural network (NN) training, an overparameterized NN often finds a low-frequency fit of the training data while higher frequencies are learned in later epochs [2]. This phenomenon is called “frequency bias." It partially explains why NN training can achieve small generalization errors.

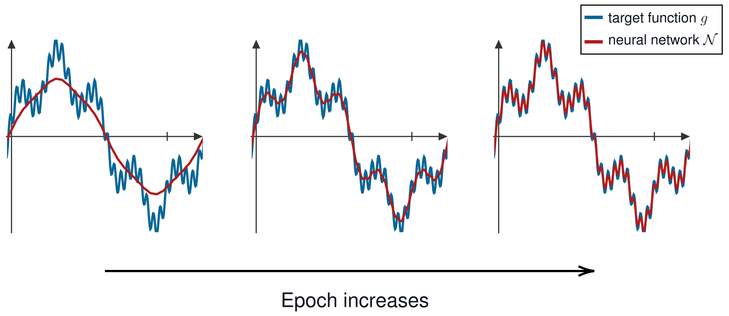

Fig. 1: Frequency bias states that the low-frequency patterns of the training data are learned faster than the high-frequency ones.

Neural Tangent Kernel

One way to theoretically understand the frequency bias is to use the so-called neural tangent kernel (NTK). Given a NN, denoted by $\mathcal{N}(\mathbf{x};\mathbf{W}(t))$, with parameters $\mathbf{W}$, the NTK is given by $K(\mathbf{x},\mathbf{x}';\mathbf{W}) = \langle(\partial/\partial\mathbf{W})\mathcal{N}(\mathbf{x};\mathbf{W}),(\partial/\partial\mathbf{W})\mathcal{N}(\mathbf{x}';\mathbf{W})\rangle$. Assume we want to learn a target function g on the unit hypersphere $\mathbb{S}^{d-1}$ using the loss function

$$\Phi(\mathbf{W}) = \frac{1}{2}\int_{\mathbb{S}^{d-1}} |g(\mathbf{x}) - \mathcal{N}(\mathbf{x};\mathbf{W})|^2 d\mu(\mathbf{x}) \tag{1}$$

and the gradient descent training algorithm with step size η. If a shallow-wide ReLU NN is overparameterized and η is small enough, the dynamic of the residual $z_t(\mathbf{x}) = g(\mathbf{x}) - \mathcal{N}(\mathbf{x};\mathbf{W}(t))$ can be written as

$$z_{t+1}(\mathbf{x}) - z_t(\mathbf{x}) \approx -\eta \int_{\mathbb{S}^{d-1}} K^{\infty}(\mathbf{x},\mathbf{x}') z_t(\mathbf{x}') d\mu(\mathbf{x}'), \tag{2}$$

where K^{∞} does not depend on $\mathbf{W}$ and approximates the NTK $K(\cdot,\cdot;\mathbf{W}(t))$.

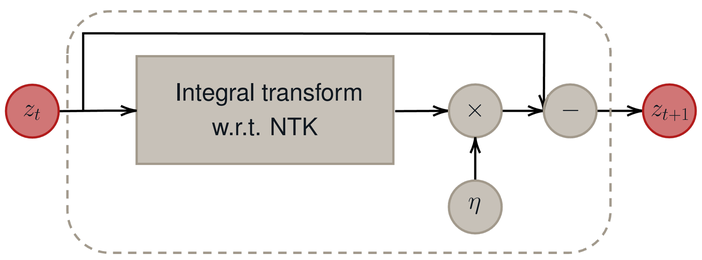

Fig. 2: The NTK characterizes the dynamic of the residual in NN training. The update in the residual is equal to η times the integral transform of the prior residual.

References

[1] R. Basri et al. “The convergence rate of neural networks for learned functions of different frequencies”. In: *Adv. Neur. Info. Proc. Syst.* 32 (2019).

[2] N. Rahaman et al. “On the spectral bias of neural networks”. In: *Inter. Conf. Mach. Learn.* PMLR. 2019, pp. 5301–5310.

Frequency Bias with Uniform Data

Suppose we are given a training dataset $\{(\mathbf{x}_i, g(\mathbf{x}_i))\}_{i=1}^N$. If $\mathbf{x}_i$ are uniformly distributed on $\mathbb{S}^{d-1}$ and μ in eq. (1) is the counting measure on $\{\mathbf{x}_i\}_{i=1}^N$, then we have

$$\Phi(\mathbf{W}) = \frac{1}{2}\sum_{i=1}^{N} |g(\mathbf{x}_i) - \mathcal{N}(\mathbf{x}_i;\mathbf{W})|^2 \approx \frac{N}{2\text{vol}(\mathbb{S}^{d-1})}\int_{\mathbb{S}^{d-1}} |g(\mathbf{x}) - \mathcal{N}(\mathbf{x};\mathbf{W})|^2 d\mathbf{x}. \tag{3}$$

When $d\mu(\mathbf{x}) = d\mathbf{x}$, [1] shows that the Fourier modes are the eigenfunctions of the integral operator in eq. (2) and the higher the frequency, the smaller the eigenvalue. In particular, if Y_ℓ is a frequency-ℓ Fourier mode, then we have

$$\int_{\mathbb{S}^{d-1}} K^{\infty}(\cdot,\mathbf{x}') Y_\ell(\mathbf{x}') d\mathbf{x}' = \mathcal{O}\big(\ell^{-d}\big) Y_\ell(\cdot).$$

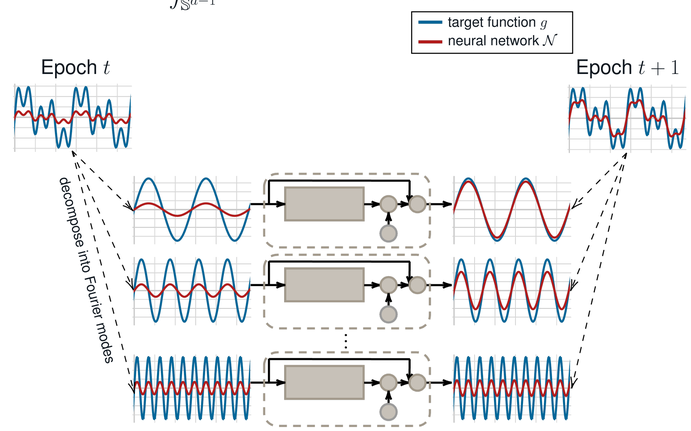

Fig. 3: The Fourier modes are the eigenfunctions of the integral transform. The higher the frequency is, the smaller the eigenvalue of the integral operator is, and the more slowly the neural network converges to the target function in the corresponding Fourier mode.

Frequency Bias with Nonuniform Data

When $\mathbf{x}_i$ are not uniform, Φ in eq. (3) is no longer an approximation of the L^2 loss with respect to the Lebesgue measure. To get frequency bias, we compute a quadrature rule $\{c_i\}_{i=1}^n$ at $\{x_i\}_{i=1}^N$ and define a quadrature-based loss function

$$\tilde{\Phi}(\mathbf{W}) = \frac{1}{2}\sum_{i=1}^{N} c_i |g(\mathbf{x}_i) - \mathcal{N}(\mathbf{x}_i;\mathbf{W})|^2 \approx \frac{1}{2}\int_{\mathbb{S}^{d-1}} |g(\mathbf{x}) - \mathcal{N}(\mathbf{x};\mathbf{W})|^2 d\mathbf{x}.$$

We can then use the spectral properties of K^{∞} to study frequency bias.

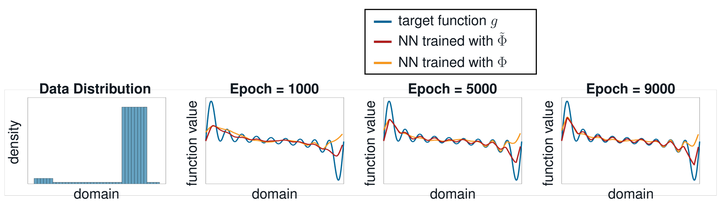

Fig. 4: Frequency bias is observed when a quadrature-based loss function is used, but not when the mean-squared error is used.

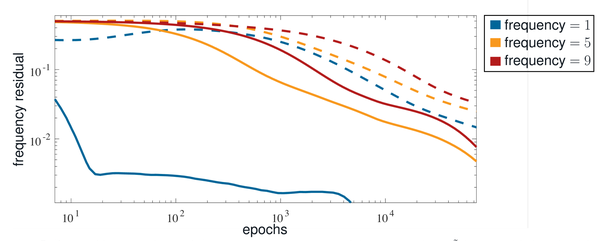

Fig. 5: Decay of the magnitudes of Fourier modes when we use Φ (dashed line) and $\tilde{\Phi}$ (solid line) to train the NN.

Tuning Frequency Bias

We obtain frequency bias by using a quadrature-based loss function $\tilde{\Phi}$. We can further tune the frequency bias by using a Sobolev-norm-based loss function

$$\Phi_s(\mathbf{W}) = \frac{1}{2}\|g - \mathcal{N}(\cdot,\mathbf{W})\|_{H^s}^2. \tag{4}$$

The parameter s determines the eigenvalues of the Fourier modes of the integral operator in eq. (2). While $\tilde{\Phi}$ is a discretization of Φ_0, frequency bias is enhanced if $s < 0$ and suppressed or reversed if $s > 0$.

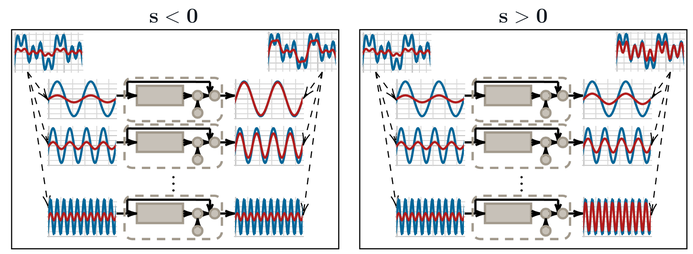

Fig. 6: If $s < 0$, the low-frequency modes are learned faster, which enhances frequency bias; if $s > 0$, the high-frequency modes are learned faster and frequency bias is suppressed or reversed.

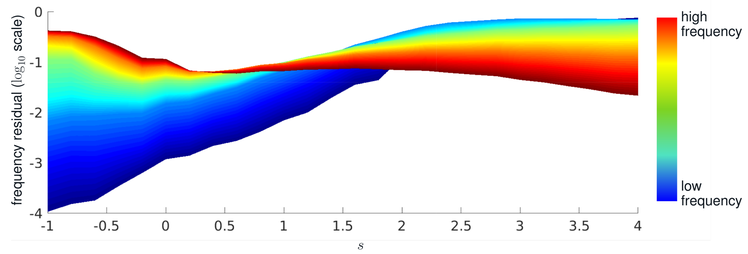

Fig. 7: We can tune frequency bias by changing s. As s increases, the low-frequency modes are learned more slowly and the high-frequency modes faster. Frequency bias is reversed at $s = 2$.

Autoencoders with Frequency Bias

Tuning frequency bias is useful in solving real-world problems. We construct an auto-encoder using NNs that deblurs MNIST images contaminated by noise. We train the auto-encoder using the Φ_s loss function with various values of s.

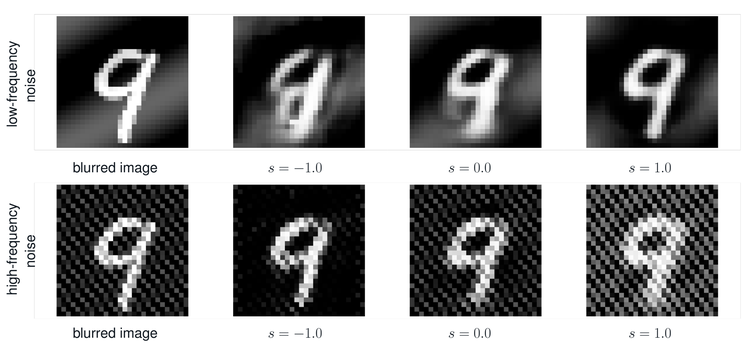

Fig. 8: When the images are contaminated by low-frequency noise, a positive s reverses the frequency bias and allows us to learn the high-frequency data faster. When the high-frequency noise prevails, a negative s enhances the frequency bias and filters out the noise.