The Spectral Bias of Polynomial Neural Networks

Moulik Choraria[1] Leello Daadi[2] Grigorios Chrysos[2]
Julien Mairal[3] Volkan Cevher[2]

[1]University of Illinois at Urbana-Champaign

[2]EPFL, Switzerland

[3]Univ. Grenoble-Alpes, Inria

Motivation

- Neural Networks (NNs) demonstrate a learning bias during training, wherein low-frequency functions are learned faster[1].
- Polynomial Neural Networks (PNNs) with multiplicative layers were recently shown to be effective at image generation and face recognition, where high-frequency information is critical[2].
- Inspired by these, we conduct a spectral analysis of PNNs, using the Π-Net parametrization.

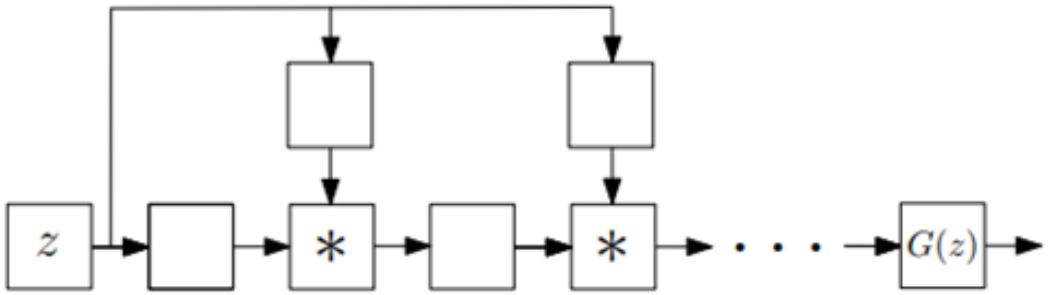

Figure: A Π-Net variant.

[1] Nasim Rahaman et al. "On the Spectral Bias of Neural Networks". In: (2019).

[2] Grigorios G. Chrysos et al. "Π-nets: Deep Polynomial Neural Networks". In: 2020.

Setup (Neural Tangent Kernel Regime)

- The 2-layer feed-forward ReLU (σ) network in infinite width limit ($m \to \infty$): $f_{\boldsymbol{W}}(\boldsymbol{x}) = \sqrt{\frac{2}{m}}\boldsymbol{W_2}\sigma(\boldsymbol{W_1}\boldsymbol{x})$
- The corresponding Π-Net includes a multiplicative layer via the Hadamard product ($*$): $f_{\boldsymbol{W}}(\boldsymbol{x}) = \sqrt{\frac{2}{m}}\boldsymbol{W_3}[\sigma(\boldsymbol{W_1}\boldsymbol{x}) * \sigma(\boldsymbol{W_2}\boldsymbol{x})]$
- The NTK for the 2-layer feed-forward network is composed of arc-cosine kernels: $\boldsymbol{\kappa}(\boldsymbol{x}, \boldsymbol{x'}) = 2\langle \boldsymbol{x}, \boldsymbol{x'}\rangle \boldsymbol{\kappa_1}(\boldsymbol{x}, \boldsymbol{x'}) + 2\boldsymbol{\kappa_2}(\boldsymbol{x}, \boldsymbol{x'}))$.
- The Π-Net NTK takes a product of arc-cosine kernels form: $\boldsymbol{\kappa_\pi}(\boldsymbol{x}, \boldsymbol{x'}) = 2(2\langle \boldsymbol{x}, \boldsymbol{x'}\rangle \boldsymbol{\kappa_1}(\boldsymbol{x}, \boldsymbol{x'}) + \boldsymbol{\kappa_2}(\boldsymbol{x}, \boldsymbol{x'}))\boldsymbol{\kappa_2}(\boldsymbol{x}, \boldsymbol{x'})$.

Result & Implications

- For the linear operator w.r.t. the 2-layer feed-forward NTK, eigenvalues $(\mu_k)_{k=1}^{\infty}$ for frequency k (even), decay as $\mu_k = \Omega(k^{-d-1})$ when $k \gg d$.[3]

Theorem (Ours)

Let $\{\mu_{\pi,1}, \mu_{\pi,2}, \dots\}$ denote the eigenvalues of the linear operator $L_{\boldsymbol{\kappa_\pi}}$ w.r.t. the to Π-kernel $\boldsymbol{\kappa_\pi}$. For $k \gg d$ ($k \neq 2$ mod 4), $\mu_{\pi,k} = \Omega(k^{-d/2-2})$.

- Slower decay leads to larger RKHS (better approximation properties)
- Slower decay leads to faster learning for higher frequencies in kernel regression (translating to NNs in NTK regime).[4]

[3] Alberto Bietti and Julien Mairal. "On the Inductive Bias of Neural Tangent Kernels". In: (2019).

[4] Yuan Cao et al. "Towards Understanding the Spectral Bias of Deep Learning". In: (2020).

Learning Sinusoids

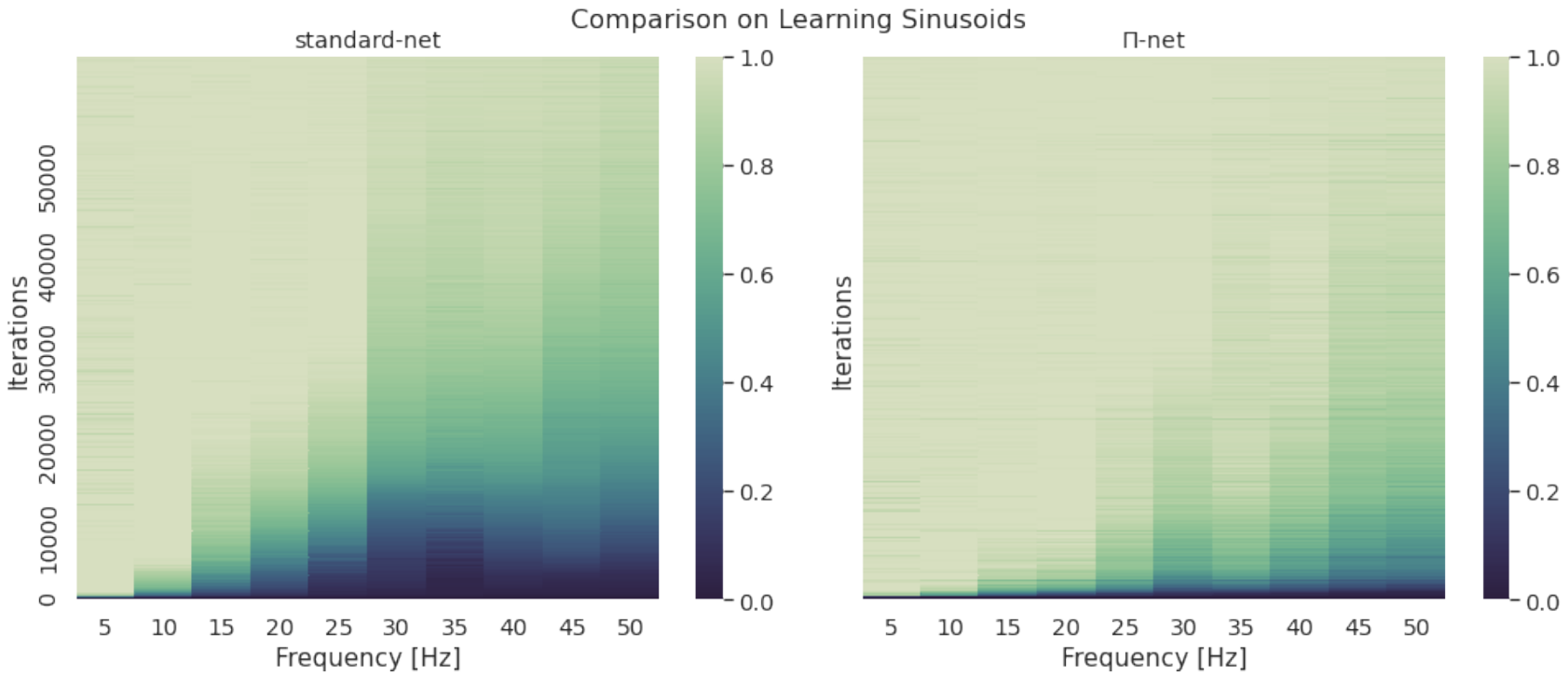

Figure: Comparison between rate of learning sinusoids with fully connected networks (heat map depicts fraction of learning for each frequency over iterations) yields that Π-Nets with multiplicative layers learn higher frequencies much faster than feed-forward networks.

Denoising Images

- We consider the task of denoising an image using a U-Net structure, in the Deep Image Prior[5] setting to observe this spectral bias in presence of noise.

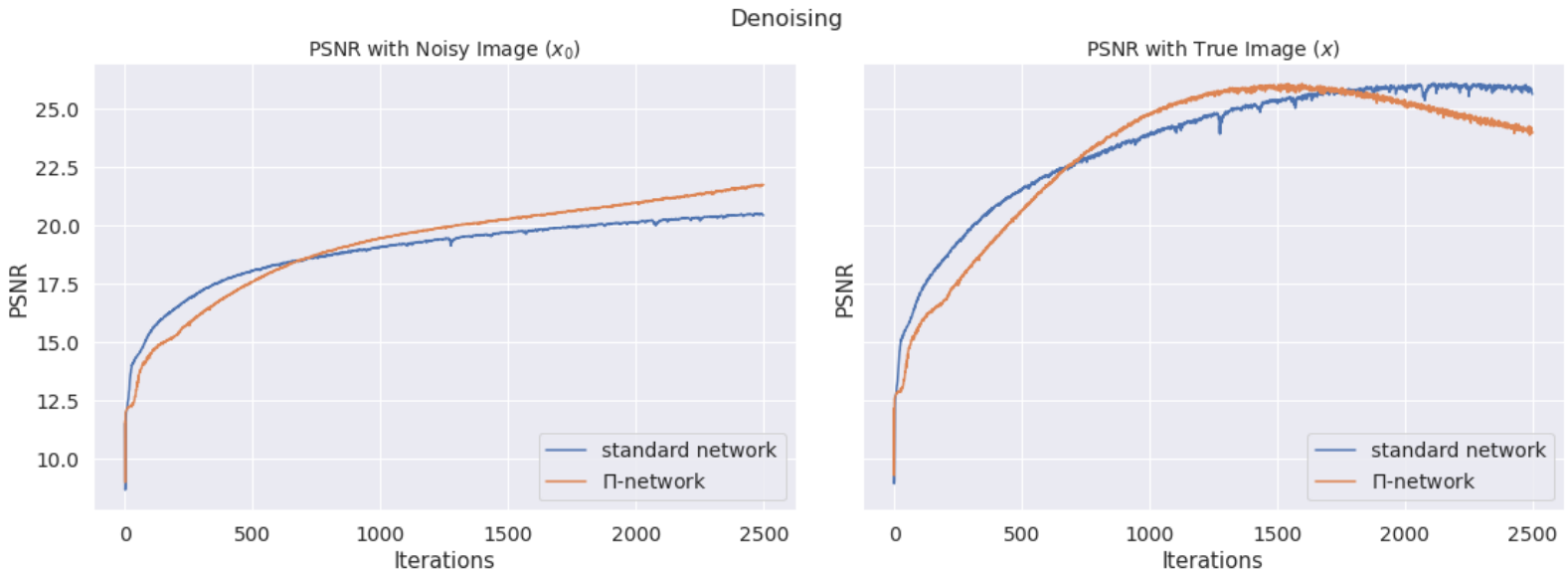

Figure: Denoising comparison shows that Π-Nets reach the peak PSNR w.r.t. the true image much faster and show a lesser impedance towards learning high frequency noise, compared to standard Deep networks.

[5] Victor Lempitsky, Andrea Vedaldi, and Dmitry Ulyanov. "Deep Image Prior". In: 2018.

Label Noise in Classification

- We consider binary classification in MNIST (with two digits) with fully connected networks in the presence of label noise, as in Rahaman et al.

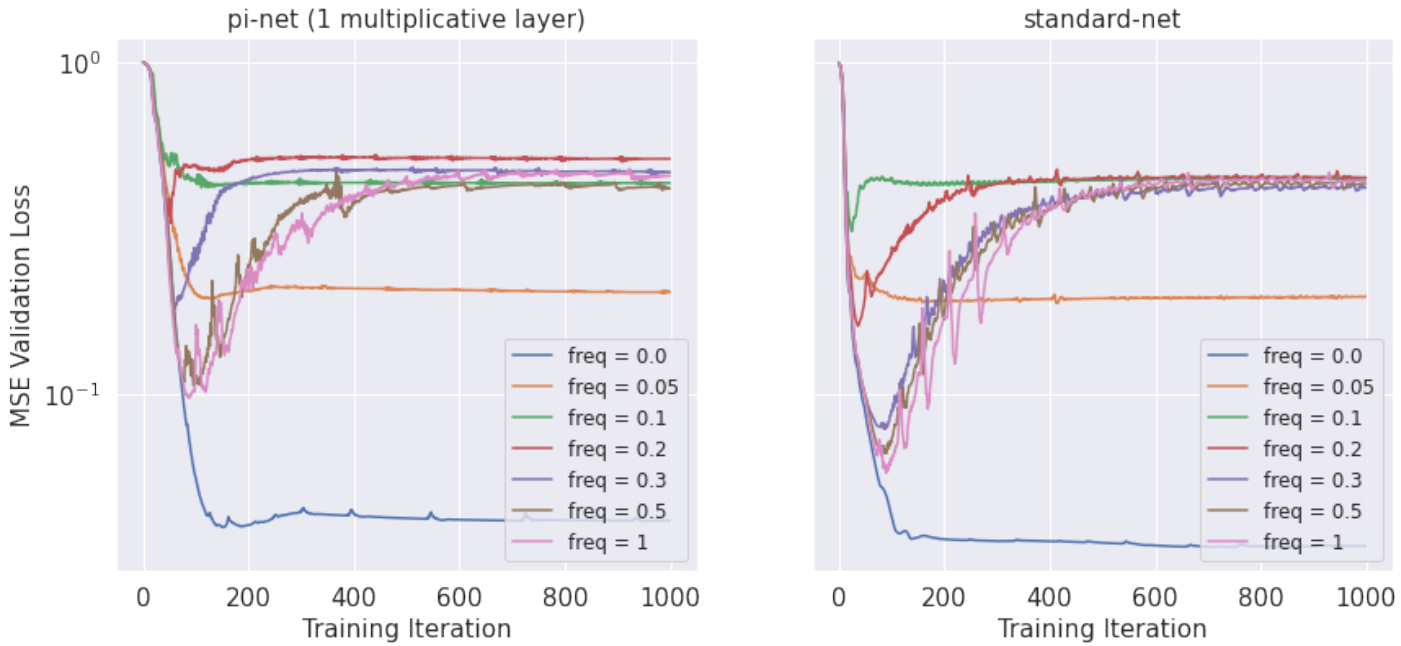

Figure: We observe that Π-Nets are much faster in picking up label noise of higher frequencies, leading to a much smaller "dip", thus demonstrating a stronger tendency to learn more complex decision boundaries.

Future Directions

- **Extensions to general polynomials:** New proof techniques (beyond NTKs) are needed for formulating more general polynomials.
- **On Generalization in PNNs**: Deviation from the "overparametrized networks learn low-complexity functions" picture of generalization or that simple decision boundaries lead to adversarial robustness.
- **On the Empirical front**: Considering SGD (instead of GD in our experiments) can lead to different results. Additionally, in training of Wasserstein GANs, it would be interesting to relate the moment-matching perspective to spectral bias in polynomial networks.

The Spectral Bias of Polynomial Neural Networks

Contact author: moulikc2@illinois.edu