Dynamics of Deep Neural Networks -- A Fourier Analysis Perspective

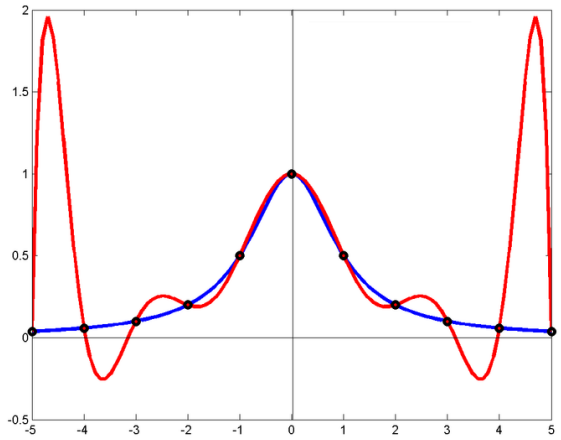
Yaoyu Zhang

Problem of fitting a 1-d function from data

Find an interpolation of $\mathcal{D}: \{(x_i, y_i)\}_{i=1}^N$ in $\mathcal{H}: \{h(\cdot; \Theta) | \Theta \in \mathbb{R}^M\}$

Example:

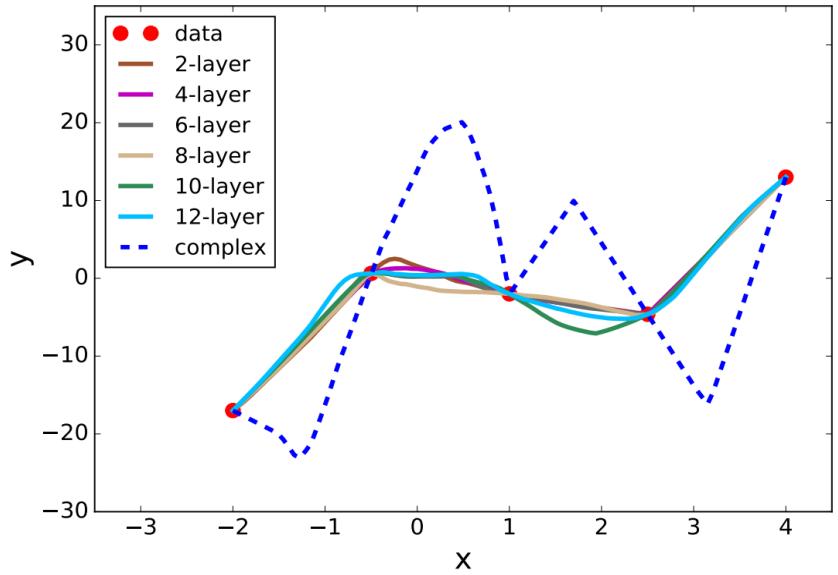
$h(x; \Theta) = \theta_1 + \theta_2 x + \cdots + \theta_M x^{M-1}$ with $M = N$



Conventional wisdom: $M < N$.

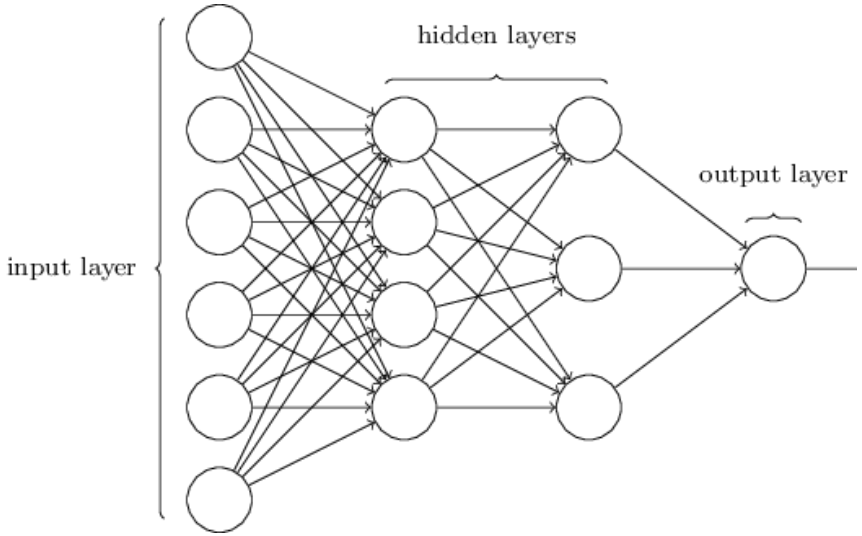
Modern wisdom?

Using neural network with $M \gg N$.



Lei Wu, Zhanxing Zhu, Weinan E, 2017

Deep Neural Network



$$h(x;\Theta) = h^{[H]}$$

$$h^{[j]} = \sigma\left(W^{[j]}h^{[j-1]} + b^{[j]}\right)$$

$$\Theta: \left[W^{[j]}, b^{[j]}\right]_{j=1,\cdots,H}$$

Example: Two-layer NN

$$h(x;\Theta) = \sum_{i=1}^{s} w_i^{[2]}\sigma(w_i^{[1]}x + b_i^{[1]})$$

Dynamics

Data: $\{(x_i, y_i)\}_{i=1}^N$

$$L(\Theta) = \sum_{i=1}^N (h(x_i;\Theta) - y_i)^2$$

$$\dot{\Theta} = -\nabla_\Theta L(\Theta)$$

Dynamics is the key to the "preference" of DNN.

$$t \mapsto \Theta(t)$$

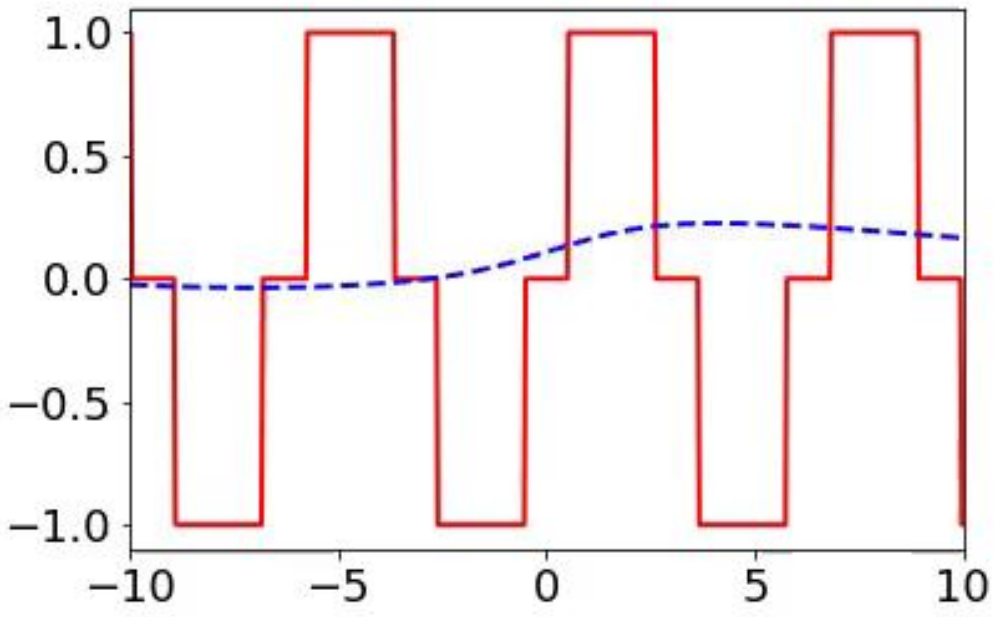
$$t \mapsto h(x,t) := h(x;\Theta(t))$$

How DNNs fit a 1-d function?

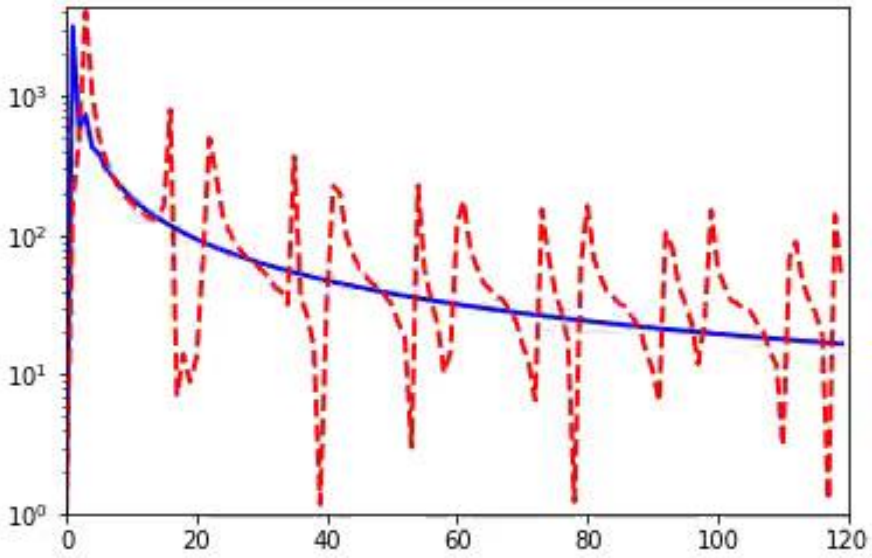
Study $h(x,t)$ through the lens of Fourier transform

- Phenomenon
- Effective model
- Analysis

Evolution of $h(x, t)$



Through the lens of Fourier transform $\widehat{h}(\xi, t)$



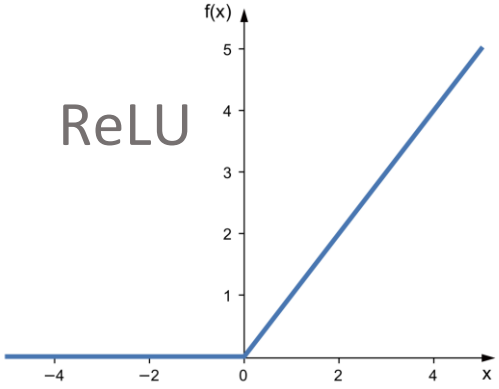
Frequency Principle (F-Principle):

DNNs often fit target functions from low to high frequencies during the training process.

Linear F-Principle dynamics

2-layer NN: $h(x;\Theta)=\sum_{i=1}^{s} w_i \mathrm{ReLU}(r_i(x+l_i))$

s sufficiently large



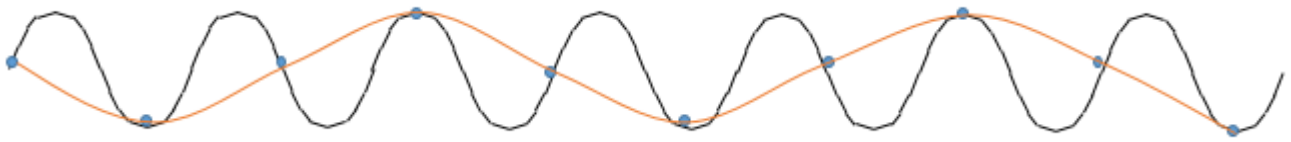
$$\partial_t \hat{h}(\xi,t) = -\left[\frac{4\pi^2\langle r^2 w^2\rangle}{\xi^2}+\frac{\langle r^2\rangle+\langle w^2\rangle}{\xi^4}\right]\left(\widehat{h_p}(\xi,t)-\hat{f}_p(\xi,t)\right)$$

$\langle r^2w^2\rangle = \mathrm{var}(r(0)w(0)), \langle r^2\rangle = \mathrm{var}(r(0)), \langle w^2\rangle = \mathrm{var}(w(0))$

f: target function; $(\cdot)_p = (\cdot)p$,where $p(x)=\frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i)$;

$\hat{\cdot}$: Fourier transform; ξ: frequency

aliasing



Preference induced by LFP dynamics

$$\partial_t \hat{h}(\xi,t) = -\left[\frac{4\pi^2\langle r^2 w^2\rangle}{\xi^2} + \frac{\langle r^2\rangle + \langle w^2\rangle}{\xi^4}\right]\left(\widehat{h_p}(\xi,t) - \hat{f}_p(\xi,t)\right)$$

low frequency preference

$$\min_{h\in F_\gamma} \int \left[\frac{4\pi^2\langle r^2 w^2\rangle}{\xi^2} + \frac{\langle r^2\rangle+\langle w^2\rangle}{\xi^4}\right]^{-1} \left|\hat{h}(\xi)\right|^2 \mathrm{d}\xi$$

$$\text{s.t. } h(x_i) = y_i \text{ for } i = 1,\cdots,N$$

Case 1: ξ^{-2} dominant

- $\min \int \xi^2 \left|\hat{h}(\xi)\right|^2 \mathrm{d}\xi \sim \min \int |h'(x)|^2 \mathrm{d}\xi \rightarrow$ **piecewise linear**

Case 2: ξ^{-4} dominant

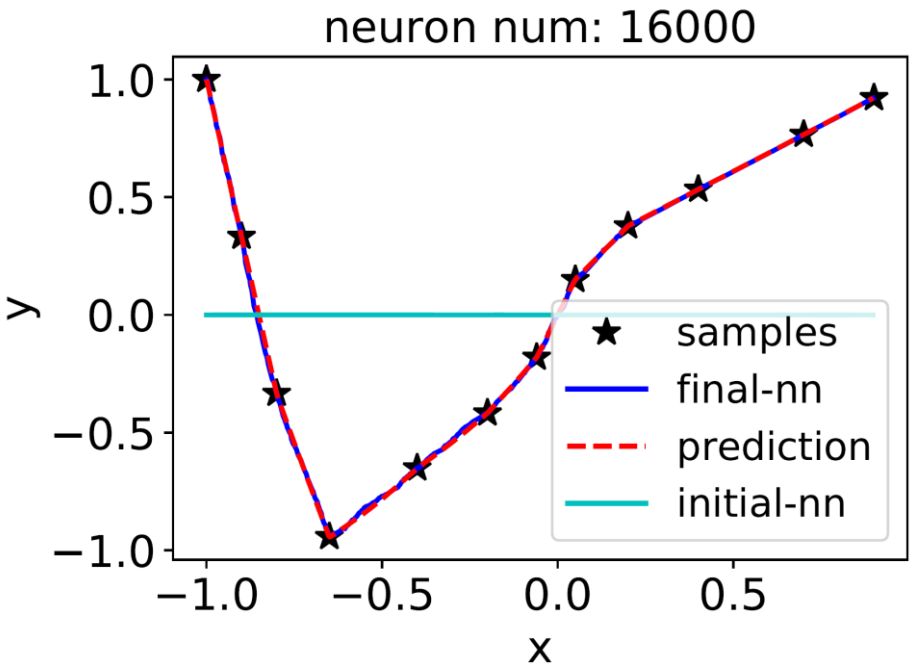
- $\min \int \xi^4 \left|\hat{h}(\xi)\right|^2 \mathrm{d}\xi \sim \min \int |h''(x)|^2 \mathrm{d}\xi \rightarrow$ **cubic spline**

Regularity can be changed through initialization

Case 1

$$\langle r^2 \rangle + \langle w^2 \rangle \gg 4\pi^2 \langle r^2 w^2 \rangle$$

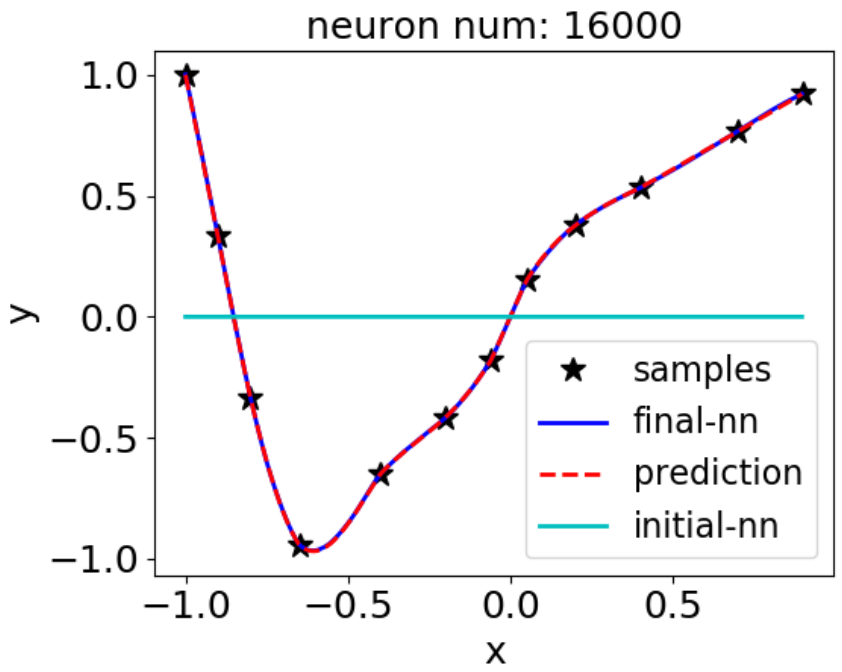
$$\min \int \xi^2 \left|\hat{h}(\xi)\right|^2 \mathrm{d}\xi$$



Case 2

$$4\pi^2 \langle r^2 w^2 \rangle \gg \langle r^2 \rangle + \langle w^2 \rangle$$

$$\min \int \xi^4 \left|\hat{h}(\xi)\right|^2 \mathrm{d}\xi$$



FP-norm and FP-space

FP-norm for all function $h \in L^2(\Omega)$:

$$\|h\|_\gamma := \left\|\hat{h}\right\|_{H_\Gamma} = \left(\sum_{k\in\mathbb{Z}^d} (\gamma(k))^{-2}|\hat{h}(k)|^2\right)^{\frac{1}{2}}.$$

$$F_\gamma(\Omega) = \{h \in L^2(\Omega) : \|h\|_\gamma < \infty\}.$$

A priori generalization error bound

Theorem *Suppose that the real-valued target function $f \in F_\gamma(\Omega)$, the training dataset $\{x_i; y_i\}_{i=1}^M$ satisfies $y_i = f(x_i)$, $i = 1, \cdots, M$, and h_M is the solution of the regularized model*

$$\min_{h-h_{\text{ini}}\in F_\gamma(\Omega)} \|h - h_{\text{ini}}\|_\gamma, \quad s.t. \quad h(x_i) = y_i, \quad i = 1, \cdots, M. \tag{17}$$

Then we have

(i) given $\gamma : \mathbb{Z}^d \to \mathbb{R}^+$, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over the random training samples, the population risk has the bound

$$L(h_M) \le \boxed{\|f - h_{\text{ini}}\|_\gamma} \|\gamma\|_{\ell^2} \left(\frac{2}{\sqrt{M}} + 4\sqrt{\frac{2\log(4/\delta)}{M}}\right). \tag{18}$$

NNs fit low frequency functions better

How NNs find a "good" interpolation of $\mathcal{D}$: $\{(x_i, y_i)\}_{i=1}^N$ in $\mathcal{H}$: $\{h(\cdot;\Theta)|\Theta \in \mathbb{R}^M\}$ for $M \gg N$?

initialization of Θ
regularity of $\sigma(\cdot)$

$$\dot{\Theta} = -\nabla_\Theta L(\Theta)$$

A function in $\mathcal{H}$ that:
1. fits the data.
2. has "specific regularity"

An ultimate answer: mathematics that explains

initialization of Θ
regularity of $\sigma(\cdot)$
architecture
(depth, width,…)
optimization alg

$$\dot{\Theta} = -\nabla_\Theta L(\Theta)$$

A function in $\mathcal{H}$ that:
1. fits the data.
2. has "specific structure"

DNNs love low frequencies!

Reference:

Zhang, Xu, Luo, Ma*, Explicitizing an Implicit Bias of the Frequency Principle in Two-layer Neural Networks, 2019*

Other works in a series:

Xu, Zhang, Xiao*, Training behavior of deep neural network in frequency domain, 2018*

Xu, Zhang, Luo, Xiao, Ma*, Frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks, 2019*

Zhang, Xu, Luo, Ma*, A type of generalization error induced by initialization in deep neural networks, 2019*

Luo, Ma, Xu, Zhang*, Theory on Frequency Principle in General Deep Neural Networks, 2019.*