II. Frequency Principle/Spectral Bias

Yaoyu Zhang

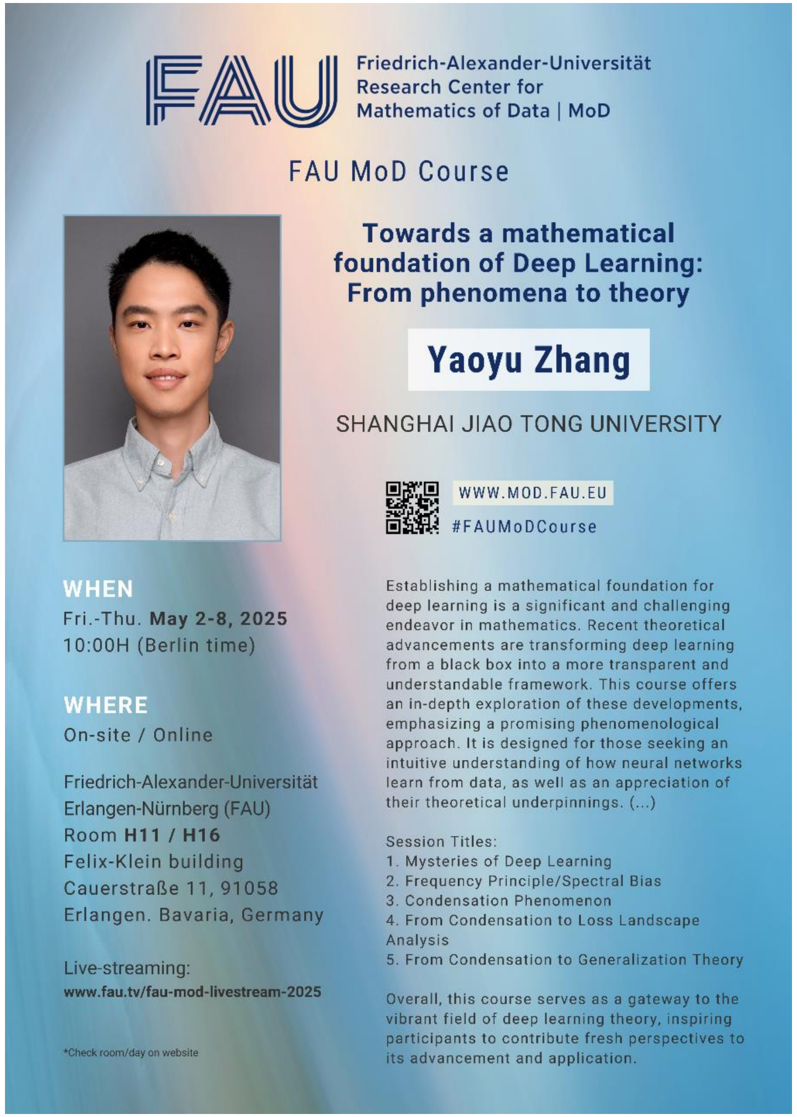
Institute of Natural Sciences & School of Mathematical Sciences

Shanghai Jiao Tong University

FAU MoD Course

饮 水 思 源 • 爱 国 荣 校

Deep learning is no longer a black-box



Towards a Mathematical Foundation of Deep Learning: From Phenomena to Theory

Date

Fri. – Thu. May 2 – 8, 2025

Session Titles

1. Mysteries of Deep Learning
2. **Frequency Principle/Spectral Bias**
3. Condensation Phenomenon
4. From Condensation to Loss Landscape Analysis
5. From Condensation to Generalization Theory

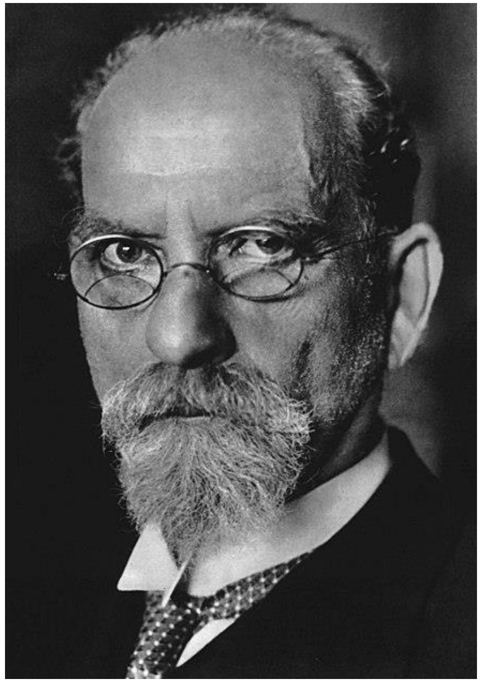
SJTU

A phenomenological methodology

1. **Suspension:** Suspend the prior and belief one may hold and focus on the facts about the object.
2. **Cumulation:** Discover and cumulate all possible facts about the object. Prioritize the more informative ones.
3. **Emergence:** A new framework shall emerge once enough pieces are uncovered.

Cumulation of phenomena from experience

> "Natural objects, for example, must be experienced before any theorizing about them can occur"
>
> \- Edmund Husserl



Husserl warns against this inversion of process, where theories can eclipse, misshape, or entirely ignore the vital qualities encountered in direct perception.

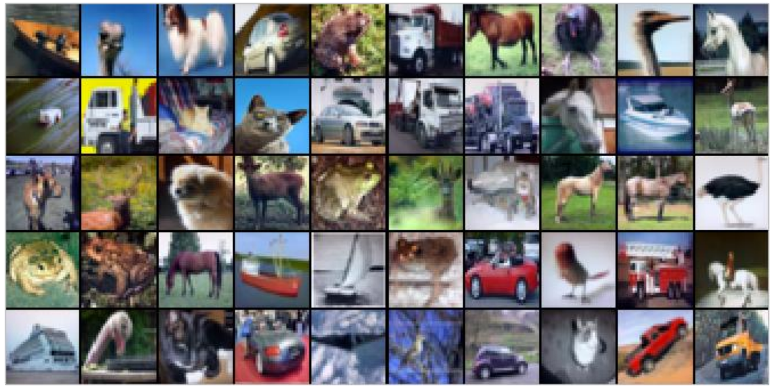
https://fixquotes.com/

https://en.wikipedia.org/wiki/Edmund_Husserl

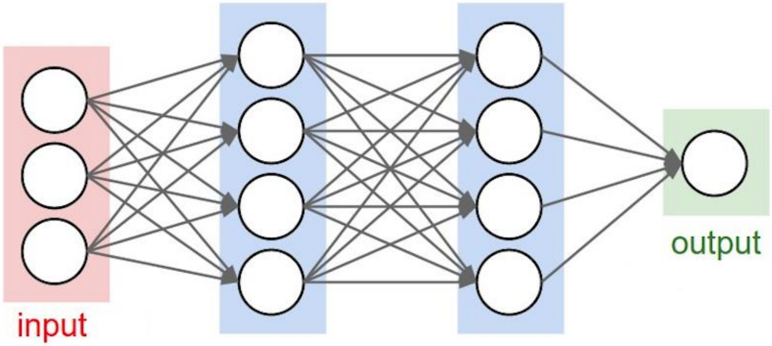
How to experience deep learning?

Problem: Given $\mathcal{D}: \{(x_i, y_i)\}_{i=1}^{n}$ and $\mathcal{H}: \{f(\cdot;\Theta) | \Theta \in \mathbb{R}^m\}$, find $f \in \mathcal{H}$ such that $f(x_i) = y_i$ for $i = 1, \cdots, n$.

$\mathcal{D}$



$\mathcal{H}$



$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{W}^{[L]}\sigma \circ (\cdots \boldsymbol{W}^{[2]}\sigma \circ (\boldsymbol{W}^{[1]}\boldsymbol{x} + \boldsymbol{b}^{[1]}) + \cdots) + \boldsymbol{b}^{[L]}$$

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

$$\Theta(0) = \Theta_0$$

$$L(\Theta) = \frac{1}{2n}\sum_{i=1}^{n}(f(x_i;\Theta) - y_i)^2$$

General observation:
$f(x_i;\Theta(\infty))$ often generalize well even when $m \gg n$.

How to experience deep learning?

Two key objects

- Trajectory in function space

$$f(\cdot, t)\colon \mathbb{R}^{+} \to \mathcal{H}$$

- Trajectory in parameter space

$$\Theta(t)\colon \mathbb{R}^{+} \to \mathbb{R}^{m}$$

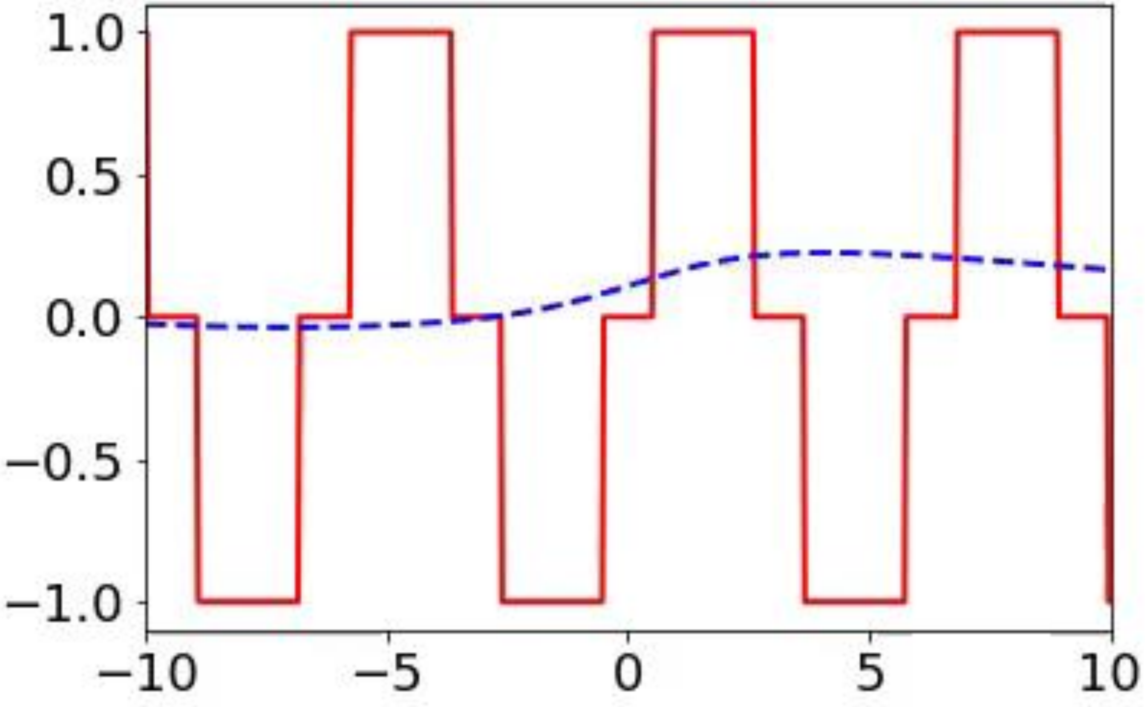
Common strategy: choose proper statistics for observation.

Limitation: choice of statistics reflect our bias, no guarantee for effectiveness.

Can we observe the whole trajectory?

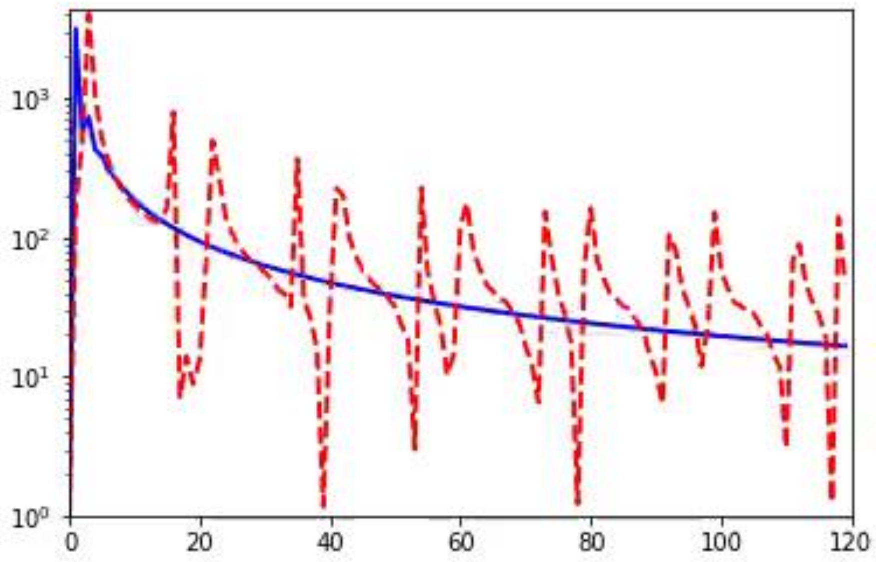
Frequency Principle

Evolution of neural network output function $f(x, \theta(t))$



tanh-DNN, 200-100-100-50

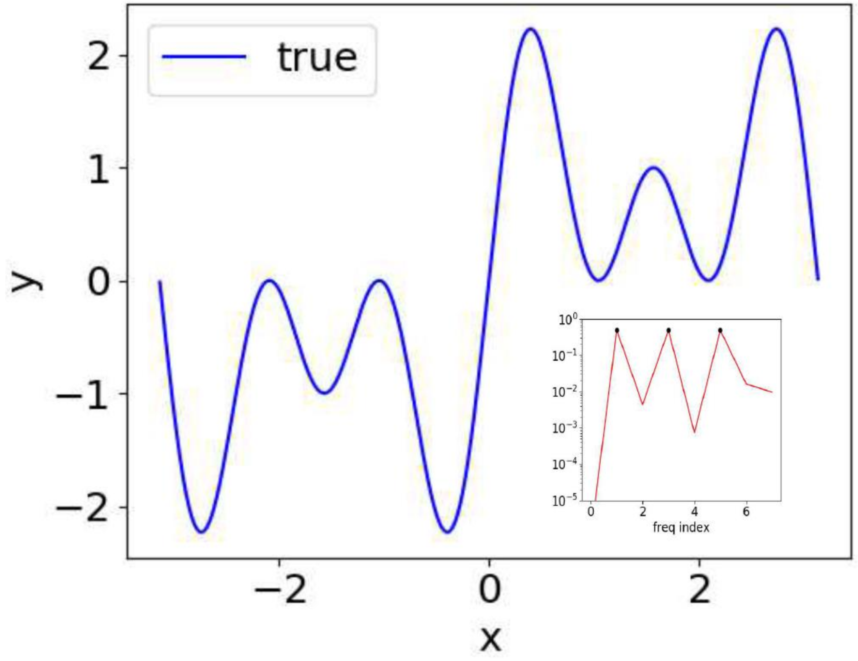
Through the lens of Fourier transform $\hat{f}(\xi, \theta(t))$

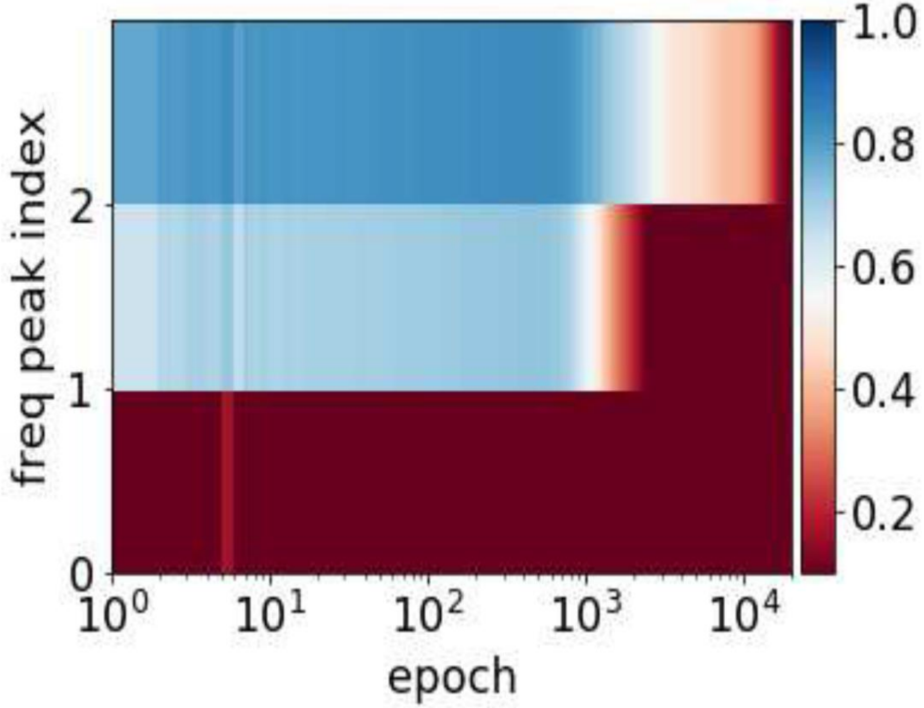


Frequency Principle (F-Principle):
DNNs often fit training data from low to high frequencies during the training.

Xu, Zhang, Xiao, *Training behavior of deep neural network in frequency domain,* 2018
Nasim Rahaman et al, *On the Spectral Bias of Neural Networks,* 2018

Synthetic curve with equal amplitude





How DNN fits a 2-d image?

(a) True image

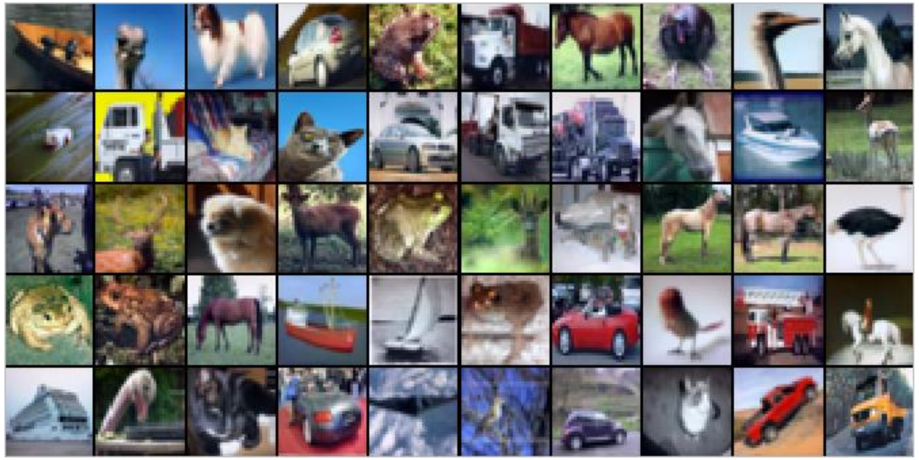
(b) DNN output

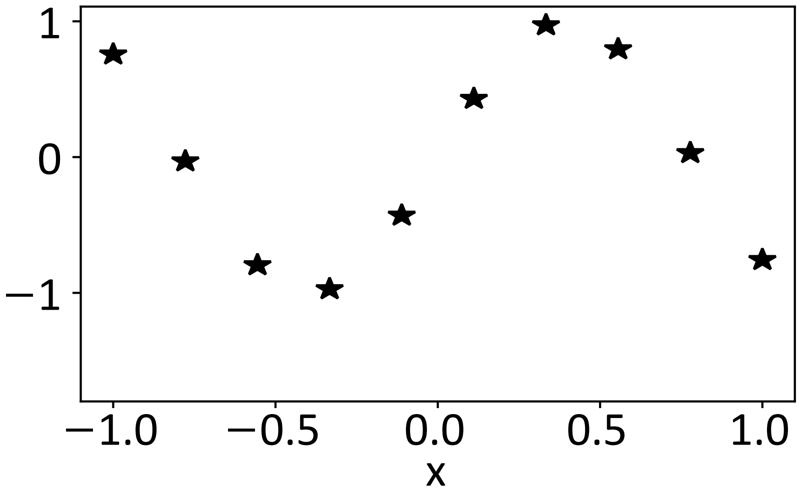
Target: image $I(\mathbf{x})\colon \mathbb{R}^2 \to \mathbb{R}$

$\mathbf{x}$: location of a pixel

$I(\mathbf{x})$: grayscale pixel value

High-dimensional real data?





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

Frequency

Image frequency (NOT USED)

- This frequency corresponds to the rate of change of intensity across neighboring pixels.

zero freq
Same color

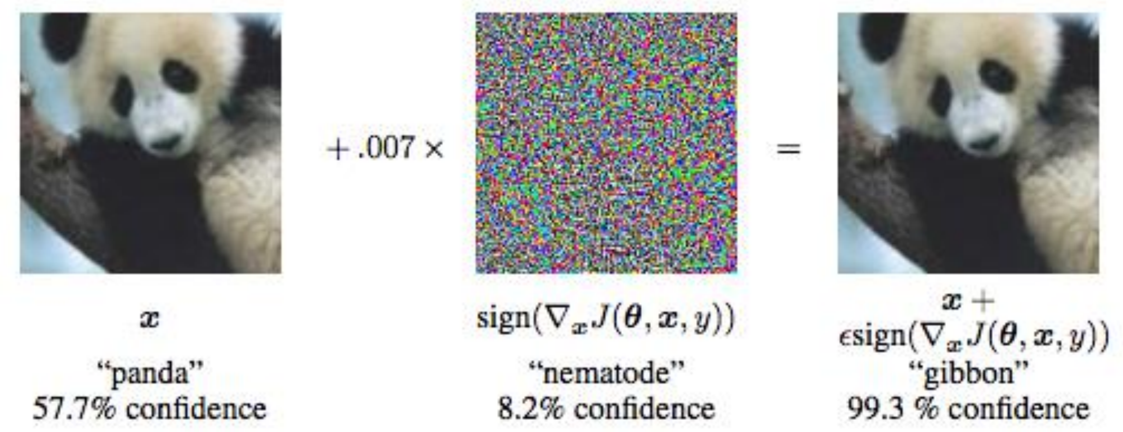
high freq
Sharp edge

★Response frequency

- Frequency of a general Input-Output mapping f.

$$\hat{f}(\mathbf{k}) = \int f(\mathbf{x}) \mathrm{e}^{-\mathrm{i}2\pi\mathbf{k}\cdot\mathbf{x}} \, \mathrm{d}\mathbf{x}$$

MNIST: $\mathbb{R}^{784} \to \mathbb{R}^{10}$, $\mathbf{k} \in \mathbb{R}^{784}$



Goodfellow et al.

high freq： Adversarial example

Examining F-Principle for high-dim real problems

Nonuniform Discrete Fourier transform (NUDFT) for training dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$:

$$\hat{y}_{\mathbf{k}} = \frac{1}{n}\sum_{i=1}^{n} y_i \mathrm{e}^{-\mathrm{i}2\pi\mathbf{k}\cdot\mathbf{x}_i}, \quad \hat{h}_{\mathbf{k}}(t) = \frac{1}{n}\sum_{i=1}^{n} h(\mathbf{x}_i, t)\mathrm{e}^{-\mathrm{i}2\pi\mathbf{k}\cdot\mathbf{x}_i}$$

Difficulty:

- Curse of dimensionality, i.e., $\#\mathbf{k}$ grows exponentially with dimension of problem d.

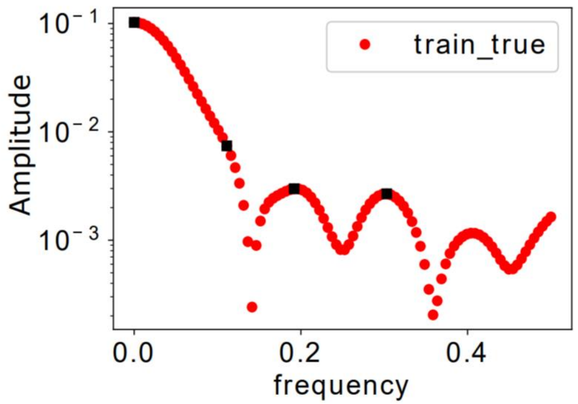
Our approaches:

- Projection, i.e., choose $\mathbf{k} = k\mathbf{p}_1$
- Filtering

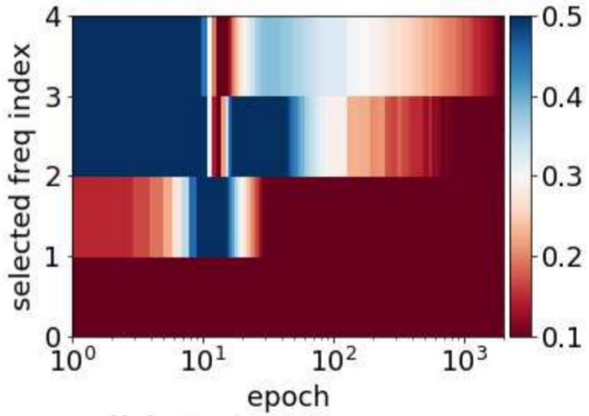
Projection approach

Relative error: $\Delta_F(k) = \left|\hat{h}_k - \hat{y}_k\right| / |\hat{y}_k|$

MNIST

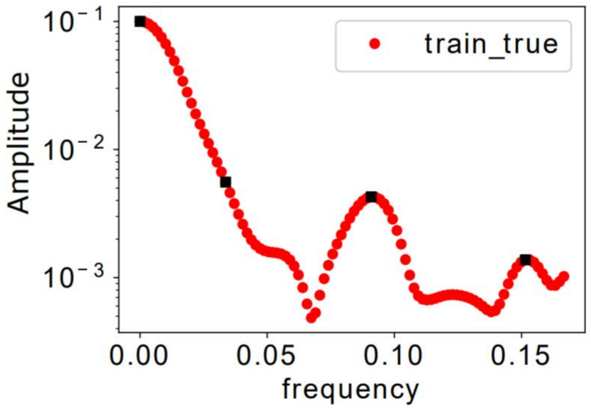


(a) Fourier domain

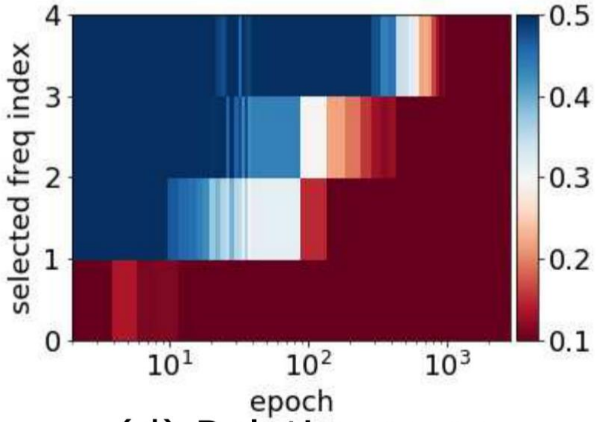


(b) Relative error

CIFAR10

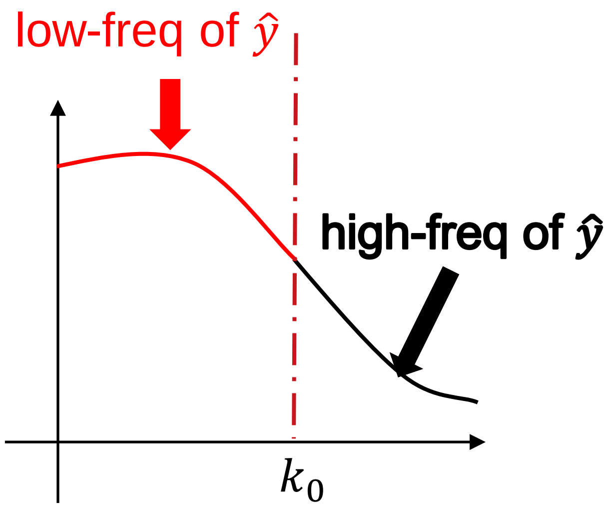


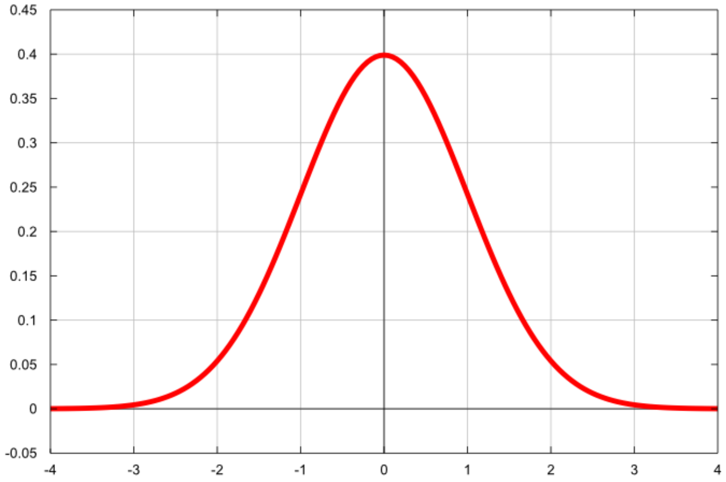
(c) Fourier domain



(d) Relative error

Decompose frequency domain by filtering





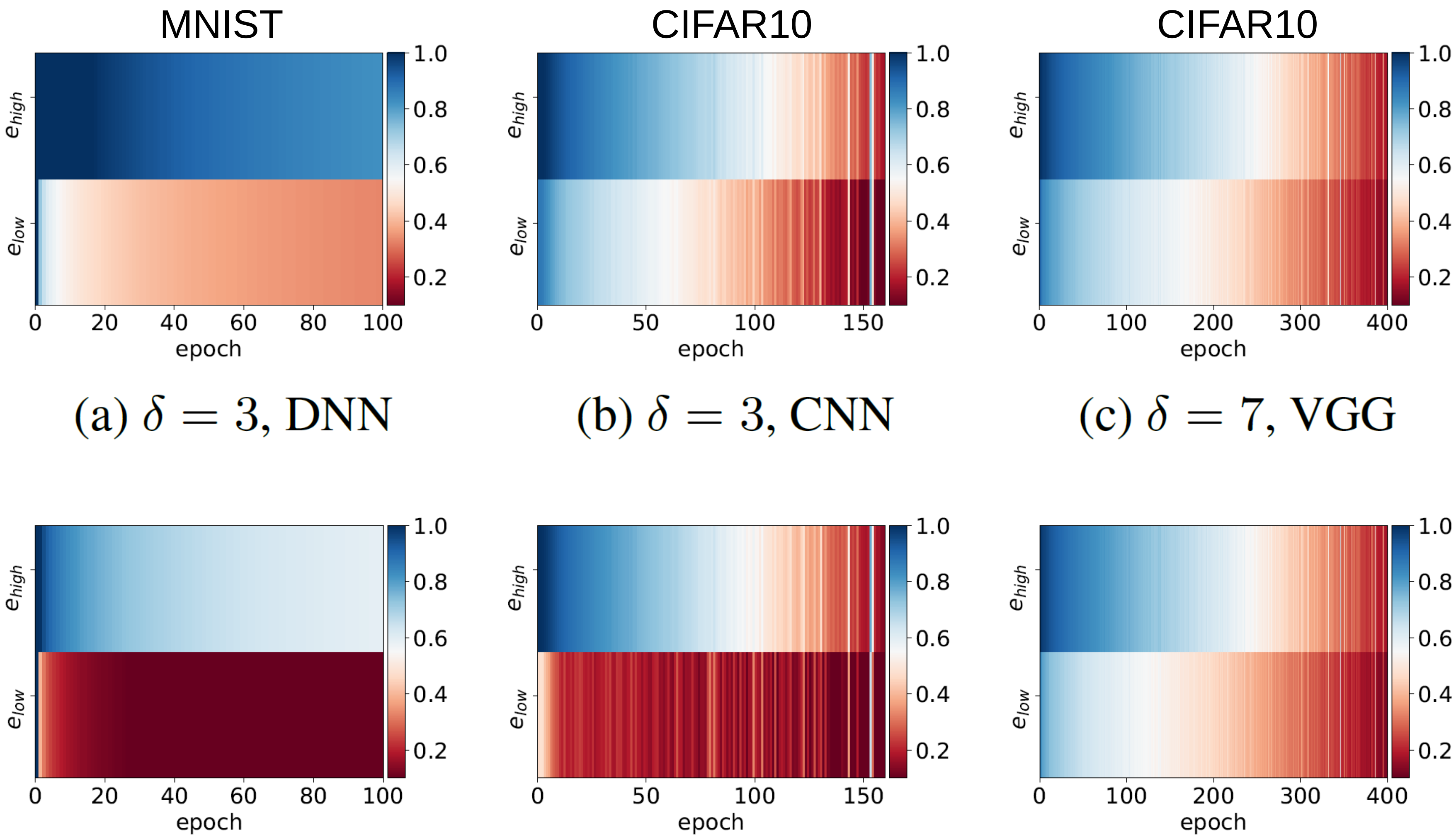
$$\mathbf{y}_i^{\text{low},\delta} = (\mathbf{y} * G^\delta)_i$$

$$\mathbf{y}_i^{\text{high},\delta} \triangleq \mathbf{y}_i - \mathbf{y}_i^{\text{low},\delta}$$

$$e_{\text{low}} = \left(\frac{\sum_i |\mathbf{y}_i^{\text{low},\delta} - \mathbf{h}_i^{\text{low},\delta}|^2}{\sum_i |\mathbf{y}_i^{\text{low},\delta}|^2} \right)^{\frac{1}{2}}$$

$$e_{\text{high}} = \left(\frac{\sum_i |\mathbf{y}_i^{\text{high},\delta} - \mathbf{h}_i^{\text{high},\delta}|^2}{\sum_i |\mathbf{y}_i^{\text{high},\delta}|^2} \right)^{\frac{1}{2}}$$

F-Principle in high-dim space



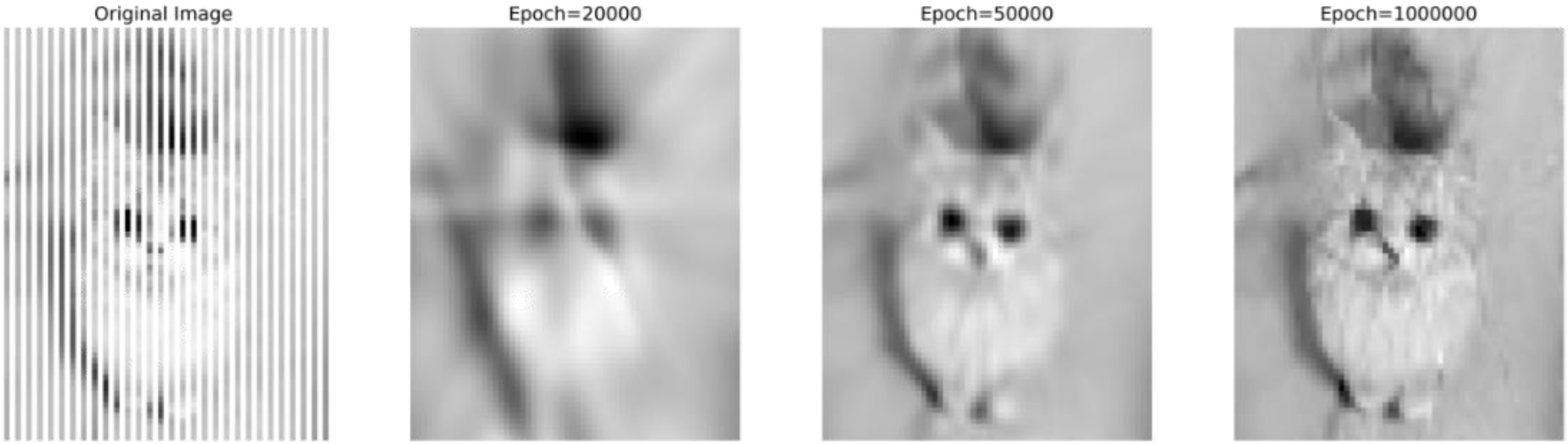
(a) $\delta = 3$, DNN (b) $\delta = 3$, CNN (c) $\delta = 7$, VGG

(d) $\delta = 7$, DNN (e) $\delta = 7$, CNN (f) $\delta = 10$, VGG

Implication of F-Principle

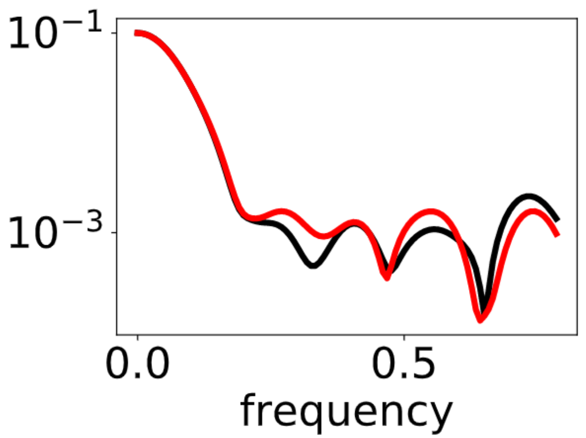
Xu, Zhang, Xiao, *Training behavior of deep neural network in frequency domain, 2018*
Xu, Zhang, Luo, Xiao, Ma, *Frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks, 2019*

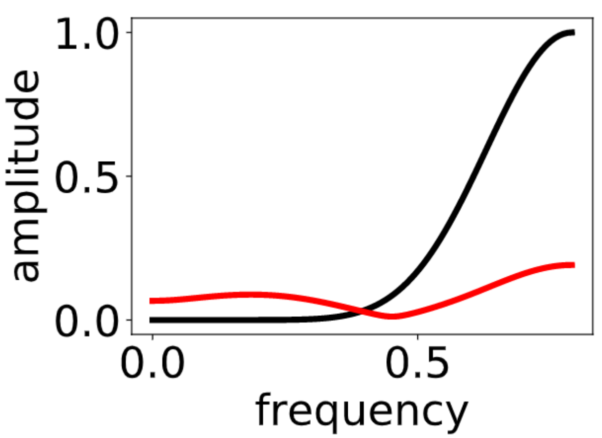
Frequency principle for 2-d image restoration



Generalization puzzle

F-Principe: DNNs prefer low frequencies





CIFAR10

parity

For $\vec{x} \in \{-1,1\}^n$

$f(\vec{x}) = \prod_{j=1}^{n} x_j$,

Even #'-1' → 1;

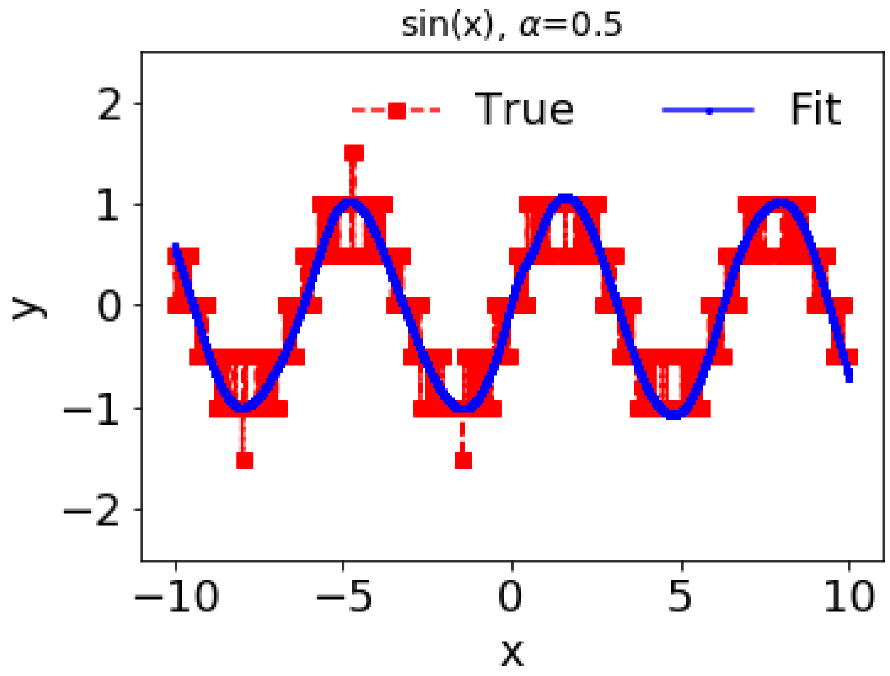
Odd #'-1' → -1.

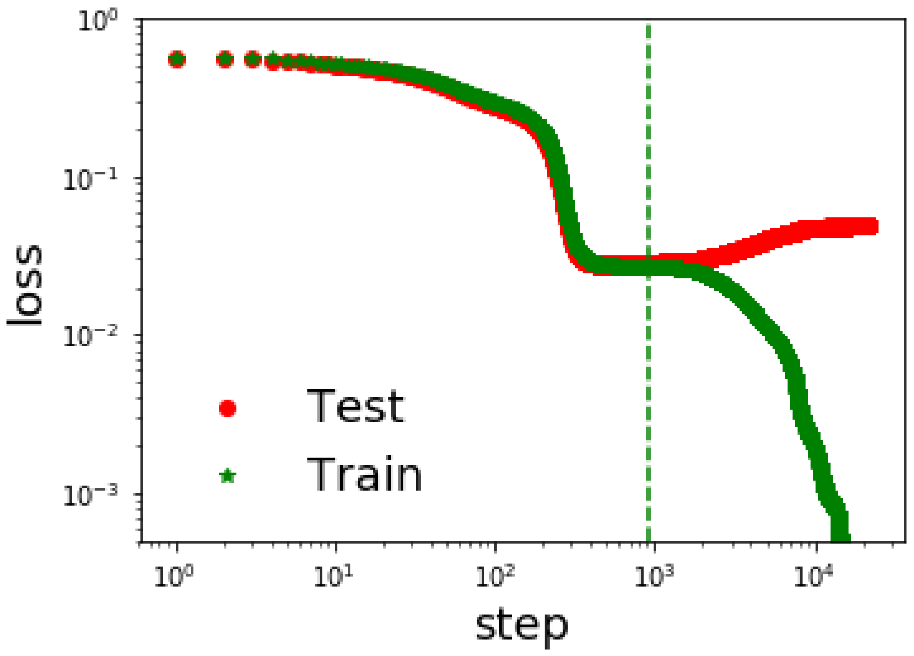
Test accuracy: 72% %>>10%

Test accuracy: ~50%, random guess

Effect of early stopping

When should one stop the backpropagation and use the current parameters?





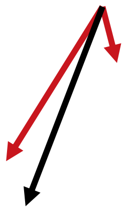
Theory of F-Principle

Theory in a idealized setting

- Consider a tanh-DNN of one-hidden layer for fitting a 1-d function f

$$h(x) = \sum_{j=1}^{m} a_j \sigma\left(w_j x + b_j\right),$$

$$\hat{h}(k) \approx \sum_{j=1}^{m} a_j \exp\left(\frac{i b_j}{w_j}\right) \exp\left(-\left|\frac{\pi k}{2 w_i}\right|\right),$$



- Define the loss at frequency k

$$L(k) = \frac{1}{2}\left|\hat{h}(k) - \hat{f}(k)\right|^2$$

By Parseval's theorem: $L = \int L(k) dk = \int \frac{1}{2} |f(x) - h(x)|^2 dx$

- Compute the gradient by the loss in Fourier domain

$$\theta \leftarrow \theta - \eta \sum \frac{\partial L(k)}{\partial \theta}$$

Analysis

$$\left|\frac{\partial L(k)}{\partial \theta}\right| \approx A(k) \exp\left(-\left|\frac{\pi k}{2w_i}\right|\right) G(\theta, k)$$

Where $A(k) = \left|\hat{h}(k) - \hat{f}(k)\right|$

- $A(k) > 0$:

If w_i is small, $\exp(-|\pi k/2w_j|)$ dominate, low frequencies dominate.
For $w_i \in B_\delta$, center at 0 with radius δ, if δ is small, contribution of high frequency loss is negligible.

- $A(k) \approx 0$:

small contribution from $L(k)$

Insight: smoothness/regularity of activation function $\sigma(\cdot)$ can be converted into F-Principle through gradient-based training.

The NTK regime

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

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

- $\partial_t h(x;\Theta) = -\sum_{i=1}^{n} K_\Theta(x,x_i)(h(x_i;\Theta) - y_i)$

Where $K_\Theta(x,x') = \nabla_\Theta h(x;\Theta) \cdot \nabla_\Theta h(x';\Theta)$

- Neural Tangent Kernel (NTK) regime:

$K_{\Theta(\mathrm{t})}(x,x') \approx K_{\Theta(0)}(x,x')$ for any t.

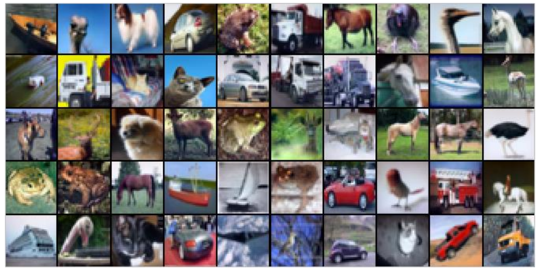
Theorem 1. *For a network of depth L at initialization, with a Lipschitz nonlinearity σ, and in the limit as the layers width $n_1, \ldots, n_{L-1} \to \infty$ sequentially, the NTK $\Theta^{(L)}$ converges in probability to a deterministic limiting kernel:*

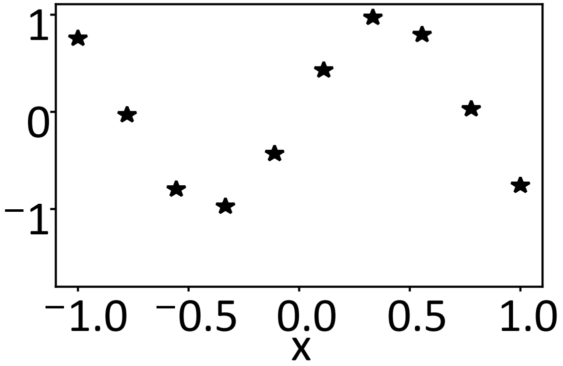
$$\Theta^{(L)} \to \Theta_\infty^{(L)} \otimes Id_{n_L}.$$

Jacot et al., 2018

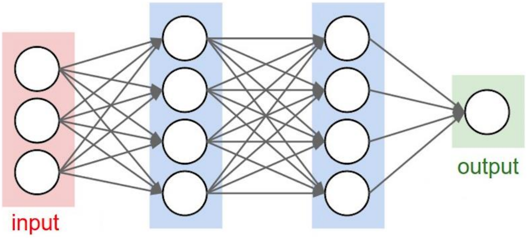
Problem simplification

$\mathcal{D}$





$\mathcal{H}$



$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{W}^{[L]}\sigma\circ(\cdots\boldsymbol{W}^{[2]}\sigma\circ(\boldsymbol{W}^{[1]}\boldsymbol{x}+\boldsymbol{b}^{[1]})+\cdots)+\boldsymbol{b}^{[L]}$$

Two-layer ReLU NN

$$h(x;\Theta) = \sum_{i=1}^{n} w_i\sigma\left(r_i(x+l_i)\right)$$

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

Initialized by special Θ_0

Kernel gradient flow

$$\partial_t f(x,t) = -\Sigma_{i=1}^{n} K_{\Theta_0}(x,x_i)(f(x_i,t)-y_i)$$

Some basics of Fourier transform

Definition 1. *Given a nonzero vector* $\boldsymbol{w} \in \mathbb{R}^d$*, we define the delta-like function* $\delta_{\boldsymbol{w}} : \mathcal{S}(\mathbb{R}^d) \to \mathbb{R}$ *such that for any* $\phi \in \mathcal{S}(\mathbb{R}^d)$,

$$\langle \delta_{\boldsymbol{w}}, \phi \rangle = \int_{\mathbb{R}} \phi(y\boldsymbol{w}) \, \mathrm{d}y. \tag{6}$$

Lemma 1 (Scaling property of delta-like function)**.** *Given any nonzero vector* $\boldsymbol{w} \in \mathbb{R}^d$ *with* $\hat{\boldsymbol{w}} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$*, we have*

$$\frac{1}{\|\boldsymbol{w}\|^d} \delta_{\hat{\boldsymbol{w}}} \left(\frac{\boldsymbol{x}}{\|\boldsymbol{w}\|} \right) = \delta_{\boldsymbol{w}}(\boldsymbol{x}). \tag{7}$$

Proof This is proved by changing of variables. In fact, for any $\phi \in \mathcal{S}(\mathbb{R}^d)$, we have

$$\begin{aligned}
\left\langle \frac{1}{\|\boldsymbol{w}\|^d} \delta_{\hat{\boldsymbol{w}}} \left(\frac{\cdot}{\|\boldsymbol{w}\|} \right), \phi(\cdot) \right\rangle_{\mathcal{S}'(\mathbb{R}^d), \mathcal{S}(\mathbb{R}^d)} &= \langle \delta_{\hat{\boldsymbol{w}}}(\cdot), \phi(\|\boldsymbol{w}\| \cdot) \rangle_{\mathcal{S}'(\mathbb{R}^d), \mathcal{S}(\mathbb{R}^d)} \\
&= \int_{\mathbb{R}} \phi\left(\|\boldsymbol{w}\| y \hat{\boldsymbol{w}}\right) \mathrm{d}y \\
&= \int_{\mathbb{R}} \phi(y\boldsymbol{w}) \, \mathrm{d}y \\
&= \langle \delta_{\boldsymbol{w}}(\cdot), \phi(\cdot) \rangle_{\mathcal{S}'(\mathbb{R}^d), \mathcal{S}(\mathbb{R}^d)} .
\end{aligned}$$

Tao Luo, Zheng Ma, Zhi-Qin John Xu, Yaoyu Zhang, “On the exact computation of linear frequency principle dynamics and its generalization”, SIAM Journal on Mathematics of Data Science 4 (4), 1272-1292, 2022.

Some basics of Fourier transform

Lemma 2 (Fourier transforms of network functions). *For any unit vector* $\boldsymbol{\nu} \in \mathbb{R}^d$, *any nonzero vector* $\boldsymbol{w} \in \mathbb{R}^d$ *with* $\hat{\boldsymbol{w}} = \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$, *and* $g \in \mathcal{S}'(\mathbb{R})$ *with* $\mathcal{F}[g] \in C(\mathbb{R})$, *we have, in the sense of distribution,*

$$(a) \quad \mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[g(\boldsymbol{\nu}^\intercal \boldsymbol{x})](\boldsymbol{\xi}) = \delta_{\boldsymbol{\nu}}(\boldsymbol{\xi})\mathcal{F}[g](\boldsymbol{\xi}^\intercal \boldsymbol{\nu}), \tag{8}$$

$$(b) \quad \mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[g(\boldsymbol{w}^\intercal \boldsymbol{x}+b)](\boldsymbol{\xi}) = \delta_{\boldsymbol{w}}(\boldsymbol{\xi})\mathcal{F}[g]\left(\frac{\boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}}{\|\boldsymbol{w}\|}\right) \mathrm{e}^{2\pi \mathrm{i}\frac{b}{\|\boldsymbol{w}\|}\boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}}, \tag{9}$$

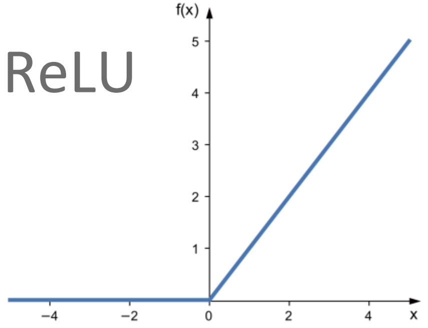
$$(c) \quad \mathcal{F}_{\boldsymbol{x}\to\boldsymbol{\xi}}[\boldsymbol{x}g(\boldsymbol{w}^\intercal \boldsymbol{x}+b)](\boldsymbol{\xi}) = \frac{\mathrm{i}}{2\pi}\nabla_{\boldsymbol{\xi}}\left[\delta_{\boldsymbol{w}}(\boldsymbol{\xi})\mathcal{F}[g]\left(\frac{\boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}}{\|\boldsymbol{w}\|}\right) \mathrm{e}^{2\pi \mathrm{i}\frac{b}{\|\boldsymbol{w}\|}\boldsymbol{\xi}^\intercal \hat{\boldsymbol{w}}}\right]. \tag{10}$$

Tao Luo, Zheng Ma, Zhi-Qin John Xu, Yaoyu Zhang, "On the exact computation of linear frequency principle dynamics and its generalization", SIAM Journal on Mathematics of Data Science 4 (4), 1272-1292, 2022.

Linear F-Principle (LFP) dynamics

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

ReLU



Assumptions:
(i) NTK regime, (ii) sufficiently wide distribution of l_i.

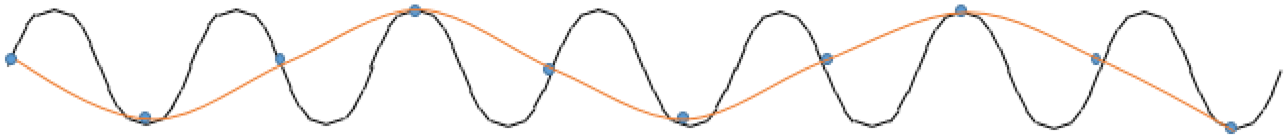
$$\boxed{\partial_t \hat{h}(\xi,t) = -\left[\frac{4\pi^2\langle r^2 w^2\rangle}{\xi^2}+\frac{\langle r^2\rangle+\langle w^2\rangle}{\xi^4}\right]\left(\widehat{h_p}(\xi,t)-\widehat{f_p}(\xi,t)\right)}$$

$\langle\cdot\rangle$: mean over all neurons at initialization

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

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

$\Updownarrow$

low frequency preference

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

s.t. $h(x_i) = y_i$ for $i = 1, \cdots, n$

Case 1: ξ^{-2} dominant

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

Case 2: ξ^{-4} dominant

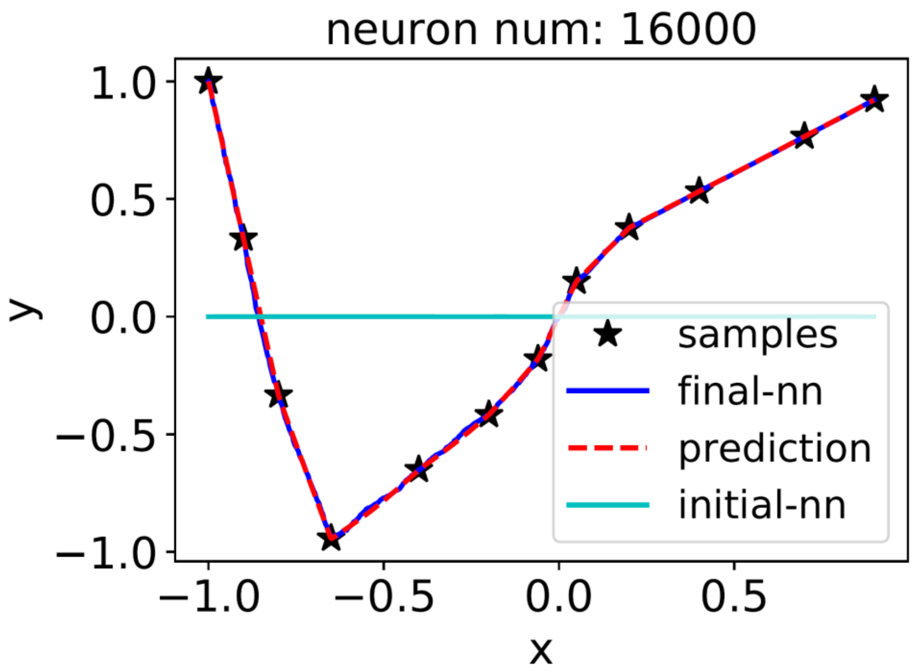
- $\min \int \xi^4 \left|\hat{h}(\xi)\right|^2 \mathrm{d}\xi \sim \min \int |h''(x)|^2 \,\mathrm{d}\xi \rightarrow$ **cubic spline**

Regularity can be changed through initialization

Case 1

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

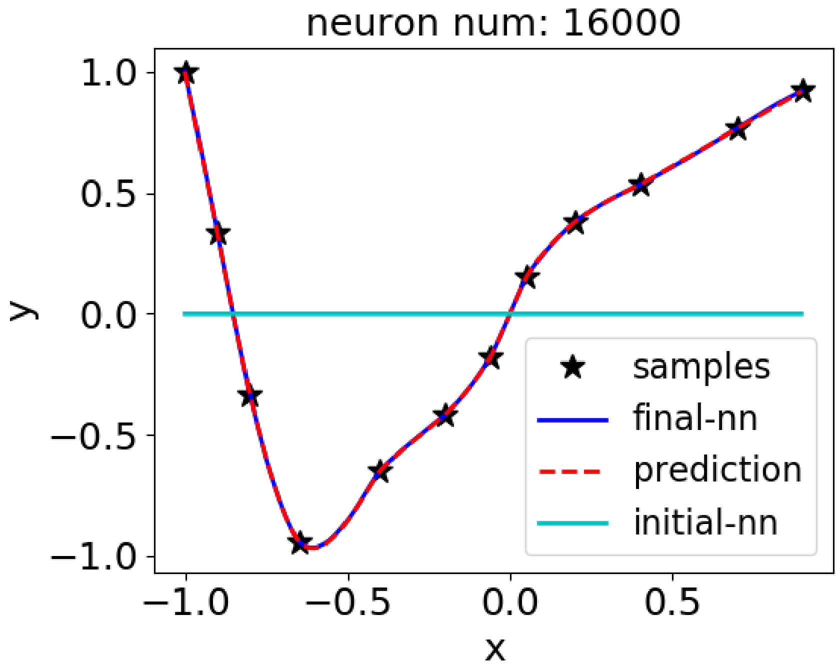
$$\min \int \xi^2 \left|\hat{h}(\xi)\right|^2 \mathrm{d}\xi$$



Case 2

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

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



High-dimensional Case

$$\partial_t \hat{h}(\xi,t) = -\left[\frac{\langle |r|^2\rangle + \langle w^2\rangle}{|\xi|^{d+3}} + \frac{4\pi^2\langle |r|^2 w^2\rangle}{|\xi|^{d+1}}\right]\left(\widehat{h_p}(\xi,t) - \widehat{f_p}(\xi,t)\right)$$

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

$\widehat{(\cdot)}$: Fourier transform; ξ: frequency.

Theorem (informal). Solution of LFP dynamics at $t\to\infty$ with initial value h_{ini} is the same as solution of the following optimization problem

$$\min_{h-h_{\mathrm{ini}}\in F_\gamma}\int\left[\frac{\langle |r|^2\rangle + \langle w^2\rangle}{|\xi|^{d+3}} + \frac{4\pi^2\langle |r|^2 w^2\rangle}{|\xi|^{d+1}}\right]^{-1}\left|\hat{h}(\xi) - \hat{h}_{\mathrm{ini}}(\xi)\right|^2 \mathrm{d}\xi$$

$$\text{s.t. } h(X) = Y.$$

FP-norm and FP-space

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

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

Next, we define the FP-space:

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

A priori generalization error bound

Theorem (informal). Suppose that the real-valued target function $f \in F_\gamma(\Omega)$, h_n is the solution of the regularized model

$$\min_{h\in F_\gamma}\|h\|_\gamma \text{ s.t. } h(X) = Y$$

Then for any $\delta \in (0,1)$ with probability at least $1-\delta$ over the random training samples, the population risk has the bound

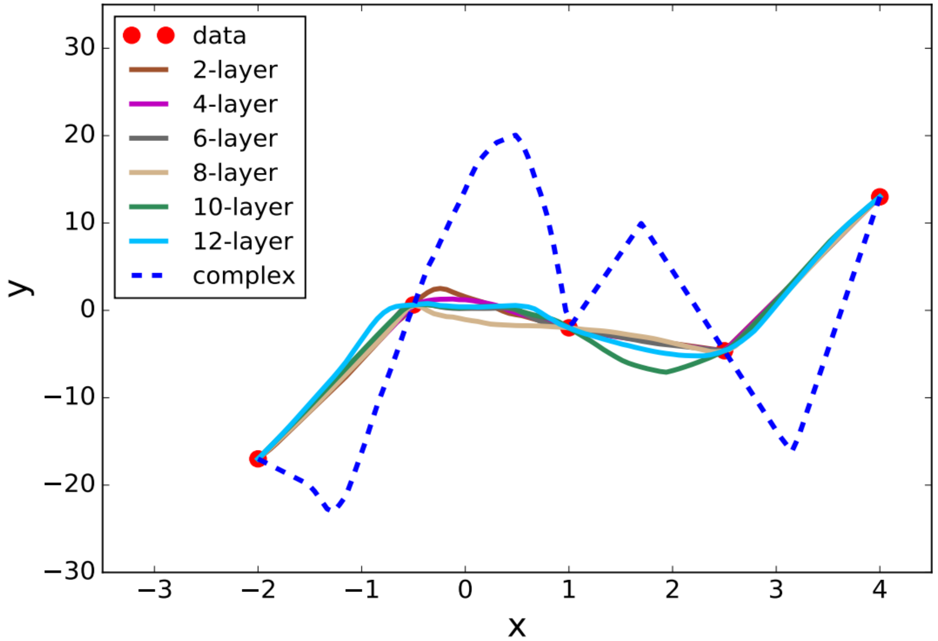
$$L(h_n) \le \left(\|f\|_\infty + 2\|f\|_\gamma\|\gamma\|_{l^2}\right)\left(\frac{2}{\sqrt{n}} + 4\sqrt{\frac{2\log(4/\delta)}{n}}\right)$$

Revisit Leo Breiman's problems (1995)

1. **Why don't heavily parameterized neural networks overfit the data?**
2. What is the effective number of parameters?
3. Why doesn't backpropagation head for a poor local minima?
4. **When should one stop the backpropagation and use the current parameters?**

Leo Breiman, Reflections After Refereeing Papers for NIPS

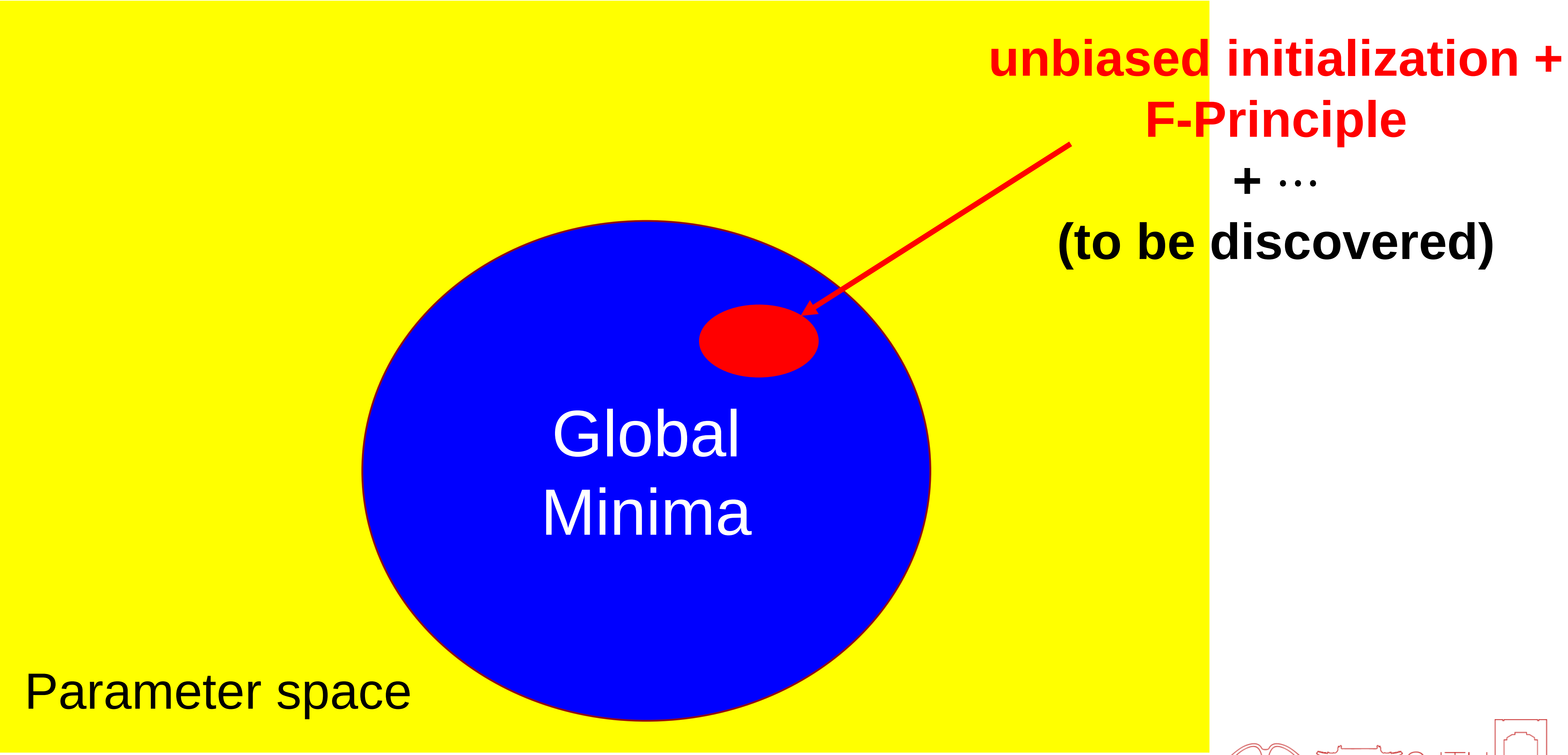
Overparameterized DNNs still generalize well



Lei Wu, Zhanxing Zhu, Weinan E, 2017

#para(~1000)>>#data: 5

A picture for the generalization puzzle



References

1. **First Paper**: Zhiqin Xu, Yaoyu Zhang, Yanyang Xiao, "Training Behavior of Deep Neural Network in Frequency Domain," ICONIP, pp. 264-274, 2019. (arXiv:1807.01251, Jul 2018)
2. **2021 World Artificial Intelligence Conference Youth Outstanding Paper Nomination Award**: Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, Zheng Ma, "Frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks," CiCP 28(5). 1746-1767, 2020.
3. **Initialization effect**: Yaoyu Zhang, Zhi-Qin John Xu, Tao Luo, Zheng Ma, "A Type of Generalization Error Induced by Initialization in Deep Neural Networks," MSML 2020.
4. **Linear Frequency Principle**: Yaoyu Zhang, Tao Luo, Zheng Ma, Zhi-Qin John Xu, "Linear Frequency Principle Model to Understand the Absence of Overfitting in Neural Networks," Chinese Physics Letters (CPL) 38(3), 038701, 2021.
5. Tao Luo, Zheng Ma, Zhi-Qin John Xu, Yaoyu Zhang, "Theory of the Frequency Principle for General Deep Neural Networks," CSIAM Trans. Appl. Math. 2 (2021), pp. 484-507.
6. **Linear Frequency Principle**: Tao Luo, Zheng Ma, Zhi-Qin John Xu, Yaoyu Zhang, "On the exact computation of linear frequency principle dynamics and its generalization", SIAM Journal on Mathematics of Data Science 4 (4), 1272-1292, 2022.
7. **Minimal decay in frequency domain**: Tao Luo, Zheng Ma, Zhiwei Wang, Zhi-Qin John Xu, Yaoyu Zhang, "An Upper Limit of Decaying Rate with Respect to Frequency in Deep Neural Network," MSML 2022.
8. **Overview**: Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, "Overview Frequency Principle/Spectral Bias in Deep Learning," Communications on Applied Mathematics and Computation (2024): 1-38.

See more works on my personal website: https://yaoyuzhang1.github.io/

DNNs are not black boxes--Problems

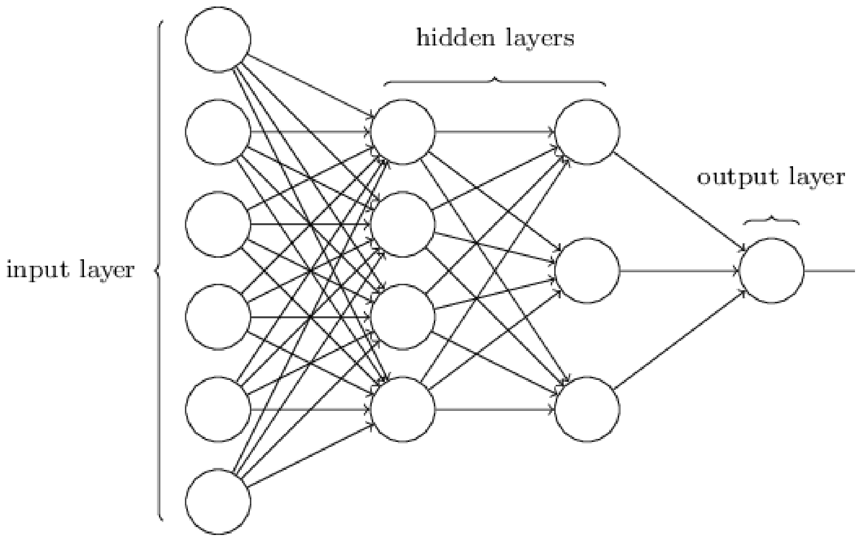
- How can neural networks be prevented from overfitting to noise in the training data?
- How can the convergence towards high-frequency components of the target function be improved during training?
- Is it possible to design a model that prioritizes fitting high-frequency components before low-frequency ones during training?
- What characteristics of a target function or dataset make it difficult for deep neural networks to generalize?
- What methods or conditions can be employed to induce deep neural networks to overfit the training data, resulting in a large generalization error?

Thanks！

饮 水 思 源　爱 国 荣 校

Impact of initialization

Deep Neural Network



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

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

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

Example: Two-layer NN

$$h_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{m_1} w_i^{[2]}\sigma(w_i^{[1]}x + b_i^{[1]})$$

Initialization

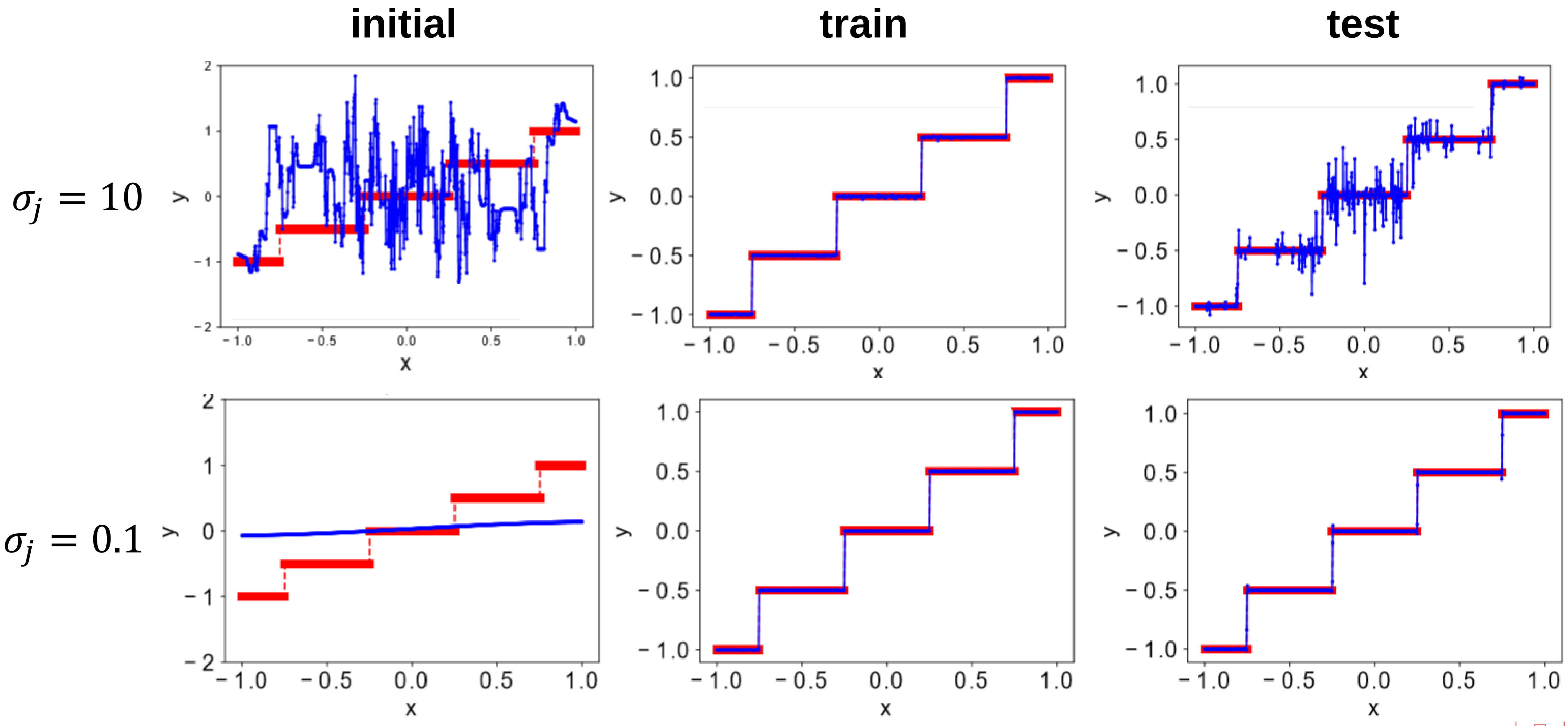
$$\boldsymbol{\theta}(0): \left[W^{[j]}(0), b^{[j]}(0)\right]_{j=1,\cdots,H}$$

$$W^{[j]}(0), b^{[j]}(0) \sim \mathcal{N}(0, \sigma_j^2)$$

$$\left[\sigma_j\right]_{j=1,\cdots,H} \overset{?}{\rightarrow} h_{\boldsymbol{\theta}(\infty)}(x)$$

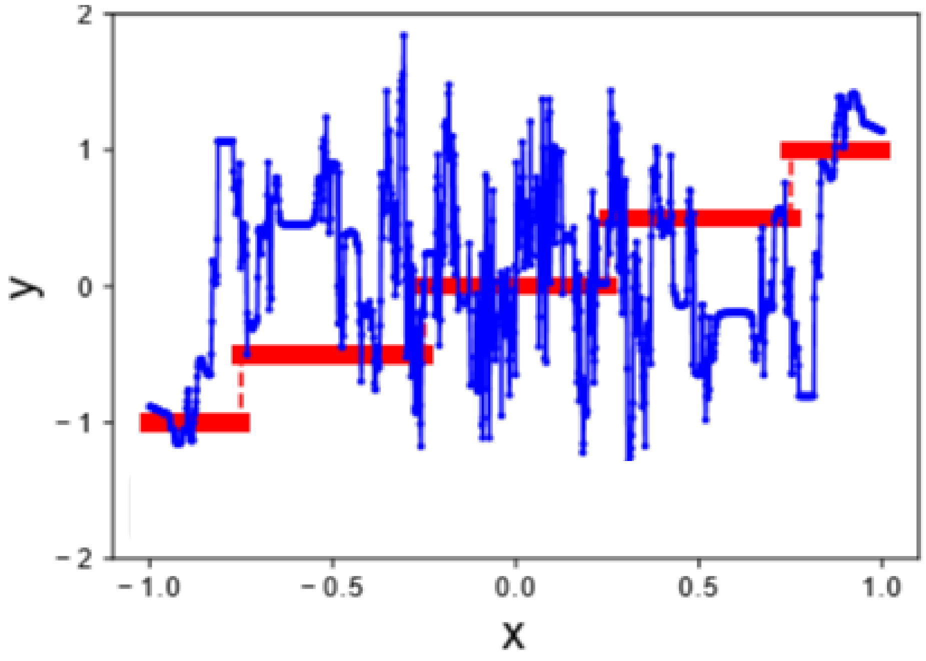
initialization generalization

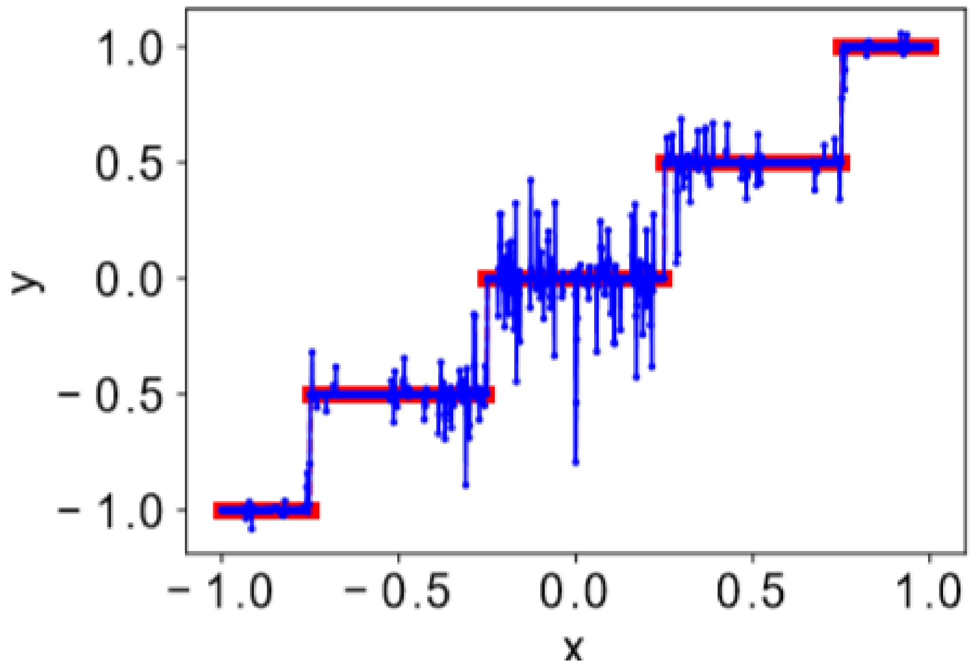
DNN can easily overfit with bad initialization



ZQJ Xu, Y Zhang, Y Xiao, *Training behavior of deep neural network in frequency domain*, ICONIP 2019.

Impact of $f(\cdot, \Theta(0))$





Setup in the linear (NTK) regime

NNs can be linearized around initialization

$$h(\boldsymbol{x}, \boldsymbol{\theta}) \approx h\left(\boldsymbol{x}, \boldsymbol{\theta}_0\right) + \nabla_{\boldsymbol{\theta}} h\left(\boldsymbol{x}, \boldsymbol{\theta}_0\right) \cdot \left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right)$$

Loss

$$R_S(\boldsymbol{\theta}) = \operatorname{dist}\left(\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{\theta}), \boldsymbol{Y}\right)$$

Neural tangent kernel (NTK)

$$k(\cdot, \cdot) = \nabla_{\boldsymbol{\theta}} h(\cdot, \boldsymbol{\theta}_0)^{\top} \nabla_{\boldsymbol{\theta}} h(\cdot, \boldsymbol{\theta}_0)$$

Kernel gradient flow

$$\partial_t h(\boldsymbol{x}, t) = -\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{X}) \nabla_{\boldsymbol{h}(\boldsymbol{X}, t)} \operatorname{dist}\left(\boldsymbol{h}(\boldsymbol{X}, t), \boldsymbol{Y}\right)$$

Equivalent optimization problems

Theorem 5 *Let $\boldsymbol{\theta}(t)$ be the solution of gradient flow dynamics*

$$\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\theta}(t) = -\nabla_{\boldsymbol{\theta}} \boldsymbol{h}(\boldsymbol{X}, \boldsymbol{\theta}_0) \nabla_{\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{\theta}(t))} \mathrm{dist}\left(\boldsymbol{h}(\boldsymbol{X}, \boldsymbol{\theta}(t)), \boldsymbol{Y}\right) \tag{8}$$

with initial value $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$, where $\nabla_{\boldsymbol{\theta}} \boldsymbol{h}(\boldsymbol{X}, \boldsymbol{\theta}_0)$ is a full rank (rank n) matrix of size $m \times n$ with $m > n$. Suppose that the limit $\boldsymbol{\theta}(\infty) = \lim_{t\to\infty} \boldsymbol{\theta}(t)$ exists. Then $\boldsymbol{\theta}(\infty)$ solves the constrained optimization problem

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2, \; s.t., \; \boldsymbol{h}(\boldsymbol{X}, \boldsymbol{\theta}) = \boldsymbol{Y}. \tag{9}$$

Theorem 6 *Let $\boldsymbol{\theta}$ be the solution of problem (9), then $h(\boldsymbol{x}, \boldsymbol{\theta})$ uniquely solves the optimization problem*

$$\min_{h - h_{\mathrm{ini}} \in H_k(\Omega)} \|h - h_{\mathrm{ini}}\|_k \quad \text{s.t.} \quad \boldsymbol{h}(\boldsymbol{X}) = \boldsymbol{Y}, \tag{12}$$

In a general class, e.g., any L^p distance, choice of loss has no impact.

In practice, different loss can be considered to accelerate convergence.

Main results

Theorem 2 *For a fixed kernel function $k \in C(\Omega \times \Omega)$, and training set $\{\boldsymbol{X}; \boldsymbol{Y}\}$, for any initial function $h_{\text{ini}} \in C(\Omega)$, $h_k(\cdot; h_{\text{ini}}, \boldsymbol{X}, \boldsymbol{Y})$ can be decomposed as*

$$h_k(\cdot; h_{\text{ini}}, \boldsymbol{X}, \boldsymbol{Y}) = h_k(\cdot; 0, \boldsymbol{X}, \boldsymbol{Y}) + \boxed{h_{\text{ini}} - h_k(\cdot; 0, \boldsymbol{X}, h_{\text{ini}}(\boldsymbol{X}))}. \tag{6}$$

unbiased fit   bias from h_{ini}

Theorem 3 *For a target function $f \in C(\Omega)$, if h_{ini} is generated from an unbiased distribution of random functions μ such that $\mathbb{E}_{h_{\text{ini}} \sim \mu} h_{\text{ini}} = 0$, then the generalization error of $h_k(\cdot; h_{\text{ini}}, \boldsymbol{X}, f(\boldsymbol{X}))$ can be decomposed as follows*

$$\begin{aligned}\mathbb{E}_{h_{\text{ini}} \sim \mu} R_S\left(h_k(\cdot; h_{\text{ini}}, \boldsymbol{X}, f(\boldsymbol{X})), f\right) = {} & R_S\left(h_k(\cdot; 0, \boldsymbol{X}, f(\boldsymbol{X})), f\right) \\ & + \boxed{\mathbb{E}_{h_{\text{ini}} \sim \mu} R_S\left(h_k(\cdot; 0, \boldsymbol{X}, h_{\text{ini}}(\boldsymbol{X})), h_{\text{ini}}\right)},\end{aligned}$$

where $R_S(h_k(\cdot; h_{\text{ini}}, \boldsymbol{X}, f(\boldsymbol{X})), f) = \|h_k(\cdot; h_{\text{ini}}, \boldsymbol{X}, f(\boldsymbol{X})) - f\|^2_{L^2(\Omega)}$.

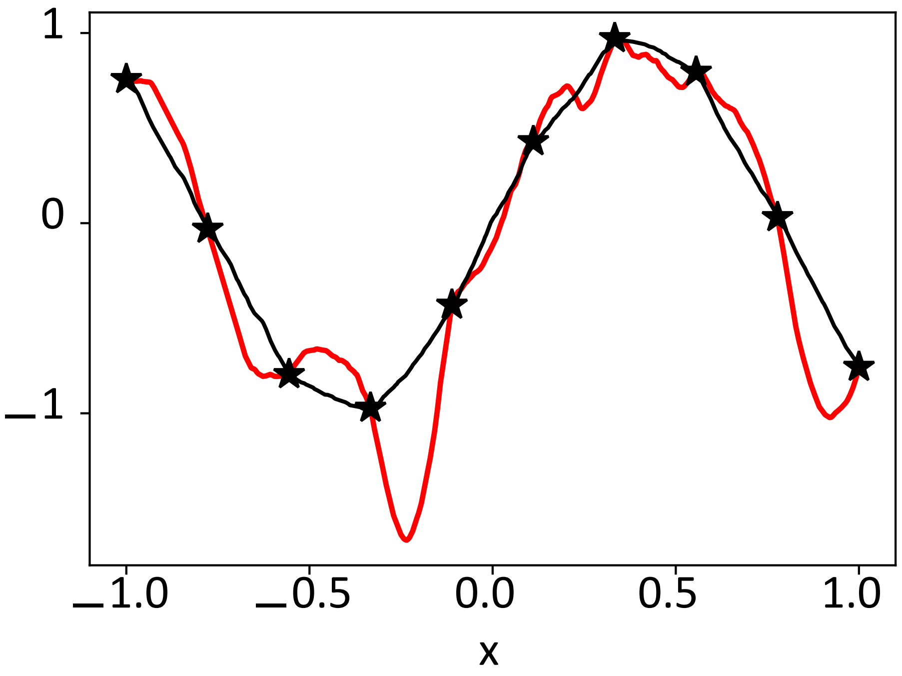
additional
generalization error

Illustration

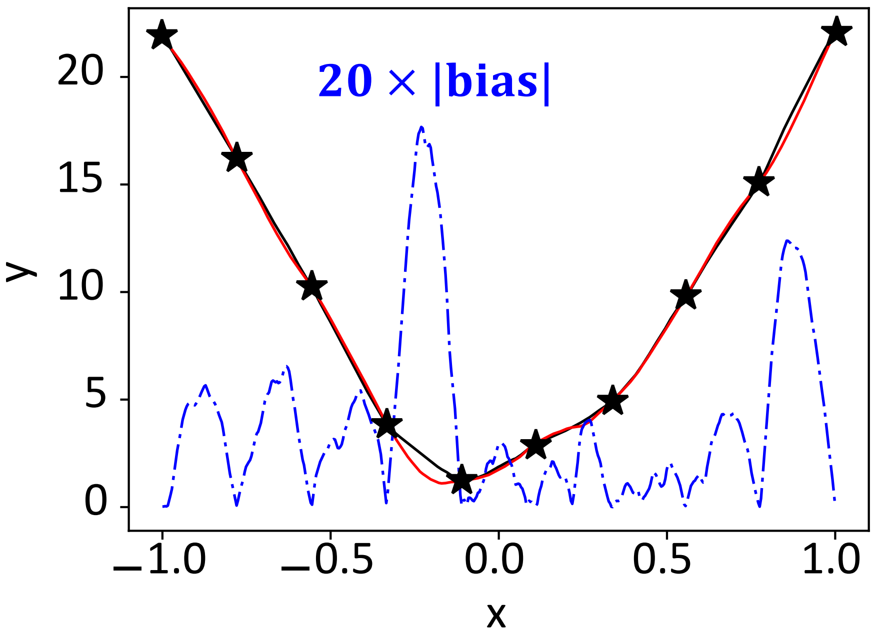
$$h_K(\cdot; h_{\text{ini}}, X, Y) - h_K(\cdot; 0, X, Y) = \boxed{h_{\text{ini}} - h_K(\cdot; 0, X, h_{\text{ini}}(X))}$$

bias: high freq of h_{ini}

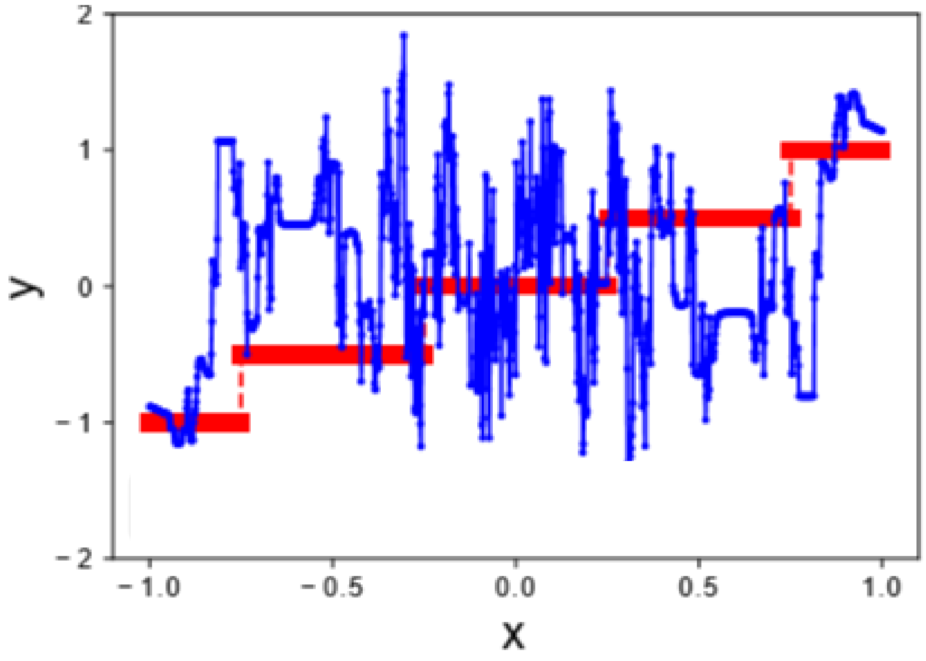
$h_K(\cdot; h_{\text{ini}}, X, Y)$-$h_K(\cdot; 0, X, Y)$

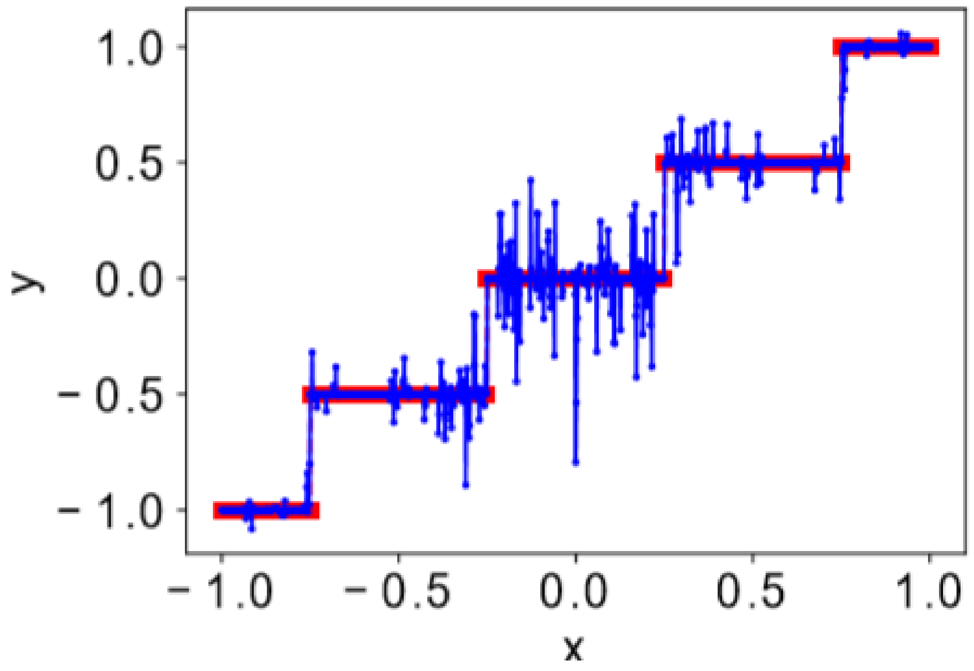


$h_{\text{ini}} - h_K(\cdot; 0, X, h_{\text{ini}}(X))$

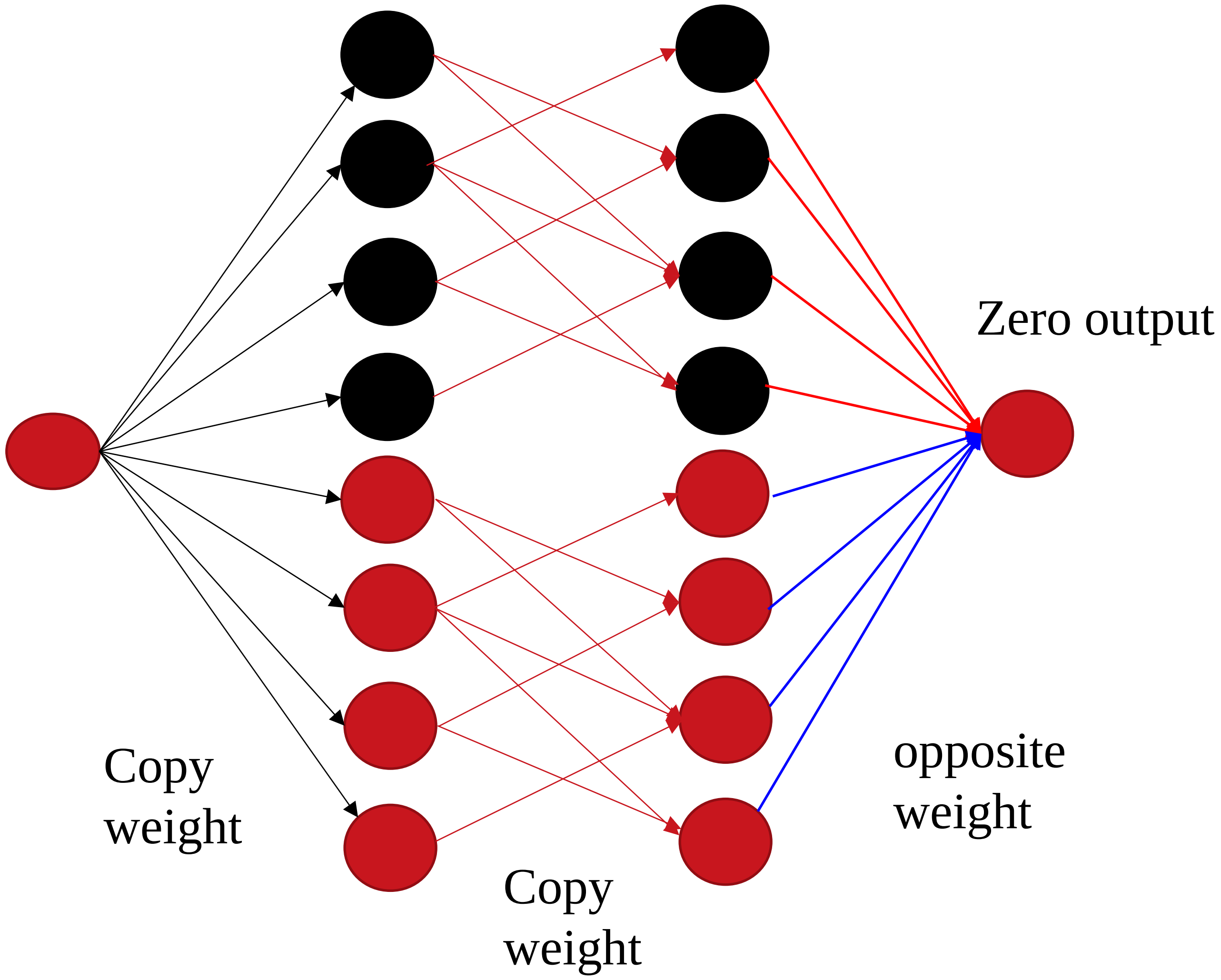


Revisit previous experiments





AntiSymmetrical Initialization (ASI) trick



Original DNN

$$h(x, \theta) \text{ with } \theta(0) = \theta_0$$

After ASI

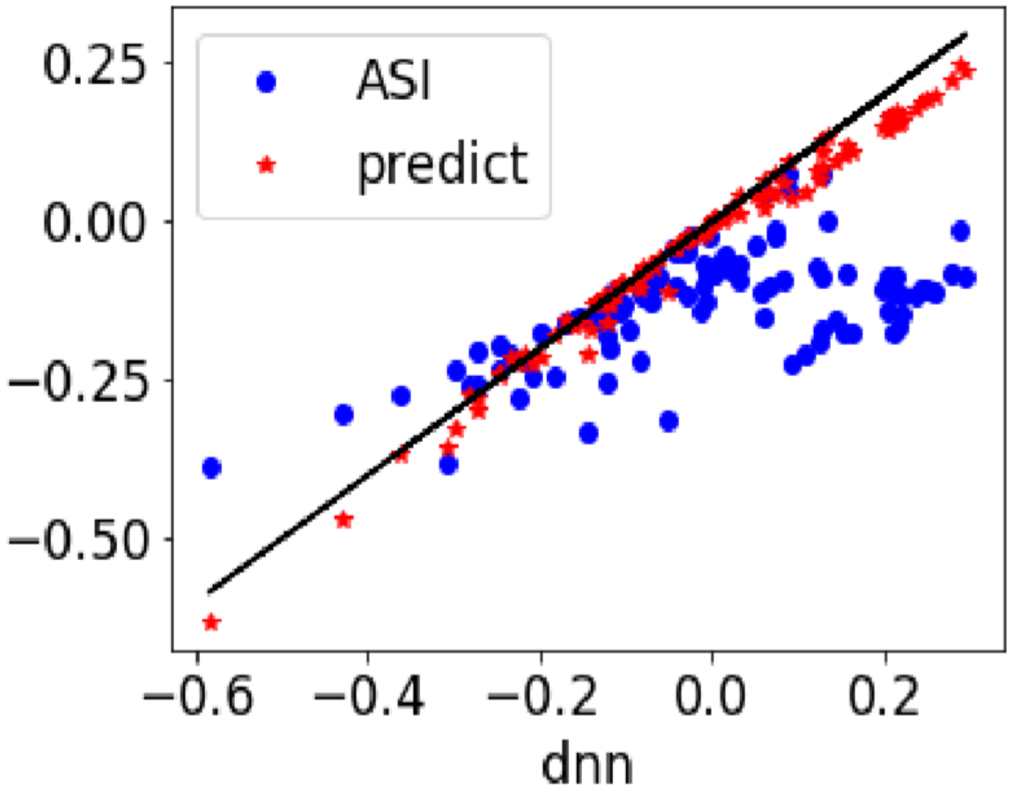
$$h_{\mathrm{ASI}} = \frac{\sqrt{2}}{2} h(x, \theta) - \frac{\sqrt{2}}{2} h(x, \theta')$$

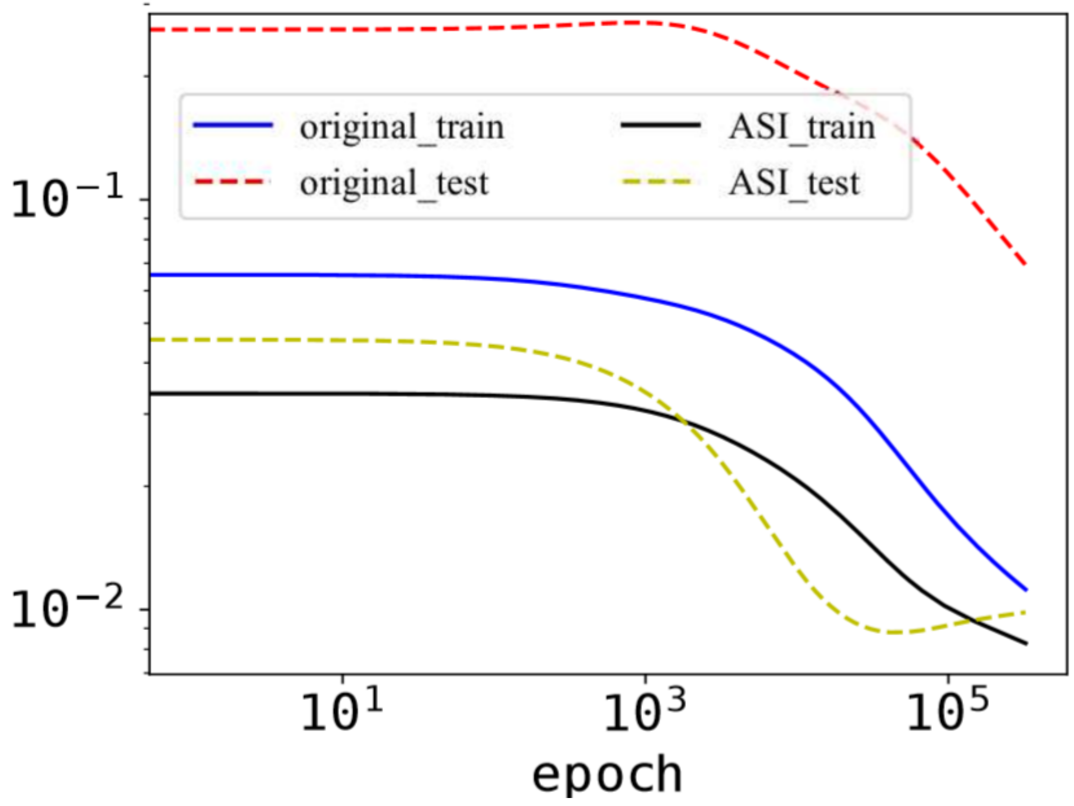
$$\theta'(0) = \theta(0) = \theta_0$$

Properties

1. $h_{\mathrm{ASI}} = 0$ at initialization
2. $k_{\mathrm{ASI}}(\cdot,\cdot) = k(\cdot,\cdot)$

Experiments--Boston house price dataset





Experiments—MNIST dataset