# Learning with norm-based neural networks: model capacity, function spaces, and computational-statistical gaps


## Fanghui LIU

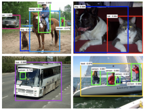
fanghui.liu@warwick.ac.uk

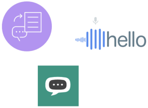
*Department of Computer Science, University of Warwick, UK*
*Centre for Discrete Mathematics and its Applications (DIMAP), Warwick*

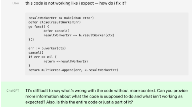
[joint work with Leello Dadi, Zhenyu Zhu, Volkan Cevher (EPFL)]

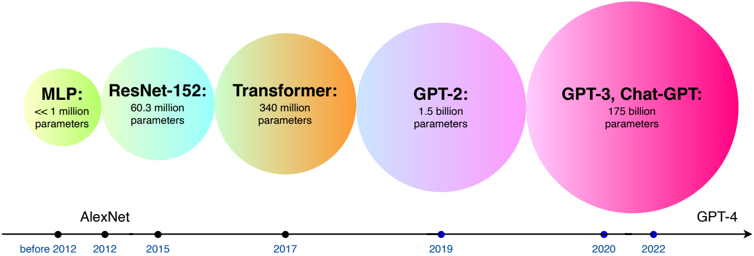

at INRIA, Paris, 2024

# In the era of deep learning





# Scaling law: under compute budget

**scaling law [13]**

$$\text{test loss} = A \quad \text{Model Size}^{a} + B \quad \text{Data Size}^{b} + C$$

under limited compute budget

- data-parameter trade-off
- time-space trade-off

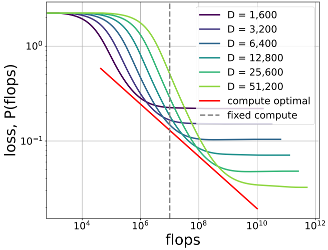
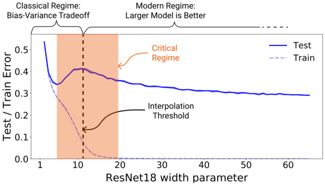


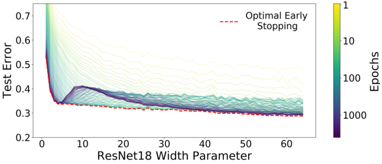

Figure 1: Scaling law under compute-optimal configuration [21].

# Model size is a "right" complexity?

- double descent [4] (Belkin, Hsu, Ma, Mandal, 2019)




(a) Results on ResNet18 [18]

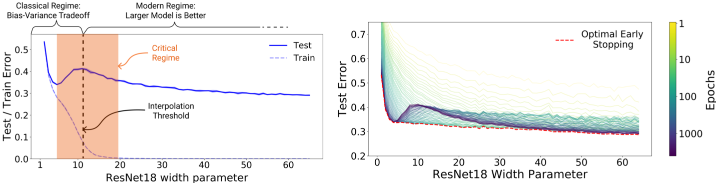



(b) Optimal early stopping [18].

- Empirically: neural network pruning [16], lottery ticket hypothesis [11], fine-tuning with large dropout [28]
- Theoretically: how much over-parameterization is sufficient? [7, 26]

# Model size is a "right" complexity?

- double descent [4] (Belkin, Hsu, Ma, Mandal, 2019)




(a) Results on ResNet18 [18] (b) Optimal early stopping [18].

- Empirically: neural network pruning [16], lottery ticket hypothesis [11], fine-tuning with large dropout [28]
- Theoretically: how much over-parameterization is sufficient? [7, 26]

# What is the "right" model complexity?

Complexity of a prediction rule, e.g.,

- number of parameters
- norm of parameters

[2] (Bartlett, 1998)

The size of the weights is more important than the size of the network!

Norm-based capacity:[19, 24, 20, 8]

| name | definition | rank correlation |
|---|---|---|
| Parameter Frobenius norm | $\sum_{i=1}^{L} \|W_i\|_F^2$ | 0.073 |
| Frobenius distance to initialization [17] | $\sum_{i=1}^{L} \|W_i - W_i^0\|_F^2$ | 0.263 |
| Spectral complexity [3] | $\prod_{i=1}^{L} \|W_i\| \left( \sum_{i=1}^{L} \frac{\|W_i\|_{2,1}^{3/2}}{\|W_i\|^{3/2}} \right)^{2/3}$ | 0.537 |
| Fisher-Rao [14] | $\frac{(L+1)^2}{n} \sum_{i=1}^{n} \langle W, \nabla_W \ell(h_W(x_i), y_i) \rangle$ | 0.078 |
| Path-norm [19] | $\sum_{(i_0,\dots,i_L)} \left( \prod_{j=1}^{L} W_{i_j, i_{j-1}} \right)^2$ | 0.373 |

Table 1: Complexity measures compared in the empirical study [12], and their correlation with generalization

# What is the “right” model complexity?

Complexity of a prediction rule, e.g.,

- number of parameters
- norm of parameters

**[2] (Bartlett, 1998)**

The size of the weights is more important than the size of the network!

Norm-based capacity:[19, 24, 20, 8]

| name | definition | rank correlation |
|---|---|---|
| Parameter Frobenius norm | $\sum_{i=1}^{L} \|W_i\|_F^2$ | 0.073 |
| Frobenius distance to initialization [17] | $\sum_{i=1}^{L} \|W_i - W_i^0\|_F^2$ | 0.263 |
| Spectral complexity [3] | $\prod_{i=1}^{L} \|W_i\| \left( \sum_{i=1}^{L} \frac{\|W_i\|_{2,1}^{3/2}}{\|W_i\|^{3/2}} \right)^{2/3}$ | 0.537 |
| Fisher-Rao [14] | $\frac{(L+1)^2}{n} \sum_{i=1}^{n} \langle W, \nabla_W \ell(h_W(x_i), y_i) \rangle$ | 0.078 |
| Path-norm [19] | $\sum_{(i_0,\dots,i_L)} \left( \prod_{j=1}^{L} W_{i_{j-1},i_j} \right)^2$ | 0.373 |

Table 1: Complexity measures compared in the empirical study [12], and their correlation with generalization

# What is the “right” model complexity?

Complexity of a prediction rule, e.g.,

- number of parameters
- norm of parameters

**[2] (Bartlett, 1998)**

The size of the weights is more important than the size of the network!

Norm-based capacity:[19, 24, 20, 8]

| name | definition | rank correlation |
|---|---|---|
| Parameter Frobenius norm | $\sum_{i=1}^{L} \|W_i\|_F^2$ | 0.073 |
| Frobenius distance to initialization [17] | $\sum_{i=1}^{L} \|W_i - W_i^0\|_F^2$ | 0.263 |
| Spectral complexity [3] | $\prod_{i=1}^{L} \|W_i\| \left( \sum_{i=1}^{L} \frac{\|W_i\|_{2,1}^{3/2}}{\|W_i\|^{3/2}} \right)^{2/3}$ | 0.537 |
| Fisher-Rao [14] | $\frac{(L+1)^2}{n} \sum_{i=1}^{n} \langle W, \nabla_W \ell(h_W(x_i), y_i) \rangle$ | 0.078 |
| Path-norm [19] | $\sum_{(i_0,\dots,i_L)} \prod_{j=1}^{L} \left( W_{i_j, i_{j-1}} \right)^2$ | 0.373 |

Table 1: Complexity measures compared in the empirical study [12], and their correlation with generalization

# What is the "right" model complexity?

Complexity of a prediction rule, e.g.,

- number of parameters
- norm of parameters

**[2] (Bartlett, 1998)**

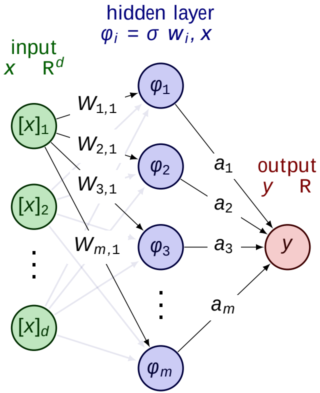
The size of the weights is more important than the size of the network!

Norm-based capacity:[19, 24, 20, 8]

| name | definition | rank correlation |
|---|---|---|
| Parameter Frobenius norm | $\sum_{i=1}^{L} \|W_i\|_F^2$ | 0.073 |
| Frobenius distance to initialization [17] | $\sum_{i=1}^{L} \|W_i - W_i^0\|_F^2$ | 0.263 |
| Spectral complexity [3] | $\prod_{i=1}^{L} \|W_i\| \left( \sum_{i=1}^{L} \frac{\|W_i\|_{2,1}^{3/2}}{\|W_i\|^{3/2}} \right)^{2/3}$ | 0.537 |
| Fisher-Rao [14] | $\frac{(L+1)^2}{n} \sum_{i=1}^{n} \langle W, \nabla_W \ell(h_W(x_i), y_i) \rangle$ | 0.078 |
| Path-norm [19] | $\sum_{(i_0,\dots,i_L)} \prod_{j=1}^{L} \left( W_{i_j, i_{j-1}} \right)^2$ | 0.373 |

Table 1: Complexity measures compared in the empirical study [12], and their correlation with generalization

# Two-layer neural networks, path norm



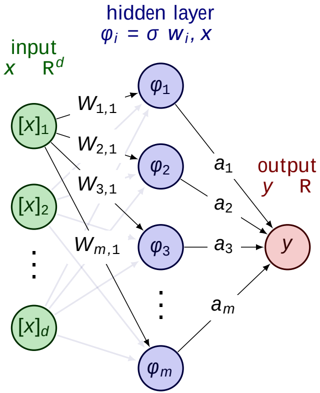

$\mathcal{P}_m = \{ f(x) := \frac{1}{m} \sum_{k=1}^{m} a_k \, \sigma(\langle w_k, x\rangle) \}$

**$\ell_1$-path norm**

$\|f\|_P := \frac{1}{m} \sum_{k=1}^{m} |a_k| \|w_k\|_1$

- semi-norm
- representation cost
- relations to Barron spaces $\mathcal{B}$ [1, 10]
- $\|f\|_{\mathcal{B}} \le \|f\|_P \le 2\|f\|_{\mathcal{B}}$

# Two-layer neural networks, path norm



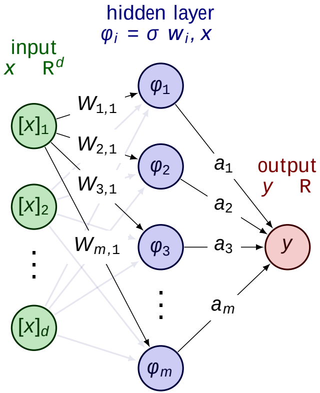

$$\mathcal{P}_m = \left\{ f(\cdot) := \frac{1}{m}\sum_{k=1}^{m} a_k \sigma\langle w_k, \cdot\rangle \right\}$$

**$\ell_1$-path norm**

$$\|\theta\|_P := \frac{1}{m}\sum_{k=1}^{m} |a_k| \|w_k\|_1$$

- semi-norm
- representation cost
- relations to Barron spaces $\mathcal{B}$ [1, 10]
- $\|f\|_{\mathcal{B}} \le \|\theta\|_P \le 2\|f\|_{\mathcal{B}}$

# Two-layer neural networks, path norm



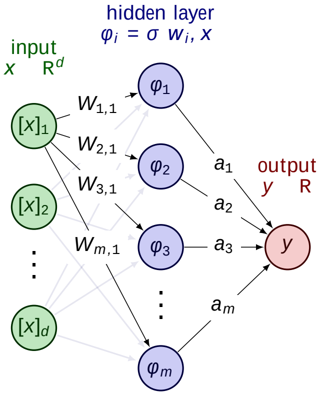

$$\mathcal{P}_m = \left\{ f(\cdot) := \frac{1}{m}\sum_{k=1}^{m} a_k\, \sigma(\langle w_k, \cdot\rangle) \right\}$$

**$\ell_1$-path norm**

$$\|\cdot\|_{\mathcal{P}} := \frac{1}{m}\sum_{k=1}^{m} |a_k|\,\|w_k\|_1$$

- semi-norm
- representation cost
- relations to Barron spaces $\mathcal{B}$ [1, 10]
- $\|f\|_{\mathcal{B}} \leq \|\cdot\|_{\mathcal{P}} \leq 2\|f\|_{\mathcal{B}}$

# Two-layer neural networks, path norm




$$\mathcal{P}_m = \left\{ f_\theta(\cdot) := \frac{1}{m}\sum_{k=1}^{m} a_k \, \sigma(\langle w_k, \cdot\rangle) \right\}$$

**$\ell_1$-path norm**

$$\|\theta\|_{\mathcal{P}} := \frac{1}{m}\sum_{k=1}^{m} |a_k| \|w_k\|_1$$

- semi-norm
- representation cost
- relations to Barron spaces $\mathcal{B}$ [1, 10]
- $\|f\|_{\mathcal{B}} \leq \|\theta\|_{\mathcal{P}} \leq 2\|f\|_{\mathcal{B}}$

Consider a random features model [22, 15]

- first layer: $w \overset{iid}{\sim} \pi \in P(\mathcal{W})$; only train the second layer

infinite many features $f_a(x) = \int_{\mathcal{W}} a(w)\phi(x; w) \mathrm{d}\pi(w)$

**Definition (RKHS and Barron space [9, 5])**

$$\mathcal{F}_{p,\pi} := \{f_a : \|a\|_{L^p(\pi)} < \infty\}, \quad \|f\|_{\mathcal{F}_{p,\pi}} := \inf_{f = f_a} \|a\|_{L^p(\pi)}$$

For any $1 \le p \le \infty$, we have

$$\mathcal{B} = \cup_{\pi \in P(\mathcal{W})} \mathcal{F}_{p,\pi}, \quad \|f\|_{\mathcal{B}} = \inf_{\pi \in P(\mathcal{W})} \|f\|_{\mathcal{F}_{p,\pi}}$$

- RFMs $\equiv$ kernel methods by taking $p = 2$ using Representer theorem [23]
- RFMs $\not\equiv$ kernel methods if $p < 2$
- function space: $\mathcal{F}_{\infty,\pi} \subseteq \mathcal{F}_{p,\pi} \subseteq \mathcal{F}_{q,\pi} \subseteq \mathcal{F}_{1,\pi}$ if $p \ge q$

Consider a random features model [22, 15]

- first layer: $w \overset{iid}{\sim} \pi \in P(W)$; only train the second layer

infinite many features $f_a(x) = \int_W a(w) \phi(x; w) \mathrm{d}\pi(w)$

**Definition (RKHS and Barron space [9, 5])**

$$\mathcal{F}_{p,\pi} := \{ f_a : \|a\|_{L^p(\pi)} < \infty \}, \quad \|f\|_{\mathcal{F}_{p,\pi}} := \inf_{f = f_a} \|a\|_{L^p(\pi)}$$

For any $1 \leq p \leq \infty$, we have

$$\mathcal{B} = \bigcup_{\pi \in P(W)} \mathcal{F}_{p,\pi}, \quad \|f\|_{\mathcal{B}} = \inf_{\pi \in P(W)} \|f\|_{\mathcal{F}_{p,\pi}}$$

- RFMs $\equiv$ kernel methods by taking $p = 2$ using Representer theorem [23]
- RFMs $\neq$ kernel methods if $p < 2$
- function space: $\mathcal{F}_{\infty,\pi} \subseteq \mathcal{F}_{p,\pi} \subseteq \mathcal{F}_{q,\pi} \subseteq \mathcal{F}_{1,\pi}$ if $p \geq q$

# Path norm, Barron spaces, RKHS

Consider a random features model [22, 15]

- first layer: $w \overset{iid}{\sim} \pi \in P(W)$; only train the second layer

infinite many features $f_a(x) = \int_W a(w)\phi(x;w)\,d\pi(w)$

**Definition (RKHS and Barron space [9, 5])**

$$\mathcal{F}_{p,\pi} := \{f_a : \|a\|_{L^p(\pi)} < \infty\}; \quad \|f\|_{\mathcal{F}_{p,\pi}} := \inf_{f=f_a} \|a\|_{L^p(\pi)}$$

For any $1 \leq p \leq \infty$, we have

$$\mathcal{B} = \bigcup_{\pi \in P(W)} \mathcal{F}_{p,\pi}; \quad \|f\|_{\mathcal{B}} = \inf_{\pi \in P(W)} \|f\|_{\mathcal{F}_{p,\pi}}$$

- RFMs $\Rightarrow$ kernel methods by taking $p = 2$ using Representer theorem [23]
- RFMs $\not\Rightarrow$ kernel methods if $p < 2$
- function space: $\mathcal{F}_{\infty,\pi} \subset \mathcal{F}_{p,\pi} \subset \mathcal{F}_{q,\pi} \subset \mathcal{F}_{1,\pi}$ if $p \geq q$

## Path norm, Barron spaces, RKHS

Consider a random features model [22, 15]

- first layer: $w \overset{iid}{\sim} \pi \in \mathcal{P}(\mathcal{W})$; only train the second layer

infinite many features $f_a(x) = \int_{\mathcal{W}} a(w)\,\phi(x;w)\,\mathrm{d}\pi(w)$

**Definition (RKHS and Barron space [9, 5])**

$$\mathcal{F}_{p,\pi} := \{f_a : \|a\|_{L^p(\pi)} < \infty\}; \quad \|f\|_{\mathcal{F}_{p,\pi}} := \inf_{f=f_a} \|a\|_{L^p(\pi)}$$

For any $1 \leq p \leq \infty$, we have

$$\mathcal{B} = \bigcup_{\pi \in \mathcal{P}(\mathcal{W})} \mathcal{F}_{p,\pi}; \quad \|f\|_{\mathcal{B}} = \inf_{\pi \in \mathcal{P}(\mathcal{W})} \|f\|_{\mathcal{F}_{p,\pi}}$$

- RFMs $\subseteq$ kernel methods by taking $p = 2$ using Representer theorem [23]
- RFMs $\not\subseteq$ kernel methods if $p < 2$
- function space: $\mathcal{F}_{\infty,\pi} \subseteq \mathcal{F}_{p,\pi} \subseteq \mathcal{F}_{q,\pi} \subseteq \mathcal{F}_{1,\pi}$ if $p \geq q$

Consider a random features model [22, 15]

- first layer: $w \overset{iid}{\sim} \pi \in \mathcal{P}(\mathcal{W})$; only train the second layer

infinite many features $f_a(x) = \int_{\mathcal{W}} a(w)\sigma(x; w)\mathrm{d}\pi(w)$

**Definition (RKHS and Barron space [9, 5])**

$$\mathcal{F}_{p;\pi} := \{f_a : \|a\|_{L^p(\pi)} < \infty\}; \quad \|f\|_{\mathcal{F}_{p;\pi}} := \inf_{f = f_a} \|a\|_{L^p(\pi)}$$

For any $1 \leq p \leq \infty$, we have

$$\mathcal{B} = \bigcup_{\pi \in \mathcal{P}(\mathcal{W})} \mathcal{F}_{p;\pi}; \quad \|f\|_{\mathcal{B}} = \inf_{\pi \in \mathcal{P}(\mathcal{W})} \|f\|_{\mathcal{F}_{p;\pi}}$$

- RFMs $\Leftrightarrow$ kernel methods by taking $p = 2$ using Representer theorem [23]
- RFMs $\not\Leftrightarrow$ kernel methods if $p < 2$
- function space: $\mathcal{F}_{\infty;\pi} \subset \mathcal{F}_{p;\pi} \subset \mathcal{F}_{q;\pi} \subset \mathcal{F}_{1;\pi}$ if $p \geq q$

# Our results: statistical guarantees

For the class of two-layer neural networks $\mathcal{G}_R = \{f \in \mathcal{P}_m : \|f\|_{\mathcal{P}} \le R\}$

$$\hat{f} := \operatorname*{argmin}_{f \in \mathcal{G}_R} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 .$$

**Theorem (Liu, Dadi, Cevher, JMLR 2024)**

*Under standard assumptions (bounded data, $f^\star \in \mathcal{B}$), for two-layer over-parameterized neural networks, we have*

$$\|\hat{f} - f^\star\|^2_{L^2_{\rho_X}} \lesssim \frac{R^2}{m} + R^2 d^{\frac{1}{3}} n^{-\frac{d+2}{2d+2}} \quad w.h.p.$$

$n^{-\frac{d+2}{2d+2}}$ is always faster than $n^{-\frac{1}{2}}$: No curse of dimensionality!

# Our results: statistical guarantees

For the class of two-layer neural networks $\mathcal{G}_R = \{f \in \mathcal{P}_m : \|f\|_{\mathcal{P}} \leqslant R\}$

$$\hat{f} := \operatorname*{argmin}_{f \in \mathcal{G}_R} \frac{1}{n}\sum_{i=1}^{n} (y_i - f(x_i))^2 .$$

**Theorem (Liu, Dadi, Cevher, JMLR 2024)**

*Under standard assumptions (bounded data, $f^\star \in \mathcal{B}$), for two-layer over-parameterized neural networks, we have*

$$\|\hat{f} - f^\star\|^2_{L^2_X} \lesssim \frac{R^2}{m} + R^2 d^{\frac{1}{3}} n^{-\frac{d+2}{2d+2}} \quad w.h.p.$$

$n^{-\frac{d+2}{2d+2}}$ is always faster than $n^{-\frac{1}{2}}$: No curse of dimensionality!

# Our results: statistical guarantees

For the class of two-layer neural networks $\mathcal{G}_R = \{f \in \mathcal{P}_m : \|f\|_{\mathcal{P}} \leq R\}$

$$\hat{f} := \operatorname*{argmin}_{f \in \mathcal{G}_R} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 .$$

**Theorem (Liu, Dadi, Cevher, JMLR 2024)**

*Under standard assumptions (bounded data, $f^\star \in \mathcal{B}$), for two-layer over-parameterized neural networks, we have*

$$\|\hat{f} - f^\star\|^2_{L^2_{\rho_X}} \lesssim \frac{R^2}{m} + R^2 d^{\frac{1}{3}} n^{-\frac{d+2}{2d+2}} \quad w.h.p.$$

$n^{-\frac{d+2}{2d+2}}$ is always faster than $n^{-\frac{1}{2}}$: No curse of dimensionality!

**Proposition (metric entropy)**

*For bounded data* $\|x\|_\infty \leq 1$, *denote* $G_R = \{f \in P_m : \|f\|_P \leqslant R\}$, *the metric entropy of* $G_1$ *can be bounded by*

$$\log N_2(G_1, \epsilon) \leqslant C d\, \epsilon^{-\frac{2d}{d+2}}, \quad \forall \epsilon > 0 \quad \text{and} \quad d \geq 5,$$

*with some universal constant* $C$ *independent of* $d$.

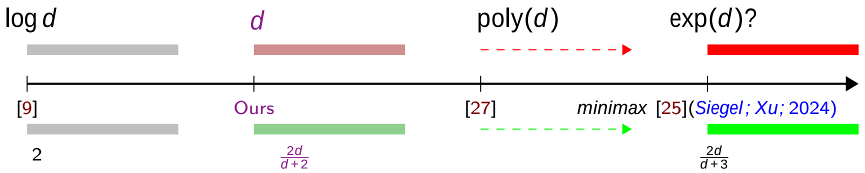
# Sample complexity

**Proposition (metric entropy)**

*For bounded data $\|x\|_\infty \leq 1$, denote $G_R = \{f \in P_m : \|f\|_P \leq R\}$, the metric entropy of $G_1$ can be bounded by*

$$\log N_2(G_1; \epsilon) \leq C d \epsilon^{-\frac{2d}{d+2}}, \quad \forall \epsilon > 0 \text{ and } d \geq 5,$$

*with some universal constant $C$ independent of $d$.*



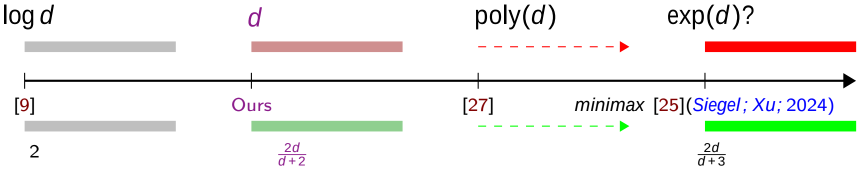
# Sample complexity

**Proposition (metric entropy)**

*For bounded data $\|x\|_\infty \leq 1$, denote $G_R = \{f \in \mathcal{P}_m : \|f\|_{\mathcal{P}} \leqslant R\}$, the metric entropy of $G_1$ can be bounded by*

$$\log \mathcal{N}_2(G_1;\epsilon) \leqslant C d\, \epsilon^{-\frac{2d}{d+2}}, \quad \forall \epsilon > 0 \ \text{and} \ d \geq 5,$$

*with some universal constant $C$ independent of $d$.*

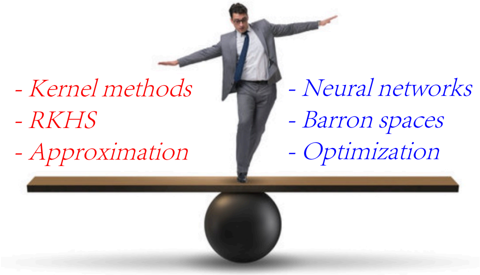



The "best" trade-off between $\epsilon$ and $d$.

# Computational-to-statistical gaps

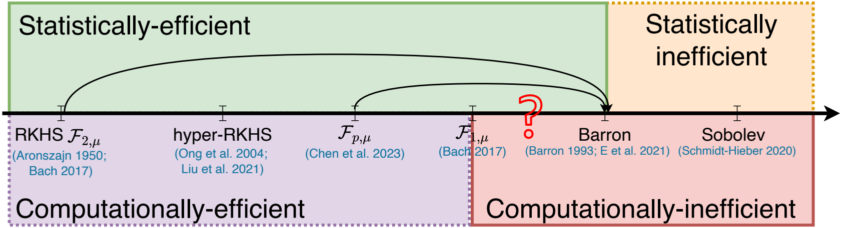
Optimization in Barron spaces is NP hard: curse of dimensionality!

Optimization in Barron spaces is NP hard: curse of dimensionality!



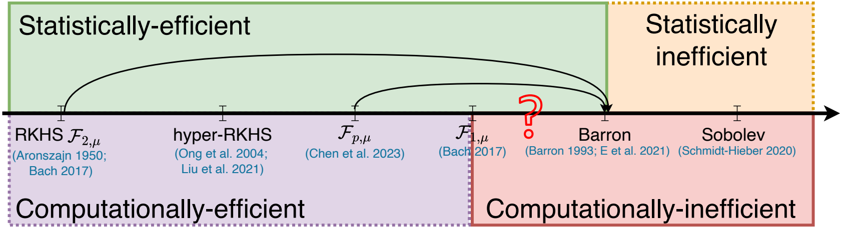
# Computational-to-statistical gaps

Optimization in Barron spaces is NP hard: curse of dimensionality!



# Computational-to-statistical gaps

Optimization in Barron spaces is NP hard: curse of dimensionality!




Do some Barron functions can be learned by two-layer NNs, both statistically and computationally e□cient?

# Learning with multiple ReLU neurons

Can we learn multiple ReLU neurons by two-layer NNs, both statistically and computationally efficient?

$$f^\star(x) = \sum_{j=1}^{k} a_j \sigma(\langle v_j, x \rangle), k = O(1)$$

$\|\hat{f} - f^\star\|_{L^2(\gamma_d)} \leq \epsilon$ from $\{x_i, f^\star(x_i)\}_{i=1}^n$ with $x_i \sim \mathcal{N}(0, I_d)$

**Theorem ([6] PAC learning $f^\star$ under Gaussian measure)**

*There exists an algorithm that requires time/samples at $(d/\epsilon)^{O(k^2)}$*

- correlational statistical query (CSQ): $|q - \mathbb{E}_{x,y}[\phi(x)y]| \leq \tau$

# Learning with multiple ReLU neurons

Can we learn multiple ReLU neurons by two-layer NNs, both statistically and computationally efficient?

$$f^\star(x) = \sum_{j=1}^{k} a_j \sigma(\langle v_j, x\rangle), \; k = O(1)$$

$\|\hat{f} - f^\star\|_{L^2(d)} \leq \epsilon$ from $\{x_i, f^\star(x_i)\}_{i=1}^n$ with $x_i \sim N(0, I_d)$

**Theorem ([6] PAC learning $f^\star$ under Gaussian measure)**

*There exists an algorithm that requires time/samples at $(d/\epsilon)^{O(k^2)}$*

- correlational statistical query (CSQ): $|q - \mathbb{E}_{x,y}[\phi(x)y]| \leq \tau$

# Learning with multiple ReLU neurons

Can we learn multiple ReLU neurons by two-layer NNs, both statistically and computationally efficient?

$$f^\star(x) = \sum_{j=1}^{k} a_j \sigma(\langle v_j, x\rangle), \; k = O(1)$$

$\|\hat{f} - f^\star\|_{L^2(\gamma_d)} \le \epsilon$ from $\{x_i, f^\star(x_i)\}_{i=1}^n$ with $x_i \sim N(0, I_d)$

**Theorem ([6] PAC learning $f^\star$ under Gaussian measure)**

*There exists an algorithm that requires time/samples at $(d/\epsilon)^{O(k^2)}$*

- correlational statistical query (CSQ): $|q - \mathbb{E}_{x,y}[\phi(x)y]| \le \tau$

# Learning with multiple ReLU neurons

Can we learn multiple ReLU neurons by two-layer NNs, both statistically and computationally e□cient?

$$f^{\star}(x) = \sum_{j=1}^{k} a_j (\langle v_j, x \rangle);\ k = O(1)$$

$\|\hat{f} - f^{\star}\|_{L^2(d)}$ from $\{x_i, f^{\star}(x_i)\}_{i=1}^{n}$ with $x_i \quad N(0, I_d)$

**Theorem ([6] PAC learning $f^{\star}$ under Gaussian measure)**

*There exists an algorithm that requires time/samples at $(d/\ )^{O(k^2)}$*

- correlational statistical query (CSQ): $|q \quad \mathbb{E}_{x,y}[\ (x) y]|$

## Learning with multiple ReLU neurons

Can we learn multiple ReLU neurons by two-layer NNs, both statistically and computationally efficient?

$$f^\star(x) = \sum_{j=1}^{k} a_j \sigma(\langle v_j, x \rangle), \quad k = O(1)$$

$\|\hat{f} - f^\star\|_{L^2(\mu_d)} \leq \epsilon$ from $\{x_i, f^\star(x_i)\}_{i=1}^n$ with $x_i \sim \mathcal{N}(0, I_d)$

**Theorem ([6] PAC learning $f^\star$ under Gaussian measure)**

*There exists an algorithm that requires time/samples at* $(d/\epsilon)^{O(k^2)}$

- correlational statistical query (CSQ): $|q - \mathbb{E}_{x,y}[\phi(x)y]| \leq \tau$

# How does student(s) become teacher(s) under GD training?

**Learning multi ReLU neurons by two-layer NN via online SGD**

$$L(W) = \frac{1}{2}\mathbb{E}_{x \sim N(0, I_d)} \left( \sum_{i=1}^{m} \sigma(\langle w_i, x \rangle) - f^{\star}(x) \right)^2$$

- Gaussian initialization $w_i \sim N(0, \sigma^2 I_d)$
- angle: $\theta_{ij} := \angle(w_i, v_j)$

**Assumption**

- *diverse teacher neurons: $\{v_j\}_{j=1}^{k}$ are orthogonal and $\|v_j\|_2 = \mathtt{const}$*
- *warm start: the smallest angle not close to orthogonal*
  $\Rightarrow$ *weak recovery: $\langle w_i, v_{i^\star} \rangle \gg \langle w_i, v_j \rangle$*

# How does student(s) become teacher(s) under GD training?

## Learning multi ReLU neurons by two-layer NN via online SGD

$$L(W) = \frac{1}{2}\mathbb{E}_{x \quad N(0;I_d)} \left( \sum_{i=1}^{m} (\langle w_i ; x \rangle) \quad f^{?}(x) \right)^2$$

- Gaussian initialization $w_i \quad N(0; \ {}^2 I_d)$
- angle: $_{ij} \ , \ \angle(w_i; v_j)$

## Assumption

- *diverse teacher neurons:* $\{v_j\}_{j=1}^{k}$ *are* orthogonal *and* $\|v_j\|_2$ = const
- *warm start: the* smallest *angle not close to orthogonal*
  *weak recovery:* $\langle w_i ; v_{i^?} \rangle \quad \langle w_i ; v_j \rangle$

# How does student(s) become teacher(s) under GD training?

**Learning multi ReLU neurons by two-layer NN via online SGD**

$$L(W) = \frac{1}{2}\mathbb{E}_{x \sim N(0, I_d)} \left( \sum_{i=1}^{m} \sigma(\langle w_i, x\rangle) - f^\star(x) \right)^2$$

- Gaussian initialization $w_i \sim N(0, \sigma^2 I_d)$
- angle: $\theta_{ij} := \angle(w_i, v_j)$

**Assumption**

- *diverse teacher neurons:* $\{v_j\}_{j=1}^{k}$ *are orthogonal and* $\|v_j\|_2 = \texttt{const}$
- *warm start: the smallest angle not close to orthogonal*

  *weak recovery:* $\langle w_{i_j^\star}, v_j\rangle \gg \langle w_i, v_j\rangle$

# How does student(s) become teacher(s) under GD training?

**Learning multi ReLU neurons by two-layer NN via online SGD**

$$L(W) = \frac{1}{2}\mathbb{E}_{x \quad N(0, I_d)} \left[ \sum_{i=1}^{m} \quad (\langle w_i, x\rangle) \quad f^\star(x) \right]^2$$

- Gaussian initialization $w_i \quad N(0, \quad ^2 I_d)$
- angle: $\quad_{ij} := \angle(w_i, v_j)$

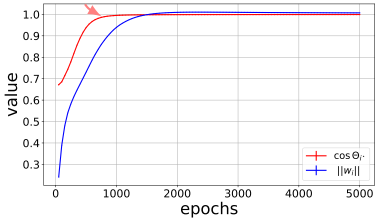
**Assumption**

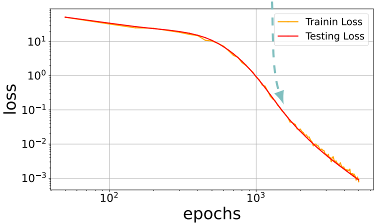
- *diverse teacher neurons:* $\{v_j\}_{j=1}^k$ *are* orthogonal *and* $\|v_j\|_2 =$ `const`
- *warm start: the* smallest *angle not close to orthogonal*

  *weak recovery:* $\langle w_i, v_{i^\star}\rangle \quad \langle w_i, v_j\rangle$

# How does student(s) become teacher(s) under GD training?

- align $_i$ ! 0 norm converge then fit



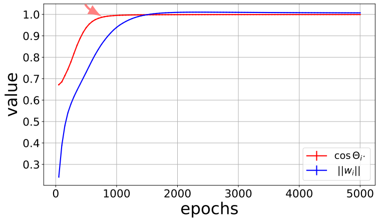

**Theorem (Zhu, Liu, Cevher, 2024)**

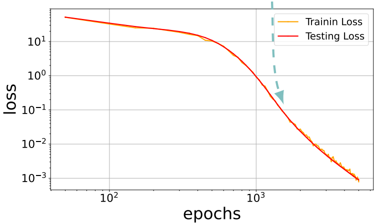
*For su ciently small initialization and step-size ; $= o(m \;\; ^{k^2})$, then there exists a time $T_2 = \frac{1}{}$ such that 8T 2 N and i 2 [m],*

$$L(W(T+T_2)) \quad O \; \frac{1}{T^3} \quad ; \|w_i(T+T_2)\|_2 = \frac{\|\|v\|_2}{m} \quad w.h.p.$$

# How does student(s) become teacher(s) under GD training?

- align $_i$ ! 0 norm converge then fit




**Theorem (Zhu, Liu, Cevher, 2024)**

*For su ciently small initialization and step-size ; = o(m $k^2$), then there exists a time $T_2$ = $\frac{1}{}$ such that 8T 2 N and i 2 [m],*

$$L(W(T+T_2)) \quad O \frac{1}{T^3} \; ; \|w_i(T+T_2)\|_2 = \frac{\|v\|_2}{m} \quad w{:}h{:}p{:}$$

# Take-away messages

- model size -> size of weights -> path norm -> Barron spaces
- statistical guarantees with improved sample complexity
- computational-statistical gap -> learning with multiple ReLU neurons

**We're organizing one workshop at NeurIPS 2024!**

Fine-Tuning in Modern Machine Learning: Principles and Scalability

https://sites.google.com/view/neurips2024-ftw/home

# Take-away messages

- model size -> size of weights -> path norm -> Barron spaces
- statistical guarantees with improved sample complexity
- computational-statistical gap -> learning with multiple ReLU neurons

[redacted]

Fine-Tuning in Modern Machine Learning: Principles and Scalability
https://sites.google.com/view/neurips2024-ftw/home

# Take-away messages

- model size -> size of weights -> path norm -> Barron spaces
- statistical guarantees with improved sample complexity
- computational-statistical gap -> learning with multiple ReLU neurons

[redacted]

Fine-Tuning in Modern Machine Learning: Principles and Scalability
https://sites.google.com/view/neurips2024-ftw/home

# Take-away messages

- model size -> size of weights -> path norm -> Barron spaces
- statistical guarantees with improved sample complexity
- computational-statistical gap -> learning with multiple ReLU neurons

[redacted]

Fine-Tuning in Modern Machine Learning: Principles and Scalability
https://sites.google.com/view/neurips2024-ftw/home

# References i


Andrew R Barron.
Universal approximation bounds for superpositions of a sigmoidal function.
IEEE Transactions on Information theory, 39(3):930 945, 1993.

Peter Bartlett.
The sample complexity of pattern classi cation with neural networks: the size of the weights is more important than the size of the network.
IEEE Transactions on Information Theory, 44(2):525 536, 1998.

Peter Bartlett, Dylan Foster, and Matus Telgarsky.
Spectrally-normalized margin bounds for neural networks.
In Advances in Neural Information Processing Systems, pages 6241 6250, 2017.

# References ii


Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal.
Reconciling modern machine-learning practice and the classical bias variance trade-o .
the National Academy of Sciences, 116(32):15849 15854, 2019.

Hongrui Chen, Jihao Long, and Lei Wu.
A duality framework for generalization analysis of random feature models and two-layer neural networks.
arXiv preprint arXiv:2305.05642, 2023.

Sitan Chen and Shyam Narayanan.
A faster and simpler algorithm for learning shallow networks.
arXiv preprint arXiv:2307.12496, 2023.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu.
How much over-parameterization is su cient to learn deep relu networks?
In International Conference on Learning Representations, 2020.

Carles Domingo-Enrich and Youssef Mroueh.
Tighter sparse approximation bounds for relu neural networks.
In International Conference on Learning Representations, 2022.

Weinan E, Chao Ma, and Lei Wu.
A priori estimates of the population risk for two-layer neural networks.
Communications in Mathematical Sciences, 17(5):1407 1425, 2019.

Weinan E, Chao Ma, and Lei Wu.
The barron space and the  ow-induced function spaces for neural network models.
*Constructive Approximation*, pages 1 38, 2021.

Jonathan Frankle and Michael Carbin.
The lottery ticket hypothesis: Finding sparse, trainable neural networks.
In *International Conference on Learning Representations*, 2019.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio.
Fantastic generalization measures and where to  nd them.
In *International Conference on Learning Representations*, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Je rey Wu, and Dario Amodei.
Scaling laws for neural language models.
arXiv preprint arXiv:2001.08361, 2020.

Tengyuan Liang, Tomaso Poggio, Alexander Rakhlin, and James Stokes.
Fisher-rao metric, geometry, and complexity of neural networks.
In International conference on Arti cial Intelligence and Statistics, pages 888 896, 2019.

Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens.
Random features for kernel approximation: A survey on algorithms, theory, and beyond.
IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10):7128 7148, 2021.

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz.
Pruning convolutional neural networks for resource e cient inference.
In International Conference on Learning Representations, 2017.

Vaishnavh Nagarajan and J Zico Kolter.
Generalization in deep networks: The role of distance from initialization.
arXiv preprint arXiv:1901.01672, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever.
Deep double descent: Where bigger models and more data hurt.
In International Conference on Learning Representations, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro.
Norm-based capacity control in neural networks.
In Conference on Learning Theory, pages 1376 1401. PMLR, 2015.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro.
A function space view of bounded norm in nite width relu nets:
The multivariate case.
In International Conference on Learning Representations, 2020.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Je rey Pennington.
4+3 phases of compute-optimal neural scaling laws.
arXiv preprint arXiv:2405.15074, 2024.

Ali Rahimi and Benjamin Recht.
Random features for large-scale kernel machines.
In Advances in Neural Information Processing Systems, pages 1177 1184, 2007.

Ali Rahimi and Benjamin Recht.
Uniform approximation of functions with random bases.
In Annual Allerton Conference on Communication, Control, and Computing pages 555 561. IEEE, 2008.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro.
How do in nite width bounded norm networks look in function space?
In Conference on Learning Theory, pages 2667 2690. PMLR, 2019.

Jonathan W Siegel and Jinchao Xu.
Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks.
arXiv preprint arXiv:2101.12365, 2021.

# References x


Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda.
Feature learning via mean- eld langevin dynamics: classifying sparse parities and beyond.
In Advances in Neural Information Processing Systems, 2023.

Lei Wu and Jihao Long.
A spectral-based analysis of the separation between two-layer neural networks and linear methods.
Journal of Machine Learning Research, 119:1 34, 2022.

Jianyu Zhang and Léon Bottou.
Fine-tuning with very large dropout.
arXiv preprint arXiv:2403.00946, 2024.