The φ curve: Generalization under scaling via norm-based capacities

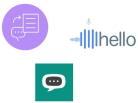
Fanghui Liu

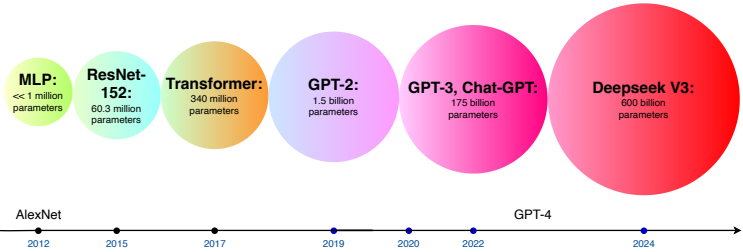
fanghui.liu@sjtu.edu.cn

Institute of Natural Sciences, School of Mathematical Sciences
Shanghai Jiao Tong University (SJTU)

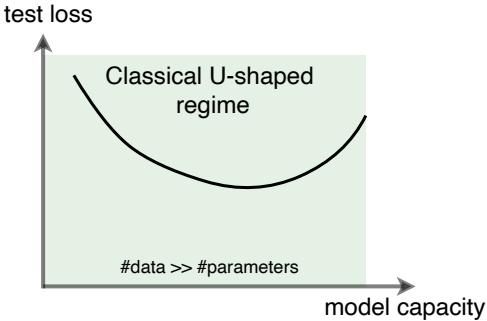

In the era of machine learning

Prefer more data and larger model to obtain better performance...



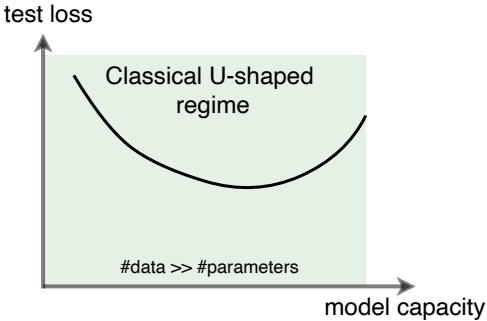


ML textbooks: Larger models tend to overfit!

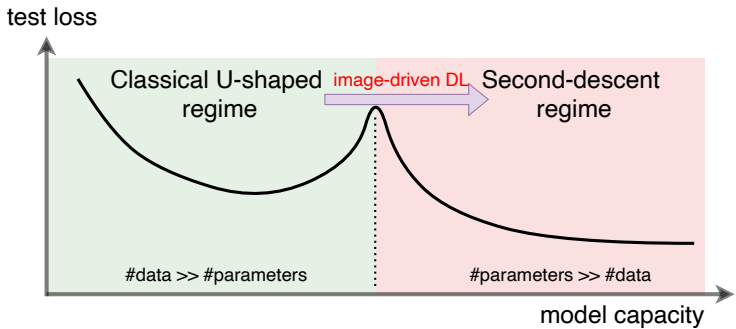


ML textbooks: Larger models tend to overfit!

Practice of deep learning: bigger models perform better!



Practice of deep learning: bigger models perform better!



Proposed explanation: double descent (Belkin et al., 2019)

Learning paradigm in the past twenty years

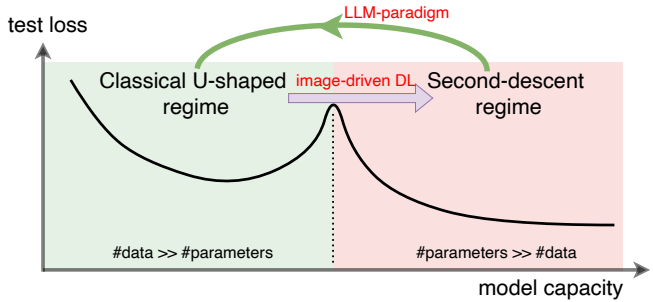


Figure 1: Paradigm among test loss, data, and model capacity.

Scaling law (Kaplan et al., 2020) in the era of LLMs

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

Learning paradigm in the past twenty years

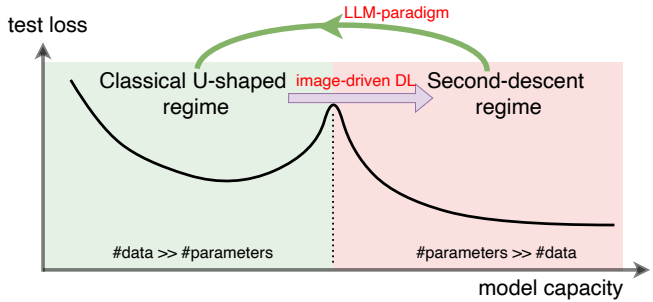


Figure 1: Paradigm among test loss, data, and model capacity.

Scaling law (Kaplan et al., 2020) in the era of LLMs

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

A fundamental concept in machine learning: model capacity

Too many learning curves...

- U-shaped curve (bias-variance trade-offs) (Vapnik, 1995; Hastie et al., 2009)
- double (multiple) descent (Belkin et al., 2019; Liang et al., 2020)
- scaling law (Kaplan et al., 2020; Paquette et al., 2024)

A fundamental concept in machine learning: model capacity

Too many learning curves...

- U-shaped curve (bias-variance trade-offs) (Vapnik, 1995; Hastie et al., 2009)
- double (multiple) descent (Belkin et al., 2019; Liang et al., 2020)
- scaling law (Kaplan et al., 2020; Paquette et al., 2024)

"Remove bias-variance trade-offs from ML textbooks"

I can define **model capacity** at random and see whatever curve I want to see.

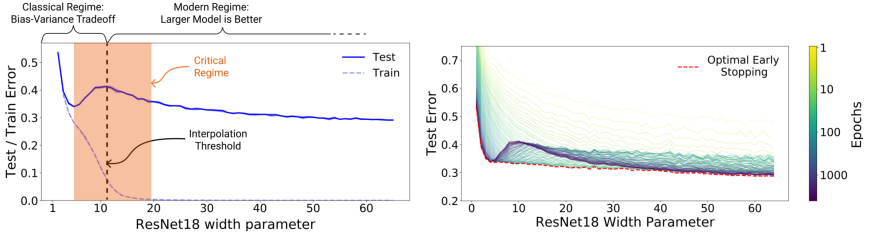
— Ben Recht, 2025

A fundamental concept in machine learning: model capacity

Too many learning curves...

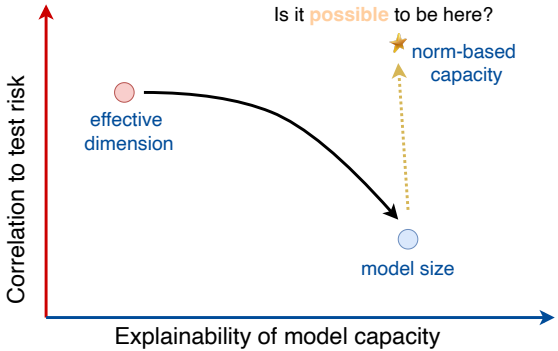
- U-shaped curve (bias-variance trade-offs) (Vapnik, 1995; Hastie et al., 2009)
- double (multiple) descent (Belkin et al., 2019; Liang et al., 2020)
- scaling law (Kaplan et al., 2020; Paquette et al., 2024)

Double descent can disappear for the same architecture!



(a) Results on ResNet18 (Nakkiran et al., 2019) **(b)** Optimal early stopping (Nakkiran et al., 2019).

Today's talk: Norm-based capacity via deterministic equivalence



Today's talk: Norm-based capacity via deterministic equivalence

(Bartlett, 1998)

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

Today's talk: Norm-based capacity via deterministic equivalence

(Bartlett, 1998)

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

- Theoretical studies (Neyshabur et al., 2015; Savarese et al., 2019)
- Min-norm solution (Hastie et al., 2022)
- Applications: neural networks pruning (Molchanov et al., 2017)
  layerNorm, RMS Norm, mu-transfer , MoE...

Today's talk: Norm-based capacity via deterministic equivalence

(Bartlett, 1998)

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

- Theoretical studies (Neyshabur et al., 2015; Savarese et al., 2019)
- Min-norm solution (Hastie et al., 2022)
- Applications: neural networks pruning (Molchanov et al., 2017)
  layerNorm, RMS Norm, mu-transfer , MoE...

How these learning curves behave under a more suitable model capacity?

Today's talk: Norm-based capacity via deterministic equivalence

(Bartlett, 1998)

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

- Theoretical studies (Neyshabur et al., 2015; Savarese et al., 2019)
- Min-norm solution (Hastie et al., 2022)
- Applications: neural networks pruning (Molchanov et al., 2017)
  layerNorm, RMS Norm, mu-transfer , MoE...

How these learning curves behave under a more suitable model capacity?

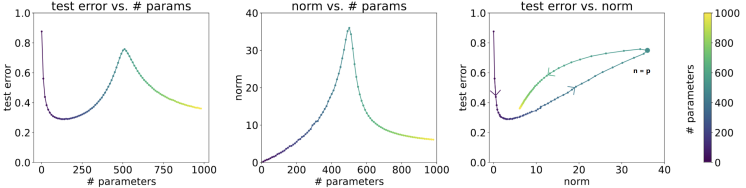


Figure 3: Stanford CS229 lecture notes (Ng and Ma, 2023, Figure 8.12).

Today's talk: Norm-based capacity via deterministic equivalence

(Bartlett, 1998)

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

❒ How to **precisely** characterize the relationship under norm-based model capacity?

- Reshape bias-variance trade-offs, double descent, scaling law under ℓ_2 norm-based capacity!
- Yichen Wang, Yudong Chen, Lorenzo Rosasco, **Fanghui Liu**. *The φ curve: The shape of generalization through the lens of norm-based capacity control.* NeurIPS'25

Today's talk: Norm-based capacity via deterministic equivalence

(Bartlett, 1998)

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

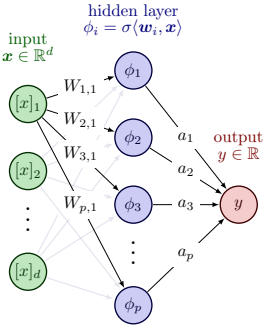
❒ How to **precisely** characterize the relationship under norm-based model capacity?

- Reshape bias-variance trade-offs, double descent, scaling law under ℓ_2 norm-based capacity!
- Yichen Wang, Yudong Chen, Lorenzo Rosasco, **Fanghui Liu**. *The φ curve: The shape of generalization through the lens of norm-based capacity control.* NeurIPS'25

❒ What is the induced function space and statistical/computational efficiency under norm-based capacity?

- Which function class can be **efficiently** learned by neural networks?
- **Fanghui Liu**, Leello Dadi, and Volkan Cevher. *Learning with norm constrained, over-parameterised, two-layer neural networks.* JMLR 2024.

Background: Random features ridge regression



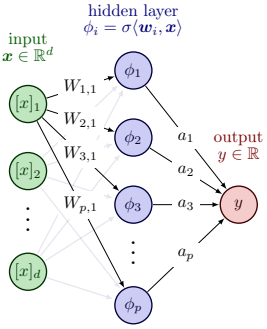
$$f_p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{p} a_i \phi(\boldsymbol{x}, \boldsymbol{w}_i), \quad \boldsymbol{\theta} := \{(a_i, \boldsymbol{w}_i)\}_{i=1}^{p}$$

- $\phi : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$, e.g., ReLU: $\phi(\boldsymbol{x}, \boldsymbol{w}) = \max(\langle \boldsymbol{x}, \boldsymbol{w}\rangle, 0)$
- Random features models (RFMs) (Rahimi and Recht, 2007; Liu et al., 2021):
  - $\{\boldsymbol{w}_i\}_{i=1}^{p} \overset{iid}{\sim} \mu$ for a given $\mu \in \mathcal{P}(\mathcal{W})$
  - only train the second layer

$$\hat{\boldsymbol{a}} := \operatorname*{argmin}_{\boldsymbol{a} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i; \boldsymbol{a}))^2 + \lambda \|\boldsymbol{a}\|_2^2 \right\} = \left(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_p \right)^{-1} \boldsymbol{Z}^\top \boldsymbol{y} \,.$$

  - $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ with $[\boldsymbol{Z}]_{ij} = \frac{1}{\sqrt{m}} \phi(\boldsymbol{x}_i, \boldsymbol{w}_j)$.
- Norm over the first-layer (untrained) $\|\boldsymbol{W}\|_{\mathrm{F}}$
- Norm over the second-layer $\|\hat{\boldsymbol{a}}\|_2^2$

Background: Random features ridge regression



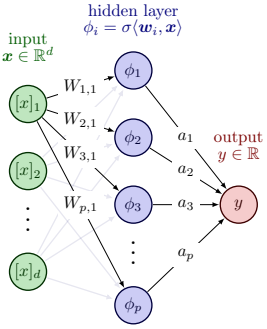
$$f_p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{p} a_i \phi(\boldsymbol{x}, \boldsymbol{w}_i), \quad \boldsymbol{\theta} := \{(a_i, \boldsymbol{w}_i)\}_{i=1}^{p}$$

- $\phi : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$, e.g., ReLU: $\phi(\boldsymbol{x}, \boldsymbol{w}) = \max(\langle \boldsymbol{x}, \boldsymbol{w}\rangle, 0)$
- Random features models (RFMs) (Rahimi and Recht, 2007; Liu et al., 2021):
  - $\{\boldsymbol{w}_i\}_{i=1}^{p} \overset{iid}{\sim} \mu$ for a given $\mu \in \mathcal{P}(\mathcal{W})$
  - only train the second layer

$$\hat{\boldsymbol{a}} := \operatorname*{argmin}_{\boldsymbol{a} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i; \boldsymbol{a}))^2 + \lambda \|\boldsymbol{a}\|_2^2 \right\} = \left(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_p \right)^{-1} \boldsymbol{Z}^\top \boldsymbol{y} \,.$$

- $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ with $[\boldsymbol{Z}]_{ij} = \frac{1}{\sqrt{m}} \phi(\boldsymbol{x}_i, \boldsymbol{w}_j)$.
- Norm over the first-layer (untrained) $\|\boldsymbol{W}\|_{\mathrm{F}}$
- Norm over the second-layer $\|\hat{\boldsymbol{a}}\|_2^2$

Background: Random features ridge regression



$$f_p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{p} a_i \phi(\boldsymbol{x}, \boldsymbol{w}_i), \quad \boldsymbol{\theta} := \{(a_i, \boldsymbol{w}_i)\}_{i=1}^{p}$$

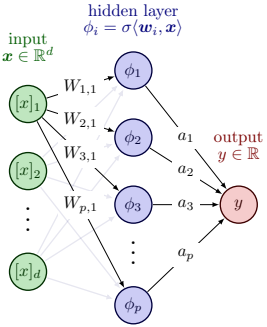
- $\phi : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$, e.g., ReLU: $\phi(\boldsymbol{x}, \boldsymbol{w}) = \max(\langle \boldsymbol{x}, \boldsymbol{w}\rangle, 0)$
- Random features models (RFMs) (Rahimi and Recht, 2007; Liu et al., 2021):
  - $\{\boldsymbol{w}_i\}_{i=1}^{p} \overset{iid}{\sim} \mu$ for a given $\mu \in \mathcal{P}(\mathcal{W})$
  - only train the second layer

$$\hat{\boldsymbol{a}} := \operatorname*{argmin}_{\boldsymbol{a} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i; \boldsymbol{a}))^2 + \lambda \|\boldsymbol{a}\|_2^2 \right\} = \left(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_p \right)^{-1} \boldsymbol{Z}^\top \boldsymbol{y} \,.$$

- $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ with $[\boldsymbol{Z}]_{ij} = \frac{1}{\sqrt{m}} \phi(\boldsymbol{x}_i, \boldsymbol{w}_j)$.

- Norm over the first-layer (untrained) $\|\boldsymbol{W}\|_{\mathrm{F}}$
- Norm over the second-layer $\|\hat{\boldsymbol{a}}\|_2^2$

Background: Random features ridge regression



$$f_p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i=1}^{p} a_i \phi(\boldsymbol{x}, \boldsymbol{w}_i), \quad \boldsymbol{\theta} := \{(a_i, \boldsymbol{w}_i)\}_{i=1}^{p}$$

- $\phi : \mathcal{X} \times \mathcal{W} \to \mathbb{R}$, e.g., ReLU: $\phi(\boldsymbol{x}, \boldsymbol{w}) = \max(\langle \boldsymbol{x}, \boldsymbol{w}\rangle, 0)$
- Random features models (RFMs) (Rahimi and Recht, 2007; Liu et al., 2021):
  - $\{\boldsymbol{w}_i\}_{i=1}^{p} \overset{iid}{\sim} \mu$ for a given $\mu \in \mathcal{P}(\mathcal{W})$
  - only train the second layer

$$\hat{\boldsymbol{a}} := \operatorname*{argmin}_{\boldsymbol{a} \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i; \boldsymbol{a}))^2 + \lambda \|\boldsymbol{a}\|_2^2 \right\} = \left(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I}_p \right)^{-1} \boldsymbol{Z}^\top \boldsymbol{y} \,.$$

- $\boldsymbol{Z} \in \mathbb{R}^{n \times p}$ with $[\boldsymbol{Z}]_{ij} = \frac{1}{\sqrt{m}} \phi(\boldsymbol{x}_i, \boldsymbol{w}_j)$.

- Norm over the first-layer (untrained) $\|\boldsymbol{W}\|_{\mathrm{F}}$
- Norm over the second-layer $\|\hat{\boldsymbol{a}}\|_2^2$

Background: Test risk of random features model

- A **compact** integral operator $\mathbb{T} : L^2(\rho_X) \to L^2(\mu_{\mathcal{W}})$ for any $f \in L^2(\rho_X)$ (Defilippis et al., 2024)

$$(\mathbb{T}f)(\boldsymbol{w}) := \int_{\mathbb{R}^d} \phi(\boldsymbol{x}, \boldsymbol{w}) f(\boldsymbol{x}) \mathrm{d}\rho(\boldsymbol{x}), \quad \mathbb{T} = \sum_{k=1}^{\infty} \xi_k \psi_k \varphi_k^* .$$

- Covariate feature matrix $\boldsymbol{G} := [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n]^\top \in \mathbb{R}^{n \times \infty}$ with $\boldsymbol{g}_i := (\psi_k(\boldsymbol{x}_i))_{k \geq 1}$
- Weight feature matrix $\boldsymbol{H} := [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p]^\top \in \mathbb{R}^{p \times \infty}$ with $\boldsymbol{h}_j := (\xi_k \varphi_k(\boldsymbol{w}_j))_{k \geq 1}$
- target function: $f_*(\boldsymbol{x}) = \sum_{k \geq 1} \theta_{*,k} \psi_k(\boldsymbol{x})$

$$\mathcal{R}^{\mathrm{RFM}} := \mathbb{E}_\varepsilon \left\| \theta_* - \frac{1}{\sqrt{p}} \boldsymbol{H}^\top \hat{\boldsymbol{a}} \right\|_2^2$$

Background: Test risk of random features model

- A **compact** integral operator $\mathbb{T} : L^2(\rho_X) \to L^2(\mu_{\mathcal{W}})$ for any $f \in L^2(\rho_X)$ (Defilippis et al., 2024)

$$(\mathbb{T}f)(\boldsymbol{w}) := \int_{\mathbb{R}^d} \phi(\boldsymbol{x}, \boldsymbol{w}) f(\boldsymbol{x}) \mathrm{d}\rho(\boldsymbol{x}), \quad \mathbb{T} = \sum_{k=1}^{\infty} \xi_k \psi_k \varphi_k^* .$$

- Covariate feature matrix $\boldsymbol{G} := [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n]^\top \in \mathbb{R}^{n \times \infty}$ with $\boldsymbol{g}_i := (\psi_k(\boldsymbol{x}_i))_{k \geq 1}$
- Weight feature matrix $\boldsymbol{H} := [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p]^\top \in \mathbb{R}^{p \times \infty}$ with $\boldsymbol{h}_j := (\xi_k \varphi_k(\boldsymbol{w}_j))_{k \geq 1}$
- target function: $f_*(\boldsymbol{x}) = \sum_{k \geq 1} \theta_{*,k} \psi_k(\boldsymbol{x})$

$$\mathcal{R}^{\mathrm{RFM}} := \mathbb{E}_\varepsilon \left\| \theta_* - \frac{1}{\sqrt{p}} \boldsymbol{H}^\top \hat{\boldsymbol{a}} \right\|_2^2$$

Background: Test risk of random features model

- A **compact** integral operator $\mathbb{T} : L^2(\rho_X) \to L^2(\mu_{\mathcal{W}})$ for any $f \in L^2(\rho_X)$ (Defilippis et al., 2024)

$$(\mathbb{T}f)(\boldsymbol{w}) := \int_{\mathbb{R}^d} \phi(\boldsymbol{x}, \boldsymbol{w}) f(\boldsymbol{x}) \mathrm{d}\rho(\boldsymbol{x}), \quad \mathbb{T} = \sum_{k=1}^{\infty} \xi_k \psi_k \varphi_k^* .$$

- Covariate feature matrix $\boldsymbol{G} := [\boldsymbol{g}_1, \dots, \boldsymbol{g}_n]^\top \in \mathbb{R}^{n \times \infty}$ with $\boldsymbol{g}_i := (\psi_k(\boldsymbol{x}_i))_{k \geq 1}$
- Weight feature matrix $\boldsymbol{H} := [\boldsymbol{h}_1, \dots, \boldsymbol{h}_p]^\top \in \mathbb{R}^{p \times \infty}$ with $\boldsymbol{h}_j := (\xi_k \varphi_k(\boldsymbol{w}_j))_{k \geq 1}$
- target function: $f_*(\boldsymbol{x}) = \sum_{k \geq 1} \boldsymbol{\theta}_{*,k} \psi_k(\boldsymbol{x})$

$$\mathcal{R}^{\mathrm{RFM}} := \mathbb{E}_\varepsilon \left\| \boldsymbol{\theta}_* - \frac{1}{\sqrt{p}} \boldsymbol{H}^\top \hat{\boldsymbol{a}} \right\|_2^2$$

Background: Test risk of random features model

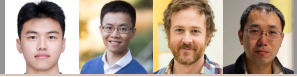
- A **compact** integral operator $\mathbb{T} : L^2(\rho_X) \to L^2(\mu_{\mathcal{W}})$ for any $f \in L^2(\rho_X)$ (Defilippis et al., 2024)

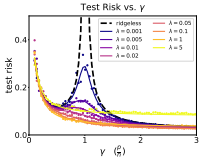
$$(\mathbb{T}f)(\boldsymbol{w}) := \int_{\mathbb{R}^d} \phi(\boldsymbol{x}, \boldsymbol{w}) f(\boldsymbol{x}) \mathrm{d}\rho(\boldsymbol{x}), \quad \mathbb{T} = \sum_{k=1}^{\infty} \xi_k \psi_k \varphi_k^* .$$

- Covariate feature matrix $\boldsymbol{G} := [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_n]^\top \in \mathbb{R}^{n \times \infty}$ with $\boldsymbol{g}_i := (\psi_k(\boldsymbol{x}_i))_{k \geq 1}$
- Weight feature matrix $\boldsymbol{H} := [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_p]^\top \in \mathbb{R}^{p \times \infty}$ with $\boldsymbol{h}_j := (\xi_k \varphi_k(\boldsymbol{w}_j))_{k \geq 1}$
- target function: $f_*(\boldsymbol{x}) = \sum_{k \geq 1} \boldsymbol{\theta}_{*,k} \psi_k(\boldsymbol{x})$

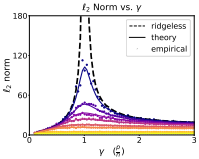
$$\mathcal{R}^{\mathtt{RFM}} := \mathbb{E}_\varepsilon \left\| \boldsymbol{\theta}_* - \frac{1}{\sqrt{p}} \boldsymbol{H}^\top \hat{\boldsymbol{a}} \right\|_2^2$$

Our results under well-behaved data

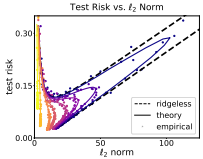




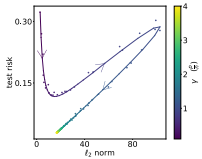
(a) Test Risk vs. γ



(b) ℓ_2 norm vs. γ



(c) Test Risk vs. norm

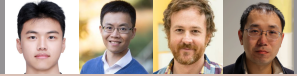


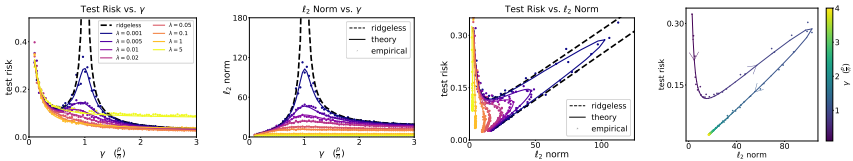
(d) $\lambda = 0.001$

- $\gamma := p/n$, p : model size (width), n : data size

p ($n = 100$)	10	50	100	150	200	250	300	350	400
Test Risk	0.32	0.12	0.29	0.08	0.05	0.04	0.03	0.03	0.03
Norm	2.93	14.66	102.89	35.26	24.76	20.67	18.63	17.47	16.69

Our results under well-behaved data



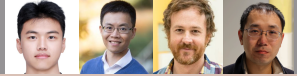


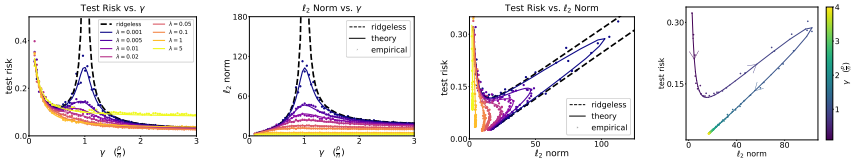
(a) Test Risk vs. γ **(b)** ℓ_2 norm vs. γ **(c)** Test Risk vs. norm **(d)** $\lambda = 0.001$

- $\gamma := p/n$, p : model size (width), n : data size

- Phase transition exists but double descent does not exist
- More close to **U-shaped** instead of double descent: A φ paradigm
- Over-parameterization is still **better than** under-parameterization

Our results under well-behaved data

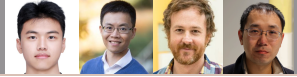


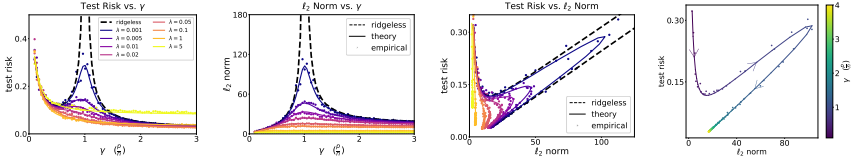


(a) Test Risk vs. γ **(b)** ℓ_2 norm vs. γ **(c)** Test Risk vs. norm **(d)** $\lambda = 0.001$

- $\gamma := p/n$, p : model size (width), n : data size
- Reshape scaling-law:
  test loss = A $\times$ Data Size^{-a} + B $\times$ **Model Size**$^{-b}$ + C with $a, b > 0$

Our results under well-behaved data





(a) Test Risk vs. γ **(b)** ℓ_2 norm vs. γ **(c)** Test Risk vs. norm **(d)** $\lambda = 0.001$

- $\gamma := p/n$, p : model size (width), n : data size
- Reshape scaling-law:
  test loss = A $\times$ Data Size^{-a} + B $\times$ **Model Size**$^{-b}$ + C with $a, b > 0$
  test loss = A $\times$ Data Size^{-a} $\times$ **Norm Capacity**$^{-b}$ with $a > 0$ and $b \in \mathbb{R}$

Control norm by tuning λ: L-curve (Hansen, 1992)

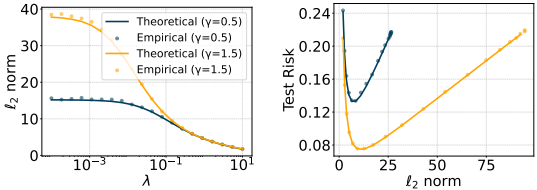
Explicit (model size) *vs.* Implicit (norm)

One-to-one mapping between norm and λ

Control norm by tuning λ: L-curve (Hansen, 1992)

Explicit (model size) *vs.* Implicit (norm)

One-to-one mapping between norm and λ



(a) Norm vs. λ (varying λ) **(b)** Risk vs. Norm (varying λ)

Control norm by tuning λ: L-curve (Hansen, 1992)

Explicit (model size) *vs.* Implicit (norm)

One-to-one mapping between norm and λ

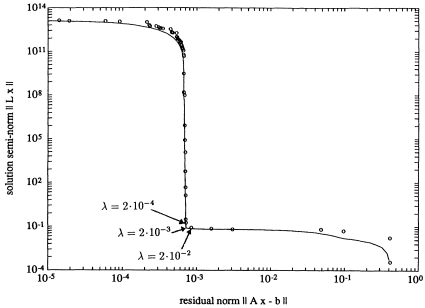


Figure 4: Source from Hansen (1992).

An example of linear regression: Textbook level and beyond

- n i.i.d. samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- $y = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2$, covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$
- ridge regression: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

Target: precise analysis

The expected test risk $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_{\boldsymbol{\Sigma}}^2$ vs. the norm $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}}\|_2^2$

An example of linear regression: Textbook level and beyond

- n i.i.d. samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- $y = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2$, covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$
- ridge regression: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

Target: precise analysis

The expected test risk $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_{\boldsymbol{\Sigma}}^2$ vs. the norm $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}}\|_2^2$

An example of linear regression: Textbook level and beyond

- n i.i.d. samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- $y = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2$, covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$
- ridge regression: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

Target: precise analysis

The expected test risk $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_{\boldsymbol{\Sigma}}^2$ vs. the norm $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}}\|_2^2$

❒ Deterministic equivalence (Cheng and Montanari, 2024; Misiakiewicz and Saeed, 2024; Bach, 2024)

The empirical spectral measure converges to a deterministic limit.

An example of linear regression: Textbook level and beyond

- n i.i.d. samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- $y = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2$, covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$
- ridge regression: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

Target: precise analysis

The expected test risk $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_{\boldsymbol{\Sigma}}^2$ vs. the norm $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}}\|_2^2$

❒ Deterministic equivalence (Cheng and Montanari, 2024; Misiakiewicz and Saeed, 2024; Bach, 2024)

$$\operatorname{Tr}\left(\boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-1}\right) \sim \operatorname{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_*)^{-1}\right), w.h.p.$$

- $\sim$ can be **asymptotic** or **non-asymptotic** at the rate of $\mathcal{O}(1/\sqrt{n})$.
- λ_* is the non-negative solution to the self-consistent equation $n - \frac{\lambda}{\lambda_*} = \operatorname{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_*)^{-1})$.

An example of linear regression: Textbook level and beyond

- n i.i.d. samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ with $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$
- $y = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2$, covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$
- ridge regression: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$

Target: precise analysis

The expected test risk $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_{\boldsymbol{\Sigma}}^2$ vs. the norm $\mathbb{E}_\varepsilon \|\hat{\boldsymbol{\beta}}\|_2^2$

- ❒ Deterministic equivalence (Cheng and Montanari, 2024; Misiakiewicz and Saeed, 2024; Bach, 2024)
- ❒ Bias-variance decomposition on the test risk

- $\mathcal{B}_{\mathcal{R},\lambda}^{\mathrm{LS}} = \lambda^2 \langle \boldsymbol{\beta}_*, (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Sigma} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\beta}_* \rangle$
- $\mathcal{V}_{\mathcal{R},\lambda}^{\mathrm{LS}} = \sigma^2 \mathrm{Tr}(\boldsymbol{\Sigma} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-2})$

Our results

Theorem (asymptotic/non-asymptotic results)

We have a bias-variance decomposition $\mathbb{E}_\varepsilon\|\hat{\boldsymbol{\beta}}\|_2^2 = \mathcal{B}_{\mathcal{N},\lambda} + \mathcal{V}_{\mathcal{N},\lambda}$.

For well-behaved data and Σ, *we have* $\mathcal{B}_{\mathcal{N},\lambda} \sim \mathrm{B}_{\mathrm{N},\lambda}$ *and* $\mathcal{V}_{\mathcal{N},\lambda} \sim \mathrm{V}_{\mathrm{N},\lambda}$, *w.h.p.*

$$\mathrm{B}_{\mathrm{N},\lambda} := \langle \boldsymbol{\beta}_*, \Sigma^2(\Sigma+\lambda_*)^{-2}\boldsymbol{\beta}_*\rangle + \frac{\mathrm{Tr}(\Sigma(\Sigma+\lambda_*)^{-2})}{n} \underbrace{\frac{\lambda_*^2\langle \boldsymbol{\beta}_*, \Sigma(\Sigma+\lambda_*)^{-2}\boldsymbol{\beta}_*\rangle}{1-\frac{1}{n}\mathrm{Tr}(\Sigma^2(\Sigma+\lambda_*)^{-2})}}_{\mathcal{B}_{\mathcal{R},\lambda}},$$

$$\mathrm{V}_{\mathrm{N},\lambda} := \frac{\sigma^2\,\mathrm{Tr}(\Sigma(\Sigma+\lambda_*)^{-2})}{n-\mathrm{Tr}(\Sigma^2(\Sigma+\lambda_*)^{-2})}.$$

Remark: Which model capacity suffices to characterize the test risk?

- Norm-based capacity: ✓ ☺
- effective dimension-style $\mathrm{Tr}(\Sigma(\Sigma+\lambda I)^{-1})$: ✗ ☹

Our results

Theorem (asymptotic/non-asymptotic results)

We have a bias-variance decomposition $\mathbb{E}_\varepsilon\|\hat{\boldsymbol{\beta}}\|_2^2 = \mathcal{B}_{\mathcal{N},\lambda} + \mathcal{V}_{\mathcal{N},\lambda}$.
For well-behaved data and $\boldsymbol{\Sigma}$, *we have* $\mathcal{B}_{\mathcal{N},\lambda} \sim \mathsf{B}_{\mathsf{N},\lambda}$ *and* $\mathcal{V}_{\mathcal{N},\lambda} \sim \mathsf{V}_{\mathsf{N},\lambda}$, *w.h.p.*

$$\mathsf{B}_{\mathsf{N},\lambda} := \langle \boldsymbol{\beta}_*, \boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*)^{-2}\boldsymbol{\beta}_*\rangle + \frac{\mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_*)^{-2})}{n} \underbrace{\frac{\lambda_*^2\langle \boldsymbol{\beta}_*, \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_*)^{-2}\boldsymbol{\beta}_*\rangle}{1-\frac{1}{n}\mathrm{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*)^{-2})}}_{\mathcal{B}_{\mathcal{R},\lambda}},$$

$$\mathsf{V}_{\mathsf{N},\lambda} := \frac{\sigma^2\,\mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_*)^{-2})}{n-\mathrm{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*)^{-2})}.$$

Remark: Which model capacity suffices to characterize the test risk?

- Norm-based capacity: ✓ ☹
- effective dimension-style $\mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda I)^{-1})$: ✗ ☹

Our results

Theorem (asymptotic/non-asymptotic results)

We have a bias-variance decomposition $\mathbb{E}_\varepsilon\|\hat{\boldsymbol{\beta}}\|_2^2 = \mathcal{B}_{\mathcal{N},\lambda} + \mathcal{V}_{\mathcal{N},\lambda}$.
For well-behaved data and $\boldsymbol{\Sigma}$, *we have* $\mathcal{B}_{\mathcal{N},\lambda} \sim \mathsf{B}_{\mathsf{N},\lambda}$ *and* $\mathcal{V}_{\mathcal{N},\lambda} \sim \mathsf{V}_{\mathsf{N},\lambda}$, *w.h.p.*

$$\mathsf{B}_{\mathsf{N},\lambda} := \langle \boldsymbol{\beta}_*, \boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*)^{-2}\boldsymbol{\beta}_*\rangle + \frac{\mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_*)^{-2})}{n} \underbrace{\frac{\lambda_*^2\langle \boldsymbol{\beta}_*, \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_*)^{-2}\boldsymbol{\beta}_*\rangle}{1-\frac{1}{n}\mathrm{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*)^{-2})}}_{\mathcal{B}_{\mathcal{R},\lambda}},$$

$$\mathsf{V}_{\mathsf{N},\lambda} := \frac{\sigma^2\,\mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_*)^{-2})}{n-\mathrm{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*)^{-2})}.$$

Remark: Which model capacity suffices to characterize the test risk?

- Norm-based capacity: ✓ ☺
- effective dimension-style $\mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda I)^{-1})$: ✗ ☹

Example: Relationship under isotropic features ($\Sigma = I_d$)

❒ Test risk R_λ and norm N_λ formulates a cubic curve (complex but precise).

- min-norm interpolator ($\lambda = 0$):

$$R_0 = \begin{cases} N_0 - \|\beta_*\|_2^2; \text{in under-parameterized regimes} \\ \sqrt{\left[N_0 - (\|\beta_*\|_2^2 - \sigma^2)\right]^2 + 4\|\beta_*\|_2^2\sigma^2} - \sigma^2. \end{cases}$$

- optimal regularization $\lambda = \frac{d\sigma^2}{\|\beta^*\|_2^2}$ (Wu and Xu, 2020): $R_\lambda = \|\beta_*\|_2^2 - N_\lambda$

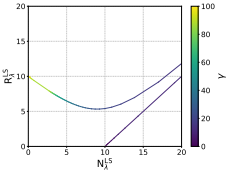
- $\lambda \to \infty$: $R_\lambda = \left(\|\beta_*\|_2 - \sqrt{N_\lambda}\right)^2$

Example: Relationship under isotropic features ($\Sigma = I_d$)

❒ Test risk R_λ and norm N_λ formulates a cubic curve (complex but precise).

- min-norm interpolator ($\lambda = 0$):

$$R_0 = \begin{cases} N_0 - \|\boldsymbol{\beta}_*\|_2^2; \text{in under-parameterized regimes} \\ \sqrt{\left[N_0 - (\|\boldsymbol{\beta}_*\|_2^2 - \sigma^2)\right]^2 + 4\|\boldsymbol{\beta}_*\|_2^2\sigma^2} - \sigma^2 . \end{cases}$$



- optimal regularization $\lambda = \frac{d\sigma^2}{\|\boldsymbol{\beta}^*\|_2^2}$ (Wu and Xu, 2020): $R_\lambda = \|\boldsymbol{\beta}_*\|_2^2 - N_\lambda$
- $\lambda \to \infty$: $R_\lambda = \left(\|\boldsymbol{\beta}_*\|_2 - \sqrt{N_\lambda}\right)^2$

Example: Relationship under isotropic features ($\Sigma = I_d$)

❒ Test risk R_λ and norm N_λ formulates a cubic curve (complex but precise).

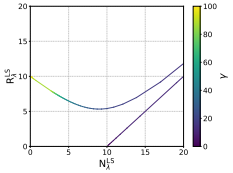
- min-norm interpolator ($\lambda = 0$):

$$R_0 = \begin{cases} N_0 - \|\beta_*\|_2^2; \text{in under-parameterized regimes} \\ \sqrt{\left[N_0 - (\|\beta_*\|_2^2 - \sigma^2)\right]^2 + 4\|\beta_*\|_2^2\sigma^2} - \sigma^2 . \end{cases}$$

Why? ◦ Variance is the same

◦ $\lambda_* = 0$ (under-parameterized)

◦ $\lambda_* = \frac{d-n}{n}$ (over-parameterized)



- optimal regularization $\lambda = \frac{d\sigma^2}{\|\beta^*\|_2^2}$ (Wu and Xu, 2020): $R_\lambda = \|\beta_*\|_2^2 - N_\lambda$

- $\lambda \to \infty$: $R_\lambda = \left(\|\beta_*\|_2 - \sqrt{N_\lambda}\right)^2$

Example: Relationship under isotropic features ($\Sigma = I_d$)

❒ Test risk R_λ and norm N_λ formulates a cubic curve (complex but precise).

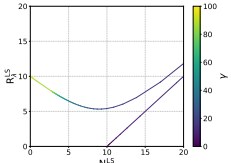
- min-norm interpolator ($\lambda = 0$):

$$R_0 = \begin{cases} N_0 - \|\boldsymbol{\beta}_*\|_2^2; \text{in under-parameterized regimes} \\ \sqrt{\left[N_0 - (\|\boldsymbol{\beta}_*\|_2^2 - \sigma^2)\right]^2 + 4\|\boldsymbol{\beta}_*\|_2^2\sigma^2} - \sigma^2 . \end{cases}$$

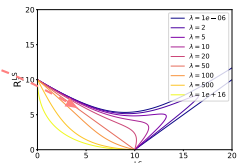
Why? ◦ Variance is the same

◦ $\lambda_* = 0$ (under-parameterized)

◦ $\lambda_* = \frac{d-n}{n}$ (over-parameterized)



- optimal regularization $\lambda = \frac{d\sigma^2}{\|\boldsymbol{\beta}^*\|_2^2}$ (Wu and Xu, 2020): $R_\lambda = \|\boldsymbol{\beta}_*\|_2^2 - N_\lambda$



- $\lambda \to \infty$: $R_\lambda = \left(\|\boldsymbol{\beta}_*\|_2 - \sqrt{N_\lambda}\right)^2$

Example: Relationship under isotropic features ($\Sigma = I_d$)

❒ Test risk R_λ and norm N_λ formulates a cubic curve (complex but precise).

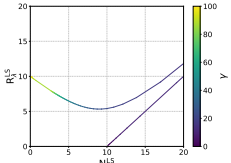
- min-norm interpolator ($\lambda = 0$):

$$R_0 = \begin{cases} N_0 - \|\boldsymbol{\beta}_*\|_2^2; \text{in under-parameterized regimes} \\ \sqrt{\left[N_0 - (\|\boldsymbol{\beta}_*\|_2^2 - \sigma^2)\right]^2 + 4\|\boldsymbol{\beta}_*\|_2^2\sigma^2} - \sigma^2 . \end{cases}$$

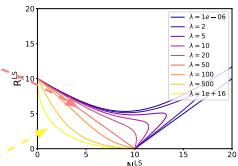
Why? ◦ Variance is the same

◦ $\lambda_* = 0$ (under-parameterized)

◦ $\lambda_* = \frac{d-n}{n}$ (over-parameterized)



- optimal regularization $\lambda = \frac{d\sigma^2}{\|\boldsymbol{\beta}^*\|_2^2}$ (Wu and Xu, 2020): $R_\lambda = \|\boldsymbol{\beta}_*\|_2^2 - N_\lambda$



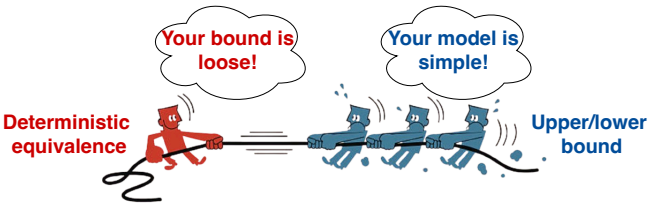
- $\lambda \to \infty$: $R_\lambda = \left(\|\boldsymbol{\beta}_*\|_2 - \sqrt{N_\lambda}\right)^2$

Precise analysis via deterministic equivalence

- Precisely describe the learning curve.
  - phase transitions, (non-)monotonicity, etc.
- Enables *accurate comparison* between estimators/algorithms.
  - Foundations of scaling law: data or parameter under the same budget, etc.

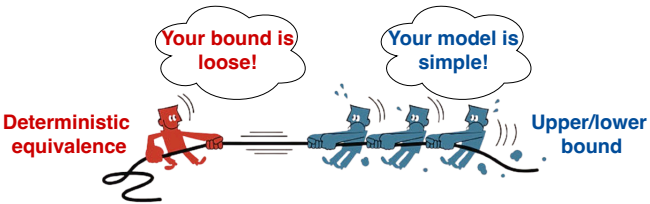
Precise analysis via deterministic equivalence

- Precisely describe the learning curve.
  - phase transitions, (non-)monotonicity, etc.
- Enables *accurate comparison* between estimators/algorithms.
  - Foundations of scaling law: data or parameter under the same budget, etc.



Precise analysis via deterministic equivalence

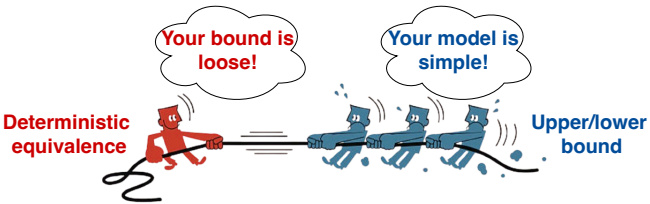
- Precisely describe the learning curve.
  - phase transitions, (non-)monotonicity, etc.
- Enables *accurate comparison* between estimators/algorithms.
  - Foundations of scaling law: data or parameter under the same budget, etc.



$(\gamma, \|\cdot\|_2)$ is sufficient to characterize generalization!

Precise analysis via deterministic equivalence

- Precisely describe the learning curve.
  - phase transitions, (non-)monotonicity, etc.
- Enables *accurate comparison* between estimators/algorithms.
  - Foundations of scaling law: data or parameter under the same budget, etc.



$(\gamma, \|\cdot\|_2)$ is sufficient to characterize generalization!

Is ℓ_2 norm-based capacity **best** for characterizing generalization?

Which model capacity is suitable (for neural networks)?

Table 1: Complexity measures compared in the empirical study (Jiang et al., 2020), and their correlation with generalization.

name	definition	rank correlation
Parameter Frobenius norm	$\sum_{i=1}^{L} \|\boldsymbol{W}_i\|_F^2$	0.073
Frobenius distance to initialization	$\sum_{i=1}^{L} \|\boldsymbol{W}_i - \boldsymbol{W}_i^0\|_F^2$	−0.263
Spectral complexity	$\prod_{i=1}^{L} \|\boldsymbol{W}_i\| \left(\sum_{i=1}^{L} \frac{\|\boldsymbol{W}_i\|_{2,1}^{3/2}}{\|\boldsymbol{W}_i\|^{3/2}} \right)^{2/3}$	−0.537
Fisher-Rao	$\frac{(L+1)^2}{n} \sum_{i=1}^{n} \langle \boldsymbol{W}, \nabla_{\boldsymbol{W}} \ell(h_{\boldsymbol{W}}(\boldsymbol{x}_i), y_i) \rangle$	0.078
Path-norm	$\sum_{(i_0,\ldots,i_L)} \prod_{j=1}^{L} \left(\boldsymbol{W}_{i_j, i_{j-1}} \right)^2$	0.373

Which model capacity is suitable (for neural networks)?

Table 1: Complexity measures compared in the empirical study (Jiang et al., 2020), and their correlation with generalization.

name	definition	rank correlation
Parameter Frobenius norm	$\sum_{i=1}^{L} \|\boldsymbol{W}_i\|_F^2$	0.073
Frobenius distance to initialization	$\sum_{i=1}^{L} \|\boldsymbol{W}_i - \boldsymbol{W}_i^0\|_{\mathrm{F}}^2$	-0.263
Spectral complexity	$\prod_{i=1}^{L} \|\boldsymbol{W}_i\| \left(\sum_{i=1}^{L} \frac{\|\boldsymbol{W}_i\|_{2,1}^{3/2}}{\|\boldsymbol{W}_i\|^{3/2}} \right)^{2/3}$	-0.537
Fisher-Rao	$\frac{(L+1)^2}{n} \sum_{i=1}^{n} \langle \boldsymbol{W}, \nabla_{\boldsymbol{W}} \ell(h_{\boldsymbol{W}}(\boldsymbol{x}_i), y_i) \rangle$	0.078
Path-norm	$\sum_{(i_0,\ldots,i_L)} \prod_{j=1}^{L} \left(W_{j,i_j,i_{j-1}} \right)^2$	0.373

Which model capacity is suitable (for neural networks)?

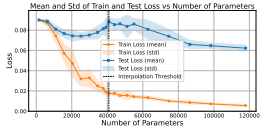
Table 1: Complexity measures compared in the empirical study (Jiang et al., 2020), and their correlation with generalization.

name	definition	rank correlation
Parameter Frobenius norm	$\sum_{i=1}^{L} \|\boldsymbol{W}_i\|_F^2$	0.073
Frobenius distance to initialization	$\sum_{i=1}^{L} \|\boldsymbol{W}_i - \boldsymbol{W}_i^0\|_{\mathrm{F}}^2$	-0.263
Spectral complexity	$\prod_{i=1}^{L} \|\boldsymbol{W}_i\| \left(\sum_{i=1}^{L} \frac{\|\boldsymbol{W}_i\|_{2,1}^{3/2}}{\|\boldsymbol{W}_i\|^{3/2}} \right)^{2/3}$	-0.537
Fisher-Rao	$\frac{(L+1)^2}{n} \sum_{i=1}^{n} \langle \boldsymbol{W}, \nabla_{\boldsymbol{W}} \ell(h_{\boldsymbol{W}}(\boldsymbol{x}_i), y_i) \rangle$	0.078
Path-norm	$\sum_{(i_0,\ldots,i_L)} \prod_{j=1}^{L} \left(\boldsymbol{W}_{i_j, i_{j-1}} \right)^2$	0.373

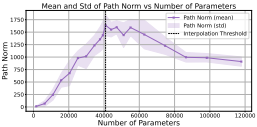
Which model capacity is suitable (for neural networks)?

Table 1: Complexity measures compared in the empirical study (Jiang et al., 2020), and their correlation with generalization.

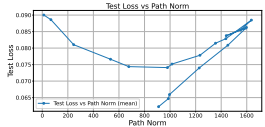
name	definition	rank correlation
Parameter Frobenius norm	$\sum_{i=\mathbf{1}}^{L} \|\boldsymbol{W}_i\|_F^2$	0.073
Frobenius distance to initialization	$\sum_{i=\mathbf{1}}^{L} \|\boldsymbol{W}_i - \boldsymbol{W}_i^{\mathbf{0}}\|_{\mathrm{F}}^2$	-0.263
Spectral complexity	$\prod_{i=\mathbf{1}}^{L} \|\boldsymbol{W}_i\| \left(\sum_{i=\mathbf{1}}^{L} \frac{\|\boldsymbol{W}_i\|_{2,1}^{3/2}}{\|\boldsymbol{W}_i\|^{3/2}} \right)^{2/3}$	-0.537
Fisher-Rao	$\frac{(L+\mathbf{1})^2}{n} \sum_{i=\mathbf{1}}^{n} \langle \boldsymbol{W}, \nabla_{\boldsymbol{W}} \ell(h_{\boldsymbol{W}}(\boldsymbol{x}_i), y_i) \rangle$	0.078
Path-norm	$\sum_{(i_{\mathbf{0}},\ldots,i_L)} \prod_{j=\mathbf{1}}^{L} \left(\boldsymbol{W}_{i_j, i_{j-\mathbf{1}}} \right)^2$	0.373



(a) Test (training) Loss vs. p



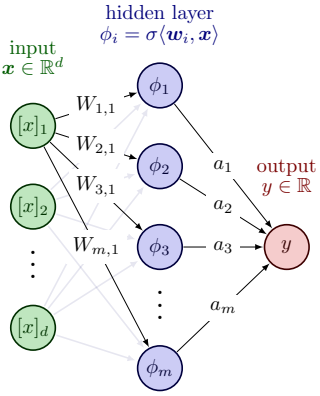
(b) Path-norm vs. p



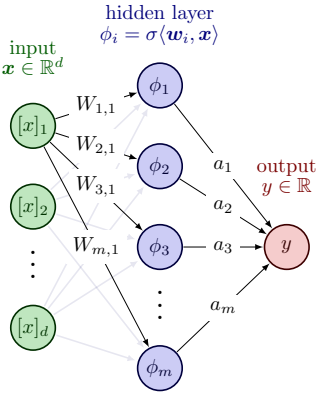
(c) Test Loss vs. Path-norm

Figure 5: Experiments on two-layer neural networks.

Two-layer neural networks, path norm



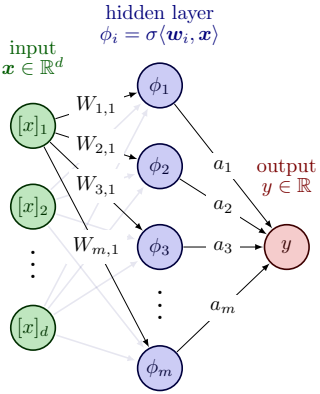
Two-layer neural networks, path norm



ℓ_1-path norm (Neyshabur et al., 2015)

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

Two-layer neural networks, path norm



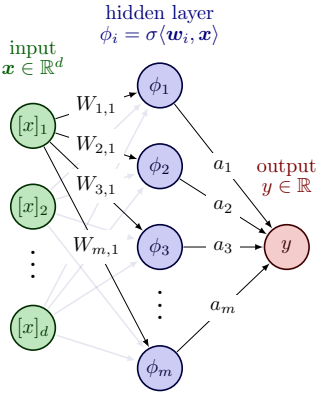
ℓ_1-path norm (Neyshabur et al., 2015)

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

- equivalent to Barron spaces $\mathcal{B}$ (Barron, 1993; E et al., 2021)

$$\mathcal{B} := \cup_{\mu \in \mathcal{P}(\mathcal{W})} \{ f_a : \|\boldsymbol{a}\|_{L^2(\mu)} < \infty \}$$

Two-layer neural networks, path norm



ℓ_1-path norm (Neyshabur et al., 2015)

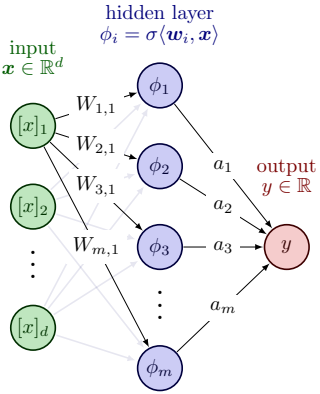
$$\|\boldsymbol{\theta}\|_{\mathcal{P}} := \frac{1}{m}\sum_{k=1}^{m} |a_k| \|\boldsymbol{w}_k\|_1$$

- equivalent to Barron spaces $\mathcal{B}$ (Barron, 1993; E et al., 2021)

$$\mathcal{B} := \cup_{\mu \in \mathcal{P}(\mathcal{W})} \{f_a : \|\boldsymbol{a}\|_{L^2(\mu)} < \infty\}$$

- Variation in only a few directions (Parhi and Nowak, 2022)

Two-layer neural networks, path norm



ℓ_1-path norm (Neyshabur et al., 2015)

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

- equivalent to Barron spaces $\mathcal{B}$ (Barron, 1993; E et al., 2021)

$$\mathcal{B} := \cup_{\mu \in \mathcal{P}(\mathcal{W})} \{ f_a : \|\boldsymbol{a}\|_{L^2(\mu)} < \infty \}$$

- Variation in only a few directions (Parhi and Nowak, 2022)

Can neural networks identify this structure?

(Liu, Dadi, Cevher, JMLR'24)

Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

- *Kernel methods* require $\Omega(\epsilon^{-d})$ samples.
- *Two-layer neural networks* require $\mathcal{O}_d(\epsilon^{-\frac{2d+2}{d+2}})$ samples. smaller than ϵ^{-2}

(Liu, Dadi, Cevher, JMLR'24)

Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

- *Kernel methods* require $\Omega(\epsilon^{-d})$ samples.
- *Two-layer neural networks* require $\mathcal{O}_d(\epsilon^{-\frac{2d+2}{d+2}})$ samples. *smaller than* ϵ^{-2}

(Liu, Dadi, Cevher, JMLR'24)

Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

- *Kernel methods* *require* $\Omega(\epsilon^{-d})$ *samples.*
- *Two-layer neural networks* *require* $\mathcal{O}_d(\epsilon^{-\frac{2d+2}{d+2}})$ *samples. smaller than* ϵ^{-2}

No **Curse of Dimensionality**: NNs adapt to directional smoothness.

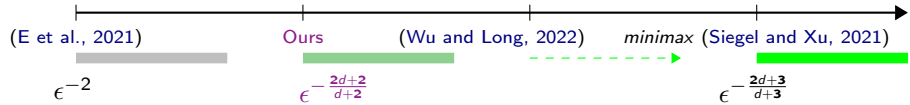
Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

- *Kernel methods* *require* $\Omega(\epsilon^{-d})$ *samples.*
- *Two-layer neural networks* *require* $\mathcal{O}_d(\epsilon^{-\frac{2d+2}{d+2}})$ *samples. smaller than* ϵ^{-2}

No **Curse of Dimensionality**: NNs adapt to directional smoothness.

❒ Track sample complexity (via metric entropy) and dimension dependence



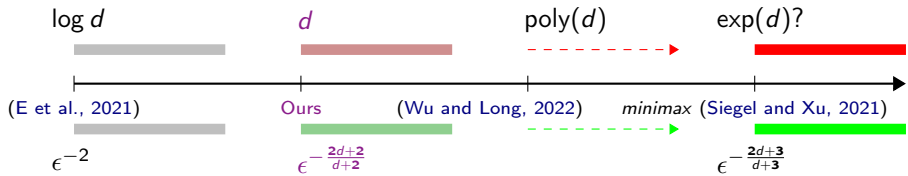
Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

- *Kernel methods* require $\Omega(\epsilon^{-d})$ samples.
- *Two-layer neural networks* require $\mathcal{O}_d(\epsilon^{-\frac{2d+2}{d+2}})$ samples. *smaller than* ϵ^{-2}

No **Curse of Dimensionality**: NNs adapt to directional smoothness.

❒ Track sample complexity (via metric entropy) and dimension dependence



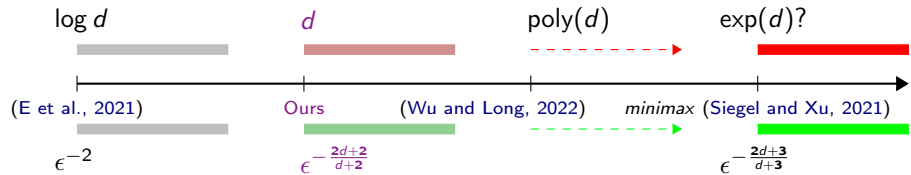
Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

- *Kernel methods* require $\Omega(\epsilon^{-d})$ samples.
- *Two-layer neural networks* require $\mathcal{O}_d(\epsilon^{-\frac{2d+2}{d+2}})$ samples. *smaller than* ϵ^{-2}

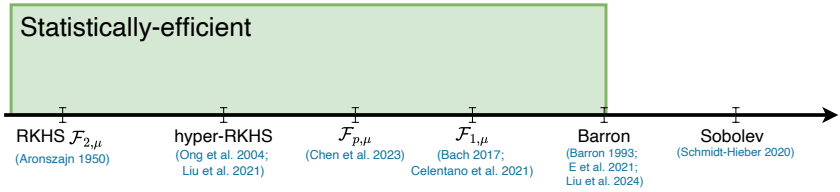
No **Curse of Dimensionality**: NNs adapt to directional smoothness.

❒ Track sample complexity (via metric entropy) and dimension dependence

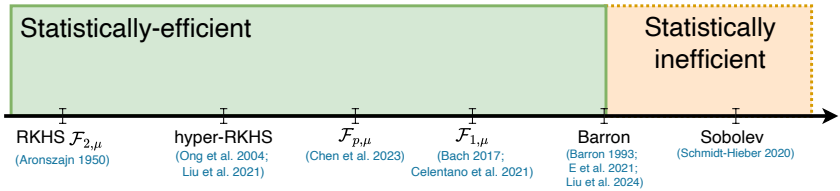


The "best" trade-off between ϵ and d.

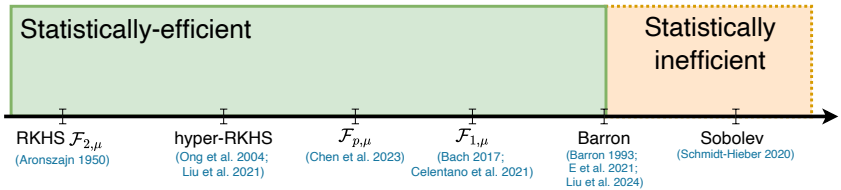
Which function class can be efficiently learned by neural networks



Which function class can be efficiently learned by neural networks



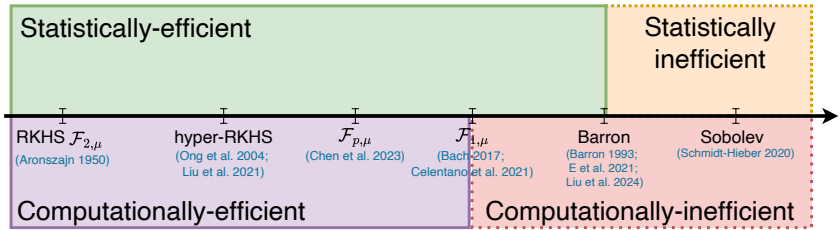
Which function class can be efficiently learned by neural networks



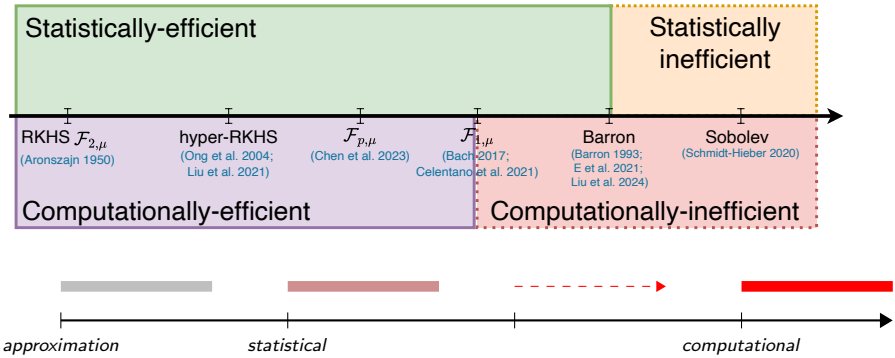
Optimization in Barron spaces is NP hard: curse of dimensionality!
(Bach, 2017)

$$\min_{f \in \mathcal{F}_1} \|f - f^*\|^2_{L^2(\mathrm{d}\mu)} + \lambda \|f\|_{\mathcal{F}_1} .$$

Which function class can be efficiently learned by neural networks



Which function class can be efficiently learned by neural networks

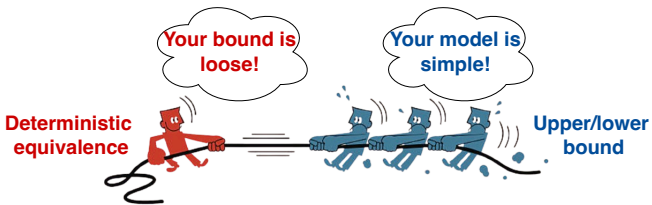


- ReLU neurons (Chen and Narayanan, 2023)
- Low-dimensional polynomials (Arous et al., 2021; Lee et al., 2024)

Takeaway messages

Deep learning phenomena $\Rightarrow$ interesting mathematical problems

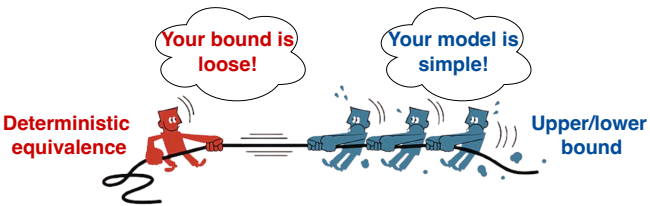
- Be aware of model capacity! A new paradigm of φ curve!
  - Reshape bias-variance trade-offs, double descent, scaling law under proper ℓ_2 norm-based capacity via deterministic equivalence.



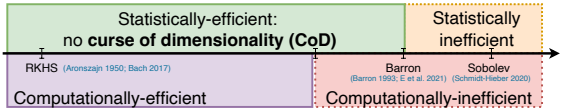
Takeaway messages

Deep learning phenomena $\Rightarrow$ interesting mathematical problems

- Be aware of model capacity! A new paradigm of φ curve!
  - Reshape bias-variance trade-offs, double descent, scaling law under proper ℓ_2 norm-based capacity via deterministic equivalence.



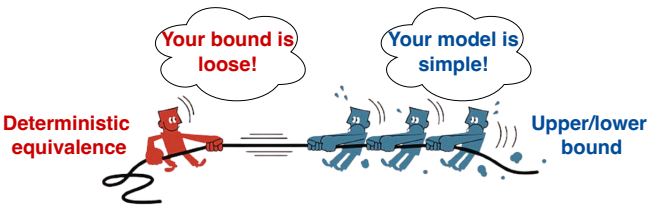
- Which function class can be **efficiently** learned by neural networks?
  - Neural networks can adapt to low-dimensional structure and avoid CoD!



Takeaway messages

Deep learning phenomena ⇒ interesting mathematical problems

- Be aware of model capacity! A new paradigm of φ curve!
  - Reshape bias-variance trade-offs, double descent, scaling law under proper ℓ_2 norm-based capacity via deterministic equivalence.



- Which function class can be **efficiently** learned by neural networks?
  - Neural networks can adapt to low-dimensional structure and avoid CoD!

Theoretical advances ⇒ principled guidance in practical problems

- How does theory contribute to practical fine-tuning problems?
  - One-step full gradient can be sufficient! [ICML'25 oral]

The cherry on top...

WARWICK Berkeley

Statistical Learning Theory in Lean 4: Empirical Processes from Scratch

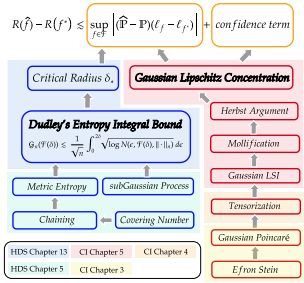
Yuanhe Zhang* Jason D. Lee† Fanghui Liu‡

February 3, 2026

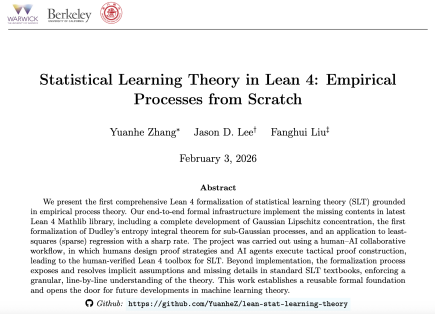
Abstract

We present the first comprehensive Lean 4 formalization of statistical learning theory (SLT) grounded in empirical process theory. Our end-to-end formal infrastructure implement the missing contents in latest Lean 4 Mathlib library, including a complete development of Gaussian Lipschitz concentration, the first formalization of Dudley's entropy integral theorem for sub-Gaussian processes, and an application to least-squares (sparse) regression with a sharp rate. The project was carried out using a human–AI collaborative workflow, in which humans design proof strategies and AI agents execute tactical proof construction, leading to the human-verified Lean 4 toolbox for SLT. Beyond implementation, the formalization process exposes and resolves implicit assumptions and missing details in standard SLT textbooks, enforcing a granular, line-by-line understanding of the theory. This work establishes a reusable formal foundation and opens the door for future developments in machine learning theory.

Github: `https://github.com/YuanheZ/lean-stat-learning-theory`



The cherry on top...



Statistical Learning Theory in Lean 4: Empirical Processes from Scratch

Yuanhe Zhang* Jason D. Lee† Fanghui Liu‡

February 3, 2026

Abstract

We present the first comprehensive Lean 4 formalization of statistical learning theory (SLT) grounded in empirical process theory. Our end-to-end formal infrastructure implement the missing contents in latest Lean 4 Mathlib library, including a complete development of Gaussian Lipschitz concentration, the first formalization of Dudley's entropy integral theorem for sub-Gaussian processes, and an application to least-squares (sparse) regression with a sharp rate. The project was carried out using a human–AI collaborative workflow, in which humans design proof strategies and AI agents execute tactical proof construction, leading to the human-verified Lean 4 toolbox for SLT. Beyond implementation, the formalization process exposes and resolves implicit assumptions and missing details in standard SLT textbooks, enforcing a granular, line-by-line understanding of the theory. This work establishes a reusable formal foundation and opens the door for future developments in machine learning theory.

Github: https://github.com/YuanheZ/lean-stat-learning-theory

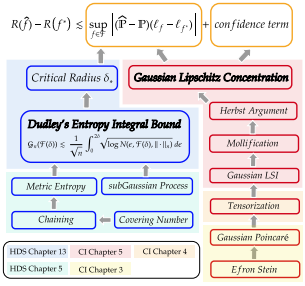


Figure credit: Canva

References

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *Journal of Machine Learning Research*, 22(106):1–51, 2021.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(1):629–681, 2017.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6 (1):26–50, 2024.

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 classification 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.

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

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

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52(6):2879–2912, 2024.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents for random feature regression. In *Advances in Neural Information Processing Systems*, 2024.

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

Per Christian Hansen. Analysis of discrete ill-posed problems by means of the l-curve. *SIAM Review*, 34(4):561–580, 1992.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find 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, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711, 2020.

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.

Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and gcv estimator. *arXiv preprint arXiv:2403.08938*, 2024.

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

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.

Andrew Ng and Tengyu Ma. CS229 lecture notes. 2023. URL https://cs229.stanford.edu/main_notes.pdf.

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

Rahul Parhi and Robert D Nowak. Near-minimax optimal estimation with shallow ReLU neural networks. *IEEE Transactions on Information Theory*, 2022.

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

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite 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.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, pages 10112–10123, 2020.

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.