Be aware of model capacity when talking about generalization in machine learning

Fanghui Liu

fanghui.liu@warwick.ac.uk

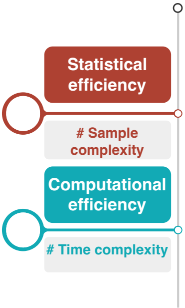
Department of Computer Science, University of Warwick, UK
Centre for Discrete Mathematics and its Applications (DIMAP), Warwick

at HKUST MATH@5470

My research

❒ Research interests

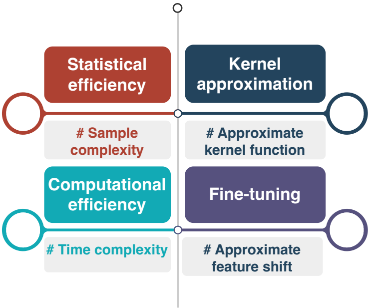
- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML



My research

❒ Research interests

- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML



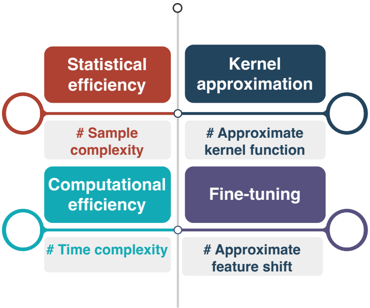
My research

❐ Research interests

- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML

❐ Research goal

- characterize **learning efficiency** in theory
- contribute to practice



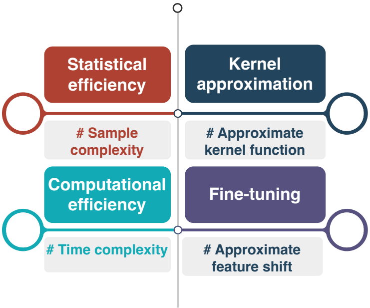
My research

❒ Research interests

- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML

❒ Research goal

- characterize **learning efficiency** in theory
- contribute to practice



Learning efficiency (Curse of Dimensionality, CoD)

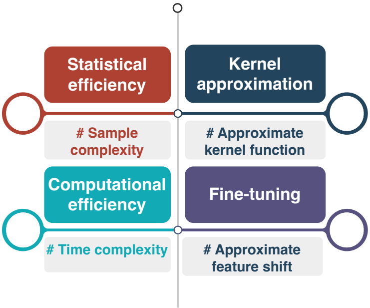
Machine learning works in **high dimensions** that can be a **curse**!

— David Donoho, 2000. (Richard E. Bellman, 1957)

My research

❒ Research interests

- Foundations of machine learning (ML)
- Theory-grounded efficient algorithm design
- Trustworthy ML



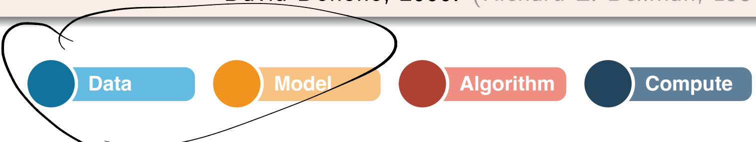
❒ Research goal

- characterize **learning efficiency** in theory
- contribute to practice

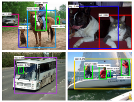
Learning efficiency (Curse of Dimensionality, CoD)

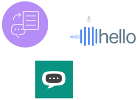
Machine learning works in **high dimensions** that can be a **curse**!

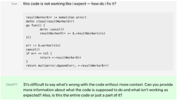
— David Donoho, 2000. (Richard E. Bellman, 1957)

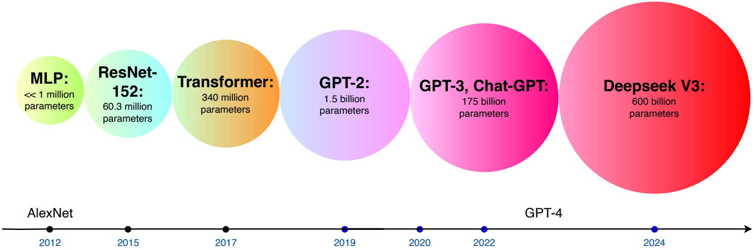


Prefer more data and larger model to obtain better performance...

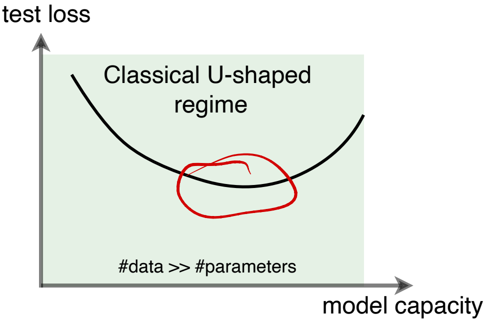






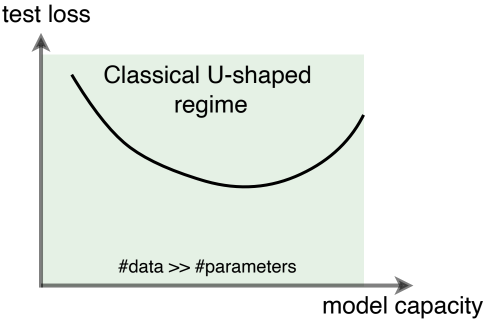


ML textbooks: Larger models tend to overfit!

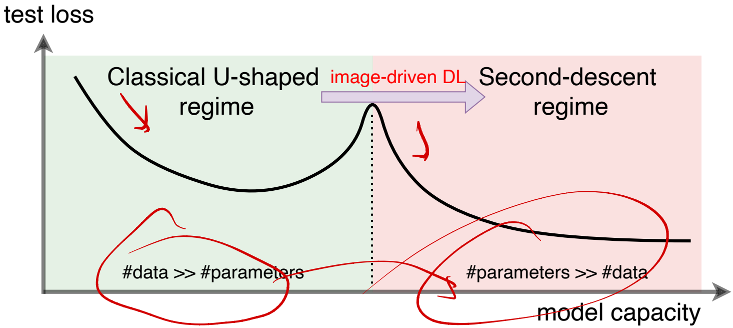


ML textbooks: Larger models tend to overfit!

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

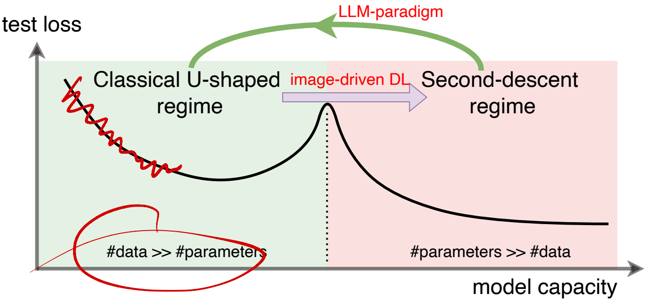


Figure 1: Paradigm among test loss, data, and model capacity.

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

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

Learning paradigm in the past twenty years

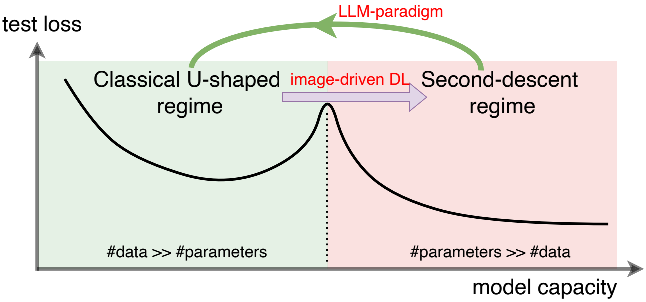


Figure 1: Paradigm among test loss, data, and model capacity.

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

test loss = A × Model Size^{-a} + B × Data Size^{-b} + C

A fundamental concept in machine learning: model capacity

Too many learning curves...

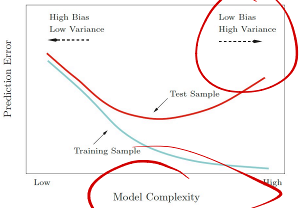
- U-shaped curve (bias-variance trade-offs) (Vapnik, 1995; Hastie et al., 2009)
- double (multiple) descent (Belkin et al., 2019; Liang et al., 2020)
- scaling law (Kaplan et al., 2020; Paquette et al., 2024)

Too many learning curves...

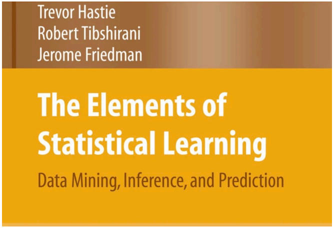
- U-shaped curve (bias-variance trade-offs) (Vapnik, 1995; Hastie et al., 2009)
- double (multiple) descent (Belkin et al., 2019; Liang et al., 2020)
- scaling law (Kaplan et al., 2020; Paquette et al., 2024)

Bias-variance decomposition

$$\text{Test error} = \text{Bias}^2 + \text{Variance}$$



(Hastie et al., 2009, Figure 2.11)



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)

Bias-variance decomposition

$$\text{Test error} = \text{Bias}^2 + \text{Variance}$$

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

Trade-off is a **misnomer**, by Geman et al. (1992); Neal (2019); Wilson (2025).

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

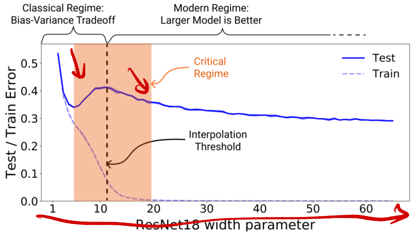
— Ben Recht, 2025

A fundamental concept in machine learning: model capacity

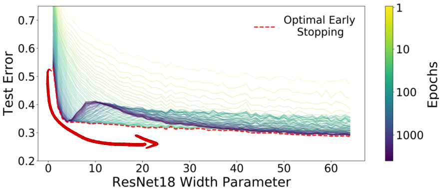
Too many learning curves...

- 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: Learning with norm-based capacity

$W \in \mathbb{R}^{d \times m}$

$\|W\|$

Today's talk: Learning with norm-based capacity

(Bartlett, 1998)

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

$||W||$

Today's talk: Learning with norm-based capacity

(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), lottery ticket hypothesis (Frankle and Carbin, 2019)

Today's talk: Learning with norm-based capacity

(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), lottery ticket hypothesis (Frankle and Carbin, 2019)

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

(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), lottery ticket hypothesis (Frankle and Carbin, 2019)

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

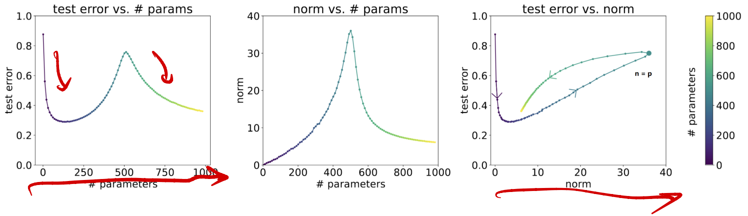


Figure 3: Stanford CS229 lecture notes (Ng and Ma, 2023, Figure 8.12).

Today's talk: Learning with norm-based capacity

(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**. *Re-examining double descent and scaling laws under norm-based capacity via deterministic equivalence.* 2025. arXiv

Today's talk: Learning with norm-based capacity

(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**. *Re-examining double descent and scaling laws under norm-based capacity via deterministic equivalence.* 2025. arXiv

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

Today's talk: Learning with norm-based capacity

(Bartlett, 1998)

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

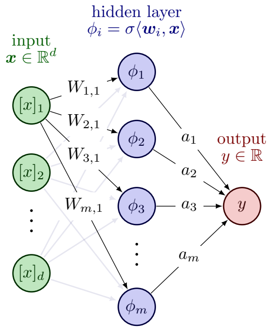
❒ How to **precisely** characterize the relationship under norm-based model capacity?

- Reshape bias-variance trade-offs, double descent, scaling law under ℓ_2 norm-based capacity!
- Yichen Wang, Yudong Chen, Lorenzo Rosasco, **Fanghui Liu**. *Re-examining double descent and scaling laws under norm-based capacity via deterministic equivalence.* 2025. arXiv

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

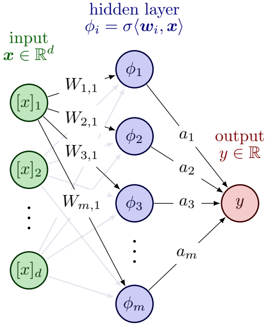
- Which function class can be **efficiently** learned by neural networks?
- **Fanghui Liu**, Leello Dadi, and Volkan Cevher. *Learning with norm constrained, over-parameterised, two-layer neural networks.* JMLR 2024.

Background: Random features model, two-layer neural networks



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

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

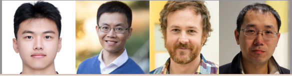


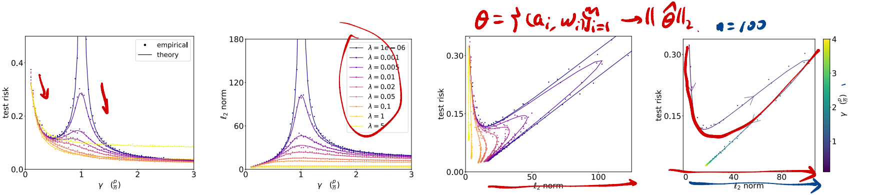
$m < n$.

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

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

(Wang, Chen, Rosasco, Liu, 2025)

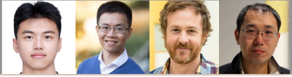


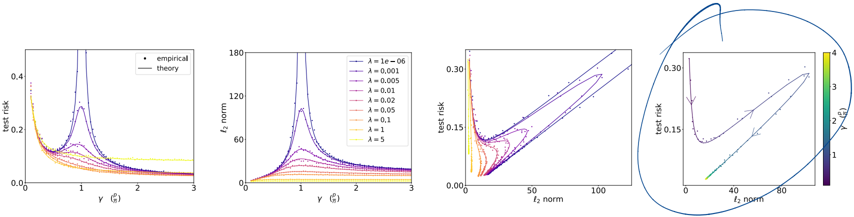


(a) Test Risk vs. γ **(b)** ℓ_2 norm vs. γ **(c)** Test Risk vs. norm **(d)** $\lambda = 0.001$

- $\gamma := p/n$, p : model size (width), n : data size
- Phase transition exists but double descent does not exist
- 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 $b \in \mathbb{R}$
- Over-parameterization is still **better than** under-parameterization

(Wang, Chen, Rosasco, Liu, 2025)





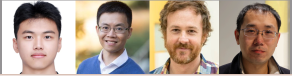
(a) Test Risk vs. γ **(b)** ℓ_2 norm vs. γ **(c)** Test Risk vs. norm **(d)** $\lambda = 0.001$

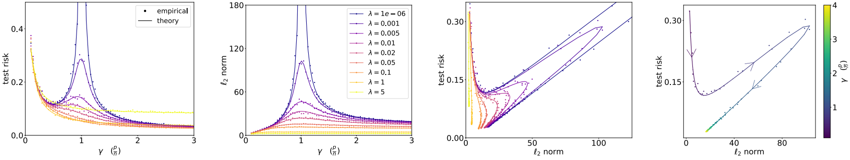
- $\gamma := p/n$, p : model size (width), n : data size
- Phase transition exists but double descent does not exist
- Reshape scaling-law:

  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 $b \in \mathbb{R}$
- Over-parameterization is still **better than** under-parameterization

(Wang, Chen, Rosasco, Liu, 2025)





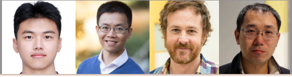
(a) Test Risk vs. γ **(b)** ℓ_2 norm vs. γ **(c)** Test Risk vs. norm **(d)** $\lambda = 0.001$

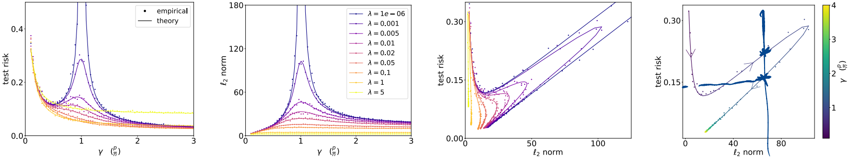
- $\gamma := p/n$, p : model size (width), n : data size
- Phase transition exists but double descent does not exist
- Reshape scaling-law:
  test loss = A $\times$ Data Size^{-a} + B $\times$ **Model Size**$^{-b}$ + C with $a, b > 0$
  test loss = A $\times$ Data Size^{-a} $\times$ **Norm Capacity**$^{-b}$ with $a > 0$ and $b \in \mathbb{R}$
- Over-parameterization is still **better than** under-parameterization

$\alpha \in (0, +\infty)$

$p = n^{\alpha}$

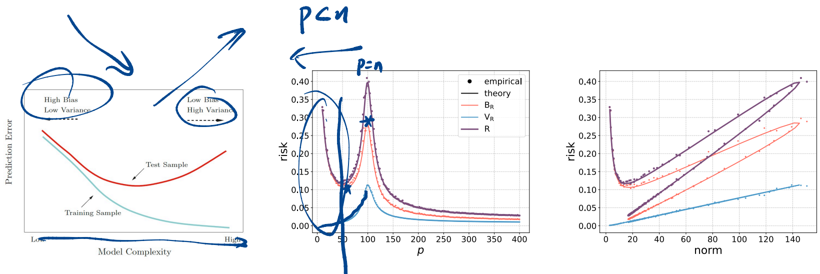
(Wang, Chen, Rosasco, Liu, 2025)





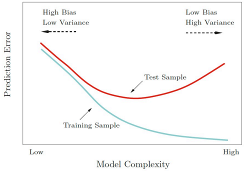
(a) Test Risk vs. γ **(b)** ℓ_2 norm vs. γ **(c)** Test Risk vs. norm **(d)** $\lambda = 0.001$

- $\gamma := p/n$, p : model size (width), n : data size
- Phase transition exists but double descent does not exist
- 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 $b \in \mathbb{R}$
- Over-parameterization is still **better than** under-parameterization

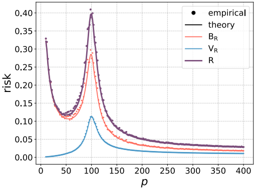


(a) Bias-variance trade-offs **(b)** Test risk vs. p **(c)** Test risk vs. Norm

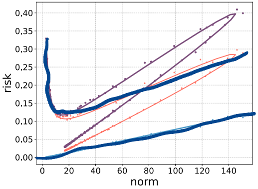
Under **small** model capacity (e.g., model size and ℓ_2 norm-based capacity), bias $\searrow$ and variance $\nearrow$.



(a) Bias-variance trade-offs



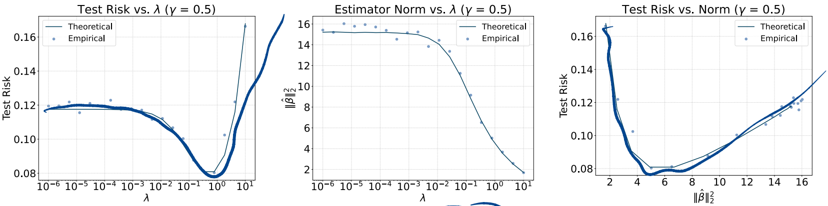
(b) Test risk vs. p



(c) Test risk vs. Norm

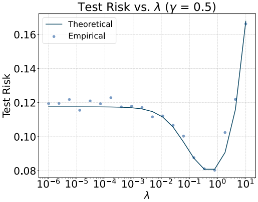
Under **small** model capacity (e.g., model size and ℓ_2 norm-based capacity), bias $\searrow$ and variance $\nearrow$.

Precise analysis: L-curve (Hansen, 1992)

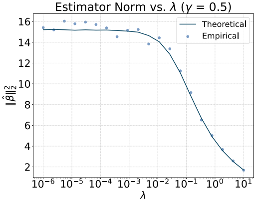


(a) Test risk vs. λ **(b)** Norm vs. λ **(c)** Test risk vs. Norm

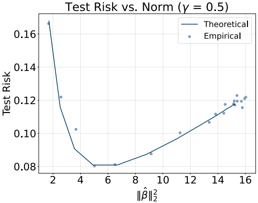
Precise analysis: L-curve (Hansen, 1992)



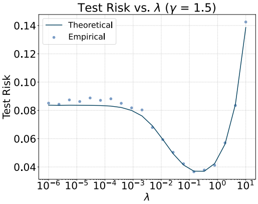
(a) Test risk vs. λ



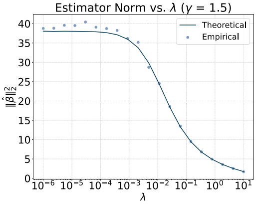
(b) Norm vs. λ



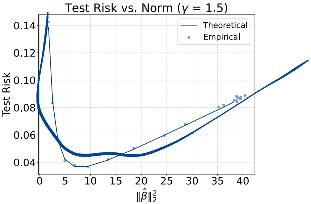
(c) Test risk vs. Norm



(d) Test risk vs. λ



(e) Norm vs. λ



(f) Test risk vs. Norm

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 $\mathbb{E}_\varepsilon \|\hat{\beta} - \beta_*\|_{\Sigma}^2$ vs. the norm $\mathbb{E}_\varepsilon \|\hat{\beta}\|_2^2$

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): law of large samples/dimensions in random matrix theory

The empirical spectral measure converges to a deterministic limit.

$x_i \sim \mathcal{N}(0, I_d)$

$\sum_{i=1}^{d} \lambda_i(X^\top X)$

$\| \frac{1}{n} X^\top X - I_d \|_{op} = O(\sqrt{\frac{d}{n}})$ w.h.p.

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): law of large samples/dimensions in random matrix theory

$V = \sigma^2 \mathrm{Tr}(\Sigma X^\top X (X^\top X + \lambda I)^{-2})$

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

- $\sim$ 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 $n - \frac{\lambda}{\lambda_*} = \mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_*)^{-1})$.

$\lambda = 0 \to \lambda_* = 0$

$\lambda_* \to C$

Our results

Theorem (Deterministic equivalence of estimator's norm)

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

For well-behaved data, we have

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

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

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

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

Theorem (Deterministic equivalence of estimator's norm)

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

For well-behaved data, we have

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

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

$B_{R,\lambda}$

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

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

Theorem (Deterministic equivalence of estimator's norm)

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

For well-behaved data, we have

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

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

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 ($\lambda = 0$):

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

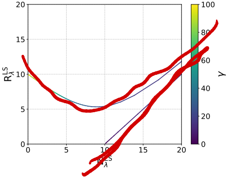
- optimal regularization $\lambda = \frac{d\sigma^2}{\|\beta^*\|_2^2}$ (Wu and Xu, 2020): $R_\lambda = \|\beta_*\|_2^2 - N_\lambda$

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

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

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

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



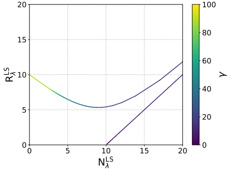
- optimal regularization $\lambda = \frac{d\sigma^2}{\|\beta^*\|_2^2}$ (Wu and Xu, 2020): $R_\lambda = \|\beta_*\|_2^2 - N_\lambda$

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

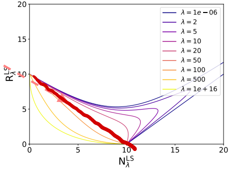
❒ Test risk R_λ and norm N_λ formulates a cubic curve (complex but precise).

- 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): $R_\lambda = \|\boldsymbol{\beta}_*\|_2^2 - N_\lambda$



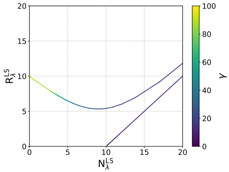
- $\lambda \to \infty$: $R_\lambda = \left(\|\boldsymbol{\beta}_*\|_2 - \sqrt{N_\lambda}\right)^2$

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

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

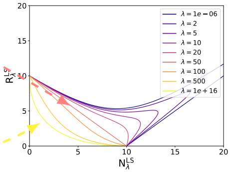
- min-norm interpolator ($\lambda = 0$):

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



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

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



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.

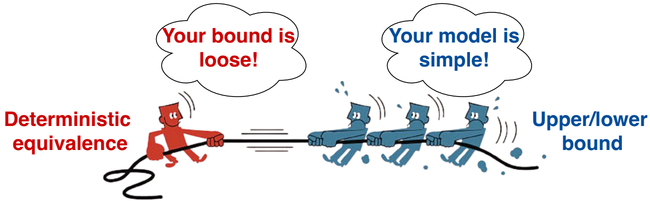


Table 1: Complexity measures compared in the empirical study (Jiang et al., 2020), and their correlation with generalization.

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	$\sum_{(i_0,\dots,i_L)} \prod_{j=1}^{L} \left(\boldsymbol{W}_{i_j, i_{j-1}} \right)^2$	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	$\sum_{i=1}^{L} \lVert \boldsymbol{W}_i \rVert_F^2$	0.073
Frobenius distance to initialization	$\sum_{i=1}^{L} \lVert \boldsymbol{W}_i - \boldsymbol{W}_i^0 \rVert_{\mathrm{F}}^2$	-0.263
Spectral complexity	$\prod_{i=1}^{L} \lVert \boldsymbol{W}_i \rVert \left(\sum_{i=1}^{L} \frac{\lVert \boldsymbol{W}_i \rVert_{2,1}^{3/2}}{\lVert \boldsymbol{W}_i \rVert^{3/2}} \right)^{2/3}$	-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

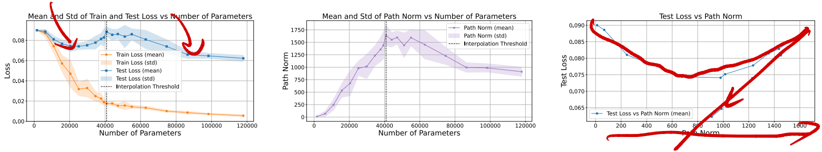
Which model capacity is suitable (for neural networks)?

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	$\sum_{i=1}^{L} \lVert \boldsymbol{W}_i \rVert_F^2$	0.073
Frobenius distance to initialization	$\sum_{i=1}^{L} \lVert \boldsymbol{W}_i - \boldsymbol{W}_i^0 \rVert_F^2$	-0.263
Spectral complexity	$\prod_{i=1}^{L} \lVert \boldsymbol{W}_i \rVert \left(\sum_{i=1}^{L} \frac{\lVert \boldsymbol{W}_i \rVert_{2,1}^{3/2}}{\lVert \boldsymbol{W}_i \rVert^{3/2}} \right)^{2/3}$	-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	$\sum_{(i_0,\dots,i_L)} \prod_{j=1}^{L} \left(\boldsymbol{W}_{i_j, i_{j-1}} \right)^2$	0.373

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

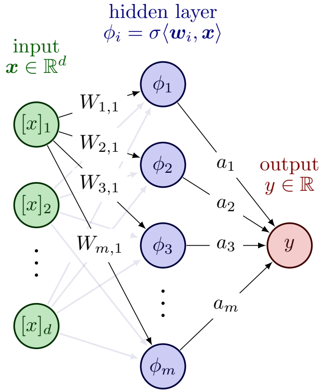
name	definition	rank correlation
Parameter Frobenius norm	$\sum_{i=1}^{L} \lVert \boldsymbol{W}_i \rVert_F^2$	0.073
Frobenius distance to initialization	$\sum_{i=1}^{L} \lVert \boldsymbol{W}_i - \boldsymbol{W}_i^0 \rVert_{\mathrm{F}}^2$	−0.263
Spectral complexity	$\prod_{i=1}^{L} \lVert \boldsymbol{W}_i \rVert \left(\sum_{i=1}^{L} \frac{\lVert \boldsymbol{W}_i \rVert_{2,1}^{3/2}}{\lVert \boldsymbol{W}_i \rVert^{3/2}} \right)^{2/3}$	−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



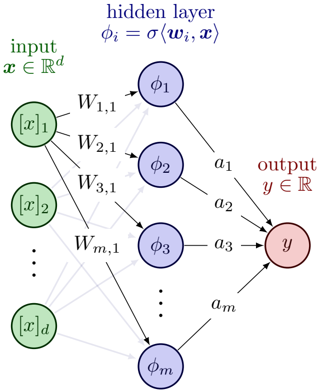
(a) Test (training) Loss vs. p **(b)** Path-norm vs. p **(c)** Test Loss vs. Path-norm

Figure 6: Experiments on two-layer neural networks.

Two-layer neural networks, path norm



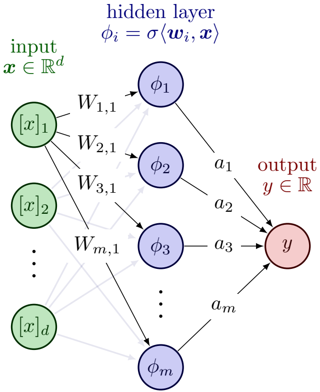
Two-layer neural networks, path norm



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)

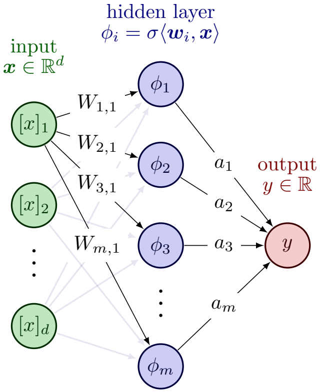
$\mathcal{B} := \cup_{\mu \in \mathcal{P}(\mathcal{W})} \{f_a : \|\boldsymbol{a}\|_{L^2(\mu)} < \infty\}$

$f_a(x) = \int a(\omega)\, \sigma(\langle \omega, x\rangle)\, d\mu(\omega)$

$\mathcal{F}_{p,\mu} = \{f_a : \|a\|_{L^p(\mu)} < +\infty\}$

$p=2$. $\mathcal{F}_{2,\mu}$

$p=1$. $\mathcal{F}_{1,\mu}$



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

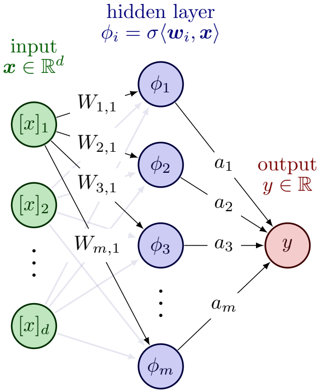
$\|\boldsymbol{\theta}\|_{\mathcal{P}} := \frac{1}{m} \sum_{k=1}^{m} |a_k| \|\boldsymbol{w}_k\|_1$

- 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)

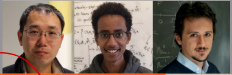
$\|\boldsymbol{\theta}\|_{\mathcal{P}} := \frac{1}{m}\sum_{k=1}^{m} |a_k| \|\boldsymbol{w}_k\|_1$

- 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)

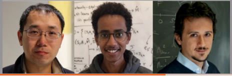
Can neural networks identify this structure?



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

To achieve ϵ-excess risk,

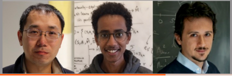
- *Kernel methods* require $\Omega(\epsilon^{-d})$ samples.
- *Two-layer neural networks* require $\Omega(\epsilon^{-\frac{2d+2}{d+2}})$ samples. smaller than ϵ^{-2}



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 $\Omega(\epsilon^{-\frac{2d+2}{d+2}})$ samples. smaller than ϵ^{-2}*

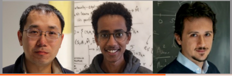


Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

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

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



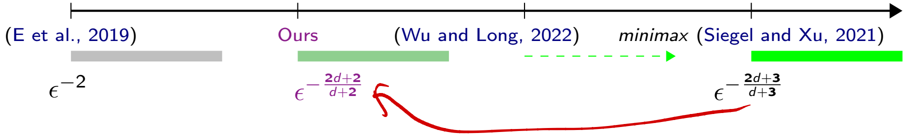
Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

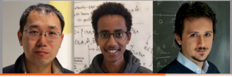
- *Kernel methods require $\Omega(\epsilon^{-d})$ samples.*
- *Two-layer neural networks require $\Omega(\epsilon^{-\frac{2d+2}{d+2}})$ samples. smaller than ϵ^{-2}*

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

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



(Liu, Dadi, Cevher, JMLR'24)



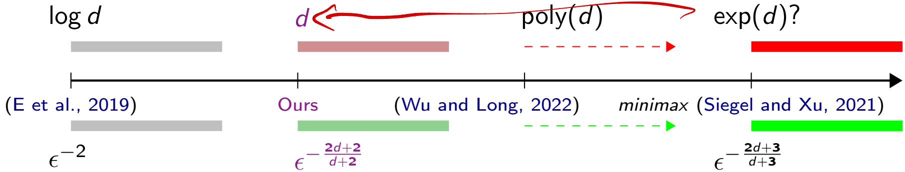
Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

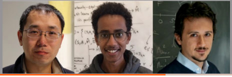
- *Kernel methods require $\Omega(\epsilon^{-d})$ samples.*
- *Two-layer neural networks require $\Omega(\epsilon^{-\frac{2d+2}{d+2}})$ samples.* *smaller than ϵ^{-2}*

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

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



(Liu, Dadi, Cevher, JMLR'24)



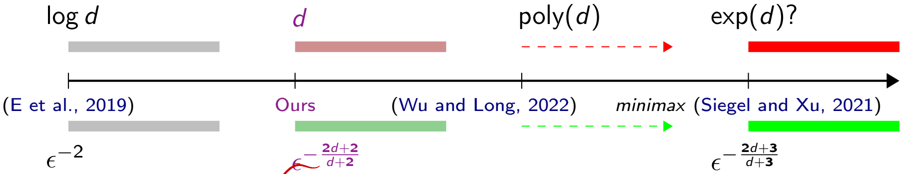
Theorem (Informal, sample complexity of learning $f^\star \in \mathcal{B}$)

To achieve ϵ-excess risk,

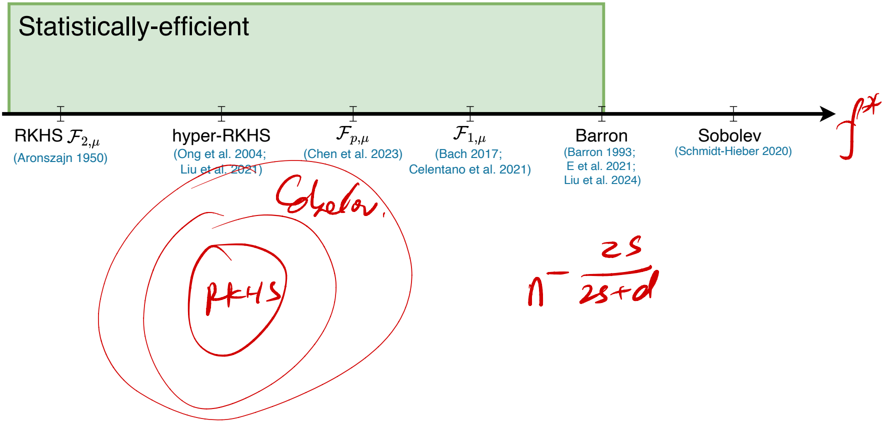
- *Kernel methods require $\Omega(\epsilon^{-d})$ samples.*
- *Two-layer neural networks require $\Omega(\epsilon^{-\frac{2d+2}{d+2}})$ samples.* *smaller than ϵ^{-2}*

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

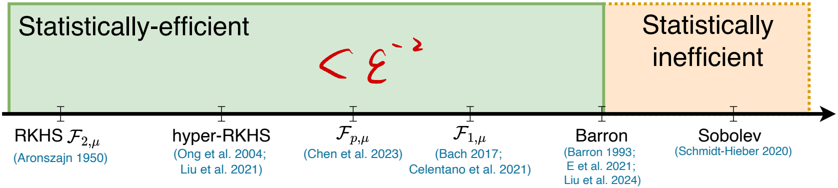
❒ Track sample complexity (via metric entropy) and dimension dependence



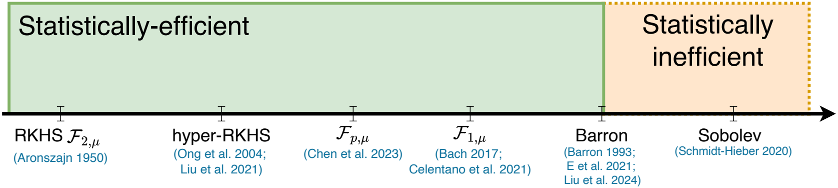
The "best" trade-off between ϵ and d.


Statistically-efficient
RKHS $\mathcal{F}_{2,\mu}$
(Aronszajn 1950)
hyper-RKHS
(Ong et al. 2004;
Liu et al. 2021)
$\mathcal{F}_{p,\mu}$
(Chen et al. 2023)
$\mathcal{F}_{1,\mu}$
(Bach 2017;
Celentano et al. 2021)
Barron
(Barron 1993;
E et al. 2021;
Liu et al. 2024)
Sobolev
(Schmidt-Hieber 2020)
f^*
Sobolev.
RKHS
$n^{-\frac{2s}{2s+d}}$

Which function class can be efficiently learned by neural networks



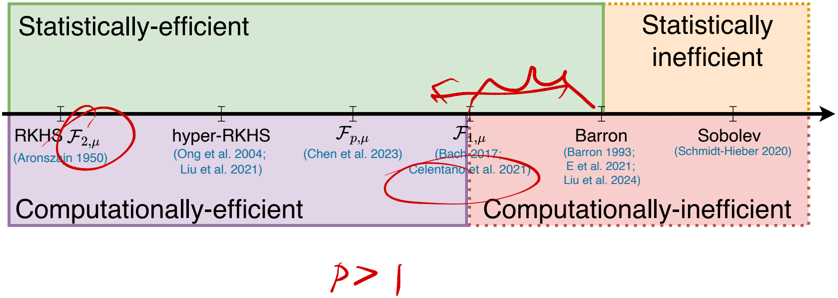
Which function class can be efficiently learned by neural networks



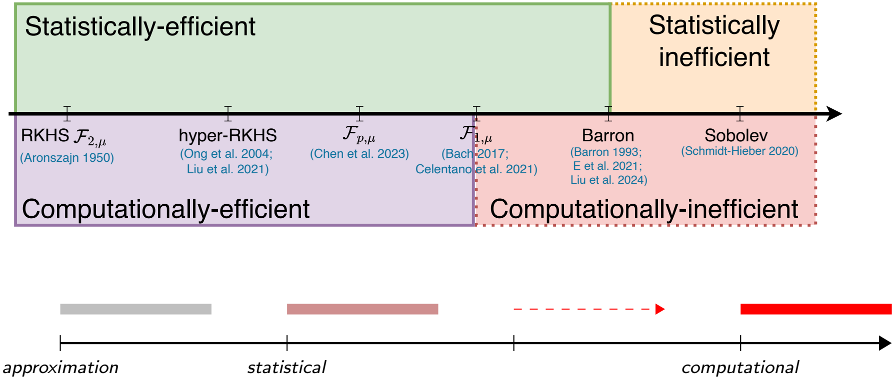
Optimization in Barron spaces is NP hard: curse of dimensionality!
(Bach, 2017)

2^d

Which function class can be efficiently learned by neural networks



Which function class can be efficiently learned by neural networks



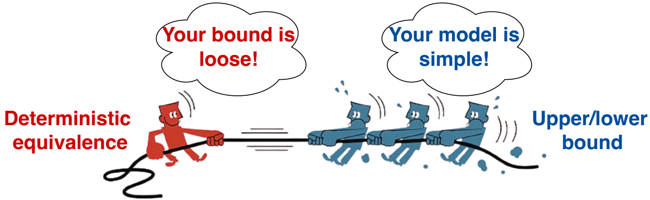
- ReLU neurons (Chen and Narayanan, 2023)
- Low-dimensional polynomials (Arous et al., 2021; Lee et al., 2024)

Takeaway messages

Deep learning phenomena ⇒ interesting mathematical problems

❒ Be aware of model capacity!

- Reshape bias-variance trade-offs, double descent, scaling law under proper ℓ_2 norm-based capacity via deterministic equivalence.

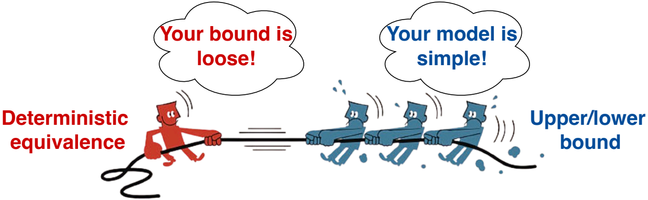


Takeaway messages

Deep learning phenomena $\Rightarrow$ interesting mathematical problems

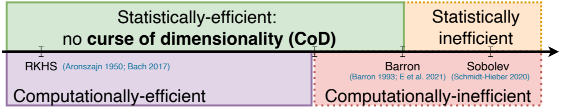
❒ Be aware of model capacity!

- 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!

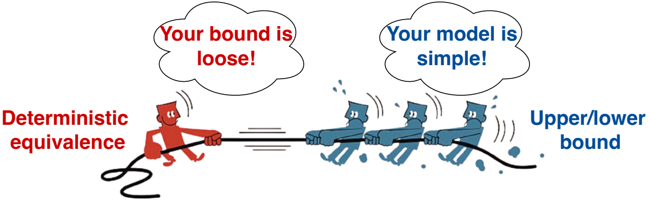


Takeaway messages

Deep learning phenomena ⇒ interesting mathematical problems

❒ Be aware of model capacity!

- 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 our theory contribute to practical fine-tuning problems?

- One-step full gradient can be sufficient! [GitHub]

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.

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.

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.

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

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 flow-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.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

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.

References v

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.

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.

Brady Neal. On the bias-variance tradeoff: Textbooks need an update. *arXiv preprint arXiv:1912.08286*, 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.

Aad W Van Der Vaart, Adrianus Willem van der Vaart, Aad van der Vaart, and Jon Wellner. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

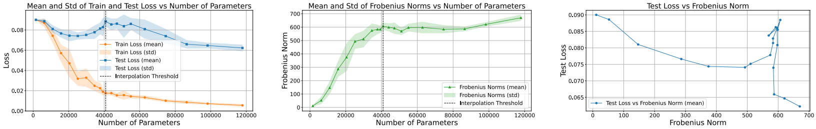
Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

Andrew Gordon Wilson. Deep learning is not so mysterious or different. *arXiv preprint arXiv:2503.02113*, 2025.

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.

Experimental results



(a) Test (training) Loss vs. p **(b)** Fro-norm vs. p **(c)** Test Loss vs. Fro-norm

Figure 7: Experiments on two-layer fully connected neural networks with noise level $\eta = 0.2$. The **left** figure shows the relationship between test (training) loss and the number of the parameters p. The **middle** figure shows the relationship between the Frobenius norm and p. The **right** figure shows the relationship between the test loss and Fro-norm.

An example of linear model: a textbook level

- $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$
- $y = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2$
- ridge regression: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$
- min-ℓ_2-norm interpolation: $\hat{\boldsymbol{\beta}}_{\min} = \operatorname{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2, \text{s.t. } \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$
- expected test risk: bias-variance decomposition

$$\mathcal{R}^{\mathrm{LS}} := \mathbb{E}_\varepsilon \|\beta_* - \hat{\beta}\|_{\Sigma}^2 = \underbrace{\|\beta_* - \mathbb{E}_\varepsilon[\hat{\beta}]\|_{\Sigma}^2}_{:=\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda}} + \underbrace{\mathrm{tr}(\boldsymbol{\Sigma}\mathrm{Cov}_\varepsilon(\hat{\beta}))}_{:=\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda}} .$$

- $\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda} = \lambda^2 \langle \beta_*, (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \Sigma (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \beta_* \rangle$
- $\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda} = \sigma^2 \mathrm{Tr}(\Sigma \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-2})$
- *Intuitive fact: for i.i.d. sub-Gaussian data $\boldsymbol{X}$, we have

$$\|\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} - \Sigma\|_{op} = \Theta(\sqrt{d/n}) \, , w.h.p \, .$$

An example of linear model: a textbook level

- $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$
- $y = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2$
- ridge regression: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$
- min-ℓ_2-norm interpolation: $\hat{\boldsymbol{\beta}}_{\min} = \operatorname{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2$, s.t. $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$
- expected test risk: bias-variance decomposition

$$\mathcal{R}^{\mathrm{LS}} := \mathbb{E}_\varepsilon \|\boldsymbol{\beta}_* - \hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}}^2 = \underbrace{\|\boldsymbol{\beta}_* - \mathbb{E}_\varepsilon[\hat{\boldsymbol{\beta}}]\|_{\boldsymbol{\Sigma}}^2}_{:=\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda}} + \underbrace{\mathtt{tr}(\boldsymbol{\Sigma}\mathtt{Cov}_\varepsilon(\hat{\boldsymbol{\beta}}))}_{:=\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda}} .$$

- $\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda} = \lambda^2 \langle \beta_*, (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \Sigma (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \beta_* \rangle$
- $\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda} = \sigma^2 \mathrm{Tr}(\Sigma \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-2})$
- *Intuitive fact: for i.i.d. sub-Gaussian data $\boldsymbol{X}$, we have

$$\|\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} - \Sigma\|_{op} = \Theta(\sqrt{d/n}), w.h.p.$$

An example of linear model: a textbook level

- $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$
- $y = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2$
- ridge regression: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$
- min-ℓ_2-norm interpolation: $\hat{\boldsymbol{\beta}}_{\min} = \mathtt{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2$, s.t. $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$
- expected test risk: bias-variance decomposition

$$\mathcal{R}^{\mathrm{LS}} := \mathbb{E}_\varepsilon \|\boldsymbol{\beta}_* - \hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}}^2 = \underbrace{\|\boldsymbol{\beta}_* - \mathbb{E}_\varepsilon[\hat{\boldsymbol{\beta}}]\|_{\boldsymbol{\Sigma}}^2}_{:=\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda}} + \underbrace{\mathtt{tr}(\boldsymbol{\Sigma}\mathtt{Cov}_\varepsilon(\hat{\boldsymbol{\beta}}))}_{:=\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda}} .$$

- $\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda} = \lambda^2 \langle \boldsymbol{\beta}_*, (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Sigma} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\beta}_* \rangle$
- $\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda} = \sigma^2 \mathtt{Tr}(\boldsymbol{\Sigma} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-2})$
- *Intuitive fact: for i.i.d. sub-Gaussian data $\boldsymbol{X}$, we have

$$\|\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} - \Sigma\|_{op} = \Theta(\sqrt{d/n}) \,, w.h.p \,.$$

An example of linear model: a textbook level

- $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \overset{i.i.d.}{\sim} \mu$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$
- $y = \langle \boldsymbol{\beta}_*, \boldsymbol{x} \rangle + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$ and $\mathbb{V}(\varepsilon) = \sigma^2$
- ridge regression: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$
- min-ℓ_2-norm interpolation: $\hat{\boldsymbol{\beta}}_{\min} = \mathtt{argmin}_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_2$, s.t. $\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{y}$
- expected test risk: bias-variance decomposition

$$\mathcal{R}^{\mathrm{LS}} := \mathbb{E}_\varepsilon \|\boldsymbol{\beta}_* - \hat{\boldsymbol{\beta}}\|_{\boldsymbol{\Sigma}}^2 = \underbrace{\|\boldsymbol{\beta}_* - \mathbb{E}_\varepsilon[\hat{\boldsymbol{\beta}}]\|_{\boldsymbol{\Sigma}}^2}_{:=\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda}} + \underbrace{\mathtt{tr}(\boldsymbol{\Sigma}\mathtt{Cov}_\varepsilon(\hat{\boldsymbol{\beta}}))}_{:=\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda}}.$$

- $\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda} = \lambda^2 \langle \boldsymbol{\beta}_*, (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Sigma} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\beta}_* \rangle$
- $\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda} = \sigma^2 \mathtt{Tr}(\boldsymbol{\Sigma} \boldsymbol{X}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I})^{-2})$
- *Intuitive fact: for i.i.d. sub-Gaussian data $\boldsymbol{X}$, we have

$$\|\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{\Sigma}\|_{op} = \Theta(\sqrt{d/n}),\ w.h.p.$$

Beyond textbook level: deterministic equivalence (Cheng and Montanari, 2024)

$$\mathtt{Tr}\left(\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-1}\right) \sim \mathtt{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I})^{-1}\right).$$

- $\sim$ 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 $n - \frac{\lambda}{\lambda_*} = \mathtt{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-1})$.

Theorem (Deterministic equivalence (Misiakiewicz and Saeed, 2024))

For sub-Gaussian data, assume Σ *is* **well-behaved**, *w.h.p.*

$$\underbrace{\|\boldsymbol{\beta}_* - \mathbb{E}_\varepsilon[\hat{\boldsymbol{\beta}}]\|_{\boldsymbol{\Sigma}}^2}_{:=\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda}} \sim \mathrm{B}^{\mathrm{LS}}_{\mathrm{R},\lambda} := \frac{\lambda_*^2 \langle \boldsymbol{\beta}_*, \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-2} \boldsymbol{\beta}_* \rangle}{1 - n^{-1}\mathtt{tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-2})}$$

$$\underbrace{\mathtt{tr}(\boldsymbol{\Sigma}\mathtt{Cov}_\varepsilon(\hat{\boldsymbol{\beta}}))}_{:=\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda}} \sim \mathrm{V}^{\mathrm{LS}}_{\mathrm{R},\lambda} := \frac{\sigma^2 \mathtt{tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-2})}{n - \mathtt{tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-2})}.$$

Beyond textbook level: deterministic equivalence (Cheng and Montanari, 2024)

$$\mathrm{Tr}\Big(\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-1}\Big) \sim \mathrm{Tr}\big(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I})^{-1}\big)\,.$$

- $\sim$ 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 $n - \frac{\lambda}{\lambda_*} = \mathrm{Tr}(\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-1})$.

Theorem (Deterministic equivalence (Misiakiewicz and Saeed, 2024))

For sub-Gaussian data, assume $\boldsymbol{\Sigma}$ *is* **well-behaved**, *w.h.p.*

$$\underbrace{\|\boldsymbol{\beta}_* - \mathbb{E}_\varepsilon[\hat{\boldsymbol{\beta}}]\|_{\boldsymbol{\Sigma}}^2}_{:=\mathcal{B}^{\mathrm{LS}}_{\mathcal{R},\lambda}} \sim \mathrm{B}^{\mathrm{LS}}_{\mathrm{R},\lambda} := \frac{\lambda_*^2 \langle \boldsymbol{\beta}_*, \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-2} \boldsymbol{\beta}_* \rangle}{1 - n^{-1}\mathrm{tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-2})}$$

$$\underbrace{\mathrm{tr}(\boldsymbol{\Sigma}\mathrm{Cov}_\varepsilon(\hat{\boldsymbol{\beta}}))}_{:=\mathcal{V}^{\mathrm{LS}}_{\mathcal{R},\lambda}} \sim \mathrm{V}^{\mathrm{LS}}_{\mathrm{R},\lambda} := \frac{\sigma^2 \mathrm{tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-2})}{n - \mathrm{tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I}_d)^{-2})}\,.$$

Proof of sketch on bias

$$\mathcal{B}^{\mathtt{LS}}_{\mathcal{N},\lambda} = \mathtt{Tr}\Big(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-1}\Big) - \lambda \mathtt{Tr}\Big(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-2}\Big)$$

- first term

$$\mathtt{Tr}\Big(\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-1}\Big) \sim \mathtt{Tr}\big(\boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_*)^{-1}\big)$$

- second term

$$\lambda \mathtt{tr}\Big(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-2}\Big) \sim \lambda \cdot \frac{\mathtt{Tr}\big(\boldsymbol{A}\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I})^{-2}\big)}{n - \mathtt{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I})^{-2})}$$

$$\leq \mathtt{Tr}(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I})^{-1}) - \mathtt{Tr}(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_* \boldsymbol{I})^{-2})$$

$$\leq \left(1 - \frac{1}{C}\right)\mathtt{Tr}(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_*)^{-1})$$

Proof of sketch on bias

$$\mathcal{B}^{\mathrm{LS}}_{\mathcal{N},\lambda} = \mathrm{Tr}\Big(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-1}\Big) - \lambda \mathrm{Tr}\Big(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-2}\Big)$$

- first term

$$\mathrm{Tr}\Big(\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-1}\Big) \sim \mathrm{Tr}\big(\boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_*)^{-1}\big)$$

- second term

$$\lambda \mathrm{tr}\Big(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda)^{-2}\Big) \sim \lambda \cdot \frac{\mathrm{Tr}\big(\boldsymbol{A}\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_*\boldsymbol{I})^{-2}\big)}{n - \mathrm{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_*\boldsymbol{I})^{-2})}$$

$$\leq \mathrm{Tr}(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_*\boldsymbol{I})^{-1}) - \mathrm{Tr}(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + \lambda_*\boldsymbol{I})^{-2})$$

$$\leq \left(1 - \frac{1}{C}\right)\mathrm{Tr}(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \lambda_*)^{-1})$$

Proof of sketch on bias

$$\mathcal{B}^{\texttt{LS}}_{\mathcal{N},\lambda} = \texttt{Tr}\Big(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda)^{-1}\Big) - \lambda\texttt{Tr}\Big(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda)^{-2}\Big)$$

- first term

$$\texttt{Tr}\Big(\boldsymbol{A}\boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda)^{-1}\Big) \sim \texttt{Tr}\big(\boldsymbol{A}\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_*)^{-1}\big)$$

- second term

$$\lambda\texttt{tr}\Big(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top \boldsymbol{X}^\top\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X}+\lambda)^{-2}\Big) \sim \lambda\cdot\frac{\texttt{Tr}\big(\boldsymbol{A}\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*\boldsymbol{I})^{-2}\big)}{n-\texttt{Tr}(\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*\boldsymbol{I})^{-2})}$$

$$\le \texttt{Tr}(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_*\boldsymbol{I})^{-1}) - \texttt{Tr}(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+\lambda_*\boldsymbol{I})^{-2})$$

$$\le \left(1-\frac{1}{C}\right)\texttt{Tr}(\boldsymbol{\beta}_*\boldsymbol{\beta}_*^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\lambda_*)^{-1})$$

*Path norm, Barron spaces, RKHS (Chen et al., 2023)

Consider a random features model (RFM) (Rahimi and Recht, 2007)

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

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

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

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

For any $1 \leq p \leq \infty$, define

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

◦ largest
◦ data-adaptive

*Path norm, Barron spaces, RKHS (Chen et al., 2023)

Consider a random features model (RFM) (Rahimi and Recht, 2007)

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

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

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

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

For any $1 \leq p \leq \infty$, define

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

- largest
- data-adaptive

*Path norm, Barron spaces, RKHS (Chen et al., 2023)

Consider a random features model (RFM) (Rahimi and Recht, 2007)

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

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

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

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

For any $1 \leq p \leq \infty$, define

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

- ◦ largest
- ◦ data-adaptive

*Path norm, Barron spaces, RKHS (Chen et al., 2023)

Consider a random features model (RFM) (Rahimi and Recht, 2007)

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

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

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

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

For any $1 \leq p \leq \infty$, define

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

- largest
- data-adaptive

*Path norm, Barron spaces, RKHS (Chen et al., 2023)

Consider a random features model (RFM) (Rahimi and Recht, 2007)

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

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

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

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

For any $1 \leq p \leq \infty$, define

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

- largest
- data-adaptive

*Path norm, Barron spaces, RKHS (Chen et al., 2023)

Consider a random features model (RFM) (Rahimi and Recht, 2007)

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

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

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

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

For any $1 \leq p \leq \infty$, define

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

◦ largest

◦ data-adaptive

Proof sketch: convex hull technique and its constant!

- Consider the following function space

  $\mathcal{F} = \{\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathcal{W}\} \cup \{0\} \cup \{-\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathbb{S}_1^{d-1} \text{ with the } \ell_1 \text{ ball}\}$

- the convex hull of $\mathcal{F}$ is

$$\overline{\mathrm{conv}}\mathcal{F} = \left\{ \sum_{i=1}^{m} \alpha_i f_i \middle| f_i \in \mathcal{F}, \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geqslant 0, m \in \mathbb{N} \right\}.$$

- convex hull technique (Van Der Vaart et al., 1996, Theorem 2.6.9)

$$\log \mathcal{N}_2(\mathcal{G}_1, \epsilon) \leqslant \log \mathcal{N}_2(\overline{\setminus}\mathcal{F}, \epsilon, \mu) \leqslant C \left(\frac{1}{\epsilon}\right)^{\frac{2d}{d+2}}.$$

- control the constant C

$$C := \underbrace{D_k}_{=\Theta(d)} [\underbrace{C_k}_{=\Theta(1)} (2^{d+1} + 1)^{\frac{1}{d}}]^{\frac{2d}{d+2}} \leq 10^7 d \quad \text{if } d > 5$$

Proof sketch: convex hull technique and its constant!

- Consider the following function space

$$\mathcal{F} = \{\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathcal{W}\} \cup \{0\} \cup \{-\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathbb{S}_1^{d-1} \text{ with the } \ell_1 \text{ ball}\}$$

- the convex hull of $\mathcal{F}$ is

$$\overline{\mathrm{conv}}\mathcal{F} = \left\{ \sum_{i=1}^{m} \alpha_i f_i \middle| f_i \in \mathcal{F}, \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geqslant 0, m \in \mathbb{N} \right\}.$$

- convex hull technique (Van Der Vaart et al., 1996, Theorem 2.6.9)

$$\log \mathcal{N}_2(\mathcal{G}_1, \epsilon) \leqslant \log \mathcal{N}_2(\overline{\mathrm{conv}}\mathcal{F}, \epsilon, \mu) \leqslant C \left(\frac{1}{\epsilon} \right)^{\frac{2d}{d+2}}.$$

- control the constant C

$$C := \underbrace{D_k}_{=\Theta(d)} [\underbrace{C_k}_{=\Theta(1)} (2^{d+1} + 1)^{\frac{1}{d}}]^{\frac{2d}{d+2}} \leq 10^7 d \quad \text{if } d > 5$$

Proof sketch: convex hull technique and its constant!

- Consider the following function space

  $$\mathcal{F} = \{\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathcal{W}\} \cup \{0\} \cup \{-\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathbb{S}_1^{d-1} \text{ with the } \ell_1 \text{ ball}\}$$

- the convex hull of $\mathcal{F}$ is

  $$\overline{\mathrm{conv}}\mathcal{F} = \left\{ \sum_{i=1}^{m} \alpha_i f_i \middle| f_i \in \mathcal{F}, \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geqslant 0, m \in \mathbb{N} \right\}.$$

- convex hull technique (Van Der Vaart et al., 1996, Theorem 2.6.9)

  $$\log \mathcal{N}_2(\mathcal{G}_1, \epsilon) \leqslant \log \mathcal{N}_2(\overline{\mathrm{conv}}\mathcal{F}, \epsilon, \mu) \leqslant C \left(\frac{1}{\epsilon}\right)^{\frac{2d}{d+2}}.$$

- control the constant C

  $$C := \underbrace{D_k}_{=\Theta(d)} [\underbrace{C_k}_{=\Theta(1)} (2^{d+1} + 1)^{\frac{1}{d}}]^{\frac{2d}{d+2}} \leq 10^7 d \quad \text{if } d > 5$$

Proof sketch: convex hull technique and its constant!

- Consider the following function space

  $\mathcal{F} = \{\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathcal{W}\} \cup \{0\} \cup \{-\sigma(\langle \widetilde{\boldsymbol{w}}, \cdot \rangle) : \widetilde{\boldsymbol{w}} \in \mathbb{S}_1^{d-1} \text{ with the } \ell_1 \text{ ball}\}$

- the convex hull of $\mathcal{F}$ is

$$\overline{\mathrm{conv}}\mathcal{F} = \left\{ \sum_{i=1}^{m} \alpha_i f_i \,\middle|\, f_i \in \mathcal{F}, \sum_{i=1}^{m} \alpha_i = 1, \alpha_i \geqslant 0, m \in \mathbb{N} \right\}.$$

- convex hull technique (Van Der Vaart et al., 1996, Theorem 2.6.9)

$$\log \mathcal{N}_2(\mathcal{G}_1, \epsilon) \leqslant \log \mathcal{N}_2(\overline{\mathrm{conv}}\mathcal{F}, \epsilon, \mu) \leqslant C \left(\frac{1}{\epsilon}\right)^{\frac{2d}{d+2}}.$$

- control the constant C

$$C := \underbrace{D_k}_{=\Theta(d)} [\underbrace{C_k}_{=\Theta(1)} (2^{d+1} + 1)^{\frac{1}{d}}]^{\frac{2d}{d+2}} \leq 10^7 d \quad \text{if } d > 5$$