A duality framework for analyzing random feature and two-layer neural networks

Lei Wu
School of Mathematical Sciences
Center for Machine Learning Research
Peking University

February 10, 2025 @ Gatsby Unit

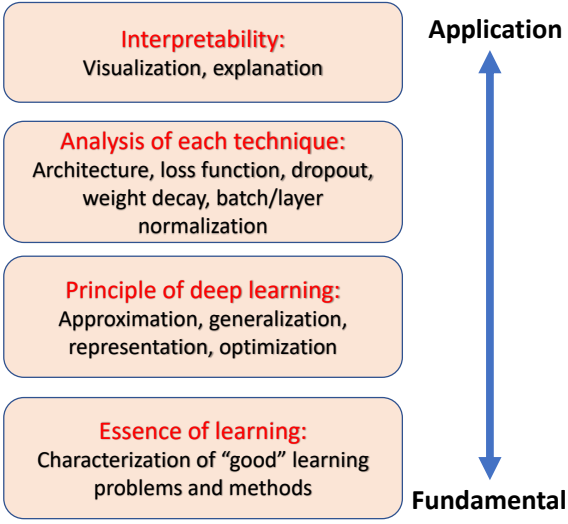


Joint work with Hongrui Chen and Jihao Long, accepted in *Annals of Statistics*.

Outline

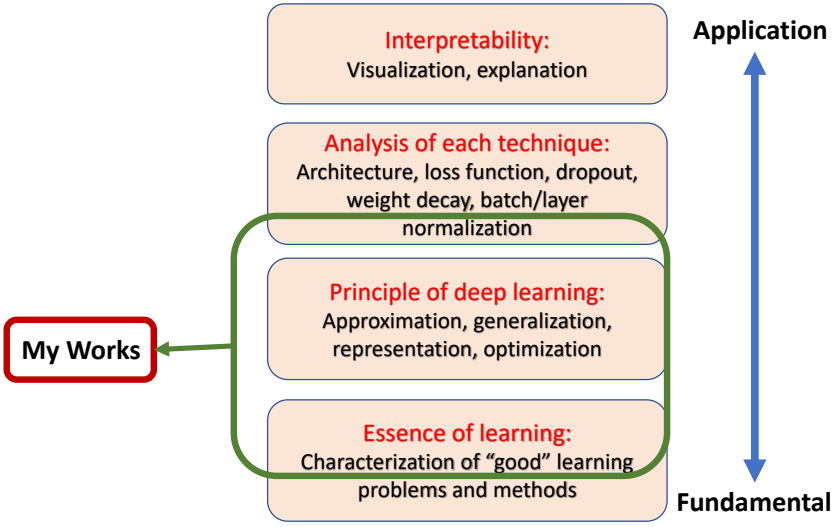
1. Motivation
2. Setup and problems
3. The duality framework via I-complexity
4. The duality for the $\mathcal{F}_{p,\pi}$ and Barron spaces
5. Application 1: Random feature learning beyond kernel regime
6. Application 2: Improved rates for KRR in the noiseless regime

The layers of deep learning theory



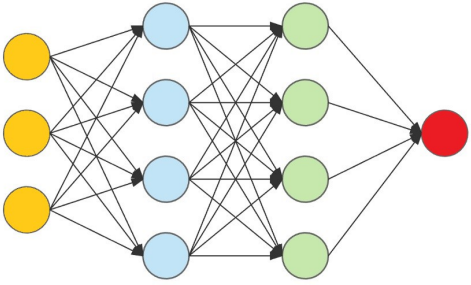
Taken from Taiji Suzuki's slide.

My focus


Interpretability:
Visualization, explanation
Analysis of each technique:
Architecture, loss function, dropout, weight decay, batch/layer normalization
Principle of deep learning:
Approximation, generalization, representation, optimization
Essence of learning:
Characterization of “good” learning problems and methods
My Works
Application
Fundamental

Question 1: Why do NNs perform unreasonably well in complex problem modeling?

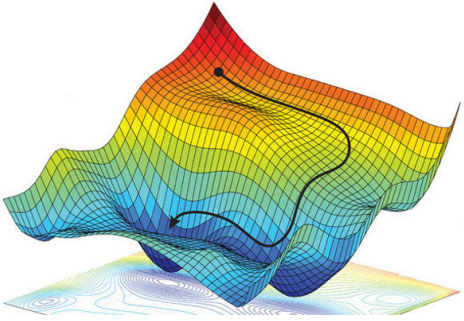
- **deep NN vs. shallow NN vs. traditional method.**
  [EMW, Comm. Math. Sci. 2019], [EMW, Constr. Appr. 2022], [WL, JMLR 2022], [W, JML 2023].
- **MLP vs. CNN vs. RNN vs. Transformer.**
  [WW, NeurIPS 2023]



The research motivation

Question 2: Why are SGD and its variants so efficient in NN training?

- **How does SGD find well-generalizing solutions despite overparameterization?**
  [WME, NeurIPS 2018] [WWS, NeurIPS 2022] [WS, ICML 2023] [WW, NeurIPS 2024][LWLW, NeurIPS 2024]
- **How does the implicit bias of SGD impact the convergence?**
  [WWM, NeurIPS 2019], [MWE, MSML 2022] [W, AISTATS 2022]



The motivation of this work

- Current theoretical works are heavily influenced by the application of DL in CV and NLP. For instance, flat minima hypothesis, double descent, benign overfitting etc.
- **What are the "distinctive" theoretical problems inspired by the application of DL in scientific computing (aka AI for Science)?**
  - Prominent applications include molecular modeling and the numerical solution of partial differential equations, where **data are often generated via simulations rather than direct observations**.

The motivation of this work

- Current theoretical works are heavily influenced by the application of DL in CV and NLP. For instance, flat minima hypothesis, double descent, benign overfitting etc.
- **What are the "distinctive" theoretical problems inspired by the application of DL in scientific computing (aka AI for Science)?**
  - Prominent applications include molecular modeling and the numerical solution of partial differential equations, where **data are often generated via simulations rather than direct observations**.
  - **Label noise is typically negligible**, making traditional assumptions of random noise in statistical ML less appropriate.

The motivation of this work

- Current theoretical works are heavily influenced by the application of DL in CV and NLP. For instance, flat minima hypothesis, double descent, benign overfitting etc.
- **What are the "distinctive" theoretical problems inspired by the application of DL in scientific computing (aka AI for Science)?**
  - Prominent applications include molecular modeling and the numerical solution of partial differential equations, where **data are often generated via simulations rather than direct observations**.
  - **Label noise is typically negligible**, making traditional assumptions of random noise in statistical ML less appropriate.
  - These characteristics necessitate a rethinking of generalization issue in statistical ML.

Setup: function regression

- Let $f^* : \mathbb{R}^d \mapsto \mathbb{R}$ be a target function and $S_n = \{(x_i, y_i = f^*(x_i) + \epsilon_i)\}_{i=1}^n$ be the training data.

Setup: function regression

- Let $f^* : \mathbb{R}^d \mapsto \mathbb{R}$ be a target function and $S_n = \{(x_i, y_i = f^*(x_i) + \epsilon_i)\}_{i=1}^n$ be the training data.
- We aim to learn f^* from S_n by
  - choosing an appropriate model $f(\cdot; \theta)$ with $\theta \in \mathbb{R}^m$;
  - minimize the empirical risk

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i) + \lambda \|\theta\|.$$

Setup: function regression

- Let $f^* : \mathbb{R}^d \mapsto \mathbb{R}$ be a target function and $S_n = \{(x_i, y_i = f^*(x_i) + \epsilon_i)\}_{i=1}^n$ be the training data.
- We aim to learn f^* from S_n by
  - choosing an appropriate model $f(\cdot; \theta)$ with $\theta \in \mathbb{R}^m$;
  - minimize the empirical risk

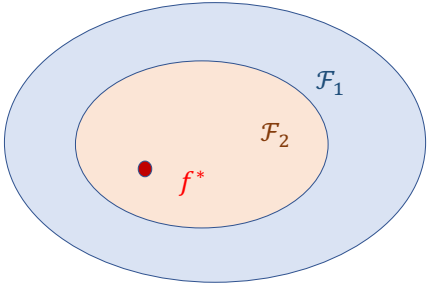
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i) + \lambda \|\theta\|.$$

Question

How many samples are required for learning f^*?

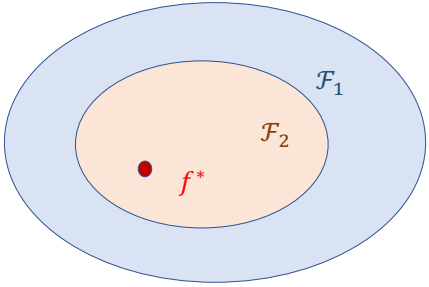
Uniform bound: Identifying the right structure

- Appropriately choose a function class $\mathcal{F}$ such that $f^* \in \mathcal{F}$.



Uniform bound: Identifying the right structure

- Appropriately choose a function class $\mathcal{F}$ such that $f^* \in \mathcal{F}$.

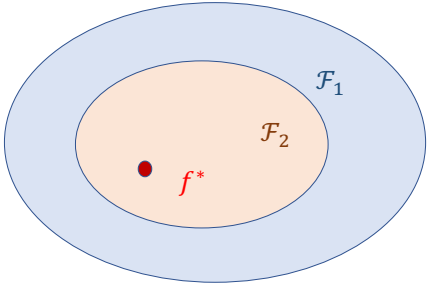


- Consider the following worst-case analysis:

$$\|f(\cdot;\hat{\theta}) - f^*\| \leq \sup_{f^* \in \mathcal{F}} \|f(\cdot;\hat{\theta}) - f^*\|$$

Uniform bound: Identifying the right structure

- Appropriately choose a function class $\mathcal{F}$ such that $f^* \in \mathcal{F}$.



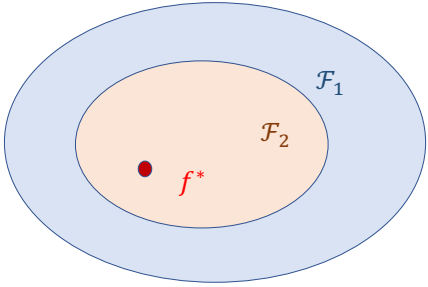
- Consider the following worst-case analysis:

$$\|f(\cdot;\hat{\theta}) - f^*\| \leq \sup_{f^* \in \mathcal{F}} \|f(\cdot;\hat{\theta}) - f^*\|$$

- Choosing $\mathcal{F} =$ identify the structure of f^*:

Uniform bound: Identifying the right structure

- Appropriately choose a function class $\mathcal{F}$ such that $f^* \in \mathcal{F}$.



- Consider the following worst-case analysis:

$$\|f(\cdot;\hat{\theta}) - f^*\| \leq \sup_{f^* \in \mathcal{F}} \|f(\cdot;\hat{\theta}) - f^*\|$$

- Choosing $\mathcal{F} =$ identify the structure of f^*:
  - Sobolev/Besov spaces; reproducing kernel Hilbert space (RKHS), Barron spaces, etc.

Formulation of estimation and approximation error

- (information-theoretical optimal) Estimation Error:

$$\mathcal{E}(n;\mathcal{F}) = \inf_{T\in\mathcal{A}_n} \sup_{f^*\in\mathcal{F}} \mathbb{E}\,\|T(S_n) - f^*\|,$$

where $\mathcal{A}_n$ denotes the set of all estimator: $(\mathcal{X}\times\mathbb{R})^d \mapsto \mathcal{F}$.

Formulation of estimation and approximation error

- (information-theoretical optimal) Estimation Error:

$$\mathcal{E}(n;\mathcal{F}) = \inf_{T\in\mathcal{A}_n} \sup_{f^*\in\mathcal{F}} \mathbb{E}\,\|T(S_n) - f^*\|,$$

  where $\mathcal{A}_n$ denotes the set of all estimator: $(\mathcal{X}\times\mathbb{R})^d \mapsto \mathcal{F}$.

- Approximation Error:

$$\mathcal{A}(m;\mathcal{F}) = \sup_{f^*\in\mathcal{F}} \inf_{\theta\in\mathbb{R}^m} \|f(\cdot;\theta) - f^*\|,$$

  depending on the specific model choice.

Formulation of estimation and approximation error

- (information-theoretical optimal) Estimation Error:

$$\mathcal{E}(n;\mathcal{F}) = \inf_{T\in\mathcal{A}_n} \sup_{f^*\in\mathcal{F}} \mathbb{E}\,\|T(S_n) - f^*\|,$$

  where $\mathcal{A}_n$ denotes the set of all estimator: $(\mathcal{X}\times\mathbb{R})^d \mapsto \mathcal{F}$.

- Approximation Error:

$$\mathcal{A}(m;\mathcal{F}) = \sup_{f^*\in\mathcal{F}} \inf_{\theta\in\mathbb{R}^m} \|f(\cdot;\theta) - f^*\|,$$

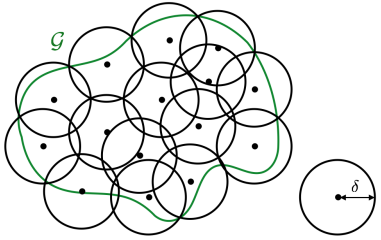
  depending on the specific model choice.

- What are the right complexity measures that determine the estimation and approximation error, respectively?

Answer from statistician

- Covering number

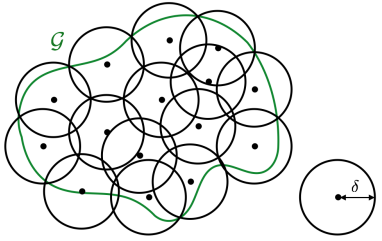
$$\mathcal{N}(\mathcal{G}, \delta) = \min \left\{ M \in \mathbb{N} : \exists z_1, \ldots, z_M \text{ such that } \mathcal{G} \subset \cup_{k=1}^{M} B(z_k, \delta) \right\}$$



Answer from statistician

- Covering number

$$\mathcal{N}(\mathcal{G}, \delta) = \min \left\{ M \in \mathbb{N} : \exists z_1, \ldots, z_M \text{ such that } \mathcal{G} \subset \cup_{k=1}^{M} B(z_k, \delta) \right\}$$



- Statisticians (e.g., [Barron and Yang, 1999; Wainwright, 2019]): covering number is **"sharp"** in quantifying estimation error:

$$\mathcal{E}(n; \mathcal{F}) \approx \int_0^{\mathrm{Diam}(\mathcal{F})} \sqrt{\frac{\log \mathcal{N}(\mathcal{F}, \delta)}{n}} \, \mathrm{d}\delta.$$

Statisticians are right only when noise is on the constant level

- Consider linear class: $\mathcal{F} = \{x \mapsto w^\top x : \|w\| = 1\}$, for which $\mathcal{N}(\mathcal{F}, \delta) = (3/\delta)^d$.

Statisticians are right only when noise is on the constant level

- Consider linear class: $\mathcal{F} = \{x \mapsto w^\top x : \|w\| = 1\}$, for which $\mathcal{N}(\mathcal{F}, \delta) = (3/\delta)^d$.
  - **The noisy case:** When $y_i = f^*(x_i) + \epsilon_i$ with $\mathbb{E}[\epsilon_i^2] = \sigma^2 \gtrsim 1$, we have

$$\mathcal{E}(n; \mathcal{F}) \asymp \int_0^{\mathrm{Diam}(\mathcal{F})} \sqrt{\frac{\log(3/\delta)^d}{n}}\, \mathrm{d}\delta \approx \sqrt{\frac{d}{n}}.$$

Statisticians are right only when noise is on the constant level

- Consider linear class: $\mathcal{F} = \{x \mapsto w^\top x : \|w\| = 1\}$, for which $\mathcal{N}(\mathcal{F}, \delta) = (3/\delta)^d$.
  - **The noisy case:** When $y_i = f^*(x_i) + \epsilon_i$ with $\mathbb{E}[\epsilon_i^2] = \sigma^2 \gtrsim 1$, we have

$$\mathcal{E}(n; \mathcal{F}) \asymp \int_0^{\mathrm{Diam}(\mathcal{F})} \sqrt{\frac{\log(3/\delta)^d}{n}}\, \mathrm{d}\delta \approx \sqrt{\frac{d}{n}}.$$

  - **The noiseless case:** Trivially, we have almost surely

$$\mathcal{E}(n; \mathcal{F}) = 0, \text{ for } n \geq d.$$

Statisticians are right only when noise is on the constant level

- Consider linear class: $\mathcal{F} = \{x \mapsto w^\top x : \|w\| = 1\}$, for which $\mathcal{N}(\mathcal{F}, \delta) = (3/\delta)^d$.
  - **The noisy case:** When $y_i = f^*(x_i) + \epsilon_i$ with $\mathbb{E}[\epsilon_i^2] = \sigma^2 \gtrsim 1$, we have

$$\mathcal{E}(n; \mathcal{F}) \asymp \int_0^{\mathrm{Diam}(\mathcal{F})} \sqrt{\frac{\log(3/\delta)^d}{n}}\, \mathrm{d}\delta \approx \sqrt{\frac{d}{n}}.$$

  - **The noiseless case:** Trivially, we have almost surely

$$\mathcal{E}(n; \mathcal{F}) = 0, \text{ for } n \geq d.$$

- This phenomenon becomes non-trivial for non-linear functions.

Statisticians are right only when noise is on the constant level

- Consider linear class: $\mathcal{F} = \{x \mapsto w^\top x : \|w\| = 1\}$, for which $\mathcal{N}(\mathcal{F}, \delta) = (3/\delta)^d$.
  - **The noisy case:** When $y_i = f^*(x_i) + \epsilon_i$ with $\mathbb{E}[\epsilon_i^2] = \sigma^2 \gtrsim 1$, we have

$$\mathcal{E}(n; \mathcal{F}) \asymp \int_0^{\mathrm{Diam}(\mathcal{F})} \sqrt{\frac{\log(3/\delta)^d}{n}}\, \mathrm{d}\delta \approx \sqrt{\frac{d}{n}}.$$

  - **The noiseless case:** Trivially, we have almost surely

$$\mathcal{E}(n; \mathcal{F}) = 0, \text{ for } n \geq d.$$

- This phenomenon becomes non-trivial for non-linear functions.
  - Since we always have $\mathcal{N}(\mathcal{F}, \delta) \gtrsim 1$, consequently

$$\mathcal{E}(n; \mathcal{F}) \gtrsim \frac{1}{\sqrt{n}} \quad \text{(aka } \textbf{fast rate! Monte-Carlo rate.}\text{)}.$$

Statisticians are right only when noise is on the constant level

- Consider linear class: $\mathcal{F} = \{x \mapsto w^\top x : \|w\| = 1\}$, for which $\mathcal{N}(\mathcal{F}, \delta) = (3/\delta)^d$.
  - **The noisy case:** When $y_i = f^*(x_i) + \epsilon_i$ with $\mathbb{E}[\epsilon_i^2] = \sigma^2 \gtrsim 1$, we have

$$\mathcal{E}(n; \mathcal{F}) \asymp \int_0^{\mathrm{Diam}(\mathcal{F})} \sqrt{\frac{\log(3/\delta)^d}{n}}\, \mathrm{d}\delta \approx \sqrt{\frac{d}{n}}.$$

  - **The noiseless case:** Trivially, we have almost surely

$$\mathcal{E}(n; \mathcal{F}) = 0, \text{ for } n \geq d.$$

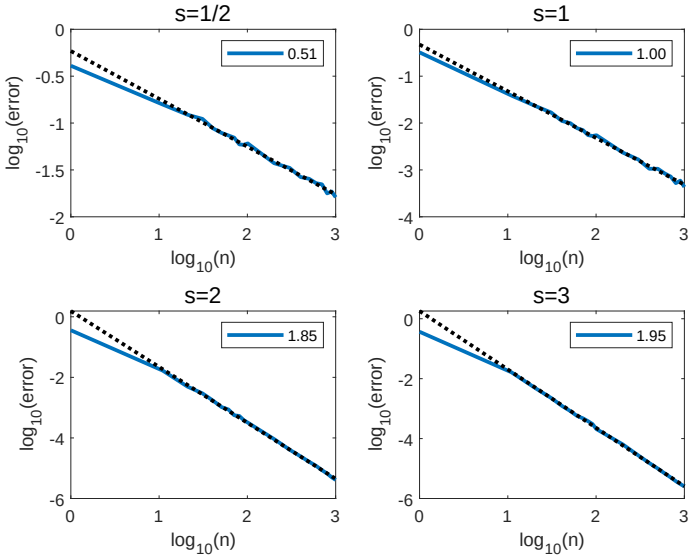
- This phenomenon becomes non-trivial for non-linear functions.
  - Since we always have $\mathcal{N}(\mathcal{F}, \delta) \gtrsim 1$, consequently

$$\mathcal{E}(n; \mathcal{F}) \gtrsim \frac{1}{\sqrt{n}} \quad \text{(aka } \textbf{fast rate! Monte-Carlo rate.}\text{)}.$$

  - Practical performance of ML methods can exhibit **much faster rates** in the noiseless case.

Kernel ridge regression (KRR) in the noiseless setting

- KRR achieves rates much faster than $O(1/\sqrt{n})$ in the noiseless setting.
- This phenomenon is important for scientific computing problems.



An explanation of why the covering number yields a loss bound

We turn to look at the closely Rademacher complexity:

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) = \mathbb{E}_{\xi \sim \mathrm{Unif}(\{\pm 1\}^n)} \left[\frac{1}{n} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(x_i)\xi_i \right]$$

- Rademacher complexity measures the size of a function class by quantifying **its ability of fitting random noise**.

An explanation of why the covering number yields a loss bound

We turn to look at the closely Rademacher complexity:

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}) = \mathbb{E}_{\xi \sim \mathrm{Unif}(\{\pm 1\}^n)} \left[\frac{1}{n} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f(x_i)\xi_i \right]$$

- Rademacher complexity measures the size of a function class by quantifying **its ability of fitting random noise**.
- Both covering number and Rademacher complexity are not right complexity measure, which may lead suboptimal estimates in noiseless settings.
- Can we develop some tools that provides tight characterization for the noiseless setting?

Summary of the motivation

Assume the noise can be decomposed into: $\xi = \xi_1 + \xi_2$ with

$$\mathbb{E}[\xi_1] = 0, \mathbb{E}[\xi_1^2] = \sigma^2, \quad |\xi_2| \leq \epsilon.$$

Summary of the motivation

Assume the noise can be decomposed into: $\xi = \xi_1 + \xi_2$ with

$$\mathbb{E}[\xi_1] = 0, \mathbb{E}[\xi_1^2] = \sigma^2, \quad |\xi_2| \leq \epsilon.$$

Then, the total error roughly follows:

$$\mathcal{E}(n; \mathcal{F}) \approx \underbrace{\epsilon}_{\text{uniform noise}} + \underbrace{\frac{1}{n^{r_1}}}_{\text{clean finite sample}} + \underbrace{\frac{\sigma}{n^{r_2}}}_{\text{random noise}}$$

Summary of the motivation

Assume the noise can be decomposed into: $\xi = \xi_1 + \xi_2$ with

$$\mathbb{E}[\xi_1] = 0, \mathbb{E}[\xi_1^2] = \sigma^2, \quad |\xi_2| \leq \epsilon.$$

Then, the total error roughly follows:

$$\mathcal{E}(n; \mathcal{F}) \approx \underbrace{\epsilon}_{\text{uniform noise}} + \underbrace{\frac{1}{n^{r_1}}}_{\text{clean finite sample}} + \underbrace{\frac{\sigma}{n^{r_2}}}_{\text{random noise}}$$

Consider $f^* \in W^{s,2}$, we observe:

- For **noisy** case $\sigma \asymp 1$, the L^2 error rate is $O(n^{-\frac{s}{2s+d}})$.
- For **noise-free** data ($\sigma = 0$), the L^2 error rate improves to $O(n^{-s/d})$, achievable via local polynomial methods (with grid points).

Summary of the motivation

Assume the noise can be decomposed into: $\xi = \xi_1 + \xi_2$ with

$$\mathbb{E}[\xi_1] = 0, \mathbb{E}[\xi_1^2] = \sigma^2, \quad |\xi_2| \leq \epsilon.$$

Then, the total error roughly follows:

$$\mathcal{E}(n; \mathcal{F}) \approx \underbrace{\epsilon}_{\text{uniform noise}} + \underbrace{\frac{1}{n^{r_1}}}_{\text{clean finite sample}} + \underbrace{\frac{\sigma}{n^{r_2}}}_{\text{random noise}}$$

Consider $f^* \in W^{s,2}$, we observe:

- For **noisy** case $\sigma \asymp 1$, the L^2 error rate is $O(n^{-\frac{s}{2s+d}})$.
- For **noise-free** data ($\sigma = 0$), the L^2 error rate improves to $O(n^{-s/d})$, achievable via local polynomial methods (with grid points).

Implications for ML:

- Models: Kernel methods, random feature models, neural networks.
- Function spaces: Reproducing Kernel Hilbert Space (RKHS), Barron space.

The framework of information-based complexity (IBC)

- IBC is an abstract framework to study approximating to solutions of problems, where information is partial, contaminated, and priced.

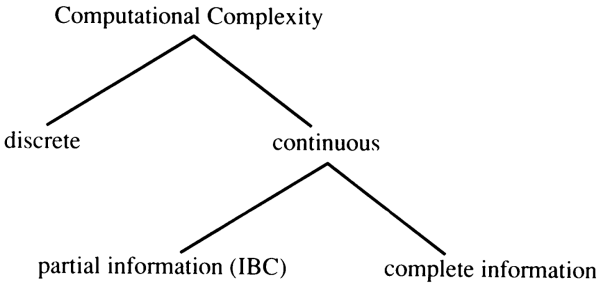


FIGURE 1

The framework of information-based complexity (IBC)

- IBC is an abstract framework to study approximating to solutions of problems, where information is partial, contaminated, and priced.

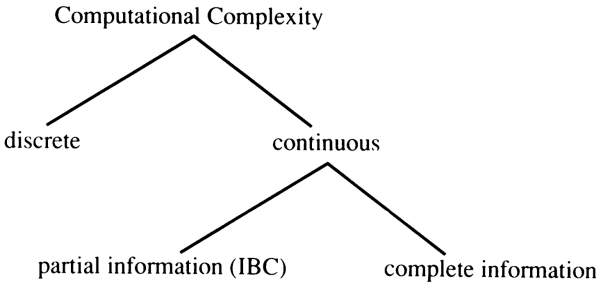


FIGURE 1

- IBC has been widely applied across numerous areas, including solving differential equations, high-dimensional integration, and function/operator approximation.

The framework of information-based complexity (IBC)

- IBC is an abstract framework to study approximating to solutions of problems, where information is partial, contaminated, and priced.

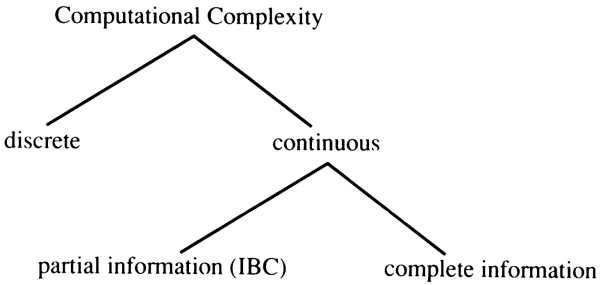


FIGURE 1

- IBC has been widely applied across numerous areas, including solving differential equations, high-dimensional integration, and function/operator approximation.
- In machine learning, "partial information" refers to the availability of only "finite samples".

The information-based complexity

Setup:

- $\mathcal{F}$: The target function space.
- $\mathcal{Q}$: The norm used to measure training error.
- $\mathcal{M}$: The norm used to measure test performance.

Definition 1 (I-complexity)

Assume $\mathcal{F}, \mathcal{Q}, \mathcal{M}$ are Banach spaces of functions over $\mathcal{X}$, and $\mathcal{F} \hookrightarrow \mathcal{M}$ and $\mathcal{F} \hookrightarrow \mathcal{Q}$. Then, for $\epsilon \geq 0$, define

$$\mathbb{I}_{\mathcal{Q},\mathcal{M}}(\mathcal{F}, \epsilon) = \sup_{f \in \mathcal{F},\, \|f\|_{\mathcal{Q}} \leq \epsilon} \|f\|_{\mathcal{M}}.$$

The information-based complexity

Setup:

- $\mathcal{F}$: The target function space.
- $\mathcal{Q}$: The norm used to measure training error.
- $\mathcal{M}$: The norm used to measure test performance.

Definition 1 (I-complexity)

Assume $\mathcal{F}, \mathcal{Q}, \mathcal{M}$ are Banach spaces of functions over $\mathcal{X}$, and $\mathcal{F} \hookrightarrow \mathcal{M}$ and $\mathcal{F} \hookrightarrow \mathcal{Q}$. Then, for $\epsilon \geq 0$, define

$$\mathbb{I}_{\mathcal{Q},\mathcal{M}}(\mathcal{F}, \epsilon) = \sup_{f \in \mathcal{F},\, \|f\|_{\mathcal{Q}} \leq \epsilon} \|f\|_{\mathcal{M}}.$$

Remark

This definition is a modification of the I-complexity proposed in (Novak, 1988):

$$I_n(\mathcal{F}) = \inf_{x_1,\ldots,x_n \in \mathcal{X}} \sup_{f \in \mathcal{F}, f(x_i)=0} \|f\|_{\infty}.$$

Tight bound in the noiseless setting

Proposition 1

Let $\mathcal{F}(r) = \{f : \|f\|_{\mathcal{F}} \le r\}$. For any $x_1, \cdots, x_n \in \mathcal{X}$, we have

$$\inf_{T \in \mathcal{A}_n} \sup_{\|f\|_{\mathcal{F}} \le 1} \|T(\{(x_i, f(x_i))\}_{i=1}^n) - f\|_{\mathcal{M}} \asymp \mathbb{I}_{L^2(\hat{\rho}_n), \mathcal{M}}(\mathcal{F}(1), 0)$$

$$= \sup_{f \in \mathcal{F}(1), f(x_i) = 0, \text{ for } i \in [n]} \|f\|_{\mathcal{M}}$$

Moreover, The optimal error can be achieved by the minimum-norm interpolation estimator:

$$S(\{(x_i, y_i)\}_{i=1}^n) = \operatorname*{argmin}_{h \in \mathcal{F},\, h(x_i) = y_i} \|h\|_{\mathcal{F}}.$$

- A similar lower bound was established in (Novak, 1988).
- **The minimum-norm interpolation may be not feasible in practice**. For instance, the model is neural networks but the target function is Sobolev spaces.

Theorem 2

It holds for any $\epsilon \geq 0$ *that*

$$\mathbb{I}_{\mathcal{Q},\mathcal{M}}(\mathcal{F}(1),\epsilon) = \sup_{\|g\|_{\mathcal{M}^*}\leq 1} \inf_{h\in\mathcal{Q}^*} \left[\|g-h\|_{\mathcal{F}^*} + \epsilon\|h\|_{\mathcal{Q}^*}\right]. \tag{1}$$

The duality framework: translating estimation into approximation

Theorem 2

It holds for any $\epsilon \geq 0$ *that*

$$\mathbb{I}_{\mathcal{Q},\mathcal{M}}(\mathcal{F}(1),\epsilon) = \sup_{\|g\|_{\mathcal{M}^*}\leq 1} \inf_{h\in\mathcal{Q}^*} \left[\|g-h\|_{\mathcal{F}^*} + \epsilon\|h\|_{\mathcal{Q}^*}\right]. \tag{1}$$

- LHS: Estimation error.

Theorem 2

It holds for any $\epsilon \geq 0$ *that*

$$\mathbb{I}_{\mathcal{Q},\mathcal{M}}(\mathcal{F}(1),\epsilon) = \sup_{\|g\|_{\mathcal{M}^*}\leq 1} \inf_{h\in\mathcal{Q}^*} \left[\|g-h\|_{\mathcal{F}^*} + \epsilon\|h\|_{\mathcal{Q}^*}\right]. \tag{1}$$

- LHS: Estimation error.
- RHS: Approximation error.

Theorem 2

It holds for any $\epsilon \geq 0$ *that*

$$\mathbb{I}_{\mathcal{Q},\mathcal{M}}(\mathcal{F}(1),\epsilon) = \sup_{\|g\|_{\mathcal{M}^*}\leq 1} \inf_{h\in\mathcal{Q}^*} \left[\|g-h\|_{\mathcal{F}^*} + \epsilon\|h\|_{\mathcal{Q}^*}\right]. \tag{1}$$

- LHS: Estimation error.
- RHS: Approximation error.
- **This theorem establishes a general duality between approximation and estimation**.

Theorem 2

It holds for any $\epsilon \geq 0$ *that*

$$\mathbb{I}_{\mathcal{Q},\mathcal{M}}(\mathcal{F}(1),\epsilon) = \sup_{\|g\|_{\mathcal{M}^*}\leq 1} \inf_{h\in\mathcal{Q}^*} \left[\|g-h\|_{\mathcal{F}^*} + \epsilon\|h\|_{\mathcal{Q}^*}\right]. \tag{1}$$

- LHS: Estimation error.
- RHS: Approximation error.
- **This theorem establishes a general duality between approximation and estimation**.
- We can focus on the more tractable one between them.

Models

- Let $\phi : \mathcal{X} \times \mathcal{V} \mapsto \mathbb{R}$ be a parametric feature function and consider the model:

$$f(x; \theta) = \frac{1}{m} \sum_{j=1}^{m} a_j \phi(x, v_j),$$

Models

- Let $\phi : \mathcal{X} \times \mathcal{V} \mapsto \mathbb{R}$ be a parametric feature function and consider the model:

$$f(x; \theta) = \frac{1}{m} \sum_{j=1}^{m} a_j \phi(x, v_j),$$

- When $\{(a_j, v_j)\}_{j=1}^{m}$ are learnable, it recovers two-layer neural net.

Models

- Let $\phi : \mathcal{X} \times \mathcal{V} \mapsto \mathbb{R}$ be a parametric feature function and consider the model:

$$f(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \phi(x, v_j),$$

- When $\{(a_j, v_j)\}_{j=1}^m$ are learnable, it recovers two-layer neural net.
- When only $\{a_j\}_{j=1}^m$ are learnable and $v_j \overset{iid}{\sim} \pi$, it recovers random feature models (RFMs).

The $\mathcal{F}_{p,\pi}$ spaces for RFMs

We shall use $\|f\|_{q,\rho} = \|f\|_{L^q(\rho)}$ for simplicity.

Definition 3

Let $\pi \in \mathcal{P}(\mathcal{V})$ be a probability distribution over the weight space. For $1 \leq p \leq \infty$, the $\mathcal{F}_{p,\pi}$ space on $\mathcal{X}$ is defined as

$$\mathcal{F}_{p,\pi} := \left\{ f = \int_{\mathcal{V}} a(v)\phi(\cdot, v)\, \mathrm{d}\, \pi(v) : \, a \in L^p(\pi) \right\}$$

equipped with the norm

$$\|f\|_{\mathcal{F}_{p,\pi}} = \inf_{a \in A_f} \|a\|_{p,\pi}, \quad A_f = \left\{ a \in L^p(\pi) : \, f = \int_{\mathcal{V}} a(v)\phi(\cdot, v)\, \mathrm{d}\, \pi(v) \right\}.$$

The $\mathcal{F}_{p,\pi}$ spaces for RFMs

We shall use $\|f\|_{q,\rho} = \|f\|_{L^q(\rho)}$ for simplicity.

Definition 3

Let $\pi \in \mathcal{P}(\mathcal{V})$ be a probability distribution over the weight space. For $1 \leq p \leq \infty$, the $\mathcal{F}_{p,\pi}$ space on $\mathcal{X}$ is defined as

$$\mathcal{F}_{p,\pi} := \left\{ f = \int_{\mathcal{V}} a(v)\phi(\cdot, v)\, \mathrm{d}\, \pi(v) : a \in L^p(\pi) \right\}$$

- When $p = 2$, $\mathcal{F}_{2,\pi} = \mathcal{H}_{k_\pi}$ is the reproducing kernel Hilbert space (RKHS) associated to the kernel (Rahimi et al.,2008):

$$k_\pi(x, x') = \int_{\mathcal{V}} \phi(x, v)\phi(x', v)\, \mathrm{d}\pi(v).$$

- $\mathcal{F}_{\infty,\pi} \subset \mathcal{F}_{q,\pi} \subset \mathcal{F}_{p,\pi} \subset \mathcal{F}_{1,\pi}$ for any $1 \leq p, q \leq \infty$.
- when $p \in [1, 2)$, $\mathcal{F}_{p,\pi}$ is strictly larger than $\mathcal{F}_{2,\pi}$ (the RKHS).

The Barron space (also called variation space in (Kurkova et al., 2001; DeVore 1998) and $\mathcal{F}_1$ in (Bach, 2017)) for studying the approximation power of two-layer neural nets.

Definition 4

Given a feature map $\phi : \mathcal{X} \times \mathcal{V} \to \mathbb{R}$, for any $x \in \mathcal{X}$, we define the Barron space $\mathcal{B}$ on $\mathcal{X}$ as

$$\mathcal{B} := \left\{ f = \int_{\mathcal{V}} \phi(\cdot, v) \mathrm{d}\, \mu(v) : \mu \in \mathcal{M}(\mathcal{V}) \right\}$$

equipped with the norm

$$\|f\|_{\mathcal{B}} = \inf_{\mu \in M_f} \|\mu\|_{\mathrm{TV}}, \quad \text{where } M_f = \left\{ \mu \in \mathcal{M}(\mathcal{V}) : f = \int_{\mathcal{V}} \phi(\cdot, v) \, \mathrm{d}\, \mu(v) \right\}.$$

The Barron spaces for two-layer neural networks

The Barron space (also called variation space in (Kurkova et al., 2001; DeVore 1998) and $\mathcal{F}_1$ in (Bach, 2017)) for studying the approximation power of two-layer neural nets.

Definition 4

Given a feature map $\phi : \mathcal{X} \times \mathcal{V} \to \mathbb{R}$, for any $x \in \mathcal{X}$, we define the Barron space $\mathcal{B}$ on $\mathcal{X}$ as

$$\mathcal{B} := \left\{ f = \int_{\mathcal{V}} \phi(\cdot, v) \mathrm{d}\, \mu(v) : \mu \in \mathcal{M}(\mathcal{V}) \right\}$$

Lemma (An alternative definition [EMW, 2019])

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

$$\mathcal{B} = \cup_{\pi \in \mathcal{P}(\mathcal{V})} \mathcal{H}_{k_\pi}, \quad \|f\|_{\mathcal{B}} = \inf_{\pi \in \mathcal{P}(\mathcal{V})} \|f\|_{\mathcal{H}_{k_\pi}}.$$

The conjugate spaces: switch the role of x and v

Take a look at the feature map:

$$(x, v) \mapsto \phi(x, v).$$

Definition 5

- For any $\gamma \in \mathcal{P}(\mathcal{X})$, let $\tilde{\mathcal{F}}_{q,\gamma}$ be a function space over the weight space $\mathcal{V}$ given by

$$\tilde{\mathcal{F}}_{q,\gamma} = \left\{ g = \int_{\mathcal{X}} b(x)\phi(x, \cdot)\,\mathrm{d}\,\gamma(x) \,:\, b \in L^q(\gamma) \right\}$$

equipped with the norm

$$\|g\|_{\tilde{\mathcal{F}}_{q,\gamma}} = \inf_{b \in L^q(\gamma)} \|b\|_{q,\gamma}, B_f = \left\{ b \in L^q(\gamma) \,:\, g = \int_{\mathcal{X}} b(x)\phi(x, \cdot)\,\mathrm{d}\,\gamma(x) \right\}.$$

The conjugate spaces: switch the role of x and v

Take a look at the feature map:

$$(x, v) \mapsto \phi(x, v).$$

Definition 5

- For any $\gamma \in \mathcal{P}(\mathcal{X})$, let $\tilde{\mathcal{F}}_{q,\gamma}$ be a function space over the weight space $\mathcal{V}$ given by

$$\tilde{\mathcal{F}}_{q,\gamma} = \left\{ g = \int_{\mathcal{X}} b(x)\phi(x, \cdot)\,\mathrm{d}\,\gamma(x) \,:\, b \in L^q(\gamma) \right\}$$

  equipped with the norm

$$\|g\|_{\tilde{\mathcal{F}}_{q,\gamma}} = \inf_{b \in L^q(\gamma)} \|b\|_{q,\gamma}, B_f = \left\{ b \in L^q(\gamma) \,:\, g = \int_{\mathcal{X}} b(x)\phi(x, \cdot)\,\mathrm{d}\,\gamma(x) \right\}.$$

- Similarly, we define the $\tilde{\mathcal{B}}$ space on $\mathcal{V}$.

The duality in the interpolation regime

Theorem 6

Let $x_1, \cdots, x_n \in \mathcal{X}$. For $1 \leq p \leq \infty$, $1 \leq q \leq \infty$, let $\frac{1}{q} + \frac{1}{q'} = 1$ and $\frac{1}{p} + \frac{1}{p'} = 1$. We have

$$\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \leq 1, f(x_i)=0} \|f\|_{q,\rho} = \sup_{\|g\|_{\tilde{\mathcal{F}}_{q',\rho}} \leq 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_{p',\pi}.$$

The duality in the interpolation regime

Theorem 6

Let $x_1, \cdots, x_n \in \mathcal{X}$. For $1 \le p \le \infty$, $1 \le q \le \infty$, let $\frac{1}{q} + \frac{1}{q'} = 1$ and $\frac{1}{p} + \frac{1}{p'} = 1$. We have

$$\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1, f(x_i)=0} \|f\|_{q,\rho} = \sup_{\|g\|_{\tilde{\mathcal{F}}_{q',\rho}} \le 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_{p',\pi}.$$

- LHS: the I-complexity for estimation within $\mathcal{F}_{p,\pi}$.
- RHS: Approximating the **conjugate** $\mathcal{F}_{q',\pi}$ with (random) features $\{\phi(x_i, \cdot)\}_{i=1}^n$.
- Informally, the $L^q(\rho)$ estimation of $\mathcal{F}_{p,\pi}$ is equivalent to the $L^{p'}(\pi)$ approximation of $\tilde{\mathcal{F}}_{q',\rho}$.

The duality in the interpolation regime (cont'd)

Theorem 7

Let $x_1, \cdots, x_n \in \mathcal{X}$ be the training data. For $1 \le p \le \infty$, $1 \le q \le \infty$ and $\frac{1}{q} + \frac{1}{q'} = 1$, let $\frac{1}{p} + \frac{1}{p'} = 1$. We have

- *For $1 \le q \le \infty$, suppose $\phi(x, \cdot) \in C_0(\mathcal{V})$, we have*

$$\sup_{\|f\|_{\mathcal{B}} \le 1, f(x_i)=0} \|f\|_{q,\rho} = \sup_{\|g\|_{\bar{\mathcal{F}}_{q',\rho}} \le 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_{C_0(\mathcal{V})}$$

- *For $1 \le p \le \infty$, suppose $\phi(\cdot, v) \in C_0(\mathcal{X})$, we have*

$$\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1, f(x_i)=0} \|f\|_{C_0(\mathcal{X})} = \sup_{\|g\|_{\bar{\mathcal{B}}} \le 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_{p',\pi}.$$

Example 8

- For $p = 2$, we have

$$\sup_{\|f\|_{\mathcal{F}_{2,\pi}} \leq 1,\, f(x_i)=0} \|f\|_\rho = \sup_{\|g\|_{\tilde{\mathcal{F}}_{2,\rho}} \leq 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_\pi$$

Example 8

- For $p = 2$, we have

$$\sup_{\|f\|_{\mathcal{F}_{2,\pi}} \le 1,\, f(x_i)=0} \|f\|_\rho = \sup_{\|g\|_{\bar{\mathcal{F}}_{2,\rho}} \le 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_\pi$$

- For the Barron space, we have

$$\sup_{\|f\|_{\mathcal{B}} \le 1,\, f(x_i)=0} \|f\|_\rho = \sup_{\|g\|_{\bar{\mathcal{F}}_{2,\rho}} \le 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_{C_0(\mathcal{V})}$$

Example 8

- For $p = 2$, we have

$$\sup_{\|f\|_{\mathcal{F}_{2,\pi}} \le 1,\, f(x_i)=0} \|f\|_\rho = \sup_{\|g\|_{\tilde{\mathcal{F}}_{2,\rho}} \le 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_\pi$$

- For the Barron space, we have

$$\sup_{\|f\|_{\mathcal{B}} \le 1,\, f(x_i)=0} \|f\|_\rho = \sup_{\|g\|_{\tilde{\mathcal{F}}_{2,\rho}} \le 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_{C_0(\mathcal{V})}$$

- *"Estimations" of RFMs and two-layer neural nets are equivalent to* ***the approximation of the same RKHS but under different norms*** *(L^2 vs. L^∞).*

Example 2: L^∞ learnability of RKHS

Example 9

When $p = 2, q = \infty$, we have

$$\sup_{\|f\|_{\mathcal{F}_{2,\pi}} \le 1,\, f(x_i)=0} \|f\|_{C_0(\mathcal{X})} = \sup_{\|g\|_{\bar{\mathcal{B}}} \le 1} \inf_{c_1,\cdots,c_n} \left\| g - \sum_{i=1}^{n} c_i \phi(x_i, \cdot) \right\|_{2,\pi}$$

- The LHS is the I-complexity that governs the L^∞ estimation error of learning $\mathcal{F}_{2,\pi}$.
- This implies that we can obtain the L^∞ estimation error through the L^2 approximation of the corresponding Barron space.

Theorem 10 (Norm-controlled approximation)

Let $\nu \in \mathcal{P}(\mathcal{X})$ and $1 \le r \le \infty$. For $1 \le p \le \infty$, $1 \le q \le \infty$, suppose that $\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1} \|f\|_{r,\nu} < \infty$, we have

$$\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1,\ \|f\|_{r,\nu} \le \epsilon} \|f\|_{q,\rho}$$

$$= \sup_{\|g\|_{\bar{\mathcal{F}}_{q',\rho}} \le 1} \inf_{c \in L^{r'}(\nu)} \left(\left\| g - \int_{\mathcal{X}} c(x)\phi(x,\cdot)\,\mathrm{d}\,\nu(x) \right\|_{p',\pi} + \epsilon \|c\|_{r',\nu} \right).$$

The duality in the non-interpolation regime

Theorem 10 (Norm-controlled approximation)

Let $\nu \in \mathcal{P}(\mathcal{X})$ and $1 \le r \le \infty$. For $1 \le p \le \infty$, $1 \le q \le \infty$, suppose that $\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1} \|f\|_{r,\nu} < \infty$, we have

$$\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1,\ \|f\|_{r,\nu} \le \epsilon} \|f\|_{q,\rho}$$

$$= \sup_{\|g\|_{\bar{\mathcal{F}}_{q',\rho}} \le 1} \inf_{c \in L^{r'}(\nu)} \left(\left\| g - \int_{\mathcal{X}} c(x)\phi(x,\cdot)\, \mathrm{d}\,\nu(x) \right\|_{p',\pi} + \epsilon \|c\|_{r',\nu} \right).$$

- Taking $\nu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$ and $\epsilon = 0$ recovers Theorem 6.
- We introduce another duality pairing:
  - The $L^r(\nu)$ constraint on fitting errors in estimation($\|f\|_{r,\nu} \le \epsilon$).
  - The $L^{r'}(\nu)$ regularization of the coefficient function in approximation($\epsilon\|c\|_{r',\nu}$).
- The choice of ν is flexible.

Theorem 10 (Norm-controlled approximation)

Let $\nu \in \mathcal{P}(\mathcal{X})$ and $1 \le r \le \infty$. For $1 \le p \le \infty$, $1 \le q \le \infty$, suppose that $\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1} \|f\|_{r,\nu} < \infty$, we have

$$\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1,\ \|f\|_{r,\nu} \le \epsilon} \|f\|_{q,\rho}$$

$$= \sup_{\|g\|_{\tilde{\mathcal{F}}_{q',\rho}} \le 1} \inf_{c \in L^{r'}(\nu)} \left(\left\| g - \int_{\mathcal{X}} c(x)\phi(x,\cdot)\,\mathrm{d}\,\nu(x) \right\|_{p',\pi} + \epsilon \|c\|_{r',\nu} \right).$$

By taking $\nu = \hat{\rho}_n$, the RHS becomes

$$\sup_{g \in \tilde{\mathcal{F}}_{q',\rho}} \inf_{c_1,\dots,c_n} \left(\left\| g - \sum_{i=1}^{n} c_i \phi(x_i,\cdot) \right\|_{p',\pi} + \epsilon \left(\frac{1}{n} \sum_{i=1}^{n} |c_j|^p \right)^{1/p} \right).$$

This allows us to study the norm-controlled approximation of RFMs.

Theorem 10 (Norm-controlled approximation)

Let $\nu \in \mathcal{P}(\mathcal{X})$ *and* $1 \le r \le \infty$.
For $1 \le q < \infty$, *suppose that* $\sup_{\|f\|_{\mathcal{B}} \le 1} \|f\|_{r,\nu} < \infty$, *we have*

$$\sup_{\|f\|_{\mathcal{B}} \le 1,\ \|f\|_{r,\nu} \le \epsilon} \|f\|_{q,\rho} = \sup_{\|g\|_{\bar{\mathcal{F}}_{q',\rho}} \le 1} \inf_{c \in L^{r'}(\nu)} \left(\left\| g - \int_{\mathcal{X}} c(x)\phi(x,\cdot)\,\mathrm{d}\,\nu(x) \right\|_{\infty} + \epsilon \|c\|_{r',\nu} \right).$$

The duality in the non-interpolation regime

Theorem 10 (Norm-controlled approximation)

Let $\nu \in \mathcal{P}(\mathcal{X})$ *and* $1 \le r \le \infty$.
For $1 < p \le \infty$, *suppose that* $\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1} \|f\|_{r,\nu} < \infty$,

$$\sup_{\|f\|_{\mathcal{F}_{p,\pi}} \le 1, \|f\|_{r,\nu} \le \epsilon} \|f\|_{\infty}$$
$$= \sup_{\|g\|_{\tilde{\mathcal{B}}} \le 1} \inf_{c \in L^{r'}(\nu)} \left(\left\| g - \int_{\mathcal{X}} c(x)\phi(x,\cdot)\,\mathrm{d}\,\nu(x) \right\|_{p',\pi} + \epsilon \|c\|_{r',\nu} \right).$$

Random feature can work beyond kernel regime

- Let $f^* \in \mathcal{F}_{p,\pi}(1)$ be the target function.
- The training data $x_i \overset{iid}{\sim} \rho$, $y_i = f^*(x_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- Let $v_1, \cdots, v_m \overset{iid}{\sim} \pi$ be the random weights.
- Consider the estimator $\hat{f} = \frac{1}{m} \sum_{j=1}^m \hat{c}_j \phi(\cdot, v_j)$ with

$$\hat{c} := \underset{\|c\|_p \le \lambda m^{1/p}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \left(y_i - \frac{1}{m} \sum_{j=1}^m c_j \phi(x_i, v_j) \right)^2.$$

Random feature can work beyond kernel regime

- Let $f^* \in \mathcal{F}_{p,\pi}(1)$ be the target function.
- The training data $x_i \overset{iid}{\sim} \rho$, $y_i = f^*(x_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.
- Let $v_1, \cdots, v_m \overset{iid}{\sim} \pi$ be the random weights.

Theorem 11

Assume $\sup_{x\in\mathcal{X}, v\in\mathcal{V}} |\phi(x, v)| \lesssim 1$. *With an appropriate* λ, *for* $\delta \in (0, 1)$, *it holds w.p. at least* $1 - \delta$ *over the sampling of data and features*

$$\|\hat{f} - f^*\|_\rho^2 \lesssim \underbrace{\sigma\sqrt{\frac{p' + \log(1/\delta)}{n}}}_{\text{Est. err I}} + \underbrace{\frac{p' \log^3 n + \log(1/\delta)}{n}}_{\text{Est. err II}} + \underbrace{\left(\frac{\log^3 m + \log(1/\delta)}{m}\right)^{2/p'}}_{\text{approx err.}}.$$

Implications

- When $p > 1$, $\mathcal{F}_{p,\pi}$ can be learned without curse of dimensionality in terms of all aspects—approximation, estimation, and optimization.
- RFM can work beyond the traditional kernel regime, where $p = 2$.

Kernel ridge regression in the noiseless regime

- Given a kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, assume it admits the following spectral decomposition

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j e_j(x) e_j(x').$$

Kernel ridge regression in the noiseless regime

- Given a kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, assume it admits the following spectral decomposition

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j e_j(x) e_j(x').$$

- Consider the induced interpolation space (a generalized Sobolev space)

$$\mathcal{H}^s = \left\{ \sum_{j=1}^{\infty} a_j \mu_j^{s/2} e_j \ : \ \sum_{j=1}^{\infty} a_j^2 < \infty \right\},$$

where $\mathcal{H}^0$ and $\mathcal{H}^1 =: \mathcal{H}$ correspond to the L^2 and reproducing kernel Hilbert space (RKHS), respectively.

Kernel ridge regression in the noiseless regime

- Given a kernel $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, assume it admits the following spectral decomposition

$$k(x, x') = \sum_{j=1}^{\infty} \mu_j e_j(x) e_j(x').$$

- Consider the induced interpolation space (a generalized Sobolev space)

$$\mathcal{H}^s = \left\{ \sum_{j=1}^{\infty} a_j \mu_j^{s/2} e_j \ : \ \sum_{j=1}^{\infty} a_j^2 < \infty \right\},$$

  where $\mathcal{H}^0$ and $\mathcal{H}^1 =: \mathcal{H}$ correspond to the L^2 and reproducing kernel Hilbert space (RKHS), respectively.

- Consider the estimator

$$\hat{f} = \operatorname*{argmin}_{f \in \mathcal{H}} \left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - f^*(x_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 \right).$$

L^2 lower bound

Theorem 12

For any $s \geq 0$, $x_1, \ldots, x_n \in \mathcal{X}$ we have that

$$\inf_{T \in \mathcal{A}_n} \sup_{\|f\|_{\mathcal{H}^s} \leq 1} \|T(\{(x_i, f(x_i))\}_{i=1}^n) - f\|_{L^2} \geq \mu_n^{s/2}, \tag{2}$$

L^2 upper bound for KRR

Theorem 13

Let $n \in \mathbb{N}^+$ and $\delta \in (0,1)$. For any $0 \leq s \leq 2$. Then, for a given f^* with $\|f^*\|_{\mathcal{H}^s} \leq 1$, there exists a λ_n such that it holds with probability $1-\delta$ that

$$\|\hat{f}_{\lambda_n} - f^*\|_{L^2} \leq C_{\delta,s}\mu_n^{s/2}, \tag{3}$$

- **Smoothness adaptivity:** When $\mu_j \asymp j^{-1-\beta}$, we have the rate becomes $O(n^{-(1+\beta)/2})$.
- **Spectral accuracy:** If $\mu_j \asymp c^{-j}$ for some $c \in (0,1)$, the convergence exhibit a spectral accuracy: **logarithmic samples suffice**.
- **Saturation:** When $s \leq 2$, no further improvement! KRR is not optimal for H^s with $s > 2$.

Experiments

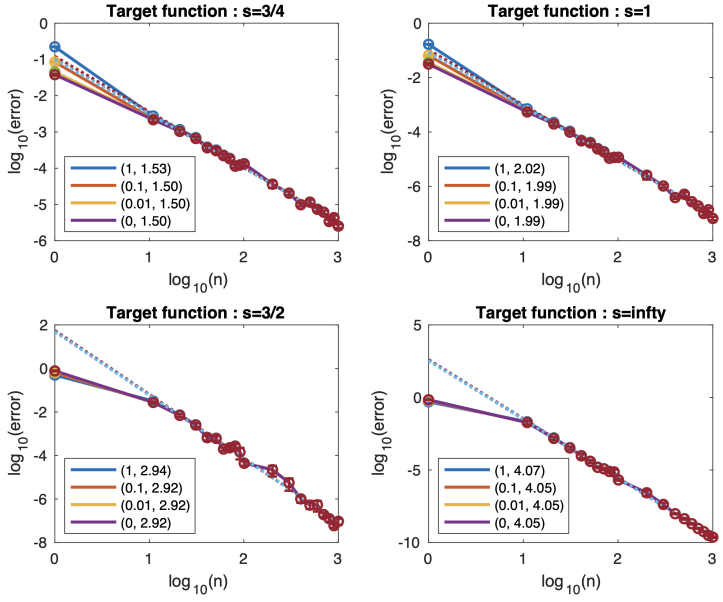


Figure 1: $k(x, x') = \min(x, x')$ for $x, x' \in [0, 1]$. In this case, $\mu_j \asymp j^{-2}$.

L^∞ lower bound

We shall focus on the case of RKHS, i.e., $s = 1$.

Theorem 14 (L^∞-estimation)

For any input data $x_1, \cdots, x_n \in \mathcal{X}$, we have

$$\inf_{\mathcal{Q} \in \mathcal{A}_n} \sup_{\|f\|_{\mathcal{H}_k} \le 1} \|T(\{(x_i, f(x_i))\}_{i=1}^n) - f\|_{C_0(\mathcal{X})} \gtrsim \sqrt{\sum_{j=n+1}^{\infty} \mu_j}. \tag{4}$$

If the eigenvalues decay at a power law $\mu_j^{(k,\gamma)} \asymp j^{-1-\beta}$, we have the rate is $\Omega(n^{-\beta/2})$.

L^∞ lower bound

We shall focus on the case of RKHS, i.e., $s = 1$.

Theorem 14 (L^∞-estimation)

For any input data $x_1, \cdots, x_n \in \mathcal{X}$, we have

$$\inf_{\mathcal{Q} \in \mathcal{A}_n} \sup_{\|f\|_{\mathcal{H}_k} \le 1} \|T(\{(x_i, f(x_i))\}_{i=1}^n) - f\|_{C_0(\mathcal{X})} \gtrsim \sqrt{\sum_{j=n+1}^{\infty} \mu_j}. \tag{4}$$

If the eigenvalues decay at a power law $\mu_j^{(k,\gamma)} \asymp j^{-1-\beta}$, we have the rate is $\Omega(n^{-\beta/2})$.

- Consider the RF kernel $k : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \mapsto \mathbb{R}$ given by

$$k(x, x') = \mathbb{E}_{w \sim \mathrm{Unif}(\mathbb{S}^{d-1})}[(w^\top x)_+^\alpha (w^\top x')_+^\alpha].$$

  The eigenvalue decay at a rate: $j^{-1-\alpha/d}$.

L^∞ lower bound

We shall focus on the case of RKHS, i.e., $s = 1$.

Theorem 14 (L^∞-estimation)

For any input data $x_1, \cdots, x_n \in \mathcal{X}$, we have

$$\inf_{\mathcal{Q} \in \mathcal{A}_n} \sup_{\|f\|_{\mathcal{H}_k} \le 1} \|T(\{(x_i, f(x_i))\}_{i=1}^n) - f\|_{C_0(\mathcal{X})} \gtrsim \sqrt{\sum_{j=n+1}^{\infty} \mu_j}. \tag{4}$$

If the eigenvalues decay at a power law $\mu_j^{(k,\gamma)} \asymp j^{-1-\beta}$, we have the rate is $\Omega(n^{-\beta/2})$.

- Consider the RF kernel $k : \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \mapsto \mathbb{R}$ given by

  $$k(x, x') = \mathbb{E}_{w \sim \mathrm{Unif}(\mathbb{S}^{d-1})}[(w^\top x)_+^\alpha (w^\top x')_+^\alpha].$$

  The eigenvalue decay at a rate: $j^{-1-\alpha/d}$.
- Learning using random ReLU features suffers from CoD under the L^∞ norm.

L^∞ upper bound

Assume there exists $\beta > 0$ and $M > 0$ such that

$$\mu_j \asymp j^{-\beta-1}, \quad \sup_{j\in\mathbb{N}^+} \|e_j\|_{C_0(\mathcal{X})} \leq M.$$

Theorem 15

Assume $\|f^*\|_{\mathcal{H}_k} \leq 1$. Under the above assumption, we have: With probability at least $1-\delta$ over the sampling of $\{(x_i, y_i)\}_{i=1}^n$, KRR have the rate

$$\|\hat{f}_0 - f^*\|_{C_0(\mathcal{X})} \lesssim n^{-\frac{\beta}{2}} (\log(n/\delta))^{\frac{\beta+1}{2}}. \tag{5}$$

Remark:

- KRR is nearly **optimal** in the noiseless regime when L^∞ learning.

Summary

Takeaways

- Learning in the noiseless regime can be **faster** than **fast rate**!
- Our duality framework provides **a principled approach** to understanding the learning in the noiseless regime.

Summary

Takeaways

- Learning in the noiseless regime can be **faster** than **fast rate**!
- Our duality framework provides **a principled approach** to understanding the learning in the noiseless regime.

Future Directions

- Apply our framework to analyze neural networks and Barron spaces.
- Beyond L^2 and L^∞ norm, such as Sobolev norms?

Thank You!

- Hongrui Chen, Jihao Long, **Lei Wu**, *A duality framework for analyzing random feature and two-layer neural networks*, arXiv: 2305.05642 (to appear in Annals of Statistics).
- Jihao Long, Xiaojun Peng, **Lei Wu**, *Optimal rates for kernel ridge regression in the noiseless regime*, in preparation.