Sharp convergence rates for spectral methods via the Feature Space Decomposition method

Zong Shang
(joint with Guillaume Lecué, and Zhifan Li)

CREST-ENSAE, Institut Polytechnique de Paris

March 12, 2026

"The profound study of nature is the most fertile source of mathematical discoveries."

— Joseph Fourier, The Analytical Theory of Heat, 1822.

Supervised learning problem and its solution

Benign Overfitting $\Rightarrow$ Feature Space Decomposition method.

- A tool to analyze the excess risk, improving the classical uniform convergence argument.
- A theoretical framework.

Supervised Regression Problems: $\mathcal{R} = (\mu_X, f^*, \xi)$, s.t.

$$Y = f^*(X) + \xi, \quad f^* \in L^2(\mu_X), \quad \mathbb{E}[\xi] = 0, \quad \xi \perp\!\!\!\perp X, \text{ and } \mathbb{E}[\xi^2] = \sigma_\xi^2.$$

Solutions: $(\mathcal{F}, \{\hat{f}_N\}_{N\in\mathbb{N}_+})$.

- Feature space: $\mathcal{F} \subset \{f : \Omega_X \to \mathbb{R},\ f \text{ measurable}\}$ (linear).
- Learning rule: $\{\hat{f}_N : (X_i, Y_i)_{i=1}^N \in (\Omega_X \times \mathbb{R})^N \mapsto \hat{f}_N((X_i, Y_i)_{i=1}^N, \cdot) \in \mathcal{F}\}_{N\in\mathbb{N}_+}$. Abbr. $\hat{f}$.

Objective: characterize $\|\hat{f} - f^*\|^2_{L^2(\mu_X)}$.

Feature Space Decomposition (FSD)

Definition

Any dirct-sum decomposition $\mathcal{F} = V_J \oplus V_{J^c}$ is called a Feature Space Decomposition (FSD).
Equivalently, $I_{\mathcal{F}} = P_{V_J} + P_{V_{J^c}}$, $f = f_J + f_{J^c}$, where $f_J = P_{V_J} f$.

For any FSD, $\hat{f} - f^* = \hat{f}_J - f^*_J + \hat{f}_{J^c} - f^*_{J^c}$. Therefore,

$$\left\|\hat{f} - f^*\right\|^2_{L^2(\mu_X)} \begin{cases} = \left\|\hat{f}_J - f^*_J\right\|^2_{L^2(\mu_X)} + \left\|\hat{f}_{J^c} - f^*_{J^c}\right\|^2_{L^2(\mu_X)}, & \text{if } V_J \perp V_{J^c}, \\ \leq 2\left\|\hat{f}_J - f^*_J\right\|^2_{L^2(\mu_X)} + 2\left\|\hat{f}_{J^c} - f^*_{J^c}\right\|^2_{L^2(\mu_X)}, & \text{otherwise.} \end{cases}$$

By triangular inequality for $\|\hat{f}_{J^c} - f^*_{J^c}\|_{L^2(\mu_X)}$,

$$\|\hat{f} - f^*\|^2_{L^2(\mu_X)} \leq 2\|\hat{f}_J - f^*_J\|^2_{L^2(\mu_X)} + 4\|\hat{f}_{J^c}\|^2_{L^2(\mu_X)} + 4\|f^*_{J^c}\|^2_{L^2(\mu_X)}. \tag{1}$$

FSD method: seeking rate function r and deviation function δ, s.t., $\forall$ FSD (V_J, V_{J^c}),

$$\mathbb{P}\left((1) \leq r^2(V_J, V_{J^c})\right) \geq 1 - \delta(V_J, V_{J^c}).$$

Two subspaces, two approaches

There are two fundamentally different ways to control $\|\hat{f} - f^*\|^2_{L^2(\mu_X)}$.

$$\|\hat{f}_J - f^*_J\|^2_{L^2(\mu_X)} + \|\hat{f}_{J^c}\|^2_{L^2(\mu_X)} + \|f^*_{J^c}\|^2_{L^2(\mu_X)}.$$

- V_J: estimation subspace.

  $\hat{f}_J$ estimates f^*_J, $\Rightarrow$ contribution to excess risk: cancellation between $\hat{f}_J$ and f^*_J

- V_{J^c}: free subspace.

  $\hat{f}_{J^c}$ fulfills certain tasks determined by the definition of $\hat{f}$, but **NOT** estimating $f^*_{J^c}$,
  $\Rightarrow$ contribution to excess risk: smallness of $\|\hat{f}_{J^c}\|^2_{L^2(\mu_X)}$ and $\|f^*_{J^c}\|^2_{L^2(\mu_X)}$.

Optimal FSD

Define

$$(V_J^*, V_{J^c}^*) \in \operatorname{argmin} \left(r(V_J, V_{J^c}) : \mathcal{F} = V_J \oplus V_{J^c}\right)$$

be the optimal FSD. Then,

$$\mathbb{P}\left(\|\hat{f} - f^*\|_{L^2(\mu_X)}^2 \leq r^2(V_J^*, V_{J^c}^*)\right) \geq 1 - \delta(V_J^*, V_{J^c}^*).$$

For some problems and solutions, e.g., ridge, gradient descent, gradient flow, $\exists 0 < c, \delta < 1$, s.t.,

$$\mathbb{P}\left(\|\hat{f} - f^*\|_{L^2(\mu_X)}^2 \geq c r^2(V_J^*, V_{J^c}^*)\right) \geq 1 - \delta.$$

Remarkable phenomenon: for those classes of (μ_X, f^*, ξ) and $(\mathcal{F}, \hat{f}_N)$, the estimation error $\|\hat{f} - f^*\|_{L^2(\mu_X)}^2$ is "characterized" by an interpolation between these two distinct approaches.

Improve the Classical Statistical Learning Theory via FSD

$\|\hat{f} - f^*\|^2_{L^2(\mu_X)} \sim r^2(V_J^*, V_{J^c}^*)$, w.h.p. $\Rightarrow \hat{f}$ estimates **ONLY** $f_{J_*}^*$ (STRIKING!).

- Classical Statistical Learning Theory.

  Belief: estimation should happen over the entire $\mathcal{F}$. $\Rightarrow V_J = \mathcal{F}$.

  Belief: If $\hat{f}$ is consistent ($\|\hat{f} - f^*\|_{L^2(\mu_X)} \to 0$), $\hat{f}$ should estimate f^*, **NOT** only a part.

Learning theory has to address the following four questions:

(i) *What are (necessary and sufficient) conditions for consistency of a learning process based on the ERM principle?*

(ii) *How fast is the rate of convergence of the learning process?*

(iii) *How can one control the rate of convergence (the generalization ability) of the learning process?*

(iv) *How can one construct algorithms that can control the generalization ability?*

Figure: Vapnik 1995. "The Nature of Statistical Learning Theory", pp. 19, Springer.

The geometry of FSD

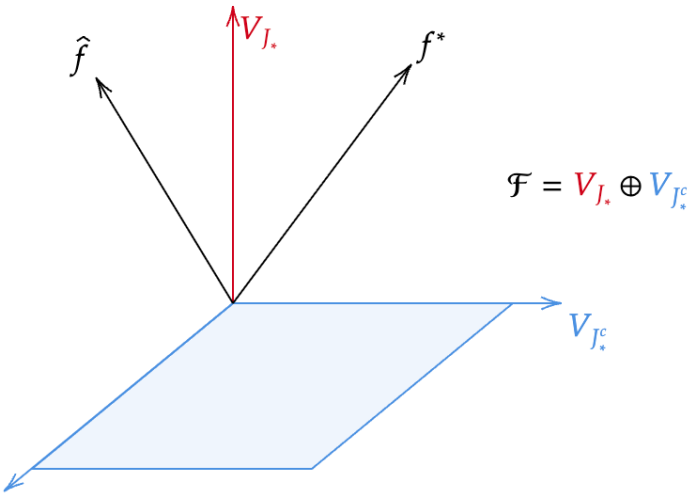


Figure: Only $f^*_{J_*}$ will be estimated by $\hat{f}_J$, $f^*_{J^c_*}$ will be regarded as the model noise of $f^*_{J_*}$.

$\mathcal{R}$: linear regression problem

- μ_X: centered, sub-Gaussian probability measure;
- $\boldsymbol{\beta}^* \in \mathbb{R}^p$: the unknown signal, s.t., $f^* : \boldsymbol{x} \in \mathbb{R}^p \mapsto f^*(\boldsymbol{x}) = \langle \boldsymbol{\beta}^*, \boldsymbol{x} \rangle \in \mathbb{R}$.
- $\Sigma = \mathbb{E}[X \otimes X] : \boldsymbol{v} \in \mathbb{R}^p \mapsto \Sigma \boldsymbol{v} = \mathbb{E}[X \langle X, \boldsymbol{v} \rangle] \in \mathbb{R}^p$.
- $\Sigma = \sum_{j=1}^p \sigma_j \boldsymbol{e}_j \otimes \boldsymbol{e}_j$: spectral decomposition, $\sigma_1 \geq \cdots \geq \sigma_p > 0$;
- $\mathcal{F} = \{f(\cdot) : f \in \mathcal{F}\} \longleftrightarrow \{\langle \boldsymbol{\beta}, \cdot \rangle : \boldsymbol{\beta} \in \mathbb{R}^p\} \cong \mathbb{R}^p$;
- $\|\hat{f} - f^*\|_{L^2(\mu_X)} = \|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2$.
- $\mathcal{R} = (\boldsymbol{\beta}^*, \Sigma, \sigma_\xi)$.

- $(X_1, Y_1), \cdots, (X_N, Y_N)$: i.i.d. copies of (X, Y).
- $\mathbb{X} : \boldsymbol{v} \in \mathbb{R}^p \mapsto \mathbb{X}\boldsymbol{v} = (\langle X_i, \boldsymbol{v} \rangle)_{i=1}^N \in \mathbb{R}^N$;
- $\mathbb{X}^\top : \boldsymbol{\lambda} \in \mathbb{R}^N \mapsto \sum_{i=1}^N \lambda_i X_i \in \mathbb{R}^p$.
- $\hat{\Sigma} = \frac{1}{N}\mathbb{X}^\top \mathbb{X} = \frac{1}{N}\sum_{i=1}^N X_i \otimes X_i$.

For any FSD, $\Sigma_J = P_J \Sigma P_J$, and $\Sigma_{J^c} = P_{J^c} \Sigma P_{J^c}$, $\mathbb{X}_J = \mathbb{X} P_J$, and $\mathbb{X}_{J^c} = \mathbb{X} P_{J^c}$.

Solution to $\mathcal{R}$: spectral methods

Definition

Let $(\varphi_t)_{t\geq 1}$ be a family of real-valued functions defined on $\mathbb{R}_+$, called the filter functions. For all $t \geq 1$, define $\hat{\boldsymbol{\beta}}$ by

$$\hat{\boldsymbol{\beta}} : \boldsymbol{y} \in \mathbb{R}^N \mapsto \hat{\boldsymbol{\beta}}(\boldsymbol{y}) = \frac{1}{N}\varphi_t(\hat{\Sigma})\mathbb{X}^\top \boldsymbol{y} = \frac{1}{N}\mathbb{X}^\top \varphi_t\left(\frac{1}{N}\mathbb{X}\mathbb{X}^\top\right)\boldsymbol{y}.$$

Here, $\varphi_t(\hat{\Sigma})$ is defined by spectral calculus. We abbreviate $\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ by $\hat{\boldsymbol{\beta}}$. Define $\psi_t(x) = 1 - x\varphi_t(x)$, called residual function.

Examples:

1. Ridge regression with tuning parameter t^{-1}, i.e., $\hat{\boldsymbol{\beta}} = \frac{1}{N}(\frac{1}{N}\mathbb{X}^\top\mathbb{X} + t^{-1}I_p)^{-1}\mathbb{X}^\top\boldsymbol{y}$ is a spectral method with

$$\varphi_t(x) = \frac{1}{t^{-1}+x}, \text{ and } \psi_t(x) = \frac{1}{xt+1}.$$

Examples of spectral methods

2. Gradient flow $\dot{\boldsymbol{\beta}}_t = -\nabla \frac{1}{2N} \|\boldsymbol{y} - \mathbb{X} \cdot\|_2^2(\boldsymbol{\beta}_t)$ for any $t \geq 1$ with $\boldsymbol{\beta}_1 = \mathbf{0}$. Then $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_t$ is a spectral method with

$$\varphi_t(x) = \frac{1 - \exp(-tx)}{x},\ \varphi_t(0) = t, \text{ and } \psi_t(x) = \exp(-tx).$$

3. Gradient Descent $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} - \eta \nabla \frac{1}{2N} \|\boldsymbol{y} - \mathbb{X} \cdot\|_2^2(\boldsymbol{\beta}_{t-1})$, $\boldsymbol{\beta}_1 = \mathbf{0}$. Then $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_t$ is a spectral method when η is small enough with

$$\varphi_t(x) = \frac{1 - (1 - \eta x)^t}{x}, \text{ and } \psi_t(x) = (1 - \eta x)^t.$$

4. Principle Components Regression (PCR) in dimension $d \leq p$, $\hat{\boldsymbol{\beta}} \in \operatorname{argmin}(\|\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 : \boldsymbol{\beta} \in \hat{V}_{\leq d})$, where $\hat{V}_{\leq d}$ is the subspace spanned by the first d eigenvectors of $\hat{\Sigma}$. Then

$$\varphi_t(x) = \frac{1}{x} \mathbb{1}(bt^{-1} \leq x), \text{ and } \psi_t(x) = \mathbb{1}(bt^{-1} > x).$$

FSD for spectral methods

Definition (estimation dimension)

Let $b > 0$, $t \geq 1$. The estimation dimension of the spectral method $\hat{\boldsymbol{\beta}}$ with filter function φ_t is defined as

$$k^* = k^*_{t^{-1},b} = \min\left\{k \in [p] : \sigma_{k+1} \leq bt^{-1}\right\}.$$

Define the optimal FSD be

$$V_J^* = \operatorname{span}(\boldsymbol{e}_j : 1 \leq j \leq k^*), \text{ and } V_{J^c}^* = \operatorname{span}(\boldsymbol{e}_j : k^* < j \leq p).$$

V_J^* is the subspace used for estimating $\boldsymbol{\beta}_J^*$, giving name for "estimation dimension".

$V_J^* \perp V_{J^c}^*$, hence

$$\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 = \|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\|_2^2 + \|\Sigma_{J^c}^{1/2}(\hat{\boldsymbol{\beta}}_{J^c} - \boldsymbol{\beta}_{J^c}^*)\|_2^2.$$

The meaning of filter function

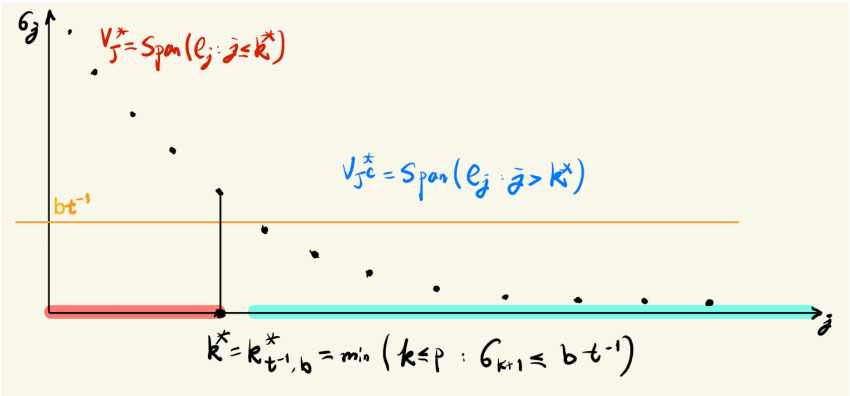


Figure: Given t^{-1}, $\hat{\boldsymbol{\beta}}$ uses $V_J^* = \mathrm{span}(\boldsymbol{e}_j : j \le k^*)$ to estimate $\boldsymbol{\beta}_J^*$, and use $V_{J^c}^*$ to absorb noise. This also explains why φ_t is called a filter function: the feature subspace corresponding to eigenvalues smaller than bt^{-1} is effectively filtered out from being used for estimation.

Definition of rate function $r(\cdot, \cdot)$

Let $J_* = \{1, \cdots, k^*\}$. Recall that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_J + \hat{\boldsymbol{\beta}}_{J^c}$.

$$r(V_J^*, V_{J^c}^*) = \|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_{J_*}^*\|_2 + \|\Sigma_{J_*^c}^{1/2}\boldsymbol{\beta}_{J_*^c}^*\|_2 \quad \text{bias } \hat{\boldsymbol{\beta}}$$

$$+ \sigma_\xi\sqrt{\frac{|J_*|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J_*^c}^2)}{N}} \quad \text{variance } \hat{\boldsymbol{\beta}}$$

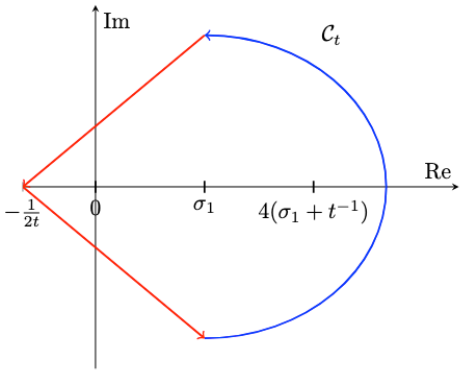
$\hat{\boldsymbol{\beta}}_{J_*}$ bias and variance $\quad$ $\hat{\boldsymbol{\beta}}_{J_*^c}$ bias and variance

1. As the terminology "residual" suggests, $\|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_{J_*}^*\|_2$ is the bias.
2. $\|\Sigma_{J_*^c}^{1/2}\boldsymbol{\beta}_{J_*^c}^*\|_2$ is the cost incurred by the un-estimated part $\boldsymbol{\beta}_{J_*^c}^*$ and also the approximation error of V_J^*.

Assumption

Suppose the family of filter functions $(\varphi_t)_{t\geq 1}$ is such that for all $t \geq 1$, φ_t has an holomorphic extension to an open subset of $\mathbb{C}$ containing some contour $\mathcal{C}_t$. Furthermore, there are two absolute constants $0 \leq c \leq C$ such that for all $t \geq 1$ and $0 \leq x \leq 8$,

$$\frac{c}{x + t^{-1}} \leq \varphi_t(x) \leq \frac{C}{x + t^{-1}}.$$



Fundamental theorem of spectral methods: informal

FSD method serves as a tool to help theorists analyze the excess risk.

Recall that $r(V_J^*, V_{J^c}^*) = \|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_{J_*}^*\|_2 + \|\Sigma_{J_*^c}^{1/2}\boldsymbol{\beta}_{J_*^c}^*\|_2 + \sigma_\xi\sqrt{\frac{|J_*|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J_*^c}^2)}{N}}$.

Under some assumptions, for any $\mathcal{R}$ characterized by $(\boldsymbol{\beta}^*, \Sigma, \sigma_\xi)$. Let $(\varphi_t)_{t\geq 1}$ be a family of filter functions satisfying the aforementioned assumption.

$$\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle\|_{L^2(\mu_X)} \lesssim r(V_J^*, V_{J^c}^*).$$

Moreover, if $X = \Sigma^{1/2}Z$ for some Z having independent and centered coordinates, then w.h.p.,

$$\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle\|_{L^2(\mu_X)} \gtrsim r(V_J^*, V_{J^c}^*).$$

Why do I call it fundamental?

Fundamental theorem is sharp

Recall that $r(V_J^*, V_{J^c}^*) = \|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_{J_*}^*\|_2 + \|\Sigma_{J_*^c}^{1/2}\boldsymbol{\beta}_{J_*^c}^*\|_2 + \sigma_\xi\sqrt{\frac{|J_*|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J_*^c}^2)}{N}}$.

Under some assumptions, for any $\mathcal{R}$ characterized by $(\boldsymbol{\beta}^*, \Sigma, \sigma_\xi)$. Let $(\varphi_t)_{t\geq 1}$ be a family of filter functions satisfying the aforementioned assumption.

$$\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle\|_{L^2(\mu_X)} \sim r(V_J^*, V_{J^c}^*).$$

1. For any $\mathcal{R} = (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, $r(V_J^*, V_{J^c}^*)$ characterizes $\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle\|_{L^2(\mu_X)}$ precisely, with no assumptions imposed on Σ nor on $\boldsymbol{\beta}^*, \sigma_\xi$.
2. Only $\|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_{J_*}^*\|_2$ depends on the choice of filter function.
3. Intrinsic drawback of spectral methods: J_* is independent with $\boldsymbol{\beta}^*$.

Partial order: comparing $\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\|^2_{L^2(\mu_X)}$ $\leftrightarrow$ comparing $r^2(V_J^*, V_{J^c}^*)$

FSD serves also as a theoretical framework.

Definition (Partial Order w.r.t. $\mathcal{R}$ on the set of spectral methods)

For any $\mathcal{R}$, and any filter functions $(\varphi_t^{(A)})_{t\geq 1}$, $(\varphi_t^{(B)})_{t\geq 1}$, any two tuning parameters t_A, t_B, we write $\hat{\boldsymbol{\beta}}_{t_A}^{(A)} \preceq_{\mathcal{R}} \hat{\boldsymbol{\beta}}_{t_B}^{(B)}$ if $r_{t_A}^{(A)}(V_J^*, V_{J^c}^*) = O(r_{t_B}^{(B)}(V_J^*, V_{J^c}^*))$ as $N, p \to \infty$.

Corollary

For any $\mathcal{R}$ and any $t \geq 1$, $\hat{\boldsymbol{\beta}}_t^{(A)} \preceq_{\mathcal{R}} \hat{\boldsymbol{\beta}}_t^{(B)}$, if and only if,
$\|\Sigma_{J_*}^{1/2} \psi_t^{(A)}(\Sigma) \boldsymbol{\beta}_{J_*}^*\|_2 = O(\|\Sigma_{J_*}^{1/2} \psi_t^{(B)}(\Sigma) \boldsymbol{\beta}_{J_*}^*\|_2)$.

Simple but important: $\psi_t^{(\mathrm{GF})}(x) = \exp(-tx)$, and $\psi_t^{(\mathrm{Ridge})}(x) = \frac{1}{xt+1}$. Then

$$\forall \mathcal{R}, \forall t \geq 1, \qquad \hat{\boldsymbol{\beta}}_t^{(\mathrm{GF})} \preceq_{\mathcal{R}} \hat{\boldsymbol{\beta}}_t^{(\mathrm{Ridge})}.$$

Is it fair to compare the same t?

Generalized saturation effect

Definition (Generalized Saturation Effect w.r.t. $\mathcal{R}$)

For any interval $I \subset [1, \infty)$, we write $\{\varphi_t^{(A)}\}_{t\in I} \preceq_{\mathcal{R}} \{\varphi_t^{(B)}\}_{t\in I}$, if as $N, p \to \infty$,

$$\inf_{t_A \in I} \left(r_{t_A}^{(A)}(V_J^*, V_{J^c}^*) \right) = O\left(\inf_{t_B \in I} \left(r_{t_B}^{(B)}(V_J^*, V_{J^c}^*) \right) \right).$$

Similarly, $\prec_{\mathcal{R}}$ if big-O is replaced by small-o.

1. Classical saturation effect in Sobolev regression setup.
   $\{\varphi_t^{(\mathrm{GF})}\}_{t\in\mathbb{R}_+} \prec_{\mathcal{R}} \{\varphi_t^{(\mathrm{Ridge})}\}_{t\in\mathbb{R}_+}$, for any $\mathcal{R} \in \mathfrak{R}_{\mathrm{Sob}}(s, \alpha)$, $s > 2$. Here, for any $\alpha > 1, s \geq 0$, define
   $\mathfrak{R}_{\mathrm{Sob}}(s, \alpha) = \{(\Sigma, \boldsymbol{\beta}^*, \sigma_\xi) : \sigma_j \sim j^{-\alpha}, \|\Sigma^{\frac{1-s}{2}} \boldsymbol{\beta}^*\|_2 < \infty, \sigma_\xi \text{ is constant}\}$ be the class of Sobolev regression problems.

Example: saturation effect in spiked covariance model

2. $\sigma_1 = \cdots = \sigma_k = \sigma$, $\sigma_{k+1} = \cdots = \sigma_p = \varepsilon$ for some $\sigma > \varepsilon$ and $k \lesssim N \lesssim p - k$. Let $|\langle \boldsymbol{\beta}^*, \boldsymbol{e}_j \rangle| = \alpha_* \mathbb{1}(j \le k)$ for some $\alpha_* > 0$. Let $\mathrm{SNR} = \frac{\|\Sigma^{1/2}\boldsymbol{\beta}^*\|_2}{\sigma_\xi} \cdot \frac{\sigma\sqrt{N}}{\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)}}$. Suppose $4 < \mathrm{SNR} \le b\frac{\sigma}{\varepsilon}$. Let $I = \{t > 1 : b^{-1}\varepsilon \le t^{-1} < \sigma\}$. Then $\{\varphi_t^{(\mathrm{GF})}\}_{t\in I} \prec_{\mathcal{R}} \{\varphi_t^{(\mathrm{Ridge})}\}_{t\in I}$ as $\mathrm{SNR}, \sigma/\varepsilon \to \infty$.

- The saturation effect is widely present.
- Developing a theory for an arbitrary $\mathcal{R}$ is important.

Let us go back to the method that produces the fundamental theorem.

How does FSD produce the main result?

$$\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim \|\Sigma_{J_*}^{1/2}(\hat{\boldsymbol{\beta}}_{J_*} - \boldsymbol{\beta}^*_{J_*})\|_2 + \|\Sigma_{J_*^c}^{1/2}\hat{\boldsymbol{\beta}}_{J_*^c}\|_2 + \|\Sigma_{J_*^c}^{1/2}\boldsymbol{\beta}^*_{J_*^c}\|_2.$$

Upper bound for $\|\Sigma_{J_*}^{1/2}(\hat{\boldsymbol{\beta}}_{J_*} - \boldsymbol{\beta}^*_{J_*})\|_2$

1. Identify new supervised regression problem.

   $\boldsymbol{y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi} = \mathbb{X}_{J_*}\boldsymbol{\beta}^*_{J_*} + \mathbb{X}_{J_*^c}\boldsymbol{\beta}^*_{J_*^c} + \boldsymbol{\xi}$, and treat $\mathbb{X}_{J_*^c}\boldsymbol{\beta}^*_{J_*^c} + \boldsymbol{\xi}$ as the noise in a new regression problem in which $\hat{\boldsymbol{\beta}}_{J_*}$ estimates $\boldsymbol{\beta}^*_{J_*}$.

$$\hat{\boldsymbol{\beta}}_{J_*}(\boldsymbol{y}) - \boldsymbol{\beta}^*_{J_*} = \underline{\hat{\boldsymbol{\beta}}_{J_*}(\mathbb{X}_{J_*}\boldsymbol{\beta}^*_{J_*}) - \boldsymbol{\beta}^*_{J_*}} + \underline{\hat{\boldsymbol{\beta}}_{J_*}(\mathbb{X}_{J_*^c}\boldsymbol{\beta}^*_{J_*^c} + \boldsymbol{\xi})}$$

2. Apply classical statistical analysis to $\hat{\boldsymbol{\beta}}_{J_*}$.

   Recall that $\hat{\boldsymbol{\beta}}_{J_*}(\mathbb{X}_J\boldsymbol{\beta}^*_{J_*}) = \varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*_{J_*}$. Let $\tilde{\boldsymbol{\beta}}_{J_*} = \varphi_t(\Sigma)\Sigma\boldsymbol{\beta}^*_{J_*}$. Then

$$\left\|\Sigma_{J_*}^{1/2}(\hat{\boldsymbol{\beta}}_{J_*}(\boldsymbol{y}) - \boldsymbol{\beta}^*_{J_*})\right\|_2 \le \|\Sigma_{J_*}^{1/2}(\tilde{\boldsymbol{\beta}}_{J_*} - \boldsymbol{\beta}^*_{J_*})\|_2$$
$$+ \|\Sigma_{J_*}^{1/2}(\hat{\boldsymbol{\beta}}_{J_*}(\mathbb{X}_{J_*}\boldsymbol{\beta}^*_{J_*}) - \tilde{\boldsymbol{\beta}}_{J_*})\|_2 + \|\Sigma_{J_*}^{1/2}\hat{\boldsymbol{\beta}}_{J_*}(\mathbb{X}_{J_*^c}\boldsymbol{\beta}^*_{J_*^c} + \boldsymbol{\xi})\|_2.$$

Classical analysis for spectral methods applied to $\hat{\boldsymbol{\beta}}_J$

1. By definition, $\tilde{\boldsymbol{\beta}}_{J_*} = \Sigma\varphi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}$, so
   $\|\Sigma^{1/2}_{J_*}(\boldsymbol{\beta}^*_{J_*} - \tilde{\boldsymbol{\beta}}_{J_*})\|_2 = \|\Sigma^{1/2}_{J_*}(I - \Sigma\varphi_t(\Sigma))\boldsymbol{\beta}^*_{J_*}\|_2 = \|\Sigma^{1/2}_{J_*}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\|_2$.
2. Using the method in Li, Gan, Shi, and Lin [2024], w.h.p.,
   $\|\Sigma^{1/2}_{J_*}(\hat{\boldsymbol{\beta}}_{J_*}(\mathbb{X}_{J_*}\boldsymbol{\beta}^*_{J_*}) - \tilde{\boldsymbol{\beta}}_{J_*})\|_2 \lesssim \|\Sigma^{1/2}_{J_*}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\|_2 +$ higher order small.
3. One may also prove that the variance part w.h.p. satisfies
   $\|\Sigma^{1/2}_{J_*}\hat{\boldsymbol{\beta}}_{J_*}(\mathbb{X}_{J^c_*}\boldsymbol{\beta}^*_{J^c_*} + \boldsymbol{\xi})\|_2 \lesssim \|\Sigma^{1/2}_{J^c_*}\boldsymbol{\beta}^*_{J^c_*}\|_2 + \sigma_\xi\sqrt{\frac{|J_*|}{N}}$.

Combining, without additional effort, w.h.p.,

$$\left\|\Sigma^{1/2}_{J_*}(\hat{\boldsymbol{\beta}}_{J_*} - \boldsymbol{\beta}^*_{J_*})\right\|_2 \lesssim \|\Sigma^{1/2}_{J_*}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\|_2 + \|\Sigma^{1/2}_{J_*{}^c}\boldsymbol{\beta}^*_{J_*{}^c}\|_2 + \sigma_\xi\sqrt{\frac{|J_*|}{N}}.$$

In spectral methods, FSD gives "correct" oracle $\boldsymbol{\beta}^*_{J_*}$ and bias $\|\Sigma^{1/2}_{J^c_*}\boldsymbol{\beta}^*_{J^c_*}\|_2$.

In this work, FSD serves as an outer analytical layer built on top of the classical analysis of spectral methods.

Upper bound for $\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$

This part remains a mystery and requires creative insight!

$$\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}(\boldsymbol{y})\|_2 \le \|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}_J^*\|_2 + \|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}_{J^c}^*\|_2 + \|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\frac{1}{N}\mathbb{X}^\top\boldsymbol{\xi}\|_2.$$

By $\varphi_t(\hat{\Sigma})\hat{\Sigma} = I_p - \psi_t(\hat{\Sigma})$, and $\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_J^* = \mathbf{0}$,

$$\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}_J^*\|_2 = \|\Sigma_{J^c}^{1/2}(\psi_t(\hat{\Sigma}) - \psi_t(\Sigma))\boldsymbol{\beta}_J^*\|_2.$$

Inserting $(\Sigma + t^{-1})(\Sigma + t^{-1})^{-1}$ to

$$\psi_t(\hat{\Sigma}) - \psi_t(\Sigma) = \frac{1}{2\pi i}\oint_{\mathcal{C}_t}(\hat{\Sigma} - zI_p)^{-1}(\hat{\Sigma} - \Sigma)(\Sigma - zI_p)^{-1}\psi_t(z)dz$$

yields $\frac{\square}{t}\|\Sigma_J^{-1/2}\boldsymbol{\beta}_J^*\|_2$. The other two terms are not difficult.

Going beyond Spectral Methods: Feature Learning Property

This part is based on joint work with G. Lecué, T. Suzuki, and T. Wakayama (on-going).

How the good features are automatically learned by a neural network?

- Given (μ_X, f^*, σ_ξ), $(X_i, Y_i)_{i=1}^N$, and an estimator $\hat{f}_N$.
- There exist a data-dependent RKHS $\mathcal{H}_{\mathrm{fea}} \subset L^2(\mu_X)$, called learned feature subspace, a latent estimator $\hat{g}_N \in \mathcal{H}_{\mathrm{fea}}$ and the oracle $g^*_{\mathcal{H}_{\mathrm{fea}}} \in \mathrm{argmin}(\|f^* - g\|_{L^2(\mu_X)} : g \in \mathcal{H}_{\mathrm{fea}})$, such that the following hold.
  - $\|f^* - g^*_{\mathcal{H}_{\mathrm{fea}}}\|_{L^2(\mu_X)}$ is small (approximation error);
  - $\|g^*_{\mathcal{H}_{\mathrm{fea}}} - \hat{g}_N\|_{L^2(\mu_X)}$ is small (estimation error);
  - $\|\hat{g}_N - \hat{f}_N\|_{L^2(\mu_X)}$ is small, ($\hat{g}_N$ can explain $\hat{f}_N$)

  and $\|\hat{g}_N - g^*_{\mathcal{H}_{\mathrm{fea}}}\|^2_{L^2(\mu_X)}$ decreases when $g^*_{\mathcal{H}_{\mathrm{fea}}}$ gets aligned with the top k eigenvectors of $\Sigma = \mathbb{E}[\phi_{\mathrm{fea}}(X) \otimes \phi_{\mathrm{fea}}(X) | (X_i, Y_i)_{i=1}^N]$ for some $k = o(N)$.

For almost any supervised regression problem, mean-field Langevin dynamics trained shallow neural network has feature learning property.

Take-home messages

- FSD as analytical tool: how features are utilized.
  - Estimation occurs only on V_J.
  - For a large class of spectral methods and almost any $\mathcal{R} = (\boldsymbol{\beta}^*, \Sigma, \sigma_\xi)$, w.h.p., $\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\|_{L^2(\mu_X)} \sim r(V_J^*, V_{J^c}^*)$, where

$$r(V_J^*, V_{J^c}^*) = \|\Sigma_{J_*}^{1/2} \psi_t(\Sigma) \boldsymbol{\beta}_{J_*}^*\|_2 + \|\Sigma_{J_*^c}^{1/2} \boldsymbol{\beta}_{J_*^c}^*\|_2 + \sigma_\xi \sqrt{\frac{|J_*|}{N}} + \sigma_\xi t \sqrt{\frac{\mathrm{Tr}(\Sigma_{J_*^c}^2)}{N}}.$$

    - For the same $\mathcal{R}$, all spectral methods use the same features for estimation.
    - For the same $\mathcal{R}$ and the same t, the estimation error of different spectral methods depends on the convergence rate of ψ_t.
    - The saturation effect is widespread and not limited to Sobolev regression.
- FSD as theoretical framework: how features are constructed (feature learning).
  - The learned feature subspace $\mathcal{H}_{\mathrm{fea}}$ approaches f^*;
  - The oracle in $\mathcal{H}_{\mathrm{fea}}$ is "aligned" with top k eigenfunctions of the integral operator of $\mathcal{H}_{\mathrm{fea}}$.
  - Mean-field Langevin dynamics has feature learning property.

End

Thank you for your attention!

FSD in classification

For any FSD and any $f_J^* \in V_J$,

$$
\begin{aligned}
& P\mathcal{L}_{\hat{f}}^{(0,1)} \\
& = \mathbb{P}\left(Y\hat{f}(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) && (2) \\
& + \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y f_J^*(X) < 0\right) && (3) \\
& + \mathbb{P}\left(Y f_J^*(X) < 0\right) - \mathbb{P}\left(Y\left(\eta(X) - \frac{1}{2}\right) < 0\right), && (4)
\end{aligned}
$$

- (2): error caused by free part $\hat{f}_{J^c}$;
- (3): prediction error caused by $\hat{f}_J$ compared to that of f_J^* ; and
- (4): approximation error compared with that of the Bayes rule.

Counterparts of $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}^2$, $\|\hat{f}_J - f_J^*\|_{L^2(\mu_X)}^2$, and $\|f_{J^c}^*\|_{L^2(\mu_X)}^2$ in regression.

NTK does not have feature learning property

Example. Consider the NTK parameterization of the shallow neural network $f_{\boldsymbol{\theta}}(\cdot) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} a_j \sigma(\langle \boldsymbol{w}_j, \cdot \rangle)$ where $\boldsymbol{\theta} = (\boldsymbol{a}, \mathbb{W})$. Define $\hat{f}_N \in \operatorname{argmin}(P_N \ell_{f_{\boldsymbol{\theta}}}^{(2)} : \boldsymbol{\theta} \in \mathbb{R}^M \times \mathbb{R}^{M \times d})$, where $P_N \ell_{f_{\boldsymbol{\theta}}}^{(2)} = \frac{1}{N} \sum_{i=1}^{N} (Y_i - f_{\boldsymbol{\theta}}(X_i))^2$, and suppose $\hat{f}_N$ is computed via gradient flow $(f_t)_{t=0}^{\infty}$ initialized at $f_0(\cdot) = \frac{1}{\sqrt{M}} \sum_{j=1}^{M} a_j^{(0)} \sigma(\langle \boldsymbol{w}_j^{(0)}, \cdot \rangle)$ where $a_j^{(0)} \sim \mathcal{N}(0,1)$ and $\boldsymbol{w}_j^{(0)} \sim \mathcal{N}(\boldsymbol{0}, I_d)$ for each $1 \le j \le M$. Then when $M \to \infty$, f_t converges in probability to the solution of the kernel gradient flow

$$\partial_t f_t(\cdot) = -\frac{2}{N} \sum_{i=1}^{N} K_{\mathrm{NTK}}(\cdot, X_i)(f_t(X_i) - Y_i),$$

where $K_{\mathrm{NTK}} : (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \Omega_X \times \Omega_X \mapsto \mathbb{E}[\langle \nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{x}_1), \nabla_{\boldsymbol{\theta}} f_0(\boldsymbol{x}_2) \rangle]$. As a result, $\hat{f}_N \in \mathcal{H}_{\mathrm{NTK}}$, where $\mathcal{H}_{\mathrm{fea}}$ is the RKHS generated by K_{NTK} and is independent with μ and with $(X_i, Y_i)_{i=1}^{N}$, hence is **not** problem-specific. Therefore, NTK parameterization shallow neural network does not have feature learning property.

Two stages trained NN has feature learning property

Example. Train NN by gradient flow $(f_t)_{t=0}^T$ in rich regime stopped at certain time $T > 0$. Denote $\boldsymbol{w}_1^{(T)}, \cdots, \boldsymbol{w}_M^{(T)}$ be the learned features of f_T.

- Define $\phi_{\text{fea}} : \boldsymbol{x} \in \Omega_X \mapsto (\sigma(\langle \boldsymbol{w}_j^{(T)}, \boldsymbol{x} \rangle))_{j=1}^M$; let $\mathcal{H}_{\text{fea}}$ be the RKHS generated by K_{fea}, called learned feature kernel.
- Let $\hat{f}_N$ be a ridge regression with some tuning parameter λ on top of the RKHS generated by $K_{\text{fea}} : (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \Omega_X \times \Omega_X \mapsto K_{\text{fea}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \phi_{\text{fea}}(\boldsymbol{x}_1), \phi_{\text{fea}}(\boldsymbol{x}_2) \rangle$.
- $\hat{f}_N \in \mathcal{H}_{\text{fea}}$, and $\hat{g}_N$ is ridge on $\mathcal{H}_{\text{fea}}$, hence alignment property.

- $\mathcal{H}_{\text{fea}}$ is problem-specific. For single- and multi-index regression problems, it can be shown that $\mathcal{H}_{\text{fea}}$ has a small approximation error.
- The $\hat{f}_N$ obtained in this way is called a two-stage trained shallow neural network, $\Rightarrow$ feature learning property for solving those supervised regression problems.

Feature Learning reduces "approximation error"

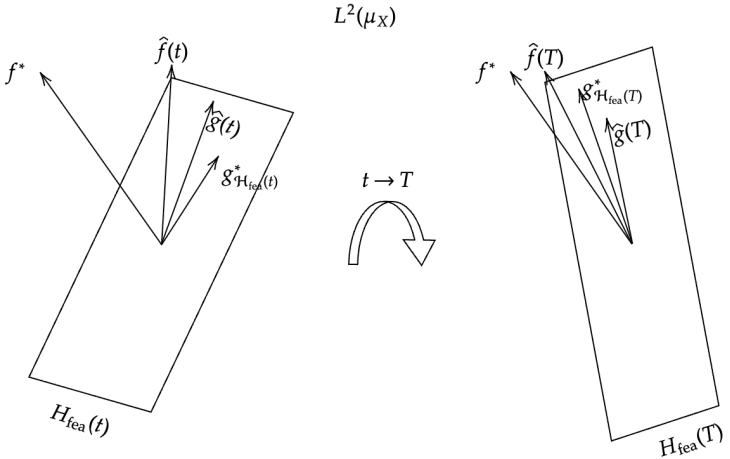


Figure: $\mathcal{H}_{\text{fea}}$ is getting closer to f^* during training: $\|f^* - g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)}$ is getting small.

Feature Learning ruduces "estimation error"

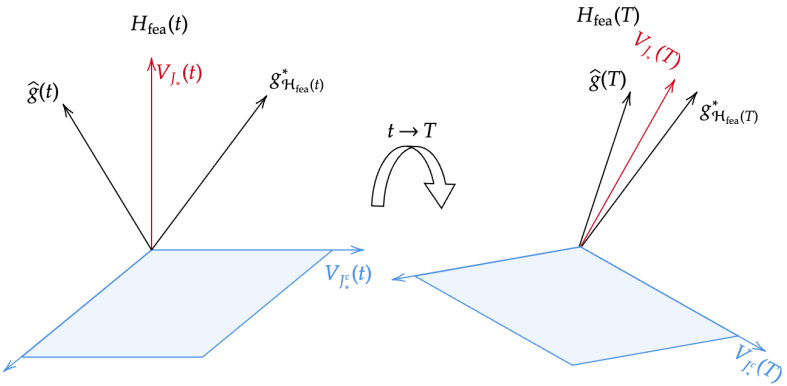


Figure: $g_{\mathcal{H}^*_{\text{fea}}}$ is getting aligned with the top k eigenvectors.

Mean-field Langevin dynamics has feature learning property

- Let $\Theta \subset \mathbb{R}^{d+1}$ compact, and $\sigma : \mathbb{R} \to \mathbb{R}$ some continuous function. For any $\boldsymbol{\theta} = (a, \boldsymbol{w}) \in \Theta$, $\phi(\cdot, \boldsymbol{\theta}) = a\sigma(\langle \boldsymbol{w}, \cdot \rangle)$, $\mathcal{F} = \{f_\nu(\cdot) = \int_\Theta \phi(\cdot, \boldsymbol{\theta}) d\nu(\boldsymbol{\theta}) : \nu \in \mathcal{P}_{\mathrm{ac}}(\Theta)\}$.
- Let $\mathrm{Ent}^-(\nu) = \int_\Theta \log(\frac{d\nu}{d\mathrm{Leb}}) d\nu$. For any $\lambda \geq 0$,

$$\hat{\nu}_\lambda \in \operatorname*{argmin}_{\nu \in \mathcal{P}_{\mathrm{ac}}(\Theta)} \left(\frac{1}{N} \sum_{i=1}^N (Y_i - f_\nu(X_i))^2 + \lambda \mathrm{Ent}^-(\nu) \right), \text{ and } \hat{f}_N = f_{\hat{\nu}_\lambda}$$

  is called mean-field Langevin dynamics.
- Let $\hat{\varphi} = (P_W)_\sharp \hat{\nu}_\lambda$, where $P_W : \boldsymbol{\theta} = (a, \boldsymbol{w}) \mapsto \boldsymbol{w}$. Let $\hat{a} : \boldsymbol{w} \mapsto \mathbb{E}_{\hat{\nu}_\lambda}[a|\boldsymbol{w}] = \int a d\hat{\nu}_\lambda(a|\boldsymbol{w})$.
- Let $\phi_{\mathrm{fea}} : \boldsymbol{x} \mapsto \sigma(\langle \boldsymbol{x}, \cdot \rangle)$, $K_{\mathrm{fea}} : (\boldsymbol{x}_1, \boldsymbol{x}_2) \mapsto \langle \phi_{\mathrm{fea}}(\boldsymbol{x}_1), \phi_{\mathrm{fea}}(\boldsymbol{x}_2) \rangle_{L^2(\hat{\varphi})}$, and

$$\mathcal{H}_{\mathrm{fea}} = \{f_g(\cdot) = \langle \phi_{\mathrm{fea}}(\cdot), g \rangle_{L^2(\hat{\varphi})} : g \in L^2(\hat{\varphi})\}.$$

- One can prove $\hat{g}_N : h \in \mathcal{H}_{\mathrm{fea}} \mapsto \langle \hat{a}, h \rangle_{L^2(\hat{\varphi})}$ satisfies $\hat{f}_N(\cdot) = \hat{g}_N(\phi_{\mathrm{fea}}(\cdot))$, and $\hat{g}_N$ is a convex regularized M-estimator.
- For "almost any" supervised regression problem, $\hat{f}_N$ has feature learning property.

Open problem: is PCR the minimal element?

Given any $\mathcal{R}$, what is the minimal element within $\preceq_{\mathcal{R}}$ (and $\preceq_{\mathcal{R}}$ for $I = [1, \infty)$ respectively)?

PCR: $\psi_t(x) = \mathbb{1}(bt^{-1} > x)$ and $k^* = \min\{k \in [p] : \sigma_{k+1} \leq bt^{-1}\}$, hence $\|\Sigma_{J_*}^{1/2} \psi_t^{(\mathrm{PCR})}(\Sigma)\boldsymbol{\beta}_{J_*}^*\|_2 = 0$. However, PCR's filter function $\varphi_t(x) = \frac{1}{x}\mathbb{1}(bt^{-1} \leq x)$ cannot be holomorphic extended over $\mathcal{C}_t$.

Conjecture: PCR is the minimal element of the partial order $\preceq_{\mathcal{R}}$ for any $\mathcal{R}$.

References I

Yicheng Li, Weiye Gan, Zuoqiang Shi, and Qian Lin. Generalization Error Curves for Analytic Spectral Algorithms under Power-law Decay, July 2024. URL http://arxiv.org/abs/2401.01599. arXiv:2401.01599.