# Feature Space Decomposition

Zong Shang

CREST-ENSAE, Institut Polytechnique de Paris

April 2, 2026

# Supervised learning problem and its solution

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 $\|\hat{f} - f^*\|^2_{L^2(\mu_X)}$. Assume $f^* \in \mathcal{F}$.

# Feature Space Decomposition (FSD)

## Definition

Any direct-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,

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

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

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

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

$$\mathbb{P}\Big(\text{R.H.S. of (1)} \leq r^2(V_J, V_{J^c})\Big) \geq 1 - \delta(V_J, V_{J^c}).$$

# 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.,

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

$\Rightarrow \|\hat{f} - f^*\|_{L^2(\mu_X)}^2 \sim r^2(V_J^*, V_{J^c}^*)$, w.h.p.: for those classes of $(\mu_X, f^*, \xi)$ 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

- Classical Statistical Learning Theory.

  Belief: An estimator should use all the features to estimate (no waste).

  Belief: Estimation should happen over the entire $\mathcal{F}$ ($V_J = \mathcal{F}$).

  Belief: If $\hat{f}$ is consistent ($\|\hat{f} - f^*\|_{L^2(\mu_X)} \to 0$), $\hat{f}$ should estimate $f^*$.

**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.

# Question (i): Consistency (Benign Overfitting)

Benign Overfitting: motivation for rethinking consistency.
Fix canonical basis $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_p\}$ of $\mathbb{R}^p$. Let $\boldsymbol{\beta}^* \in \mathbb{R}^p$, s.t.,

$$Y = \langle \boldsymbol{\beta}^*, X \rangle + \xi.$$

The minimum $\ell_q$-norm interpolant estimator is defined as

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \left( \|\boldsymbol{\beta}\|_q : \mathbb{X}\boldsymbol{\beta} = \boldsymbol{y} \right),$$

where $\mathbb{X} = [X_1 | \cdots | X_N]^\top \in \mathbb{R}^{N \times p}$, $\boldsymbol{y} = (Y_1, \cdots, Y_N)$. [Bartlett, Long, Lugosi, and Tsigler, 2020, Chinot, Loffler, and Van de Geer, 2022, Donhauser, Ruggeri, Stojanovic, and Yang, 2022, Wang, Donhauser, and Yang, 2022, Boyer, 2022].

Belief: $\hat{\boldsymbol{\beta}}$ should not be consistent, because of overfitting/interpolation.

**Definition.** We say $\hat{\boldsymbol{\beta}}$ exhibits benign overfitting (B.O.) under a given limit, if

$$\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\|_{L^2(\mu_X)} \to 0.$$

# Overfitting estimator and Benign Overfitting estimator

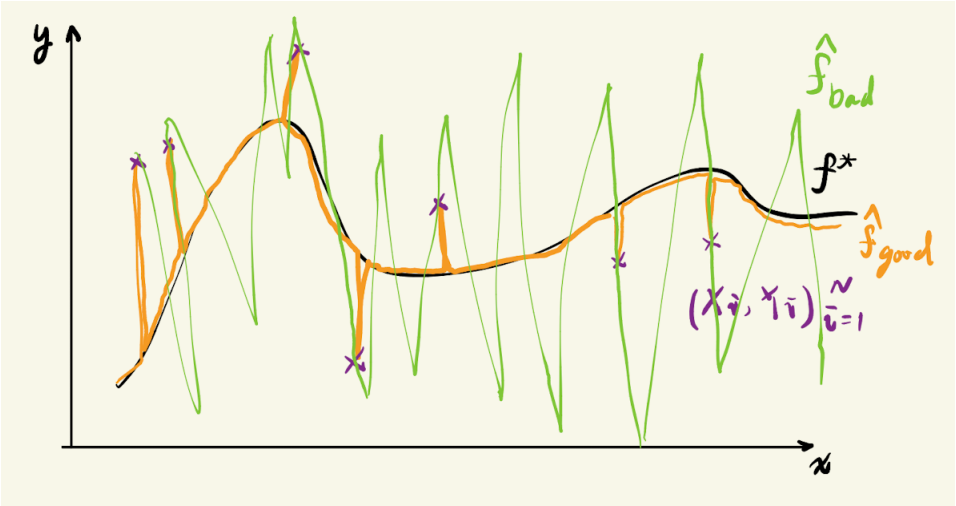


Figure: [Belkin, Rakhlin, and Tsybakov, 2019]

# Classical Statistical Learning Theory fails to explain

Notation: $P_N \ell_f$ empirical risk; $P\ell_f = \mathbb{E}[P_N \ell_f]$ population risk; $P_N \mathcal{L}_f = P_N \ell_f - P_N \ell_{f^*}$ empirical excess risk; and $P\mathcal{L}_f = P\ell_f - P\ell_{f^*}$ population excess risk.

1. Localization: Take any deterministic $\mathcal{G} \subset \mathcal{F}$, s.t. $\hat{f} \in \mathcal{G}$, w.h.p..

2. Uniform Convergence argument. [Van De Geer, 2006, Koltchinskii, 2011].

$$P\mathcal{L}_{\hat{f}} = P_N \mathcal{L}_{\hat{f}} + (P - P_N)\mathcal{L}_{\hat{f}} \leq P_N \mathcal{L}_{\hat{f}} + \sup((P - P_N)\mathcal{L}_f : f \in \mathcal{G}),$$

Take $P_N \ell_{\boldsymbol{\beta}} = \frac{1}{N} \sum_{i=1}^N (Y_i - \langle \boldsymbol{\beta}, X_i \rangle)^2$. Then $P\mathcal{L}_{\boldsymbol{\beta}} = \|\langle \boldsymbol{\beta} - \boldsymbol{\beta}^*, X \rangle\|^2_{L^2(\mu_X)}$.

- Def: $P_N \mathcal{L}_{\hat{\boldsymbol{\beta}}} = P_N \ell_{\hat{\boldsymbol{\beta}}} - P_N \ell_{\boldsymbol{\beta}^*} = -(1 + o(1))\sigma_\xi^2$, because $\boldsymbol{y} = \mathbb{X}\hat{\boldsymbol{\beta}} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$.
- We want: $P\mathcal{L}_{\hat{\boldsymbol{\beta}}} = o(1)\sigma_\xi^2$.
- We need: $\sup((P - P_N)\mathcal{L}_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathcal{G}) = (1 + o(1))\sigma_\xi^2$ to use U.C.

B.O.: a paradox that classical statistical learning theory fails to explain.
Next: non-trivial upper bound for $\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\|_{L^2(\mu_X)}$ via FSD $\Rightarrow$ suff. cond. for B.O.

## • $J$ defines new estimator: $\hat{\boldsymbol{\beta}}_J$ is a RERM.

Let $\Sigma = \mathbb{E}[X \otimes X]$. Then $\|\langle \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*, X\rangle\|_{L^2(\mu_X)} = \|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2$. By FSD,

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

We first study $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J)\|_2$. Key: $\hat{\boldsymbol{\beta}}_J$ is RERM.

Notation: For any $J \subset [p]$, let $V_J = \mathrm{span}(\boldsymbol{e}_j : j \in J)$, and $\mathbb{X}_J = \mathbb{X}P_J$, $\mathbb{X}_{J^c} = \mathbb{X}P_{J^c}$.

**Proposition** Define the non-linear metric embedding operator

$$\mathcal{A} : \boldsymbol{\mu} \in \mathbb{R}^N \mapsto \mathcal{A}[\boldsymbol{\mu}] \in \mathrm{argmin}(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{\mu}), \text{ then} \tag{2}$$
$$\hat{\boldsymbol{\beta}}_J \in \mathrm{argmin}\left(P_N \ell_{\boldsymbol{\beta}_J} + \|\boldsymbol{\beta}_J\|_q^q\right), \text{ where } P_N \ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$$

is a random loss determined by $\mathbb{X}_{J^c}$. This is called **self-regularization.**

**Implication**: $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J)\|_2^2$ is the estimation error of RERM. We need to understand this random loss function.

# Asymptotic Geometric Analysis for and from Statistical Learning Theory

**AGA for SLT**: By duality, $\|\mathcal{A}[\boldsymbol{\mu}]\|_q = \min(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{\mu}) = \max(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle : \|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\|_{q'} \le 1)$.

When $N \lesssim \left(\frac{\ell_*(\Sigma_{J^c}^{1/2} B_q)}{\mathrm{diam}(\Sigma_{J^c}^{1/2} B_q)}\right)^2$ and $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$, w.h.p.,

$$\Omega_{\mathsf{DM,reg}} = \left\{\forall \boldsymbol{\lambda} \in \mathbb{R}^N : c\ell_*(\Sigma_{J^c}^{1/2} B_q^p)\|\boldsymbol{\lambda}\|_2 \le \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \le C\ell_*(\Sigma_{J^c}^{1/2} B_q^p)\|\boldsymbol{\lambda}\|_2\right\}$$

$$\subset \left\{\forall \boldsymbol{\mu} \in \mathbb{R}^N : \frac{\|\boldsymbol{\mu}\|_2}{C\ell_*(\Sigma_{J^c}^{1/2} B_q^p)} \le \|\mathcal{A}[\boldsymbol{\mu}]\|_q \le \frac{\|\boldsymbol{\mu}\|_2}{c\ell_*(\Sigma_{J^c}^{1/2} B_q^p)}\right\}.$$

**Implication**: Recall $P_N \ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J]\|_q^q$,

$$\hat{\boldsymbol{\beta}}_J \in \mathrm{argmin}\left(P_N \ell_{\boldsymbol{\beta}_J} + \|\boldsymbol{\beta}_J\|_q^q\right) \approx \mathrm{argmin}\left(\frac{1}{N^{\frac{q}{2}}}\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J\|_2^q + \frac{\ell_*(\Sigma_{J^c}^{1/2} B_q^p)^q}{N^{\frac{q}{2}}}\|\boldsymbol{\beta}_J\|_q^q\right).$$

Upper bound of $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\|_2$ follows.

Insight: when $q = 1$, $\hat{\boldsymbol{\beta}}_J$ is almost square-root LASSO.

Rmk: in classification, $\hat{\boldsymbol{\beta}}_J$ is squared hinge loss SVM.

**AGA from SLT**: generalized Dvoretzky-Milman theorem for general probability measures.

# Free subspace and the energy of $\hat{\boldsymbol{\beta}}_{J^c}$

Recall:

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

Fact: $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]$ and $\|\Sigma^{1/2}_{J^c}\|_{\ell_q\to\ell_2} = \mathrm{diam}(\Sigma^{1/2}_{J^c}B^p_q)$. Therefore,

$$\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2 \le \|\Sigma^{1/2}_{J^c}\|_{\ell_q\to\ell_2}\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]\|_q \lesssim \frac{\mathrm{diam}(\Sigma^{1/2}_{J^c}B^p_q)}{\ell_*(\Sigma^{1/2}_{J^c}B^p_q)}\|\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J\|_2.$$

It is easy to prove

$$\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2 \lesssim \sqrt{\frac{N}{d_*(\Sigma^{1/2}_{J^c}B^p_{q'})}}\sigma_\xi. \tag{3}$$

Now we have an upper bound for $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J)\|_2 + \|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2 + \|\Sigma^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2$. Sending it to $0$ gives us sufficient conditions.

# Sufficient conditions for B.O.

**Sufficient conditions (informal)**: there exists admissible FSD, s.t.

$$\left\|\left|\boldsymbol{\beta}^*_J \odot |\boldsymbol{\beta}^*_J|^{\odot(q-2)}\right|\right\| = o\left(\frac{N^{q/2}}{\ell_*^q(\Sigma^{1/2}_{J^c} B^p_q)}\right) \quad , \quad \|\Sigma^{1/2}_{J^c}\boldsymbol{\beta}^*_{J^c}\|_2 = o(1) \,, \qquad \text{bias } \hat{\boldsymbol{\beta}}$$

$$|J| = o(N) \quad , \text{ and } \quad N = o(d_*(\Sigma^{-1/2}_{J^c} B^p_{q'})) \qquad \text{variance } \hat{\boldsymbol{\beta}}$$

$\downarrow$ $\qquad\qquad$ $\downarrow$

$\hat{\boldsymbol{\beta}}_J$ bias and variance $\qquad$ $\hat{\boldsymbol{\beta}}_{J^c}$ "bias" and "variance"

**Interpretation**:

- $\hat{\boldsymbol{\beta}}_J \approx \operatorname{argmin}(\frac{1}{N^{\frac{q}{2}}}\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J\|_2^q + \frac{\ell_*^q}{N^{\frac{q}{2}}}\|\boldsymbol{\beta}_J\|_q^q)$.
  - Bias $\left\|\left|\boldsymbol{\beta}^*_J \odot |\boldsymbol{\beta}^*_J|^{\odot(q-2)}\right|\right\| = o(\frac{N^{\frac{q}{2}}}{\ell_*^q})$,
  - Variance $|J| = o(N)$.
- $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]$.
  - $\boldsymbol{\beta}^*_{J^c}$ is not estimated $\Rightarrow$ model noise $\|\Sigma^{1/2}_{J^c}\boldsymbol{\beta}^*_{J^c}\|_2$.
  - Variance $N = o(d_*(\Sigma^{-1/2}_{J^c} B^p_{q'}))$.

Rethinking consistency: estimation happens only on $V_J$, and $\boldsymbol{\beta}^*_{J^c}$ is not estimated.

# Question (ii),(iii): Convergence Rate

## Definition (Spectral Methods)

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

$$\hat{\boldsymbol{\beta}} = \frac{1}{N}\varphi_t(\hat{\Sigma})\mathbb{X}^\top \boldsymbol{y}.$$

Here, $\varphi_t(\hat{\Sigma})$ is defined by spectral calculus, and $\hat{\Sigma} = \frac{1}{N}\mathbb{X}^\top\mathbb{X}$ is the empirical covariance matrix. Define $\psi_t(x) = 1 - x\varphi_t(x)$, called residual function. [Bauer, Pereverzev, and Rosasco, 2007, Blanchard and Mücke, 2016, Celisse and Wahl, 2021, Zhang, Li, and Lin, 2023].

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 $\eta$ 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

We directly define optimal FSD.

Let $\Sigma = \mathbb{E}[X \otimes X] = \sum_{j=1}^{p} \sigma_j \boldsymbol{v}_j \otimes \boldsymbol{v}_j$ be its spectral decomposition, where $\sigma_j$ is the $j$-th largest eigenvalue.

## Definition (estimation dimension)

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

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

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

# The meaning of filter function

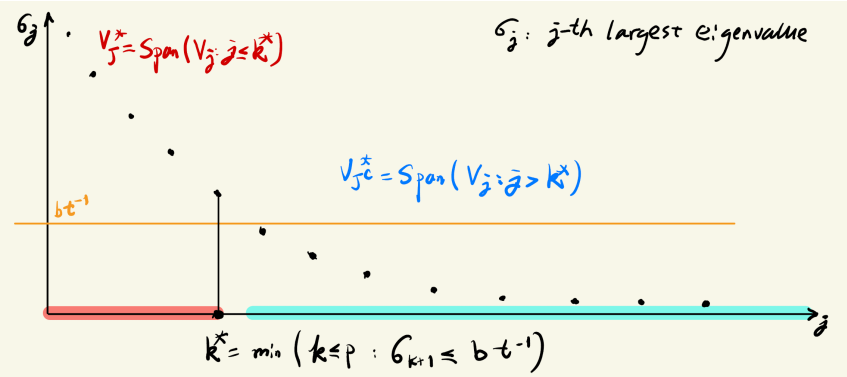


Figure: Given $t^{-1}$, $\hat{\beta}$ uses $V_J^* = \mathrm{span}(e_j : j \le k^*)$ to estimate $\beta_J^*$, and use $V_{J^c}^*$ to absorb noise. This also explains why $\varphi_t$ is called a filter function: the feature subspace corresponding to eigenvalues smaller than $bt^{-1}$ is effectively filtered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

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

Let $J_* = \{1, \cdots, k^*\}$, $P_{J_*} = \sum_{j \in J_*} \boldsymbol{v}_j \otimes \boldsymbol{v}_j$, where $\boldsymbol{v}_j$ is the eigenvector associated to $\sigma_j$.
Let $\Sigma_{J_*} = P_{J_*} \Sigma P_{J_*}$ and $\Sigma_{J_*^c} = P_{J_*^c} \Sigma P_{J_*^c}$.

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

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

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 $V_J^*$.

# Key theorem of spectral methods: informal

Recall that $r(V_J^*, V_{J^c}^*) = \|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_{J_*}^*\|_2 + \|\Sigma_{J_*^c}^{1/2}\boldsymbol{\beta}_{J_*^c}^*\|_2 + \sigma_\xi\sqrt{\frac{|J_*|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J_*^c}^2)}{N}}$, where $V_J^* = \mathrm{span}(\boldsymbol{v}_j : 1 \le j \le k^*)$ and $k^* = \max(k \le p : \sigma_{k+1} \le bt^{-1})$.

Under some assumptions, for any $\mathcal{R}$ characterized by $(\boldsymbol{\beta}^*, \Sigma, \sigma_\xi)$. Let $(\varphi_t)_{t\ge1}$ be a family of filter functions satisfying the aforementioned assumption.

$$\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim r(V_J^*, V_{J^c}^*).$$

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

$$\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \gtrsim r(V_J^*, V_{J^c}^*).$$

Comments:

1. For any $\mathcal{R} = (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, $r(V_J^*, V_{J^c}^*)$ characterizes $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2$ precisely, with no assumptions imposed on $\Sigma$ 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 $\boldsymbol{\beta}^*$.

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

## Definition (Pre-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 $\hat{\boldsymbol{\beta}}^{(A)}_{t_A} \preceq_{\mathcal{R}} \hat{\boldsymbol{\beta}}^{(B)}_{t_B}$ if $r^{(A)}_{t_A}(V_J^*, V_{J^c}^*) = O(r^{(B)}_{t_B}(V_J^*, V_{J^c}^*))$ as $N, p \to \infty$.

## Corollary

*For any $\mathcal{R}$ and any $t \geq 1$, $\hat{\boldsymbol{\beta}}^{(A)}_t \preceq_{\mathcal{R}} \hat{\boldsymbol{\beta}}^{(B)}_t$, 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

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

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

# Feature Learning Property

The central problem in neural network theory: how NN learns good features/representations? [Ben Arous, Gheissari, and Jagannath, 2021, Damian, Lee, and Soltanolkotabi, 2022, Suzuki, Wu, Oko, and Nitanda, 2023, Bietti, Bruna, and Pillaud-Vivien, 2025]

- FSD: how features are utilized for estimation.
- Feature Learning: how features are learned.

Connection: FSD tells us which features are good for estimation.

- Given $(\mu_X, f^*, \sigma_\xi)$, $(X_i, Y_i)_{i=1}^N$, and an estimator $\hat{f}_N$.
- (Informal) There exist a RKHS $\mathcal{H}_{\mathrm{fea}} \subset L^2(\mu_X)$, called learned feature subspace, with associated feature map $\phi_{\mathrm{fea}} : \Omega_X \to \mathcal{H}_{\mathrm{fea}}$, covariance $\Sigma = \mathbb{E}[\phi_{\mathrm{fea}}(X) \otimes \phi_{\mathrm{fea}}(X) | (X_i, Y_i)_{i=1}^N]$, oracle $g^*_{\mathcal{H}_{\mathrm{fea}}} \in \mathrm{argmin}(\mathbb{E}[(Y - g(\phi_{\mathrm{fea}}(X)))^2 | (X_i, Y_i)_{i=1}^N] : g \in \mathcal{H}_{\mathrm{fea}})$, and a latent estimator $\hat{g}_N : \mathcal{H}_{\mathrm{fea}} \to \mathbb{R}$, such that the following hold.
  - $\hat{f}_N(\cdot) = \hat{g}_N(\phi_{\mathrm{fea}}(\cdot))$;
  - $\|g^*_{\mathcal{H}_{\mathrm{fea}}}(\phi_{\mathrm{fea}}(\cdot)) - f^*(\cdot)\|_{L^2(\mu_X)} = o_{\mathbb{P}}(1)$;
  - There exists $V_J$, such that $\dim(V_J) = o(N)$, and $\|\Sigma_{J^c}^{1/2} g^*_{\mathcal{H}_{\mathrm{fea}}}\|_{\mathcal{H}_{\mathrm{fea}}} = o_{\mathbb{P}}(1)$, and
  - $\|\hat{g}_N(\phi_{\mathrm{fea}}(\cdot)) - g^*_{\mathcal{H}_{\mathrm{fea}}}(\phi_{\mathrm{fea}}(\cdot))\|^2_{L^2(\mu_X)}$ decreases when $g^*_{\mathcal{H}_{\mathrm{fea}}}$ gets aligned with the top $\dim(V_J)$ eigenvectors of $\Sigma$.

# Feature Learning reduces "approximation error"

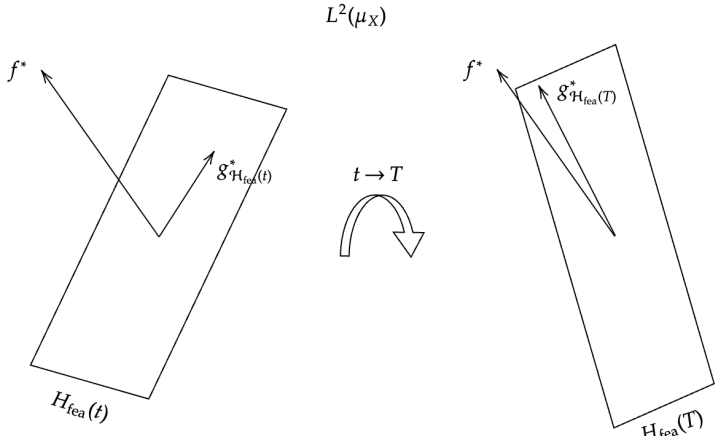


Figure: $\mathcal{H}_{\text{fea}}$ is getting closer to $f^*$ during training: $\|f^* - g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)}$ is getting small.

# Feature Learning ruduces "estimation error"

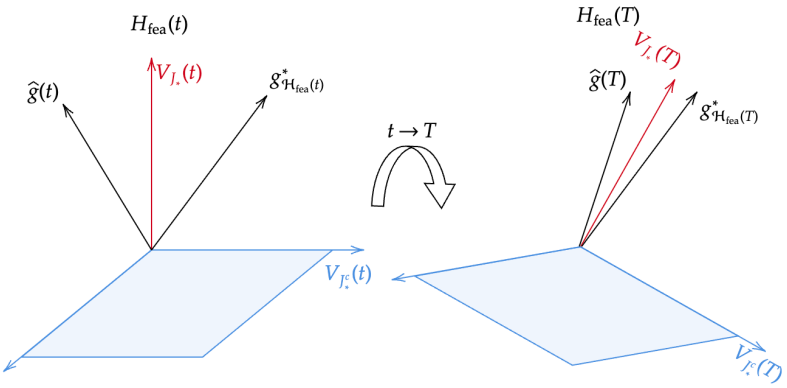


Figure: $g_{\mathcal{H}^*_{\text{fea}}}$ is getting aligned with the top $k$ eigenvectors.

# What does FSD provide?

- **FSD as analytical tool**: how features are utilized.
  $V_J$ **is an arrow.**

$$\bullet_J : (\mu_X, f^*, \xi, \mathcal{F}, \hat{f}) \longmapsto (\mu_X, f_J^*, \zeta, V_J, \hat{f}_J), \text{ preserves input-output relation,}$$

since $Y = f^*(X) + \xi = f_J^*(X) + \underline{f_{J^c}^*(X) + \xi} = f_J^*(X) + \zeta$, where $\zeta = \xi + f_{J^c}^*$.

  - The new problem-solution is either easier to study, e.g., Benign Overfitting.
  - $V_J^*$ provides the correct bias and variance.

  Classical SLT: identity map, $V_J = \mathcal{F}$.

  $V_{J^c}$ **is still a mystery.**
  - For overfitting estimator, $\hat{\boldsymbol{\beta}}_{J^c}$ is interpolating $\boldsymbol{y} - \mathbb{X}\hat{\boldsymbol{\beta}}_J$.
  - What is the true task of $\hat{\boldsymbol{\beta}}_{J^c}$ for general nonlinear estimators?

- **FSD as theoretical framework**: how features are learned.

# End

**Thank you for your attention!**

# FSD in classification

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

$$
\begin{aligned}
& P\mathcal{L}_{\hat{f}}^{(0,1)} \\
& = \mathbb{P}\left(Y\hat{f}(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) && (4) \\
& + \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y f_J^*(X) < 0\right) && (5) \\
& + \mathbb{P}\left(Y f_J^*(X) < 0\right) - \mathbb{P}\left(Y\left(\eta(X) - \frac{1}{2}\right) < 0\right), && (6)
\end{aligned}
$$

- (4): error caused by free part $\hat{f}_{J^c}$;
- (5): prediction error caused by $\hat{f}_J$ compared to that of $f_J^*$ ; and
- (6): 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.

# $\hat{\boldsymbol{\beta}}_J$ is "almost" a support vectors machine with squared hinge loss

**Proposition.** Recall $P_N \ell_{\boldsymbol{\beta}_J} = \|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}} \boldsymbol{\beta}_J]\|_2^2$ in classification. Let $P_N \ell^{(\mathrm{sh})}_{\boldsymbol{\beta}_J} = \frac{1}{N}\sum_{i=1}^N (1 - Y_i \langle \boldsymbol{\beta}_J, X_i\rangle)_+^2$ be the squared hinge loss. Then w.h.p.,

$$\hat{\boldsymbol{\beta}}_J \in \operatorname{argmin}\left(P_N \ell_{\boldsymbol{\beta}_J} + \|\boldsymbol{\beta}_J\|_2^2\right) \approx \operatorname{argmin}\left(\frac{N}{\operatorname{Tr}(\Sigma_{J^c})} P_N \ell^{(\mathrm{sh})}_{\boldsymbol{\beta}_J} + \|\boldsymbol{\beta}_J\|_2^2\right).$$

**Proof.** The dual problem of $\|\mathcal{B}[\boldsymbol{\mu}]\|_2$ is $\max(\langle \boldsymbol{\lambda}, \boldsymbol{\mu}\rangle : \boldsymbol{\lambda} \succeq \mathbf{0}, \|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\|_2 \le 1)$. Let $H(\boldsymbol{\mu}) := \{i \in [N] : \mu_i < 0\}$ and let

$$\boldsymbol{\lambda}^- \in \operatorname{argmax}(\langle \boldsymbol{\lambda}, \boldsymbol{\mu}\rangle : \boldsymbol{\lambda} \succeq \mathbf{0}, (1+\tilde{\delta})\sqrt{\operatorname{Tr}(\Sigma_{J^c})}\|\boldsymbol{\lambda}\|_2 \le 1). \tag{7}$$

Suppose $i \in H(\boldsymbol{\mu})$ but $\lambda_i^- > 0$. Let $\tilde{\boldsymbol{\lambda}}^- = (\lambda_1^-, \cdots, \lambda_{i-1}^-, 0, \lambda_{i+1}^-, \cdots, \lambda_N^-)$, then $\|\tilde{\boldsymbol{\lambda}}^-\|_2 < \|\boldsymbol{\lambda}^-\|_2 \le \frac{1}{(1+\tilde{\delta})\sqrt{\operatorname{Tr}(\Sigma_{J^c})}}$. Since $\mu_i \lambda_i^- < 0$,

$$\langle \boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}^-\rangle = \sum_{i' \neq i} \mu_{i'} \lambda_{i'}^- > \sum_{i'=1}^N \mu_{i'} \lambda_{i'}^- = \langle \boldsymbol{\mu}, \boldsymbol{\lambda}^-\rangle, \Rightarrow \boldsymbol{\lambda}^- \text{ is not the maximizer of (7).}$$

Necessarily, $\forall i \in H(\boldsymbol{\mu})$, $\lambda_i^- = 0$. Therefore, $\boldsymbol{\lambda}^- = \left(\frac{\boldsymbol{\mu}}{(\|\boldsymbol{\mu}\|_2(1+\tilde{\delta})\sqrt{\operatorname{Tr}(\Sigma_{J^c})})}\right)_+$. ■

# Free subspace and the energy of $\hat{\boldsymbol{\beta}}_{J^c}$ in classification

Let $\mathcal{F} = V_J \oplus V_{J^c}$ be any FSD and $\hat{f}_N$ be any estimator. There holds $\mu^{\otimes N}$-a.s.,

$$\begin{aligned}
&\mathbb{P}\left(Y\hat{f}(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) \\
&\leq \mathbb{P}\left(|\hat{f}_{J^c}(X)| > |\hat{f}_J(X)| \middle| (X_i, Y_i)_{i=1}^N\right).
\end{aligned}$$

# Classical theory on RERM

We say $\Psi$ has non-trivial Bregman div. around $f^*$, if $D_\Psi(\cdot, f^*)$ is bounded from below by convex, non-negative func., where $D_\Psi(f,g) = \Psi(f) - \Psi(g) - \langle \nabla\Psi(g), f-g\rangle$. For any $\rho > 0$, let $B_\Psi(f^*, \rho) = \{f \in \mathcal{F} : D_\Psi(f, f^*) \le \rho\}$.
E.g., $\Psi(\cdot) = \|\cdot\|_q^q$, then $D_\Psi(\boldsymbol{v}_1, \boldsymbol{v}_2) \ge c\alpha_q(|\boldsymbol{v}_2|, \boldsymbol{v}_1 - \boldsymbol{v}_2)$, where

$$\alpha_q(x,y) = \begin{cases} \frac{q}{2}x^{q-2}y^2, & |y| \le x, \\ |y|^q + (\frac{q}{2}-1)x^q, & \text{otherwise.} \end{cases}$$

## Theorem

*Suppose $\Psi$ has non-trivial Breg. div. around $f^*$. For any $\rho > 0$, let $\mathcal{G} = B_\Psi(f^*, \rho)$ for any $\lambda > 0$, let $r_{\mathrm{iso}}(\rho) = r_{\mathrm{iso}}(\mathcal{G}, \kappa)$. Let $r_*$ and $\rho_*$ be the smallest $r$ and its corrsponding. $\rho$, s.t.,*

$$r \ge r_{\mathrm{iso}}(\rho), \quad 3\square r^{\frac{2}{\kappa}} > \lambda \|\!|\nabla\Psi(f^*)|\!\|_{(r,\rho)}, \text{ and } \quad \rho \ge \frac{1}{\lambda}\triangle r^{\frac{2}{\kappa}}$$

*where* $\|\!|\nabla\Psi(f^*)|\!\|_{(r,\rho)} = \sup\left(\langle \nabla\Psi(f^*), f - f^*\rangle : f \in B_\Psi(f^*;\rho) \cap B_{L^2(\mu_X)}(f^*; r)\right)$.
*Then* $\|\hat{f} - f^*\|_{L^2(\mu_X)}^2 \le r_*^2$, *w.p.a.l.* $1 - \delta_{\mathrm{M}} - \delta_{\mathrm{Q}}$.

# Notations

For any $x \geq 0$ and $y \in \mathbb{R}$, let

$$\alpha_q(x,y) = \begin{cases} \frac{q}{2} x^{q-2} y^2, & \text{if } |y| \leq x \\ |y|^q + \left(\frac{q}{2} - 1\right) x^q, & \text{otherwise.} \end{cases}$$

For any $\rho > 0$, we define

$$\rho K_{\text{model}} = \begin{cases} \{\boldsymbol{v} \in V_J : \sum_{j \in J} \alpha_q(|\beta_j^*|, v_j) \leq \rho^q\} & \text{when } q < 2, \\ \rho B_q^J & \text{when } q \geq 2. \end{cases}$$

Define

$$|||\boldsymbol{\beta}||| = \sup \left( \langle \boldsymbol{\beta}, \boldsymbol{u} \rangle : \boldsymbol{u} \in \frac{C_1 \sqrt{N}}{\ell_*(\Sigma_{J^c}^{1/2} B_q^p)} K_{\text{model}} \cap \Sigma_J^{-1/2} B_2^J \right),$$

where $C_1 = C_1(q)$ is some absolute constant.

# Weak moments case

Suppose $\Sigma^{-1/2}X$ has i.i.d. coordinates that satisfy $L^{8+\varepsilon} - L^2$ norm equivalence assumption.

We say that an FSD $\mathbb{R}^p = V_J \oplus V_{J^c}$ is admissible in the heavy-tailed case if the following conditions are satisfied:

1. $V_J = \operatorname{span}(\boldsymbol{e}_j : j \in J)$, where $J \subset [p]$.
2. There exist absolute constants $0 < \kappa_{RIP}, \kappa_{DM} < 1$ and $C_2 > 1$ such that
   - when $1 < q < 2$, there holds $N^{\frac{4}{4+\varepsilon}} \lesssim |J| \lesssim N \lesssim d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p)$, and
   - when $q \geq 2$, there holds

$$N^{\frac{4}{4+\varepsilon}} \lesssim |J| \leq N \lesssim \frac{d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p)}{\log^2(|J^c|^{1/q'}/d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p))}.$$

In particular, when $X$ is sub-Gaussian, the log factor can be removed.

# Universality of Gaussian case

## Theorem (Universality, informal)

*Suppose $\Sigma$ is diagonal. For any admissible FSD in the heavy-tailed case, when $N$ is large enough, and the following hold w.h.p..*

1. *When $q \geq 2$, assume there exists an absolute constant $C_3$ such that $(\mathbb{E}[|\xi|^4])^{1/4} \leq C_3\sigma_\xi$. Then*

$$\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 \lesssim R(V_J, V_{J^c}) \log^{\frac{q}{q-1}}(|J^c|) + \sqrt{\frac{N}{d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p)}}\,\sigma_\xi.$$

   *In particular, when $X$ is sub-Gaussian, the* log *factor can be removed.*

2. *When $1 < q < 2$, assume additionally that $X_J \sim \mathcal{N}(\mathbf{0}, \Sigma_J)$, $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$, and $\boldsymbol{\beta}^*_{J^c} = \mathbf{0}$. If $X_{J^c}$ satisfies the $L^{8+\varepsilon}$-$L^2$ equivalence condition, then the same result holds.*

## $\bullet_J$ reduces $r_M$

Recall that in regression, $P_N \ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q$ and

$$\partial^- P_N \ell_{\boldsymbol{\beta}_J} = \left\{ -q\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^{q-1}\mathbb{X}_J^\top \boldsymbol{\lambda}^*[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] : \boldsymbol{\lambda}^*[\cdot] \text{ is dual solution} \right\}. \quad (8)$$

**Proposition.** Under some assumptions, $\exists\ 0 < \delta_{\mathrm{M}} < \frac{1}{100}$, $c, c' < 1$, $\ell_* > 0$ and $c'' = c''(c, c', \delta_{\mathrm{M}}) > 1$ such that for any local. subset $\mathcal{G} \subset V_J$,

$$\begin{cases} r_{\mathrm{M}}\left(\mathcal{G}, \delta_{\mathrm{M}}, \frac{2}{q}, 4c\frac{N^{\frac{q}{2}}}{\ell_*^q}\right) \le c''\sigma_\xi \left(\frac{|J|}{N}\right)^{\frac{1}{2(q-1)}}, & \text{when } q \ge 2, \text{ and} \\ r_{\mathrm{M}}\left(\mathcal{G}, \delta_{\mathrm{M}}, 1, 4c'\sigma_\xi^{q-2}\frac{N^{\frac{q}{2}}}{\ell_*^q}\right) \le c''\sigma_\xi^{q-1}\left(\frac{|J|}{N}\right)^{\frac{1}{2}}, & \text{when } 1 < q < 2. \end{cases}$$

**Proof.** For any $\boldsymbol{g} \in \partial^- P_N \ell_{\boldsymbol{\beta}_J^*}$ from (8), by $B_{L^2(\mu_X)}(\boldsymbol{\beta}_J^*, r) = \boldsymbol{\beta}_J^* + r\Sigma_J^{-1/2}B_2$,

$$\sup_{\boldsymbol{\beta}_J \in \mathcal{G} \cap B_{L^2(\mu_X)}(\boldsymbol{\beta}_J^*, r)} |\langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}_J^* \rangle| = q\|\mathcal{A}[\mathbb{X}_{J^c}\boldsymbol{\beta}_J^* + \boldsymbol{\xi}]\|_q^{q-1} \sup_{\boldsymbol{v} \in \mathcal{G} \cap r\Sigma_J^{-1/2}S_2} \sum_{i=1}^N \lambda_i^*[\mathbb{X}_{J^c}\boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}]\langle X_i, \boldsymbol{v}\rangle$$

$$\le C_q\|\mathcal{A}[\mathbb{X}_{J^c}\boldsymbol{\beta}_J^* + \boldsymbol{\xi}]\|_q^{q-1}\|\boldsymbol{\lambda}^*[\mathbb{X}_{J^c}\boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}]\|_2\gamma_2(\Sigma_J^{1/2}\mathcal{G} \cap rB_2^J, \|\cdot\|_2).$$

## Proposition

*Suppose* $|J| \le cN$. *Under the assumptions,* $\exists\ 0 < \delta_{\mathrm{Q}} < \frac{1}{100}$, $c = c(q)$, *and* $c' = c'(q)$, *s.t., when* $q \ge 2$. *Then* $\forall r > 0$, $\forall \mathcal{G}$, *w.p.a.l.* $1 - \delta_{\mathrm{Q}}$, *for any local. sub.* $\boldsymbol{\beta}_J \in \mathcal{G} \cap (\boldsymbol{\beta}^*_J + r\Sigma_J^{-1/2} S_2)$,

$$P_N \mathcal{L}_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q - \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}^*_J]\|_q^q \ge \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J \rangle + \triangle r^q, \text{ where } \triangle = c\frac{N^{\frac{q}{2}}}{\ell_*^q}.$$

*That is,* $r_{\mathrm{Q}}(\mathcal{G}, \delta_{\mathrm{Q}}, \frac{2}{q}) = 0$, $\forall \mathcal{G}$.

**Proof.**

$$P_N \mathcal{L}_{\boldsymbol{\beta}_J} \ge \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J \rangle + c_q \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}^*_J]\|_q^q$$

$$\ge \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J \rangle + c_q \|\mathcal{A}[\mathbb{X}(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J)]\|_q^q$$

$$\ge \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J \rangle + c_q \frac{\|\mathbb{X}(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J)\|_2^q}{\ell_*^q} \ge \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J \rangle + c'_q \frac{N^{\frac{q}{2}}}{\ell_*^q} \|\Sigma_J^{1/2}(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J)\|_2^q.$$

■

# 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 $K_{\mathrm{NTK}}$ and is independent with $\mu$ 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 $K_{\text{fea}}$, called learned feature kernel.
- Let $\hat{f}_N$ be a ridge regression with some tuning parameter $\lambda$ 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.

# 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


Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, December 2020. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/doi/10.1073/pnas.1907378117.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, February 2007. ISSN 0885-064X. doi: 10.1016/j.jco.2006.07.001. URL https://www.sciencedirect.com/science/article/pii/S0885064X06000781.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1611–1619. PMLR, April 2019. URL https://proceedings.mlr.press/v89/belkin19a.html.

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. ISSN 1533-7928. URL http://jmlr.org/papers/v22/20-1288.html.

# References II


Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning Gaussian multi-index models with gradient flow part I: General properties and two-timescale learning. *Communications on Pure and Applied Mathematics*, n/a(n/a), July 2025. ISSN 1097-0312. doi: 10.1002/cpa.70006. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.70006. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.70006.

Gilles Blanchard and Nicole Mücke. Kernel regression, minimax rates and effective dimensionality: beyond the regular case, November 2016. URL http://arxiv.org/abs/1611.03979. arXiv:1611.03979 [stat].

Claire Boyer. Living la vida loca: learning in interpolation regimes. Lecture notes, Sorbonne University, Paris Sacaly University, 2022.

Alain Celisse and Martin Wahl. Analyzing the discrepancy principle for kernelized spectral filter learning algorithms. *Journal of Machine Learning Research*, 22(76):1–59, 2021. URL http://jmlr.org/papers/v22/20-358.html.

# References III


Geoffrey Chinot, Matthias Loffler, and Sara Van de Geer. On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *The Annals of Statistics*, 2022. doi: 10.1214/22-aos2190.

Alex Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural Networks can Learn Representations with Gradient Descent, June 2022. URL http://arxiv.org/abs/2206.15144. arXiv:2206.15144 [cs, math, stat].

Konstantin Donhauser, Nicolò Ruggeri, Stefan Stojanovic, and Fanny Yang. Fast rates for noisy interpolation require rethinking the effect of inductive bias. In *Proceedings of the 39th International Conference on Machine Learning*, pages 5397–5428. PMLR, June 2022. URL https://proceedings.mlr.press/v162/donhauser22a.html.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École d'Été de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033 of *Lecture Notes in Mathematics*. Springer, Berlin, Heidelberg, 2011. ISBN 978-3-642-22146-0 978-3-642-22147-7. doi: 10.1007/978-3-642-22147-7. URL https://link.springer.com/10.1007/978-3-642-22147-7.

# References IV


Taiji Suzuki, Denny Wu, Kazusato Oko, and Atsushi Nitanda. Feature learning via mean-field Langevin dynamics: classifying sparse parities and beyond. *Advances in Neural Information Processing Systems*, 36:34536–34556, December 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6cc321baf0a8611b1d1bdbd18822667b-Abstract-Conference.html.

Sara Van De Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, Cambridge, 2006. ISBN 978-0-521-65002-1.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, 1995. ISBN 978-1-4419-3160-3 978-1-4757-3264-1. doi: 10.1007/978-1-4757-3264-1. URL http://link.springer.com/10.1007/978-1-4757-3264-1.

Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum $\ell_1$-norm interpolation of noisy data. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 10572–10602. PMLR, May 2022. URL https://proceedings.mlr.press/v151/wang22k.html.

# References V

Haobo Zhang, Yicheng Li, and Qian Lin. On the Optimality of Misspecified Spectral Algorithms, August 2023. URL http://arxiv.org/abs/2303.14942. arXiv:2303.14942 [math, stat].