Benign Overfitting Property of the Minimum ℓ_q-norm interpolant estimator in linear regression problem via Feature Space Decomposition

Zong Shang
(joint with Radosław Adamczak, Guillaume Lecué, and Marta Strzelecka)

CREST-ENSAE, Institut Polytechnique de Paris

March 24, 2026

"The unknown thing to be known appeared to me as some stretch of earth or hard marl, resisting penetration... the sea advances insensibly in silence, nothing seems to happen, nothing moves, the water is so far off you hardly hear it... yet it finally surrounds the resistant substance."

— Alexandre Grothendieck, Récoltes et Semailles(1985), pp. 552–553.

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 (not in this talk).

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,

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

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 δ, s.t., $\forall$ FSD (V_J, V_{J^c}),

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

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, see [LLS25, arXiv: 2512.14473], $\exists 0 < c, \delta < 1$, 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 (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

- 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.

Rethink 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 ℓ_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 $\mathbb{X} = [X_1 | \cdots | X_N]^\top \in \mathbb{R}^{N \times p}$, $\boldsymbol{y} = (Y_1, \cdots, Y_N)$.

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

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

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

Overfitting estimator and Benign Overfitting estimator

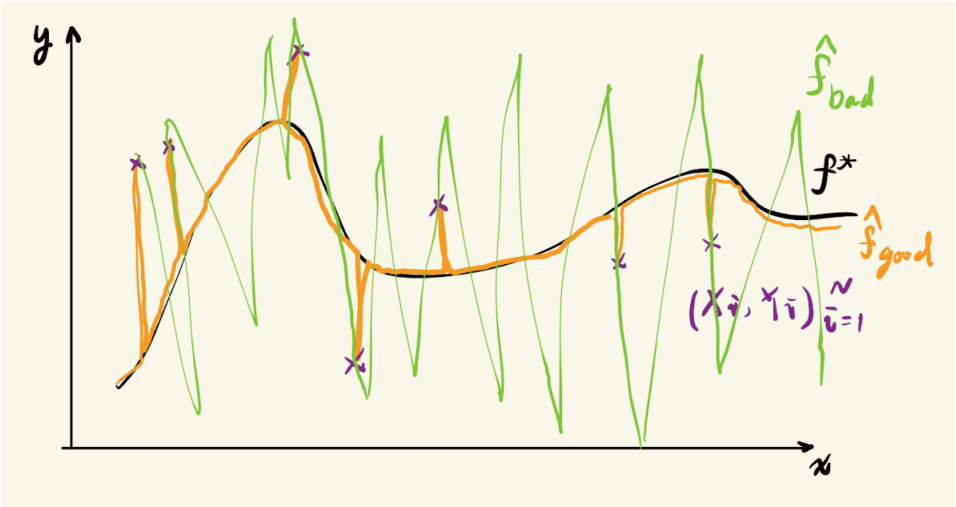

Figure: Overfitting estimator (green curve) and Benign Overfitting estimator (orange)

Classical Stats. Learning Theory fails to penetrate

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.

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

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 resistant substance that classical uniform convergence fails to penetrate.
This talk: 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.

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

$$\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J].$$

$\bullet_J$ defines new f^*_J

$\boldsymbol{\beta}^*_J$ is consistent with the random loss function $P_N\ell_\bullet$.

Proposition ($\boldsymbol{\beta}^*_J$ is oracle) Suppose $\mathbb{X}\boldsymbol{\beta}^*_{J^c} + \boldsymbol{\xi}$ is independent with $\mathbb{X}_J$, and $\mathbb{E}[\mathbb{X}_J] = \mathbf{0}$, then

$$\boldsymbol{\beta}^*_J \in \operatorname{argmin}(P\ell_{\boldsymbol{\beta}_J} : \boldsymbol{\beta}_J \in V_J), \text{ where } P\ell_{\boldsymbol{\beta}_J} = \mathbb{E}_{\mathbb{X}_J,\boldsymbol{\xi}}[P_N\ell_{\boldsymbol{\beta}_J}], \tag{3}$$

and $P_N\ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$.

Implication: $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J)\|_2^2$ is the estimation error of RERM

$$\hat{\boldsymbol{\beta}}_J \in \operatorname{argmin}\left(P_N\ell_{\boldsymbol{\beta}_J} + \|\boldsymbol{\beta}_J\|_q^q\right), \text{ where } P_N\ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q.$$

We need to understand this random loss function. This seems to be desperate!

• J^c: new tools from Asymptotic Geometric Analysis

Theorem (Dvoretzky-Milman) There exist $\kappa_{DM}, c_1 \leq 1$ s.t. the following holds. Let $\|\cdot\|$ be some norm on $\mathbb{R}^p$ and denote by B° the unit ball of the dual norm of $\|\cdot\|$. Denote by $\mathbb{G}$, the $N \times p$ standard Gaussian matrix. Given any $0 < \varepsilon_1 \leq 1$. Assume that $N \leq \kappa_{DM}\varepsilon_1^2 d_*(B)$. Then w.h.p.,

$$\forall \boldsymbol{\lambda} \in \mathbb{R}^N, \quad (1-\varepsilon_1)\,\|\boldsymbol{\lambda}\|_2\,\ell_*(B^\circ) \leq \left\|\mathbb{G}^\top \boldsymbol{\lambda}\right\| \leq (1+\varepsilon_1)\,\|\boldsymbol{\lambda}\|_2\,\ell_*(B^\circ), \text{ where} \tag{4}$$

$\ell_*(B^\circ) = \mathbb{E}\|G\|$, $\operatorname{diam}(B^\circ) = \max(\|\boldsymbol{v}\|_2 : \boldsymbol{v} \in B^\circ)$, and $d_*(B) = (\frac{\ell_*(B^\circ)}{\operatorname{diam}(B^\circ, \ell_2)})^2$.
Let $\Sigma_{J^c} = P_{J^c}\Sigma P_{J^c}$. When $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$ is Gaussian,

$$\mathbb{X}_{J^c}^\top = [\Sigma_{J^c}^{1/2} G_1 | \cdots | \Sigma_{J^c}^{1/2} G_N] = \Sigma_{J^c}^{1/2} \mathbb{G}^\top.$$

By DM applied to $\|\cdot\| = \|\Sigma_{J^c}^{1/2} \cdot\|_{q'}$, when $N \leq \kappa_{DM}\varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p)$, w.h.p.,

$$\Omega_{\mathsf{DM},\mathsf{reg}}(\varepsilon_1) = \left\{\forall \boldsymbol{\lambda} \in \mathbb{R}^N : (1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)\|\boldsymbol{\lambda}\|_2 \leq \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \leq (1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)\|\boldsymbol{\lambda}\|_2\right\}.$$

How does it help understanding $P_N \ell_\bullet$?

An intuition from Dvoretzky-Milman theorem

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

$$\Omega_{\mathsf{DM,reg}}(\varepsilon_1) = \left\{\forall \boldsymbol{\lambda} \in \mathbb{R}^N : (1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)\|\boldsymbol{\lambda}\|_2 \le \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \le (1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)\|\boldsymbol{\lambda}\|_2\right\}$$
$$\subset \left\{\forall \boldsymbol{\mu} \in \mathbb{R}^N : \frac{\|\boldsymbol{\mu}\|_2}{(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)} \le \|\mathcal{A}[\boldsymbol{\mu}]\|_q \le \frac{\|\boldsymbol{\mu}\|_2}{(1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)}\right\}.$$

Implication: Denote ℓ_* by $\ell_*(\Sigma_{J^c}^{1/2} B_q)$. From $\|\mathcal{A}[\boldsymbol{\mu}]\|_q \sim \ell_*^{-1}\|\boldsymbol{\mu}\|_2$, and $P_N \ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J]\|_q^q$,

$$\hat{\boldsymbol{\beta}}_J \in \operatorname{argmin}\left(P_N \ell_{\boldsymbol{\beta}_J} + \|\boldsymbol{\beta}_J\|_q^q\right) \approx \operatorname{argmin}\left(\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J\|_2^q + \ell_*^q \|\boldsymbol{\beta}_J\|_q^q\right).$$

We have to extend Dvoretzky-Milman theorem.

New result in AGA: Generalized Dvoretzky-Milman theorem

- Y. Gordon, A. Litvak, A. Pajor and N. Tomczak-Jaegermann. "Random epsilon-nets and embeddings in ℓ_N^∞", Studiea Mathematica, 2007.
- S. Mendelson, N. Tomczak-Jaegermann. "A subgaussian embedding theorem", Israel Journal of Mathematics, 2008.
- D. Bartl, S. Mendelson. "Random embeddings with an almost Gaussian distortion", Advances in Mathematics, 2022.
- S. Mendelson. "An isomorphic Dvoretzky-Milman Theorem using general random ensembles", Journal of Functional Analysis, 2024.
- D. Bartl, S. Mendelson. "Optimal Non-Gaussian Dvoretzky–Milman Embeddings", International Mathematics Research Notices, 2024.
- D. Bartl, S. Mendelson. "Structure preservation via the Wasserstein distance", Journal of Functional Analysis, to appear.

Theorem (this paper) Suppose $X = \Sigma^{1/2}\boldsymbol{\zeta}$ for some $\boldsymbol{\zeta}$ that has i.i.d. coordinates, satisfying $L^4(\mu_X) - L^2(\mu_X)$ norm equivalence condition, and for some diagonal Σ. The following hold w.h.p.,

$$\forall \boldsymbol{\lambda} \in \mathbb{R}^N, \begin{cases} c\ell_* \leq \left\|\mathbb{X}^\top \boldsymbol{\lambda}\right\|_{q'} \leq C\log(p)\ell_*, & \text{if } q \geq 2, \text{ and } N \leq \kappa_{\mathsf{DM}} d_* \log^{-2}(p^{\frac{1}{q'}}/d_*) \\ c\ell_* \leq \|\mathbb{X}^\top \boldsymbol{\lambda}\|_{q'} \leq C\ell_*, & \text{if } 1 < q < 2, \text{ and } N \leq \kappa_{\mathsf{DM}} d_*. \end{cases}$$

Come back to classics: fixed points

Recap (intuition): $\hat{\boldsymbol{\beta}}_J \in \operatorname{argmin}\left(P_N \ell_{\boldsymbol{\beta}_J} + \|\boldsymbol{\beta}_J\|_q^q\right) \approx \operatorname{argmin}\left(\|\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J\|_2^q + \ell_*^q \|\boldsymbol{\beta}_J\|_q^q\right)$.
Reg. $\Psi : \mathcal{F} \to \mathbb{R}$, convex; empirical risk: $P_N \ell_\bullet : f \in \mathcal{F} \mapsto P_N \ell_f \in \mathbb{R}$.

$$\hat{f}_N \in \operatorname{argmin}(P_N \ell_f + \lambda \Psi(f) : f \in \mathcal{F}).$$

Definition (Multiplier & Quadratic fixed points) Let $\kappa, \square, \triangle > 0$, $\triangle \geq 4\square$, $0 < \delta_{\mathrm{M}}, \delta_{\mathrm{Q}} < \frac{1}{2}$. For any $\mathcal{G} \subset \mathcal{F}$, s.t., $f^* \in \mathcal{G}$, define

$$r_M(\mathcal{G}, \delta_{\mathrm{M}}, \kappa, \square) = \min_{r>0} \left(\mathbb{P}\left(\sup_f \inf_{\boldsymbol{g} \in \partial^- P_N \ell_{f^*}} |\langle \boldsymbol{g}, f - f^* \rangle| \leq \square r^{\frac{2}{\kappa}} \right) \geq 1 - \delta_{\mathrm{M}} \right),$$

where sup is over $f \in \mathcal{G} \cap B_{L^2(\mu_X)}(f^*, r)$ and $r_{\mathrm{Q}}(\mathcal{G}, \delta_{\mathrm{Q}}, \kappa, \triangle)$ be

$$\min_{r>0} \left(\mathbb{P}\left(\forall f \in \mathcal{G} \cap S_{L^2(\mu_X)}(f^*, r),\ P_N \mathcal{L}_f \geq \triangle r^{\frac{2}{\kappa}} + \sup_{\boldsymbol{g} \in \partial^- P_N \ell_{f^*}} (\langle \boldsymbol{g}, f - f^* \rangle) \right) \geq 1 - \delta_{\mathrm{Q}} \right).$$

$\bullet_J$ reduces fixed points

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 $\mathcal{G} \subset V_J$ s.t. $f^* \in \mathcal{G}$, w.h.p.,

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

Proposition. Suppose $|J| \le cN$. Under some assumptions, $\exists\ 0 < \delta_{\mathrm{Q}} < \frac{1}{100}$, $c = c(q)$, and $c' = c'(q)$, s.t., when $q \ge 2$. Then $\forall r > 0$, $\forall \mathcal{G} \subset V_J$ s.t. $f^* \in \mathcal{G}$, w.p.a.l. $1 - \delta_{\mathrm{Q}}$,

$$r_{\mathrm{Q}}\left(\mathcal{G}, \delta_{\mathrm{Q}}, \frac{2}{q}\right) = 0.$$

FSD improves classical SLT: applying classical theory on V_J.

What does FSD provide?

V_J is morphism on the category of supervised regression problems and solutions.

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

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.

- $\hat{\boldsymbol{\beta}}_J$ is RERM, via self-regularization;
- $\boldsymbol{\beta}_J^*$ is new oracle;
- r_M and r_Q are reduced.

Other role: V_J^* is the true subspace where estimation happens (not this paper).

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

V_{J^c} provides stochastic properties of $\hat{\boldsymbol{\beta}}_J$.

$$\text{Dvoretzky} - \text{Milman} : \forall \boldsymbol{\mu} \in \mathbb{R}^N, \ \frac{\|\boldsymbol{\mu}\|_2}{(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)} \leq \|\mathcal{A}[\boldsymbol{\mu}]\|_q \leq \frac{\|\boldsymbol{\mu}\|_2}{(1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)}.$$

$\Rightarrow$ upper bounds for $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\|_2^2$. We are left with $\|\Sigma^{1/2}\boldsymbol{\beta}_{J^c}^*\|_2 + \|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$.

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

Remember that $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J \hat{\boldsymbol{\beta}}_J]$ in regression. By $N \lesssim d_*(\Sigma_{J^c}^{1/2} B_{q'}^p)$, and the fact $\|\Sigma_{J^c}^{1/2}\|_{\ell_q \to \ell_2} = \mathrm{diam}(\Sigma_{J^c}^{1/2} B_q^p)$, there holds

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

Since $\|\boldsymbol{y} - \mathbb{X}_J \hat{\boldsymbol{\beta}}_J\|_2 = \|\mathbb{X}_J(\boldsymbol{\beta}_J^* - \hat{\boldsymbol{\beta}}_J) + \mathbb{X}_{J^c} \boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}\|_2 \lesssim \sqrt{N}\sigma_\xi$, and $d_*(\Sigma_{J^c}^{1/2} B_{q'}^p) = (\frac{\ell_*(\Sigma_{J^c}^{1/2} B_q^p)}{\mathrm{diam}(\Sigma_{J^c}^{1/2} B_q^p)})^2$,

$$\|\Sigma^{1/2} \hat{\boldsymbol{\beta}}_{J^c}\|_2 \lesssim \sqrt{\frac{N}{d_*(\Sigma_{J^c}^{1/2} B_{q'}^p)}} \sigma_\xi. \tag{5}$$

However, this general bound is not always sharp.

Open direction: non-linear metric embedding

When $q = 2$, $\mathcal{A}[\boldsymbol{\mu}] = \operatorname{argmin}(\|\boldsymbol{\nu}\|_2 : \mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{\mu}) = \mathbb{X}_{J^c}^\top(\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top)^{-1}\boldsymbol{\mu}$. Linear!

Let $\boldsymbol{\zeta} = \mathbb{X}_J(\boldsymbol{\beta}^*_J - \hat{\boldsymbol{\beta}}_J) + \mathbb{X}_{J^c}\boldsymbol{\beta}^*_{J^c}$. Then $\|\boldsymbol{\zeta}\|_2 = o(\sqrt{N})$. As a result,

$$\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2 = \|\Sigma^{1/2}\mathcal{A}[\mathbb{X}_J(\boldsymbol{\beta}^*_J - \hat{\boldsymbol{\beta}}_J) + \mathbb{X}_{J^c}\boldsymbol{\beta}^*_{J^c} + \boldsymbol{\xi}]\|_2 \le \|\Sigma^{1/2}_{J^c}\mathcal{A}[\boldsymbol{\zeta}]\|_2 + \|\Sigma^{1/2}_{J^c}\mathcal{A}[\boldsymbol{\xi}]\|_2,$$

where $\|\Sigma^{1/2}_{J^c}\mathcal{A}[\boldsymbol{\zeta}]\|_2$ is much smaller, and

$$\mathbb{E}_{\boldsymbol{\xi}}\|\Sigma^{1/2}_{J^c}\mathcal{A}[\boldsymbol{\xi}]\|_2 \le \sigma_\xi(\operatorname{Tr}((\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top)^{-2}\mathbb{X}_{J^c}\Sigma_{J^c}\mathbb{X}_{J^c}^\top))^{1/2} \le \sigma_\xi\frac{\sqrt{N\operatorname{Tr}(\Sigma^2_{J^c})}}{\operatorname{Tr}(\Sigma_{J^c})}. \tag{6}$$

However, (5) provides only

$$\sigma_\xi\sqrt{\frac{N\|\Sigma_{J^c}\|_{\mathsf{op}}}{\operatorname{Tr}(\Sigma_{J^c})}} \ge \sigma_\xi\frac{\sqrt{N\operatorname{Tr}(\Sigma^2_{J^c})}}{\operatorname{Tr}(\Sigma_{J^c})}. \quad \Rightarrow \text{ The tool we used is not correct!}$$

What is the correct mathematical tool for studying $\mathcal{A} : (\mathbb{R}^N, \|\cdot\|_2) \to (V_{J^c}, \|\cdot\|_q)$'s $\|\Sigma^{1/2}_{J^c}\cdot\|_2$-norm, that is, $\|\Sigma^{1/2}_{J^c}\mathcal{A}[\boldsymbol{\zeta}+\boldsymbol{\xi}]\|_2^2$, so that it recovers (6) when $q = 2$?

Gaussian case

We say that an FSD is an admissible FSD in the Gaussian case (abbreviated as admissible FSD) if the following conditions are satisfied:

1. $V_J = \mathrm{span}(\boldsymbol{e}_j : j \in J)$, where $J \subset [p]$, and X_J is independent with X_{J^c}.
2. $|J| \lesssim N \lesssim d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p)$.
3. When $1 < q < 2$, $\boldsymbol{\beta}^*_{J^c} = \mathbf{0}$.

For any admissible FSD, and a positive parameter σ_ξ we define

$$R(V_J, V_{J^c}) = \begin{cases} \sigma_\xi \left(\frac{|J|}{N}\right)^{\frac{1}{2(q-1)}} + \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2 + \left|\!\left|\!\left|\boldsymbol{\beta}^*_J \odot |\boldsymbol{\beta}^*_J|^{\odot(q-2)}\right|\!\right|\!\right|^{\frac{1}{q-1}} \frac{\ell_*^{\frac{q}{q-1}}(\Sigma_{J^c}^{1/2} B_{q'}^p)}{N^{\frac{q}{2(q-1)}}} & \text{if } q \geq 2, \\ \sigma_\xi^{\frac{1}{3}} \left(\frac{|J|}{N}\right)^{\frac{1}{6}} + \sigma_\xi^{q-2} \left|\!\left|\!\left|\boldsymbol{\beta}^*_J \odot |\boldsymbol{\beta}^*_J|^{\odot(q-2)}\right|\!\right|\!\right| \frac{\ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)}{N^{\frac{q}{2}}} & \text{if } 1 < q < 2, \\ \sqrt{\frac{|J|}{N}} + \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2 + \varepsilon_1 \sigma_\xi & \text{if } q = 1, \end{cases}$$

where ε_1 is the distortion of Dvoretzky-Milman.

Upper bound of the estimation error

Theorem (Gaussian case, informal)

Suppose ξ is independent of X, $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$, and $X \sim \mathcal{N}(\mathbf{0}, \Sigma)$. Let $\hat{\boldsymbol{\beta}}$ be the minimum ℓ_q-norm interpolant estimator in the linear regression problem. For any admissible FSD in the Gaussian case, when N is large enough, the following hold w.h.p.,

1. *When $q > 1$,*

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

2. *When $q = 1$, suppose there exists an absolute constant $c_2 \in (0,1)$ such that*

$$\big(\ell^*(\Sigma_{J^c}^{1/2} B_1^p)\big)^2 \le c_2 N \Big(\sum_{j \in J \cap \mathrm{supp}(\boldsymbol{\beta}^*)} \sigma_j^{-1} \Big)^{-1}.$$

Then the same upper bound holds.

Rmk: Gaussian case is universal. **Rmk**: to obtain B.O., we need $r(V_J, V_{J^c}) \to 0$.

Sufficient conditions for B.O.

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

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

$$|J| = o(N) \qquad \text{, and} \qquad N = o(d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p)) \qquad \text{variance } \hat{\boldsymbol{\beta}}$$

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

$\hat{\boldsymbol{\beta}}_J$ 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 $\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2$.
  - Variance $N = o(d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p))$.

Take-home messages

- **Improvement of classical SLT via FSD**.
  - estimation ONLY happens on V_J instead of the entire $\mathcal{F}$.
  - the estimation error $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2$ can be controlled by interpolating cancellation $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\|_2$, and smallness $\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2 + \|\Sigma^{1/2}\boldsymbol{\beta}_J^*\|_2$.
- **Reason of B.O.: Self-regularization**, $\hat{\boldsymbol{\beta}}_J$ is a RERM with random loss function.

$$\hat{\boldsymbol{\beta}}_J \in \operatorname{argmin}\left(P_N\ell_{\boldsymbol{\beta}_J} + \|\boldsymbol{\beta}_J\|_q^q\right), \text{ where } P_N\ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q, \text{ and } \boldsymbol{\beta}_J^* \in \operatorname*{argmin}_{\boldsymbol{\beta}_J \in V_J} P\ell_{\boldsymbol{\beta}_J}.$$

- **AGA results <u>for</u> stats**: D-M theorem provides stochastic properties of $P_N\ell_\bullet$.
- **AGA results <u>from</u> stats**: Generalized Dvoretzky-Milman.
- **Implications in B.O.**
  - B.O. depends on the alignment between $\boldsymbol{\beta}^*$ and canonical basis.
  - Gaussian case is universal.
- **Mathematical challenge**: non-linear metric embedding.

End

The 'hard marl' of Benign Overfitting has finally been surrounded and dissolved by the geometric structure we built. What was a paradox in the classical uniform convergence world becomes a lucid geometric reality here.

Thank you for your attention!

FSD in classification

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

$$
\begin{aligned}
& P\mathcal{L}_{\hat{f}}^{(0,1)} \\
& = \mathbb{P}\left(Y\hat{f}(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) && (7) \\
& + \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y f_J^*(X) < 0\right) && (8) \\
& + \mathbb{P}\left(Y f_J^*(X) < 0\right) - \mathbb{P}\left(Y\left(\eta(X) - \frac{1}{2}\right) < 0\right), && (9)
\end{aligned}
$$

- (7): error caused by free part $\hat{f}_{J^c}$;
- (8): prediction error caused by $\hat{f}_J$ compared to that of f_J^* ; and
- (9): 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

$$\boldsymbol{\lambda}^- \in \operatorname{argmax}(\langle\boldsymbol{\lambda}, \boldsymbol{\mu}\rangle : \boldsymbol{\lambda} \succeq \mathbf{0}, (1+\tilde{\delta})\sqrt{\operatorname{Tr}(\Sigma_{J^c})}\|\boldsymbol{\lambda}\|_2 \le 1). \tag{10}$$

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 $\mu_i\lambda_i^- < 0$,

$$\langle\boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}^-\rangle = \sum_{i'\neq i}\mu_{i'}\lambda_{i'}^- > \sum_{i'=1}^N \mu_{i'}\lambda_{i'}^- = \langle\boldsymbol{\mu}, \boldsymbol{\lambda}^-\rangle, \Rightarrow \boldsymbol{\lambda}^- \text{ is not the maximizer of (10).}$$

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 Ψ 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 Ψ 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 ρ_* be the smallest r and its corrsponding. ρ, 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 $0 < \varepsilon_1, \kappa_{RIP}, \kappa_{DM} < 1$ 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 Σ 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.*

• 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

$$\partial^- P_N \ell_{\boldsymbol{\beta}_J} = \left\{ -q\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^{q-1} \mathbb{X}_J^\top \boldsymbol{\lambda}^*[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] : \boldsymbol{\lambda}^*[\cdot] \text{ is dual solution} \right\}. \tag{11}$$

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 (11), 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.

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

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

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

■

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.