INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2026IPPAG004

Thèse de doctorat

ENSAE

IP PARIS

# Feature Space Decomposition

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École nationale de la statistique et de l'administration économique

école doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 2 avril, 2026, par

**ZONG SHANG**

Composition du Jury :

| | |
|---|---|
| M. Alexandre Tsybakov<br>Professeur, ENSAE-CREST | Président |
| M. Vladimir Koltchinskii<br>Professeur, Georgia Institute of Technology | Rapporteur |
| Mme. Sara van de Geer<br>Professeur , ETH Zürich | Rapporteur |
| Mme. Claire Boyer<br>Professeur, Université Paris-Saclay | Examinateur |
| Mme. Marta Strzelecka<br>Professeur, University of Warsaw | Examinateur |
| M. Guillaume Lecué<br>Professeur, ESSEC Bussiness School | Directeur de thèse |
| M. Matthieu Lerasle<br>Professeur, ENSAE-CREST | Co-directeur de thèse |
| M. Taiji Suzuki<br>Professeur, University of Tokyo | Invité |

# Contents

# Acknowledgment

Standing at the end of this doctoral journey, my thoughts travel back not to the streets of Paris, but to the biting winters of Changchun, thousands of miles away, where the journey truly began.

First and foremost, I would like to express my deepest and most sincere gratitude to my supervisor, Guillaume Lecué. My life changed in the summer of 2018 because of him. Back then, I was an undergraduate who spent his days in the library self-studying high-dimensional probability, often at the risk of failing my own classes. When my academic future seemed bleak and I was told that no doors would open for me, Guillaume saw potential in the notes I sent him. He offered me a life-changing opportunity to join the PhD track at IP Paris. His rigor, brilliance, and immense kindness have transformed me from an unconventional "amateur" into a serious researcher. I owe my entire academic identity to his trust and guidance.

I am deeply grateful to Matthieu Lerasle, my co-advisor, for his constant support and insightful guidance throughout these years. I also wish to thank Johannes Schmidt-Hieber for his unwavering encouragement. I still remember our Skype call in 2021 when he first called me a "rising star"—a generous remark he reaffirmed during my visit to Twente in 2025. His consistent belief in my potential has been a powerful beacon during my many nights of self-doubt, even as I remain acutely aware of how much further I have to travel toward mathematical maturity.

I am profoundly honored to have Vladimir Koltchinskii and Sara Van de Geer as my rapporteurs. Having studied Professor Koltchinskii's Saint-Flour lectures and Professor Van de Geer's foundational works on M-estimation since my undergraduate years, I once dreamed of contributing to the methodologies they pioneered. In this thesis, I hope I have made a humble step toward that goal. This August, I will join the School of Mathematics at Georgia Tech as a Visiting Assistant Professor, and I am truly honored that Vladimir has accepted to be my mentor.

I wish to express my profound respect and gratitude to Alexandre Tsybakov. His immense scholarship is something I have long looked up to with great admiration, and I am honored to have him serve as a member of my jury. I also thank Claire Boyer for her time and for her valuable role on the jury. My gratitude also goes to Qian Lin from Tsinghua University, Taiji Suzuki from Tokyo University, Radoslaw Adamczak from the University of Warsaw, and Florentina Bunea from Cornell University for their hospitality and support during my long-term academic visits. I am also indebted to Martin Wahl, Lorenzo Rosasco, and Qiyang Han for their hospitality and the stimulating intellectual exchanges during my short-term visits. Their insights have broadened my perspective in ways that I deeply value.

A special mention must go to my collaborators and invited jury members, Taiji Suzuki and Marta Strzelecka. Taiji, your 2012 paper was the very first research article I ever read in my life; to work with you now is a surreal honor. My two-month visit to RIKEN-AIP in Tokyo—my favorite city in the world—remains the most cherished memory of my PhD. I also want to express my deep respect for Marta; our collaboration in Warsaw was a humbling experience. Her mathematical depth served as a mirror, revealing the thinness of my own foundations and reminding me of the vastness of the field I have chosen.

I am grateful to all those I have bothered with my oftentimes strange and eccentric questions; I thank you for your patience and goodwill. In particular, I must mention Gérard Ben Arous. Our encounter was an exceptionally difficult and, by his own later admission, harsh one; yet it taught me, perhaps more than anything else, the true nature of intellectual discourse and the reality of engaging with others in this field. Finally, my heartfelt thanks go to my colleagues and collaborators.

To my parents: I have always been an unconventional student. Through the years of 2017 to 2021, when I chose to walk my own path instead of following the traditional curriculum, your unconditional love and tolerance gave me the courage to explore the unknown. Even from thousands of miles away, you have been my strongest pillar.

To my girlfriend, Shanshan: As someone who constantly struggles with self-doubt and the anxiety of my "non-traditional" background, I never imagined I could achieve this. You were the one who stood firmly by me at every critical moment when I was on the verge of collapsing. Your company has been the warmest solace in this long, solitary trek.

Finally, I dedicate this thesis to the version of myself who walked through the snow in Changchun from 2017 to 2021. In those four years, I was truly "carrying my books and dragging my shoes through deep mountains and great valleys," as described by the ancient scholar Song Lian: *"When I was following my teachers, I carried my books and dragged my shoes, traveling through deep mountains and great valleys. In the depth of winter, the gale blew fiercely, and the snow was several feet deep..."* My "valleys" were the -30°C winds of Northeast China and the knee-deep snow I trudged through to reach the library. I thank and "resent" fate—for leading me to mathematics in that long winter and for bringing all these life-changing people into my path. To this day, I cannot say if this destiny has brought me more joy or sorrow, nor am I certain if I truly love mathematics enough to justify all that has been sacrificed. Yet, this thesis stands as the most authentic mark I can leave in the mist of an uncertain fate.

# Chapter 1

# Introduction

## 1.1 Statistical Supervised Learning Theory

We consider supervised learning problems under random design setup in statistical learning theory.

**Supervised Learning Problems and Their Solutions.** Let $(\Omega_X, \mu_X)$ and $(\Omega_Y, \mu_Y)$ be two probability spaces. Let $\Omega = \Omega_X \times \Omega_Y$. Let $X$ be a random element with distribution $\mu_X$ and $Y$ be a random element with distribution $\mu_Y$. We refer to $X$ as the design variable, input, or experiment, and to $Y$ as the response variable, label or output. Let $\mu$ be the joint probability measure of $(X, Y)$. Let $\ell : (\boldsymbol{y}_1, \boldsymbol{y}_2) \in \Omega_Y \times \Omega_Y \mapsto \ell(\boldsymbol{y}_1, \boldsymbol{y}_2) \in \mathbb{R}_+$ referred to as the loss function. A supervised learning problem is uniquely characterized by the pair of probability space and loss function, that is, $(\Omega, \mu, \ell)$.

A pair $(\mathcal{F}, \{\hat{f}_N\}_{N\in\mathbb{N}_+})$, consisting of a set $\mathcal{F} \subset \{f : \Omega_X \to \Omega_Y, f \text{ is measurable}\}$, called the statistical model, and a sequence of measurable maps $\{\hat{f}_N : (X_i, Y_i)_{i=1}^N \in \Omega^N \mapsto \hat{f}_N((X_i, Y_i)_{i=1}^N; \cdot) \in \mathcal{F}\}_{N\in\mathbb{N}_+}$, called decision rules, is called a solution, where $\hat{f}_N((X_i, Y_i)_{i=1}^N; \cdot) : \boldsymbol{x} \in \Omega_X \mapsto \hat{f}_N((X_i, Y_i)_{i=1}^N; \boldsymbol{x}) \in \Omega_Y$. Here, $N$ is called the sample size, $(X_i, Y_i)_{i=1}^N$ are called the training samples, and $\{\hat{f}_N\}_{N\in\mathbb{N}_+}$ is a sequence of decision rules that take the training samples and output a function $\hat{f}_N((X_i, Y_i)_{i=1}^N; \cdot)$ in $\mathcal{F}$; afterwards we abuse notation by simply writing $\hat{f}_N(\cdot)$, called estimator or learning machine. In this thesis, we always assume that the training samples $(X_i, Y_i)_{i=1}^N$ are independent copies of $(X, Y)$. Sometimes we also denote $(\mathcal{F}, \{\hat{f}_N\}_{N\in\mathbb{N}_+})$ by $(\{\hat{f}_N\}_{N\in\mathbb{N}_+}, \mathcal{F})$. Here, $N \in \mathbb{N}_+$ is for convenience; some decision rules are only defined for sufficiently large $N$, i.e., there exists $N_0 \in \mathbb{N}_+$ such that only $\{\hat{f}_N\}_{N>N_0}$ is well-defined; this case is still called a solution. When no confusion arises, we sometimes omit the subscript $N$ of $\hat{f}_N$ and simply write $\hat{f}$.

Research on supervised learning problems in statistical learning theory thus consists of two parts:

1. studying the structure of the supervised learning problem $(\Omega, \mu, \ell)$, and
2. investigating the statistical properties of a solution $(\mathcal{F}, \{\hat{f}_N\}_{N\in\mathbb{N}_+})$ of the given problem.

We provide numerous examples later in Section 1.2. Below we introduce the statistical property that is the focus of this thesis: prediction risk.

**Statistical Properties of Solutions.** The problems considered in this thesis fall under statistical prediction problems, whose core task is: given a test sample $X$, to predict its corresponding response $Y$ by using $\hat{f}_N(X)$. Thus, the quantity that evaluates the quality of a solution on the given supervised learning problem is the prediction risk incurred by the prediction $\hat{f}_N(X)$ compared to the true response $Y$. Next, we define the prediction risk mathematically.

Define $\ell_\bullet : f \in \mathcal{F} \mapsto \ell_f$ where $\ell_f : (\boldsymbol{x}, \boldsymbol{y}) \in \Omega \mapsto \ell(f(\boldsymbol{x}), \boldsymbol{y})$. We denote by $P : f \in L^1(\mu) \mapsto \mathbb{E}[f(X, Y)]$ the population mean, and we define the prediction risk/ population risk/ generalization error $P\ell_\bullet : f \in \mathcal{F} \mapsto P\ell_f$. Therefore, the prediction risk of an estimator $\hat{f}_N$ is a random variable defined by

$$P\ell_{\hat{f}_N} = \mathbb{E}\left[\ell(\hat{f}_N(X), Y) \middle| (X_i, Y_i)_{i=1}^N\right],$$

where $(X, Y)$ follows the probability distribution $\mu$ and is independent of the training samples $(X_i, Y_i)_{i=1}^N$; it is called the test sample. The predictor risk of an estimator measures its ability to generalize well on out-of-sample data.

To quantify the "smallness" of $P\ell_{\hat{f}_N}$, one typically compares it with either the minimum of the population risk over all measurable maps (assuming it exists) or the minimum over $\mathcal{F}$ (assuming it exists). Define

$$f^* \in \operatorname{argmin}(P\ell_f : f : \Omega_X \to \Omega_Y \text{ measurable}),$$

referred to as the Bayes rule; and

$$f^*_{\mathcal{F}} \in \operatorname{argmin}(P\ell_f : f \in \mathcal{F}),$$

referred to as the oracle in $\mathcal{F}$ associated with this problem.

The quantity $P\ell_{f^*}$ measures the intrinsic difficulty of the supervised learning problem $(\Omega, \mu, \ell)$, for instance its noise level. $P\ell_{f^*_{\mathcal{F}}}$ quantifies the minimal prediction risk achievable by using the function with the smallest prediction risk within the statistical model $\mathcal{F}$, i.e., the oracle. Thus, it characterizes the "approximation error" between the statistical model $\mathcal{F}$ and the problem $(\Omega, \mu, \ell)$.

**Remark 1.** *Beyond statistical prediction parameters, mathematical statistics is also cencerned with inference problems, i.e., constructing confidence intervals for statistical quantities. Such problems lie outside the scope of this thesis, but they are related to the statistical prediction problems studied herein, as seen in [BK20, Section 5]. Investigating the application of the Feature Space Decomposition method developed in this thesis to inference problems is an interesting research direction.*

**Remark 2.** *This thesis considers the random design setup, i.e., the design vectors $(X_i)_{i=1}^N$ are assumed to be random. In statistical learning theory, there is another setup called fixed design, where $(X_i)_{i=1}^N$ are assumed to be deterministic, while the response variables $(Y_i)_{i=1}^N$ are assumed to be random—the randomness stems only from the noise. In fixed design problems, it is typically assumed that $(X_i)_{i=1}^N$ are certain scattered sampling points on $\Omega_X$, as in [Tsy09, Chapter 1], see also [BvdG11], and the noise perturbs the corresponding responses at these points. In fixed design problems, there is no notion of test error; instead, the quantity used to evaluate the statistical performance of an estimator $\hat{f}_N$ is the empirical estimation error, for example, $\frac{1}{N}\sum_{i=1}^N(\hat{f}_N(X_i) - f^*(X_i))^2$. Although results obtained under fixed design hold for arbitrary $(X_i)_{i=1}^N$ and thus may appear stronger than random-design results that hold with high probability or in expectation—and sometimes the training error under fixed design is of the same order as the test error (in probability or expectation) under random design for the same estimator—we emphasize that these are two fundamentally different problems. This distinction becomes particularly evident when there is a gap between test error and training error, as with interpolant estimators; see Example 10 and Section 1.4.2.*

**Practice, Theory and the aim of this thesis.** The task of a practitioner is as follows: given $(\Omega, \ell)$ and some partial prior knowledge of $\mu$, one constructs a solution such that the prediction risk is as small as possible with high probability, or in expectation. In practice, the statistician does not know $\mu$ (otherwise one could simply set $\hat{f}_N = f^*$). Therefore, in practice, to assess the magnitude of the prediction risk, a practitioner usually needs to take multiple i.i.d. distributed test samples, say $(X_j, Y_j)_{j=1}^M$ independent with the training samples, to measure the prediction risk of $\hat{f}_N$, for instance, compute the mean of test errors $(\ell(\hat{f}_N(X_j), Y_j))_{j=1}^M$.

The aim of theorists is to assess the performance of such a solution from a theoretical viewpoint (so the theorists "know" $\mu$), thereby evaluating the effectiveness of the solution for the supervised learning problem by proving probability inequalities. A theorist still possesses the authority to choose a solution, i.e., when the statistical properties of a given solution are difficult to analyze, the theorist may opt to replace it with another solution for the same supervised learning problem whose statistical properties are easier to analyze, termed a "provable solution."

Below we denote $\mathcal{L}_\bullet : f \in \mathcal{F} \mapsto \mathcal{L}_f := \ell_f - \ell_{f^*}$ as the excess risk functional and $\mathcal{L}^{\mathcal{F}}_\bullet : f \in \mathcal{F} \mapsto \mathcal{L}^{\mathcal{F}}_f := \ell_f - \ell_{f^*_{\mathcal{F}}}$ as the excess risk functional with respect to the oracle $f^*_{\mathcal{F}}$. This thesis focuses on the following abstract mathematical problem:

> given an arbitrary problem $(\Omega, \mu, \ell)$ and arbitrary solution $(\mathcal{F}, \{\hat{f}_N\}_{N\in\mathbb{N}_+})$, investigate the properties of the random variable $P\mathcal{L}^{\mathcal{F}}_{\hat{f}_N}$ and $P\mathcal{L}_{\hat{f}_N}$.

The aim of this thesis is to

introduce a systematic methodology for tackling this abstract mathematical problem, thereby providing better mathematical strategies and tools for theorists to analyze the statistical properties of a solution.

Therefore, this thesis focuses on certain solutions that have been difficult for theorists to analyze, such as the interpolant estimators introduced in Example 10 later, and by establishing their statistical properties, it aims to develop a new set of proof strategies intended to improve the most fundamental proof methods in statistical learning theory.

Below, we review a classical task theorists usually face to analyze the statistical properties of an estimator $\hat{f}_N$.

**Oracle inequalities.** Regardless of whether the comparison is made with the Bayes rule or with the oracle, theorists may establish two different types of oracle inequalities, [Kol11, LM12].

1. **Exact oracle inequality.** One seeks a quantity $R(\mu, \ell, \hat{f}_N, \mathcal{F}, N)$ such that, with high probability or in expectation, one has $P\ell_{\hat{f}_N} - P\ell_{f^*_{\mathcal{F}}} \leq R(\mu, \ell, \hat{f}_N, \mathcal{F}, N)$, or $P\ell_{\hat{f}_N} - P\ell_{f^*} \leq R(\mu, \ell, \hat{f}_N, \mathcal{F}, N)$.

2. **Non-exact oracle inequality.** One seeks an absolute constant $C > 1$ and a quantity $R(\mu, \ell, \hat{f}_N, \mathcal{F}, N)$ such that, with high probability or in expectation, one has $P\ell_{\hat{f}_N} - CP\ell_{f^*_{\mathcal{F}}} \leq R(\mu, \ell, \hat{f}_N, \mathcal{F}, N)$, or $P\ell_{\hat{f}_N} - CP\ell_{f^*} \leq R(\mu, \ell, \hat{f}_N, \mathcal{F}, N)$.

Once either type of inequality is established, by comparing with $P\ell_{f^*_{\mathcal{F}}}$ or $P\ell_{f^*}$ (or a constant multiple thereof), one obtains an upper bound (in the high-probability or expectation sense) on the population risk/ generalization error of $P\ell_{\hat{f}_N}$.

**Remark 3.** *There are certain special situations in which one does not compare $P\ell_{\hat{f}}$ with $P\ell_{f^*_{\mathcal{F}}}$ or $P\ell_{f^*}$. Instead, for various reasons (for instance, although $f^*_{\mathcal{F}}$ or $f^*$ attains the minimal population risk, we may not favor it for some reason), one wishes to retain the freedom to choose the function (called the target function) against which the population risk is compared. This leads to a somewhat non-standard form of oracle inequality; see, for example, [P5]. See Section 1.3.4 later for details. For instance, some oracle inequalities compare with the minimizer of the regularized population risk, e.g., [Kol09]. We do not discuss them here.*

In Section 1.3 later, we detail how classical mathematical statistics establishes oracle inequalities.

## 1.2 Examples

### 1.2.1 Examples of supervised learning problems

Below we present the two most important examples of supervised learning problems.

**Example 1** (Real-valued regression)**.** *Let $f^* \in L^2(\mu_X)$ be an unknown function, also called the regression function of $Y$ given $X$, and let $\xi$ be a centered random variable independent of $X$ with variance $\mathbb{E}[\xi^2]$ denoted by $\sigma_\xi^2$. Define $Y = f^*(X) + \xi$. Taking $\ell = \ell^{(2)}$ as the squared loss, that is, $\ell^{(2)} : (y_1, y_2) \in \mathbb{R} \mapsto (y_1 - y_2)^2$. A real-valued regression problem is defined by such a triple $(\mu_X, f^*, \xi)$. In real-valued regression problems, the Bayes rule is $f^*$, satisfying $P\ell_{f^*} = \sigma_\xi^2$. Beyond this, one has $P\ell_{f^*} = \mathbb{E}[\xi^2]$ and $P\ell_{\hat{f}_N} = \|f^* - \hat{f}_N\|^2_{L^2(\mu_X)} + \mathbb{E}[\xi^2]$. If $f^* \notin \mathcal{F}$, the model is said to be misspecified. The term $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$ is referred to as the estimation error of $\hat{f}_N$ relative to the oracle $f^*_{\mathcal{F}}$; the term $\|f^*_{\mathcal{F}} - f^*\|^2_{L^2(\mu_X)}$ is referred to as the approximation error of the statistical model $\mathcal{F}$ relative to the signal $f^*$.*

*Since the approximation error $\|f^*_{\mathcal{F}} - f^*\|_{L^2(\mu_X)}$ is independent of the decision rule and depends only on the choice of the statistical model $\mathcal{F}$, and its study is essentially non-stochastic, in this thesis we focus primarily on the estimation error $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$.*

**Example 2** (Binary classification problems)**.** *Let $\Omega_Y = \{-1, 1\}$, and define the 0/1 loss function by $\ell = \ell^{(0/1)} : (y_1, y_2) \in \{-1, 1\} \times \{-1, 1\} \mapsto \mathbb{1}(\mathrm{sign}(y_1) \neq \mathrm{sign}(y_2))$, where $\mathrm{sign}(t) = 1$ if $t > 0$, $\mathrm{sign}(t) = -1$ if $t < 0$, and $\mathrm{sign}(t) \in [-1, 1]$ if $t = 0$. This is referred to as the supervised (binary) classification problem (also known as pattern recognition/ discriminant analysis). Define the function $\eta : \boldsymbol{x} \in \Omega_X \mapsto \mathbb{P}(Y = 1 \mid X = \boldsymbol{x})$, called the posterior distribution. The Bayes rule is then given by $f^* : \boldsymbol{x} \mapsto \mathrm{sign}(\eta(\boldsymbol{x}) - \frac{1}{2})$. For supervised classification problems, $(\mu, \ell)$ is uniquely characterized by the pair $(\mu_X, \eta)$. In supervised classification, for any given $f \in \mathcal{F}$, one has $P\ell_f = \mathbb{P}[\mathrm{sign}(Y) \neq \mathrm{sign}(f(X))] = \mathbb{P}[Yf(X) < 0]$.*

### 1.2.2 Examples of solutions to supervised learning problems

**Examples of statistical models in parametric statistics**

The statistical model determines the best function within that model, i.e., the population risk achievable by $f^*_{\mathcal{F}}$. Consequently, a statistical model with stronger "representational capacity" tends to have a smaller approximation error. However, such a model is typically more complex.

**Linear regression model.** For any solution $(\{\hat{f}_N\}_{N\in\mathbb{N}_+}, \mathcal{F})$ to a supervised learning problem, if $\mathcal{F}$ is parameterized by a linear algebraic structure, we call this a linear regression method. Note that functions in $\mathcal{F}$ may be allowed to be nonlinear with respect to the input.

**Example 3** (Real spaces)**.** *Let $p \in \mathbb{N}_+$. If $\mathcal{F} = \{\langle \cdot, \boldsymbol{v}\rangle : \boldsymbol{v} \in \mathbb{R}^p\}$, then clearly $\mathcal{F}$ is linearly parameterized. Therefore, on $\mathbb{R}^p$, any $\{\hat{f}_N\}_{N\in\mathbb{N}_+}$ is linear regression.*

**Example 4** (Reproducing kernel Hilbert space)**.** *If $\mathcal{F}$ is a real-valued Hilbert space of functions $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$, and $\mathcal{H} \subset L^2(\mu_X)$ satisfies that for any $\boldsymbol{x} \in \Omega_X$, there exists a constant $C_{\boldsymbol{x}}$ such that for any $f \in \mathcal{H}$, $|f(\boldsymbol{x})| \le C_{\boldsymbol{x}}\|f\|_{\mathcal{H}}$, where $\|\cdot\|_{\mathcal{H}}$ is the norm induced by the inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, then $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS). For any $\boldsymbol{x} \in \Omega_X$, there exists a continuous functional called the evaluation functional $\mathrm{ev}_{\boldsymbol{x}} : f \in \mathcal{H} \mapsto f(\boldsymbol{x}) \in \mathbb{R}$. By the Riesz representation theorem, for any $\boldsymbol{x} \in \Omega_X$, there exists $\phi(\boldsymbol{x}) \in \mathcal{H}$ such that for any $f \in \mathcal{H}$, $f(\boldsymbol{x}) = \langle f, \phi(\boldsymbol{x})\rangle_{\mathcal{H}}$; this is called the reproducing property. We call $\phi : \Omega_X \to \mathcal{H}$ the feature map. In an RKHS, $\mathcal{F}$ is linearly parameterized by $\phi$; therefore, any $\{\hat{f}_N\}_{N\in\mathbb{N}_+}$ on an RKHS is linear regression. The most classical RKHS is the Sobolev space based on $L^2(\mu_X)$ that satisfies certain Sobolev embedding theorem, when $\Omega_X$ is a compact subset of $\mathbb{R}^d$, [Bac24, Chapter 7].*

**Example 5** (Reproducing kernel Banach space)**.** *If $\mathcal{F}$ is a Banach space $(\mathcal{B}, \|\cdot\|_{\mathcal{B}})$ consisting of real-valued functions, and $\mathcal{B} \subset L^2(\mu_X)$ satisfies that for any $\boldsymbol{x} \in \Omega_X$, there exists a constant $C_{\boldsymbol{x}}$ such that for any $f \in \mathcal{B}$, $|f(\boldsymbol{x})| \le C_{\boldsymbol{x}}\|f\|_{\mathcal{B}}$, then $\mathcal{B}$ is called a Reproducing Kernel Banach Space (RKBS). Analogously (see, e.g., [BVRV24]), for any $\boldsymbol{x} \in \Omega_X$, there exists $\mathrm{ev}_{\boldsymbol{x}} \in \mathcal{B}'$ such that for any $f \in \mathcal{B}$, $f(\boldsymbol{x}) = \langle \mathrm{ev}_{\boldsymbol{x}}, f\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the dual pairing between $\mathcal{B}'$ and $\mathcal{B}$. We call $\phi : \boldsymbol{x} \in \Omega_X \mapsto \phi(\boldsymbol{x}) = \mathrm{ev}_{\boldsymbol{x}} \in \mathcal{B}'$ the feature map. In an RKBS, $\mathcal{F}$ is linearly parameterized by $\phi$; therefore, on an RKBS, any $\{\hat{f}_N\}_{N\in\mathbb{N}_+}$ is linear regression. The most classical RKBS is the Banach space $\mathcal{M}(\Theta)$ of signed measures on a compact set $\Theta$ equipped with the total variation norm, which parameterizes shallow neural networks. That is, $\mathcal{B}$ consists of bounded continuous functions on $\Omega_X$, and $\mathcal{B}'$ is identified with $C_b(\Theta)$, [BVRV24].*

Apart from the examples mentioned above, many classical function spaces, such as Hölder spaces, general Sobolev spaces, and Besov spaces, are also linear spaces. However, on these spaces, we typically do not linearly parameterize $\mathcal{F}$; therefore, they are not generally referred to as linear regression, but are typically referred to as nonparametric statistics.

**Non-linear regression model.** Statistical models that do not fall under linear regression are referred to as nonlinear regression models.

**Example 6** (Neural networks)**.** *Let $d, L$ be positive integers, and let $W_1, \dots, W_{L-1}$ be $L-1$ positive integers. Let $\mathbb{W}_1 \in \mathbb{R}^{W_1\times d}$, $\mathbb{W}_2 \in \mathbb{R}^{W_2\times W_1}, \dots, \mathbb{W}_{L-1} \in \mathbb{R}^{W_{L-1}\times W_{L-2}}$ be $L-1$ matrices, $\boldsymbol{w}_L \in \mathbb{R}^{W_{L-1}}$, and let $\sigma_1, \dots, \sigma_{L-1}$ be $L-1$ real-valued functions (typically nonlinear). The statistical model formed by fully connected feedforward neural networks is defined as*

$$\mathcal{F} = \{f_{\mathbb{W}_1,\cdots,\mathbb{W}_{L-1},\boldsymbol{w}_L}(\cdot) = \langle \boldsymbol{w}_L, \sigma_{L-1}(\mathbb{W}_{L-1}\sigma_{L-2}(\mathbb{W}_{L-2}\sigma_{L-3}(\cdots \mathbb{W}_2\sigma_1(\mathbb{W}_1\cdot))))\rangle : \mathbb{W}_1, \cdots, \mathbb{W}_{L-1}, \boldsymbol{w}_L\}.$$

*Here, $(\sigma_j)_{j=1}^{L-1}$'s act coordinate-wisely. Due to the nonlinearity of $\sigma_1, \dots, \sigma_{L-1}$, although $\mathcal{F}$ is parameterized, it is not linearly parameterized. With a bit abuse of notation, we write $\mathbb{W}_L = \boldsymbol{w}_L^\top$.*

The statistical model in Example 6 is not only nonlinear but also nonconvex. In contrast, the following model (Example 7) is convex.

**Example 7** (mean-field shallow neural networks)**.** *Let $\Theta \subset \mathbb{R}^{d+1}$ be a compact set. An element $\boldsymbol{\theta} \in \Theta$ is written as $(a, \boldsymbol{w})$, where $a \in \mathbb{R}$ and $\boldsymbol{w} \in \mathbb{R}^d$. Let $\mathcal{P}(\Theta)$ denote the set of all probability measures that are absolutely continuous with respect to the Lebesgue measure $d\boldsymbol{\theta}$. For any bounded continuous function $g \in C_b(\Theta)$ and probability measure $\nu \in \mathcal{P}(\Theta)$, define $\langle g, \nu\rangle = \int_\Theta g(\boldsymbol{\theta})\, d\nu(\boldsymbol{\theta})$. Let $\phi : \boldsymbol{x} \in \mathbb{R}^d \mapsto \phi(\boldsymbol{x}) \in C_b(\Theta)$ satisfying $\sup_{\boldsymbol{x}} \|\phi(\boldsymbol{x})\|_\infty < \infty$ be the feature*

*map. Here* $\phi(\boldsymbol{x}) : (a, \boldsymbol{w}) \in \Theta \mapsto a\sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle)$ *where* $\sigma : \mathbb{R} \to \mathbb{R}$ *is a continuous bounded function. Define* $\mathcal{F} = \{f_\nu(\cdot) = \langle \phi(\cdot), \nu\rangle : \nu \in \mathcal{P}(\Theta)\}$. *Then for any* $\boldsymbol{x} \in \mathbb{R}^d$ *and any* $f_\nu \in \mathcal{F}$, *there holds* $|f_\nu(\boldsymbol{x})| \leq \|\phi(\boldsymbol{x})\|_\infty \|\nu\|_{L^1(\Theta)} = \|\phi(\boldsymbol{x})\|_\infty$ *where we used the* reproducing property $f_\nu(\boldsymbol{x}) = \langle \phi(\boldsymbol{x}), \nu\rangle$ *that holds for all* $\boldsymbol{x}$. *Therefore, the statistical model* $\mathcal{F}$ *is a subset of a ball of the reproducing kernel Banach space generated by* $\phi$ *defined in Example 5.*

### Examples of Decision rules

decision rules determine how to effectively construct an estimator/predictor $\hat{f}_N$ from the training samples $(X_i, Y_i)_{i=1}^N$ within a given statistical model $\mathcal{F}$, such that the estimation error is small. If the statistical model is extremely complex, this construction can be exceedingly difficult. A decision rule uses the training samples to probe the structure of $\mu$.

**Example 8** ((Regularized) Empirical Risk Minimization)**.** *The most fundamental decision rules are empirical risk minimization (ERM) and regularized empirical risk minimization (RERM), defined as*

$$\hat{f}_N \in \operatorname{argmin}\left(L_f((X_i, Y_i)_{i=1}^N) + \lambda\Psi(f) : f \in \mathcal{F}\right), \text{ for some } L_\bullet : f \in \mathcal{F} \mapsto L_f \in \{g : \Omega^N \to \mathbb{R}\},$$

*and* $\Psi : \mathcal{F} \to \mathbb{R}$ *is some regularization functional and* $\lambda \in \mathbb{R}$ *is called the regularization parameter. When* $\lambda = 0$, $\hat{f}_N$ *is called ERM, otherwise called RERM.*

*A most classical form of* $L_f$ *is* $L_f = \frac{1}{N}\sum_{i=1}^N \ell_f$ *as the empirical mean. However,* $L_f$ *can also take other forms, e.g.,* $L_f : (X_i, Y_i)_{i=1}^N \mapsto \left(\sum_{i=1}^N |Y_i - f(X_i)|^q\right)^{1/q}$, *where* $0 \leq q \leq \infty$.

ERM and RERM are at the origin of statistical learning theory [VC68]. However, because real-valued regression problems also encompass estimation problems in mathematical statistics, this thesis and the Feature Space Decomposition introduced herein are also applicable to classical mathematical-statistics methods beyond (R)ERM, such as the following class of spectral methods.[1]

**Example 9** (Spectral methods)**.** *Spectral methods are a class of methods adapted to linear regression problems on RKHS (see Example 4)* $(\mathcal{H}, \langle\cdot,\cdot\rangle_\mathcal{H})$, *whose feature map is* $\phi$. *Denote* $\mathbb{X} : f \in \mathcal{H} \mapsto (\langle f, \phi(X_i)\rangle_\mathcal{H})_{i=1}^N \in \mathbb{R}^N$ *as the design matrix, sometimes also called the sampling operator; and denote* $\mathbb{X}^\top : \boldsymbol{\lambda} \in \mathbb{R}^N \mapsto \sum_{i=1}^N \lambda_i \phi(X_i) \in \mathcal{H}$ *by its conjugate operator. Define* $\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^N \phi(X_i) \otimes \phi(X_i)$ *as the sample covariance matrix. Let* $\boldsymbol{y} = (Y_1, \ldots, Y_N) \in \mathbb{R}^N$ *be the response vector. Let* $(\varphi_t)_{t\geq 1}$ *be a family of real-valued functions defined on* $\mathbb{R}^+$ *call the filter functions. For all* $t \geq 1$, *we define the spectral method associated with* $\varphi_t$ *by:*

$$\hat{f}_N : \boldsymbol{y} \in \mathbb{R}^N \mapsto \hat{f}_N(\boldsymbol{y}) = \frac{1}{N}\varphi_t(\hat{\Sigma})\mathbb{X}^\top \boldsymbol{y} = \frac{1}{N}\mathbb{X}^\top \varphi_t(\frac{1}{N}\mathbb{X}\mathbb{X}^\top)\boldsymbol{y} \tag{1.1}$$

*where* $\varphi_t(\hat{\Sigma})$ *and* $\varphi_t(\frac{1}{N}\mathbb{X}\mathbb{X}^\top)$ *are defined via the spectral calculus. When there is no ambiguity, we abbreviate* $\hat{f}_N(\boldsymbol{y})$ *as* $\hat{f}_N$. *Spectral methods encompass a broad class of common linear regression techniques, such as ridge regression, gradient descent, gradient flow, heavy-ball method, Nesterov's acceleration, and principal component regression, among others, see Example 16 of Chapter 3 for details.*

**Example 10** (Minimum norm interpolant estimators)**.** *In this thesis, we only consider the minimum norm interpolant estimator in linear regression problems. If the statistical model* $\mathcal{F}$ *is a RKBS* $(\mathcal{F}, \|\cdot\|)$, *see Example 5, then the minimum* $\|\cdot\|$*-norm interpolant estimator in regression is defined as*

$$\hat{f}_N \in \operatorname{argmin}\left(\|f\| : f(X_i) = Y_i, \forall 1 \leq i \leq N\right).$$

*Its existence is established in [BDVRV23]. Similarly, we define the minimum norm interpolant classifier as*

$$\hat{f}_N \in \operatorname{argmin}(\|f\| : \operatorname{sign}(f(X_i)) = Y_i, \forall 1 \leq i \leq N).$$

Minimum norm interpolant estimators often arise as an implicit regularization in a class of optimization algorithms (see Section 1.4.1 later).

[1] Many spectral methods can also be written in the form of RERM, [AKT19], but in such cases the regularization is a quadratic form that depends on $(X_i)_{i=1}^N$, rather than a fixed functional $\Psi$ as in the general definition of RERM.

## 1.3 Classical Statistical Learning Theory based on Uniform Convergence Argument

In this section, we introduce one of the most fundamental proof methods in statistical learning theory: the uniform convergence argument, [VC68, Kol11]. Most of the content in this section is a summary and synthesis of developments in statistical learning theory over the past few decades, and includes some unpublished observations by the author. The uniform convergence argument can be traced back at least to [VC68] proposed to study ERM. We first recall ERM defined in Example 8:

$$\hat{f}_N \in \operatorname{argmin}\left(L_f((X_i,Y_i)_{i=1}^N) : f \in \mathcal{F}\right), \text{ for some } L_\bullet : f \in \mathcal{F} \mapsto L_f \in \{g : \Omega^N \to \mathbb{R}\}.$$

With some abuse of notation, we still denote $L_f((X_i,Y_i)_{i=1}^N)$ as $P_N\ell_f$, even though $L_f((X_i,Y_i)_{i=1}^N)$ cannot always be written as an integral with respect to an empirical measure, e.g., $L_f((X_i,Y_i)_{i=1}^N) = (\sum_{i=1}^N |Y_i - f(X_i)|^2)^{1/2}$. We write $P\ell_f = \mathbb{E}[L_f((X_i,Y_i)_{i=1}^N)]$, $P_N\mathcal{L}_f^{\mathcal{F}} = L_f((X_i,Y_i)_{i=1}^N) - L_{f_{\mathcal{F}}^*}((X_i,Y_i)_{i=1}^N)$, and $P\mathcal{L}_f^{\mathcal{F}} = \mathbb{E}[P_N\mathcal{L}_f^{\mathcal{F}}]$. In particular, when there exists $\ell_f : \Omega_X \times \Omega_Y \to \mathbb{R}$ such that $L_f((X_i,Y_i)_{i=1}^N) = \frac{1}{N}\sum_{i=1}^N \ell_f(X_i,Y_i)$, then $P\ell_f = \mathbb{E}[\ell_f(X,Y)]$ is precisely the prediction risk of $f$.

After nearly sixty years of development since [VC68], the connotation of the uniform convergence argument has long extended beyond its initial version. In this thesis, we consider the following generalized uniform convergence argument. Broadly speaking, the uniform convergence argument refers to the method of bounding the population excess risk of an estimator by the supremum or infimum of an empirical process. The uniform convergence argument is based on the following idea: since the estimator $\hat{f}_N$ is a random function, a useful approach when analyzing its population excess risk is to carefully construct a suitable subset of $\mathcal{F}$ and prove that $\hat{f}_N$ belongs to this subset with high probability, transforming the random nature of $\hat{f}_N$ into the stochastic properties of each deterministic function in that set.

**Proposition 1.** *Let $(\Omega,\mu,\ell)$ be a supervised learning regression problem, and $(\hat{f}_N,\mathcal{F})$ be an ERM. Then $P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} \le \sup\left((P-P_N)\mathcal{L}_f : f \in \mathcal{F}\right)$. Furthermore, if there are two real numbers $R(\mu,\ell,\mathcal{F},N) > 0$ and $0 < \delta(\mu,\ell,\mathcal{F},N) < 1$ such that $\mathbb{P}\left(\sup\left((P-P_N)\mathcal{L}_f^{\mathcal{F}} : f \in \mathcal{F}\right) \le R(\mu,\ell,\mathcal{F},N)\right) \ge 1-\delta(\mu,\ell,\mathcal{F},N)$, then the oracle inequality $P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} \le R(\mu,\ell,\mathcal{F},N)$ holds with probability at least $1-\delta(\mu,\ell,\mathcal{F},N)$. If there exists a real number $R'(\mu,\ell,\mathcal{F},N) > 0$ such that $\mathbb{E}\sup\left((P-P_N)\mathcal{L}_f^{\mathcal{F}} : f \in \mathcal{F}\right) \le R'(\mu,\ell,\mathcal{F},N)$, then the oracle inequality in expectation $\mathbb{E}[P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}}] \le R'(\mu,\ell,\mathcal{F},N)$ holds.*

**Proof.** By definition, $P_N\ell_{\hat{f}_N} \le P_N\ell_{f_{\mathcal{F}}^*}$, hence

$$P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} = P_N\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} + (P-P_N)\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} \le (P-P_N)\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} \le \sup\left((P-P_N)\mathcal{L}_f^{\mathcal{F}} : f \in \mathcal{F}\right). \tag{1.2}$$

The conclusion in the "furthermore" part is evident. ∎

Proposition 1 is the most primitive form of the uniform convergence argument. In Proposition 1, $\sup((P-P_N)\mathcal{L}_f^{\mathcal{F}} : f \in \mathcal{F})$ is an (one-sided) empirical process if $P_N\ell_f = \frac{1}{N}\sum_{i=1}^N \ell_f(X_i,Y_i)$. Proposition 1 imposes almost no restrictions on the supervised learning problem or its solution, except for the following two points.

1. $\sup(\mathbb{E}\mathcal{L}_f(X,Y) - \frac{1}{N}\sum_{i=1}^N \mathcal{L}_f(X_i,Y_i) : f \in \mathcal{F})$ is finite (with high probability or in expectation). This requires $\mathcal{F}$ to satisfy certain conditions; for example, if one wishes to use Proposition 1 to prove that $P\mathcal{L}_{\hat{f}_N}$ tends to 0 in some limit as $N \to \infty$, a necessary condition is that $\mathcal{F}$ is a Glivenko–Cantelli class, [VDVW23, pp. 130];
2. $\hat{f}_N$ is a global minimizer of $P_N\ell_f$ over $f \in \mathcal{F}$.

Existing refinements of the most primitive uniform convergence argument provided by Proposition 1 focus primarily on handling (1.2).

1. The loss $L_\bullet$ used by the ERM and the loss $\ell_\bullet$ defining the supervised learning problem may differ. For instance, in binary classification, if we take $L_f = \frac{1}{N}\sum_{i=1}^N \ell_f^{(0,1)}$, i.e., $\ell_f^{(0,1)} : (\boldsymbol{x},y) \in \Omega_X \times \{-1,1\} \mapsto \mathbb{1}(\operatorname{sign}(f(\boldsymbol{x})) \neq \operatorname{sign}(y))$, then computing this ERM is often an NP-hard problem; see, e.g., [BBL05, Section 4]. Consequently, practitioners typically choose some convex surrogate to define the ERM, e.g., $\ell_f^{(\text{hinge})} : (y_1,y_2) \in \mathbb{R}\times\mathbb{R} \mapsto (1-y_1y_2)_+$, where $(x)_+ = \max(x,0)$. The resulting (R)ERM $\hat{f}_N \in \operatorname{argmin}(\frac{1}{N}\sum_{i=1}^N \ell_f^{(\text{hinge})}(X_i,Y_i) + \lambda\Psi(f) : f \in \mathcal{F})$ is called a support vector machine. If we directly use Proposition 1, we would only obtain the population excess risk defined by $\ell_f^{(\text{hinge})}$, rather than the population excess risk defined by $\ell_f^{(0,1)}$, which is our primary concern.

2. Directly applying Proposition 1 may yield an upper bound on $P\mathcal{L}^{\mathcal{F}}_{\hat{f}_N}$ that is far looser than what is actually observed. This is often because the step of bounding $(P-P_N)\mathcal{L}^{\mathcal{F}}_{\hat{f}_N}$ by $\sup\left((P-P_N)\mathcal{L}^{\mathcal{F}}_f : f\in\mathcal{F}\right)$ is too crude. In fact, in many situations, there exists a subset $\mathcal{G}\subset\mathcal{F}$ such that the supremum of the empirical process over $\mathcal{G}$ is much smaller than that over $\mathcal{F}$, and $\hat{f}_N$ belongs to $\mathcal{G}$ with high probability. This implies that, on this random event, we could have used a more precise, localized uniform convergence argument. This scenario is particularly common when $\mathcal{F}$ is a convex set.

3. For some estimators like ERM when $\mathcal{F}$ is non-convex, when the actually computed estimator $\tilde{f}_N$ is not the theoretical global minimizer $\hat{f}_N$—even if $\hat{f}_N$ might have a very small $P\mathcal{L}^{\mathcal{F}}_{\hat{f}_N}$, such an analysis is often meaningless (see Section 1.4.1). In this thesis, we do not address this situation.

In Section 1.3.1 below, we introduce some fixed points, which are used to address item *1* and item *2*. For convenience, in what follows we focus primarily on regression problems; hence our main goal is to obtain an upper bound for the estimation error $\|\hat{f}_N - f^*_{\mathcal{F}}\|_{L^2(\mu_X)}$.

### 1.3.1 The fixed points as a way to measure the statistical complexity of $\mathcal{F}$

We introduce two sets of fixed points, both of which describe a certain kind of lower isomorphy of the empirical excess risk $P_N\mathcal{L}^{\mathcal{F}}_f$ at those fixed-point levels. The first set consists of the multiplier fixed point, the quadratic fixed point, and their maximum, which defines the lower isomorphic fixed point, [LM16]. This set is mainly suitable when $\ell(\cdot,y)$ is a strongly convex function for any $y\in\mathbb{R}$. The second set is the two-sided isomorphic fixed point, which is built upon the (local) Bernstein's condition that is defined later, [BM06, CLL20]. The latter set of fixed points is primarily applicable when $\ell$ itself is not strongly convex, but due to properties of $\mu$, the function $P\ell_\bullet$ exhibits strong convexity on some localized subset. Here we recall that $P\ell_f = \mathbb{E}[\ell(f(X),Y)]$, and the convexity of $\ell(\cdot,y)$ implies that for any $0\le\alpha\le 1$ and any $f_1,f_2\in\mathcal{F}$, one has $P\ell_{\alpha f_1+(1-\alpha)f_2} = \mathbb{E}[\ell(\alpha f_1(X)+(1-\alpha)f_2(X),Y)] \le \alpha\mathbb{E}[\ell(f_1(X),Y)] + (1-\alpha)\mathbb{E}[\ell(f_2(X),Y)] = \alpha P\ell_{f_1} + (1-\alpha)P\ell_{f_2}$. Similar convexity holds for $P_N\ell_f$ as well. For any set $\mathcal{G}$, we say $\mathcal{G}$ is a localization subset if $f^*_{\mathcal{F}}\in\mathcal{G}\subset\mathcal{F}$. For any $r\ge 0$ and any $f_{\mathcal{F}}\in\mathcal{F}$, we denote $B_{L^2(\mu_X)}(f_{\mathcal{F}};r)$ by $\{f\in\mathcal{F} : \|f-f_{\mathcal{F}}\|_{L^2(\mu_X)}\le r\}$, and $S_{L^2(\mu_X)}(f_{\mathcal{F}};r)$ by $\{f\in\mathcal{F} : \|f-f_{\mathcal{F}}\|_{L^2(\mu_X)} = r\}$. We denote by $\langle\cdot,\cdot\rangle$ the dual pair of a Banach space and its dual space. Let $\partial^- f$ denote the sub-differential of a convex function $f$. For the convex analysis in Banach spaces, see [BP12].

**Definition 1** (Multiplier fixed point)**.** *Let $(\Omega,\mu,\ell)$ be a supervised learning problem. Suppose $\mathcal{F}$ is contained in a Banach space. Let $\mathcal{G}$ be any localization subset, $\kappa,\square>0$ and $0<\delta_{\mathrm{M}}<\frac{1}{2}$ real numbers. Then the multiplier fixed point $r_{\mathrm{M}}(\mathcal{G},\delta_{\mathrm{M}},\kappa,\square)$ is defined as*

$$r_M(\mathcal{G},\delta_{\mathrm{M}},\kappa,\square) = \min_{r>0}\left(\mathbb{P}\left(\sup_f \inf_{\boldsymbol{g}\in\partial^- P_N\ell_{f^*_{\mathcal{F}}}} |\langle \boldsymbol{g}, f-f^*_{\mathcal{F}}\rangle| \le \square r^{\frac{2}{\kappa}}\right)\ge 1-\delta_{\mathrm{M}}\right), \tag{1.3}$$

*where the supremum is taken over $f\in\mathcal{G}\cap B_{L^2(\mu_X)}(f^*_{\mathcal{F}};r)$.*

**Definition 2** (Quadratic fixed point)**.** *Let $(\Omega,\mu,\ell)$ be a supervised learning problem. Suppose $\mathcal{F}$ is contained in a Banach space. Let $\mathcal{G}$ be a localization subset, and let $0<\delta_{\mathrm{Q}}<\frac{1}{2}$, $\kappa>0$ be real numbers. Then the quadratic fixed point $r_{\mathrm{Q}}(\mathcal{G},\delta_{\mathrm{Q}},\kappa,\triangle)$ is defined as the solution of the following minimization problem*

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

*Abusing notation, we also denote this minimal $\triangle$ as $\triangle$.*

In [Wai19, Section 9.3], the term $P_N\mathcal{L}^{\mathcal{F}}_f - \langle \boldsymbol{g}, f-f^*_{\mathcal{F}}\rangle$ is called the error of the first-order Taylor expansion, when $\partial^- P_N\ell_{f^*_{\mathcal{F}}}$ is a singleton. The Restricted Strong Convexity (RSC) introduced in [Wai19, Definition 9.15] is similar to our subsequent treatment.

**Example 11.** *When $L_f : (X_i,Y_i)_{i=1}^N \mapsto \frac{1}{N}\sum_{i=1}^N (Y_i - f(X_i))^2$, $r_{\mathrm{Q}}(\mathcal{G},\delta_{\mathrm{Q}},1)$ is equivalent to*

$$\min_{r>0}\left(\exists\triangle>0 \text{ independent with } r,\ \text{,s.t., } \mathbb{P}\left(\inf_{f\in\mathcal{G}\cap S_{L^2(\mu_X)}(f^*_{\mathcal{F}};r)} P_N(f-f^*_{\mathcal{F}})^2 \ge \triangle r^2\right)\ge 1-\delta_{\mathrm{Q}}\right),$$

*where $\triangle$ may be taken as an absolute constant. This is called the fixed point of the one-sided Restricted Isomorphic Property (RIP), denoted by $r_{\mathrm{RIP},-}(\mathcal{G})$, [CT05]. The corresponding $\delta_{\mathrm{Q}}$ is denoted by $\delta_{\mathrm{RIP}}$. A standard result is: if $\mathcal{F}$ can be identified with linear functionals on $\mathbb{R}^p$ and $\dim(\mathcal{F}) < N$, then under very general conditions (e.g., the $L^4(\mu_X)$–$L^2(\mu_X)$ norm-equivalence property), the following holds: there exists $\delta_{\mathrm{RIP}} < \frac{1}{100}$ such that for every $\mathcal{G} \subset \mathcal{F}$, we have $r_{\mathrm{RIP}}(\mathcal{G}) = 0$; see [KM15]. See also [P6].*

**Definition 3** (Lower isomorphic fixed point)**.** *Let $(\Omega, \mu, \ell)$ be a supervised learning problem. Suppose $\mathcal{F}$ is contained in a Banach space. Let $\mathcal{G} \subset \mathcal{F}$ be a localization subset, and let $0 < \delta_{\mathrm{Q}}, \delta_{\mathrm{M}} < \frac{1}{2}$, $\kappa > 0$ be real numbers. Define $r_{\mathrm{iso}}(\mathcal{G}, \delta_{\mathrm{M}}, \delta_{\mathrm{Q}}, \kappa, \square, \triangle) = \max\{r_{\mathrm{M}}(\mathcal{G}, \delta_{\mathrm{M}}, \kappa, \square), r_{\mathrm{Q}}(\mathcal{G}, \delta_{\mathrm{Q}}, \kappa, \triangle)\}$ (which we abbreviate as $r_{\mathrm{iso}}(\mathcal{G}, \kappa)$) and $\delta_{\mathrm{iso}}(\mathcal{G}, \kappa) = \delta_{\mathrm{M}} + \delta_{\mathrm{Q}}$.*

If $\triangle > 2\square$, then with probability at least $1 - \delta_{\mathrm{iso}}(\mathcal{G}, \kappa)$, the following random event happens:

$$\forall f \in \mathcal{G} \cap B_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r_{\mathrm{iso}}(\mathcal{G}, \kappa)),\ P_N \mathcal{L}_f^{\mathcal{F}} \geq -\square r_{\mathrm{iso}}^{\frac{2}{\kappa}}(\mathcal{G}, \kappa), \tag{1.4}$$

and

$$\forall f \in \mathcal{G} \cap S_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r_{\mathrm{iso}}(\mathcal{G}, \kappa)),\ P_N \mathcal{L}_f^{\mathcal{F}} \geq \triangle r_{\mathrm{iso}}(\mathcal{G}, \kappa)^{\frac{2}{\kappa}} - \square r_{\mathrm{iso}}^{\frac{2}{\kappa}}(\mathcal{G}, \kappa). \tag{1.5}$$

**Definition 4** (Two-sided isomorphic fixed point)**.** *Let $(\Omega, \mu, \ell)$ be a supervised learning problem. Let $\mathcal{G}$ be a localization subset, and let $0 < \delta < 1$, $\kappa > 0$ and $\kappa_{\mathrm{iso},2} < 1$ be real numbers. Define*

$$r_{\mathrm{iso},2}(\mathcal{G}, \delta, \kappa) = \min\left(r > 0 : \mathbb{P}\left(\sup\left(|(P - P_N)\mathcal{L}_f^{\mathcal{F}}| : f \in \mathcal{G} \cap B_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r)\right) \leq \kappa_{\mathrm{iso},2} r^{\frac{2}{\kappa}}\right) \geq 1 - \delta\right). \tag{1.6}$$

Any one of $r_{\mathrm{iso}}$ and $r_{\mathrm{iso},2}$ can be used as an upper bound for the estimation error of RERM; the main difference lies in

1. when $P_N \ell_\bullet$ is strongly convex, one can directly compute the smallest singular value of the empirical Hessian matrix $\nabla^2 P_N \ell_f$ uniformly over a neighborhood of $f^*_{\mathcal{F}}$, c.f., [CLM24, Lemma 3], [Men15]. In this case, we should adopt $r_{\mathrm{iso}}$.
2. When $\ell_\bullet$ is not necessarily strongly convex, we instead compare $P_N \mathcal{L}_\bullet$ with $P\mathcal{L}_\bullet$, using properties of $\mu$ to obtain some strong convexity at the population level, i.e., the Bernstein condition. In this case, we should use $r_{\mathrm{iso},2}$.

We emphasize that when using $r_{\mathrm{iso},2}$, i.e., item *2.*, to establish a uniform lower bound for $P_N \mathcal{L}_f^{\mathcal{F}}$, we usually need the condition $f^*_{\mathcal{F}} \in \arg\min(P\ell_f : f \in \mathcal{F})$. In contrast, when we directly exploit the strong convexity of $P_N \ell_\bullet$, i.e., item *1.*, it is not strictly necessary to use that condition. Instead, we rely on the fact that $\hat{f}_N$ is a minimizer of the RERM, i.e., first-order and second-order optimality conditions—which in some situations allows us to handle upper bounds for $P\ell_{\hat{f}_N} - P\ell_{f_{\mathcal{F}}}$, where $f_{\mathcal{F}}$ is an arbitrary function in $\mathcal{F}$ (see Section 1.3.4). Moreover, in certain cases it is more straightforward to establish a connection between $P_N \mathcal{L}_f^{\mathcal{F}}$ and the empirical excess risk of the squared loss via route item *1.*, e.g., for the minimum norm interpolant estimator, or for the offset condition in aggregation problems (e.g., [KRV24]). The advantage of using $r_{\mathrm{iso},2}$, i.e., item *2.*, is its broader applicability; in particular, it can smooth out a loss function with poor intrinsic properties or extract second-order information through the measure $\mu$.

Below we explain how $r_{\mathrm{iso}}(\mathcal{G}, \kappa)$ and $r_{\mathrm{iso},2}(\mathcal{G}, \delta, \kappa)$ can be used to improve the uniform convergence argument.

### 1.3.2 Uniform convergence argument based on lower isomorphic fixed point

#### Convex localization: the homogeneity argument

When both $\mathcal{F}$ and $P_N \ell_\bullet$ are convex (see Lemma 1 for definition), the following homogeneity argument (Lemma 1) is the first step to improve the uniform convergence argument in existing literature. When $\mathcal{F}$ and $\ell$ may be non-convex, see, for instance, [KRV24].

In this section we use RERM as an example to introduce convex localization. We recall the definition in Example 8. Let $\lambda \geq 0$ be the tuning parameter, $\Psi : \mathcal{F} \to \mathbb{R}$ be some regularization functional. Define

$$\hat{f}_N \in \operatorname{argmin}\left(L_f((X_i, Y_i)_{i=1}^N) + \lambda\Psi(f) : f \in \mathcal{F}\right), \text{ for some } L_\bullet : f \in \mathcal{F} \mapsto L_f \in \{g : \Omega^N \to \mathbb{R}\}.$$

For any $r > 0$ and any $f_{\mathcal{F}} \in \mathcal{F}$, let $B_{P\mathcal{L}}(f_{\mathcal{F}}; r) = \{f \in \mathcal{F} : P\ell_f - P\ell_{f_{\mathcal{F}}} \leq r\}$ and $S_{P\mathcal{L}}(f_{\mathcal{F}}; r) = \{f \in \mathcal{F} : P\ell_f - P\ell_{f_{\mathcal{F}}} = r\}$. The following Lemma 1 summarizes a series of existing proofs [CLL20, CLL21]. Recall that $P_N \ell_\bullet$ denotes $L_\bullet((X_i, Y_i)_{i=1}^N)$.

**Lemma 1.** *Assume that $\mathcal{F}$ is convex, and $\ell$ is convex in its first argument in the sense that for any $0 \le \alpha \le 1$, any $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega$ and any $f, g \in \mathcal{F}$, there holds $\ell(\alpha f(\boldsymbol{x}) + (1-\alpha) g(\boldsymbol{x}), \boldsymbol{y}) \le \alpha \ell(f(\boldsymbol{x}), \boldsymbol{y}) + (1-\alpha)\ell(g(\boldsymbol{x}), \boldsymbol{y})$. Let $(\hat{f}_N, \mathcal{F})$ be a RERM for solving $(\Omega, \mu, \ell)$ with regularization term $\lambda\Psi(\cdot)$, where $\Psi : \mathcal{F} \to \mathbb{R}$ is assumed to be convex. Let $f_{\mathcal{F}} \in \mathcal{F}$ be any function and $\lambda \in \mathbb{R}$ be any real number. Suppose there exists some $r > 0$ and some convex subset $\mathcal{G}$ with $f_{\mathcal{F}} \in \mathcal{G}$, such that $\mathcal{G} \cap B_{P\mathcal{L}}(f_{\mathcal{F}}; r) \subset \mathcal{F}$, and such that for every*

$$f^\circ \in \partial(B_{P\mathcal{L}}(f_{\mathcal{F}}; r) \cap \mathcal{G}) := (S_{P\mathcal{L}}(f_{\mathcal{F}}; r) \cap \mathcal{G}) \sqcup (B_{P\mathcal{L}}(f_{\mathcal{F}}; r) \cap \partial\mathcal{G}),$$

*where $\partial\mathcal{G}$ is the boundary of $\mathcal{G}$ (see, e.g., [BP12, Section 1.1.3] or [Roc70, Section 6]), there holds $P_N\ell_{f^\circ} - P_N\ell_{f_{\mathcal{F}}} + \lambda(\Psi(f^\circ) - \Psi(f_{\mathcal{F}})) > 0$. There then holds $P\ell_{\hat{f}_N} - P\ell_{f_{\mathcal{F}}} < r$. If we replace $B_{P\mathcal{L}}$ by $B_{L^2(\mu_X)}$, then similarly we have $\|\hat{f}_N - f_{\mathcal{F}}\|_{L^2(\mu_X)} \le r$.*

**Proof.** Since $P_N\ell_f + \lambda\Psi(f)$ is minimized over $f \in \mathcal{F}$ by $\hat{f}_N$, any $f \in \mathcal{F}$ satisfying $P_N\ell_f - P_N\ell_{f_{\mathcal{F}}} + \lambda(\Psi(f) - \Psi(f_{\mathcal{F}})) > 0$ cannot be $\hat{f}_N$. Notice that $\mathcal{F}$ is convex around $f_{\mathcal{F}}$; thus for any $f \in \mathcal{F} \backslash (B_{P\mathcal{L}}(f_{\mathcal{F}}; r) \cap \mathcal{G})$, there exist $f^\circ \in \partial(B_{P\mathcal{L}}(f_{\mathcal{F}}; r) \cap \mathcal{G})$ and $\alpha > 1$ such that $f - f_{\mathcal{F}} = \alpha(f^\circ - f_{\mathcal{F}})$. Moreover, since $P_N\ell_f + \lambda\Psi(f) - (P_N\ell_{f_{\mathcal{F}}} + \lambda\Psi(f_{\mathcal{F}}))$ is convex in $f$, we have $P_N\ell_f + \lambda\Psi(f) - (P_N\ell_{f_{\mathcal{F}}} + \lambda\Psi(f_{\mathcal{F}})) \ge \alpha[(P_N\ell_{f^\circ} + \lambda\Psi(f^\circ) - (P_N\ell_{f_{\mathcal{F}}} + \lambda\Psi(f_{\mathcal{F}}))] > 0$. Therefore, it must hold that $\hat{f}_N \in (B_{P\mathcal{L}}(f_{\mathcal{F}}; r) \cap \mathcal{G})$, hence, $P\ell_{\hat{f}_N} - P\ell_{f_{\mathcal{F}}} < r$. ∎

Lemma 1 transforms an upper bound for $P\ell_{\hat{f}_N} - P\ell_{f_{\mathcal{F}}}$ of RERM into a lower bound for the regularized empirical excess risk $P_N\ell_f - P_N\ell_{f_{\mathcal{F}}} + \lambda(\Psi(f) - \Psi(f_{\mathcal{F}}))$, and this is precisely where the lower isomorphic fixed point $r_{\mathrm{iso}}(\mathcal{G}, \kappa)$ comes into play.

### RERM via $r_{\mathrm{iso}}$

In RERM, the uniform convergence argument mainly combines $r_{\mathrm{iso}}(\mathcal{G}, \kappa)$ with the subgradient of the regularization term and the Bregman divergence, thus allowing $\mathcal{F}$ to be unbounded in $L^2(\mu_X)$. In this section we assume that $\mathcal{F}$ can be identified with a subset of a Banach space. For convenience, we assume here that $\Psi$ is convex (for the case where $\Psi$ is nonconvex, see [LW15]). We call $D_\Psi : (f, g) \in \mathcal{F} \times \mathcal{F} \mapsto D_\Psi(f, g) = \Psi(f) - \Psi(g) - \langle \nabla\Psi(g), f - g\rangle$ the Bregman divergence of $\Psi$, where we recall that $\langle \cdot, \cdot \rangle$ is the dual pairing, [BP12, Section 2.2]. The approach to handling RERM depends primarily on whether $D_\Psi(f, f^*_{\mathcal{F}})$ is non-negative.

**When $\Psi$ has non-trivial Bregman divergence.** We say that $\Psi$ has a non-trivial Bregman divergence around $f^*_{\mathcal{F}}$ if $\Psi$ satisfies $D_\Psi(\cdot, f^*_{\mathcal{F}})$ is convex and non-negative, or $D_\Psi(\cdot, f^*_{\mathcal{F}})$ can be bounded from below by a non-negative convex function; in that case we abuse notation by denoting this non-negative convex function also as $D_\Psi(\cdot, f^*_{\mathcal{F}})$.

**Example 12.** *The following Bregman divergences are non-trivial. The proof of item 3. is provided in Lemma 26 in Chapter 4.*

1. *When $\mathcal{F}$ is identified by the Euclidean space $(\mathbb{R}^p, \|\cdot\|_2)$ for some $p \in \mathbb{N}_+$, and $\Psi : \boldsymbol{v} \in \mathbb{R}^p \mapsto \|\boldsymbol{v}\|_2^2$, then $D_\Psi(\boldsymbol{v}_1, \boldsymbol{v}_2) = \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2^2$.*
2. *When $\mathcal{F}$ is identified by $(\mathcal{P}(\Theta), \|\cdot\|_{L^1(\Theta)})$, where $\mathcal{P}(\Theta)$ is the manifold of probability measure on $\Theta$ which has density with respect to the Lebesgue measure $d\boldsymbol{\theta}$, see Example 7, and $\Psi : \nu \in \mathcal{P}(\Theta) \mapsto \mathrm{Ent}^-(\nu) = \int_\Theta \frac{d\nu}{d\boldsymbol{\theta}} \log(\frac{d\nu}{d\boldsymbol{\theta}}) d\boldsymbol{\theta}$, then $D_\Psi(\nu_1, \nu_2) = \mathrm{KL}(\nu_1 \| \nu_2)$ and $\mathrm{KL}(\nu_1 \| \nu_2)$ is the Kullback-Leibler divergence between $\nu_1$ and $\nu_2$.*
3. *Let $1 < q < \infty$ be a real number, and $\Psi : \boldsymbol{v} \in \mathbb{R}^p \mapsto \|\boldsymbol{v}\|_q^q$. Then there exists an absolute constant $c = \frac{q-1}{q2^q}$ such that $D_\Psi(\boldsymbol{v}_1, \boldsymbol{v}_2) \ge c\,\alpha_q(|\boldsymbol{v}_2|, \boldsymbol{v}_1 - \boldsymbol{v}_2)$, where $\alpha_q$ is a real-valued function acting coordinate-wise, defined by*

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

*Abusing notation, we redefine $D_\Psi(\cdot, \cdot)$ as $c\alpha_q(\cdot, \cdot)$. It is easy to check that $D_\Psi(\cdot, f^*_{\mathcal{F}}) = c\alpha_q(|f^*_{\mathcal{F}}|, \cdot - f^*_{\mathcal{F}})$ is a convex function for any $1 < q < \infty$.*

For any $\rho > 0$, define $B_\Psi(f^*_{\mathcal{F}}; \rho) = \{f \in \mathcal{F} : D_\Psi(f, f^*_{\mathcal{F}}) \le \rho\}$ and $S_\Psi(f^*_{\mathcal{F}}; \rho) = \{f \in \mathcal{F} : D_\Psi(f, f^*_{\mathcal{F}}) = \rho\}$. If $\Psi$ has a non-trivial Bregman divergence, then $B_\Psi(f^*_{\mathcal{F}}; \rho)$ is a (non-empty) convex set and $S_\Psi(f^*_{\mathcal{F}}; \rho)$ is its boundary, [Roc70].

Depending on whether theorists have the freedom to choose $\lambda$ or not, we present two refinements of the uniform convergence argument below in Theorem 1. To the best of our knowledge, the following result is novel.

**Theorem 1.** *Let $(\Omega, \mu, \ell)$ be a supervised learning problem and $\mathcal{F}$ be a subset of a Banach space. Suppose $\Psi$ has a non-trivial Bregman divergence. Let $\kappa > 0$ be some real number. Recall $\square$, $\triangle$ and $r_{\mathrm{iso}}$ from Definition 1, Definition 2 and Definition 3. Let $\triangle > 4\square$. For any $\rho > 0$, denote $r_{\mathrm{iso}}(\rho)$ by $r_{\mathrm{iso}}(B_\Psi(f^*_{\mathcal{F}}; \rho), \kappa)$. For any $\lambda > 0$, let $r_*$ and $\rho_*$ be the smallest $r$ and its corresponding $\rho$ such that the following system of inequalities on $(\rho, r)$ holds simultaneously:*

$$\begin{cases} r \geq r_{\mathrm{iso}}(\rho), \\ 3\square r^{\frac{2}{\kappa}} > \lambda |||\nabla\Psi(f^*_{\mathcal{F}})|||_{(r,\rho)}, \text{ and} \\ \rho \geq \dfrac{1}{\lambda}\triangle r^{\frac{2}{\kappa}}, \end{cases} \tag{1.7}$$

*where*

$$|||\nabla\Psi(f^*_{\mathcal{F}})|||_{(r,\rho)} = \sup\left(\langle \nabla\Psi(f^*_{\mathcal{F}}), f - f^*_{\mathcal{F}}\rangle : f \in B_\Psi(f^*_{\mathcal{F}}; \rho) \cap B_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r)\right).$$

*Let $(\hat{f}_N, \mathcal{F})$ be a RERM with regularization term $\lambda\Psi(\cdot)$. Then with probability at least $1 - \delta_{\mathrm{iso}}(B_\Psi(f^*_{\mathcal{F}}; \rho_*), \kappa)$, there hold $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)} \leq r_*^2$ and $D_\Psi(\hat{f}_N, f^*_{\mathcal{F}}) \leq \rho_*$.*

*Specifically, if there exists $\rho_* = \mathrm{argmin}\{\rho > 0 : \rho > \frac{\triangle}{3\square}|||\nabla\Psi(f^*_{\mathcal{F}})|||_{(r_{\mathrm{iso}}(\rho),\rho)}\}$, and we set $\lambda_* = \frac{\triangle}{\rho_*} r_{\mathrm{iso}}(\rho_*)^{\frac{2}{\kappa}} = \frac{3\square}{|||\nabla\Psi(f^*_{\mathcal{F}})|||_{(r_{\mathrm{iso}}(\rho_*),\rho_*)}} r_{\mathrm{iso}}(\rho_*)^{\frac{2}{\kappa}}$, then $\rho_*$ and $r_* = r_{\mathrm{iso}}(\rho_*)$ constitute a solution to the aforementioned system of inequalities.*

**Proof.** We work on the intersection of the random events provided in (1.4) and (1.5). Let $\mathcal{G} = B_\Psi(f^*_{\mathcal{F}}; \rho_*)$ in Lemma 1 and with $f_{\mathcal{F}} = f^*_{\mathcal{F}}$, it suffices to prove that for every

$$f \in \left(S_\Psi(f^*_{\mathcal{F}}; \rho_*) \cap B_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r_*)\right) \sqcup \left(B_\Psi(f^*_{\mathcal{F}}; \rho_*) \cap S_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r_*)\right),$$

we have $P_N\ell_f + \lambda\Psi(f) - \left(P_N\ell_{f^*_{\mathcal{F}}} + \lambda\Psi(f^*_{\mathcal{F}})\right) > 0$.

By the definition of Bregman divergence, for any $f \in \mathcal{G} \cap B_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r_*)$, there holds

$$\lambda(\Psi(f) - \Psi(f^*_{\mathcal{F}})) = \lambda\langle\nabla\Psi(f^*_{\mathcal{F}}), f - f^*_{\mathcal{F}}\rangle + \lambda D_\Psi(f, f^*_{\mathcal{F}}) \geq \lambda D_\Psi(f, f^*_{\mathcal{F}}) - \lambda|||\nabla\Psi(f^*_{\mathcal{F}})|||_{(r_*,\rho_*)}, \tag{1.8}$$

1. When $f \in B_\Psi(f^*_{\mathcal{F}}; \rho_*) \cap S_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r(\rho_*))$. By (1.8), $\lambda(\Psi(f) - \Psi(f^*_{\mathcal{F}})) \geq -\lambda|||\nabla\Psi(f^*_{\mathcal{F}})|||_{(\rho_*)}$. By (1.5), for any such $f$,
$$P_N\mathcal{L}^{\mathcal{F}}_f + \lambda(\Psi(f) - \Psi(f^*_{\mathcal{F}})) \geq \triangle r_*^{\frac{2}{\kappa}} - \square r_*^{\frac{2}{\kappa}} - \lambda|||\nabla\Psi(f^*_{\mathcal{F}})|||_{(r_*,\rho_*)}.$$
2. When $f \in S_\Psi(f^*_{\mathcal{F}}; \rho_*) \cap B_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r(\rho_*))$. By (1.8), there holds $\lambda(\Psi(f) - \Psi(f^*_{\mathcal{F}})) \geq \lambda\rho_* - \lambda|||\nabla\Psi(f^*_{\mathcal{F}})|||_{(\rho_*)}$. By (1.4), we have $P_N\mathcal{L}^{\mathcal{F}}_f \geq -\square r_*^{\frac{2}{\kappa}}$. Combining with $\rho_* \geq \frac{1}{\lambda}\triangle r_*^{\frac{2}{\kappa}}$, for any such $f$,
$$P_N\mathcal{L}^{\mathcal{F}}_f + \lambda(\Psi(f) - \Psi(f^*_{\mathcal{F}})) \geq \triangle r_*^{\frac{2}{\kappa}} - \square r_*^{\frac{2}{\kappa}} - \lambda|||\nabla\Psi(f^*_{\mathcal{F}})|||_{(r_*,\rho_*)}.$$

Finally, from the definition of $\rho_*$, we have $P_N\mathcal{L}^{\mathcal{F}}_f + \lambda(\Psi(f) - \Psi(f^*_{\mathcal{F}})) > 0$ in both cases. ∎

**When $\Psi$ is a norm.** In this case, one must resort to the sparsity equation, [LM18].

**Definition 5.** *Let $\Psi(\cdot)$ be some norm $\|\cdot\|$. Let $\partial^-\|\cdot\|(f^*_{\mathcal{F}})$ be the sub-differential of $\|\cdot\|$ evaluated at $f^*_{\mathcal{F}}$. For any $\rho, r > 0$, when the following equation is satisfied, we say that $\|\cdot\|$ satisfies the sparsity-equation at scale $(\rho, r)$.*

$$\inf\left(\sup\left(\langle \boldsymbol{g}, f - f^*_{\mathcal{F}}\rangle : \boldsymbol{g} \in \partial^-\|\cdot\|(f^*_{\mathcal{F}})\right) : f \in B_{\|\cdot\|}(f^*_{\mathcal{F}}; \rho) \cap B_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r)\right) \geq \frac{4}{5}\rho,$$

*where $B_{\|\cdot\|}(f^*_{\mathcal{F}}; \rho)$ by $\{f \in \mathcal{F} : \|f - f^*_{\mathcal{F}}\| \leq \rho\}$.*

When the sparsity equation holds, it suffices to replace the term $\lambda D_\Psi(f, f^*_{\mathcal{F}})$ provided by the Bregman divergence in item 2 of Theorem 1 with $\lambda\rho$; we omit further details here and refer to [LM17, LM18].

Before concluding this section, we emphasize that in fact, when $\mathcal{F}$ and $\ell$ are convex and $\hat{f}_N$ is a RERM, obtaining the estimation error of $\hat{f}_N$ only requires the homogeneity argument provided by Lemma 1, and the various fixed points can be combined in different ways. For instance, one could also use only the multiplier fixed point $r_{\mathrm{M}}(\mathcal{G}, \delta_{\mathrm{M}})$ together with the one-sided version of $r_{\mathrm{iso},2}$ from Definition 4, namely $r_{\mathrm{iso},-}(\mathcal{G}, \delta, \kappa) = \min\big(r > 0 : \mathbb{P}\big(\sup\left((P - P_N)\mathcal{L}^{\mathcal{F}}_f : f \in \mathcal{G} \cap B_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r)\right) \leq \kappa_{\mathrm{iso},-} r^{\frac{2}{\kappa}}\big) \geq 1 - \delta\big)$, where $\kappa_{\mathrm{iso},-}$ is an absolute constant. This would still yield conclusions analogous to (1.4) and (1.5), which can then be used to bound the estimation error.

**Computation of multiplier fixed point**

The computation of the multiplier fixed point $r_{\mathrm{M}}(\mathcal{G}, \delta_{\mathrm{M}}, \kappa, \square)$ is usually relatively straightforward; it requires an upper bound for the multiplier process. For the sake of simplicity, we assume that the sub-differential of $P_N \ell_{f_{\mathcal{F}}^*}$ is trivial, i.e., $\partial^- P_N \ell_{f_{\mathcal{F}}^*} = \nabla P_N \ell_{f_{\mathcal{F}}^*}$. Here, we refer to the following stochastic process as the multiplier process:

$$\sup\left(\left|\left\langle \nabla P_N \ell_{f_{\mathcal{F}}^*}, f - f_{\mathcal{F}}^* \right\rangle\right| : f \in \mathcal{G} \cap B_{L^2(\mu_X)}(f_{\mathcal{F}}^*; r_{\mathrm{iso}}(\mathcal{G}, \kappa))\right). \tag{1.9}$$

Consequently, we can employ various existing tools for multiplier processes, c.f., [Men16, Men17, HW19]. More specifically, if $P_N \ell_{f_{\mathcal{F}}^*} = \frac{1}{N}\sum_{i=1}^N \ell(f_{\mathcal{F}}^*(X_i) - Y_i)$ for some $\ell : \mathbb{R} \to \mathbb{R}$, then

$$(1.9) = \sup\left(\left|\frac{1}{N}\sum_{i=1}^N \ell'(Y_i - f_{\mathcal{F}}^*(X_i))(f - f_{\mathcal{F}}^*)(X_i)\right| : f \in \mathcal{G} \cap B_{L^2(\mu_X)}(f_{\mathcal{F}}^*; r_{\mathrm{iso}}(\mathcal{G}, \kappa))\right).$$

When $P_N \ell_{f_{\mathcal{F}}^*}$ is not decomposable, the situation is slightly more complicated, as illustrated by the following example.

**Example 13.** *Let $L_f : \boldsymbol{x} \in \mathbb{R}^N \mapsto \|\boldsymbol{x}\|_q$, where $1 \le q \le \infty$. Let $(\mu_X, f_{\mathcal{F}}^*, \xi)$ be an additive regression problem and let $\mathcal{F} = \{\langle \cdot, \boldsymbol{v}\rangle : \boldsymbol{v} \in \mathbb{R}^p\}$. Then $\nabla P_N \ell_{f_{\mathcal{F}}^*} = \nabla L_f(\boldsymbol{\xi})$, where $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_N)$.*

In this case the multipliers are correlated, yet independent of the $X_i$'s; then the following upper bound for the multiplier process can be used. Its proof may be found in Section 4.9.1 in Chapter 4, see also [P1].

**Lemma 2.** *Let $F \subset L^2(\mu_X)$ be a functions class with sub-Gaussian increments with respect to $\|\cdot\|_{L^2(\mu_X)}$, that is, there exists an absolute constant $\theta > 1$ such that for any $f, g \in F$, $\|f - g\|_{\psi_2} \le \theta \|f - g\|_{L^2(\mu_X)}$. Let $\boldsymbol{w} = (w_i)_{i=1}^N \in \mathbb{R}^N$ be a deterministic vector. Let $X_1, \cdots, X_N$ be i.i.d. random vectors distributed as $\mu_X$. Suppose $\mathbf{0} \in F$. Then there exists an absolute constant $C$ depending only on $\theta$ such that for any $t > 0$, with probability at least $1 - 2\exp(-t^2)$,*

$$\sup\left(\left|\sum_{i=1}^N w_i \left(f(X_i) - \mathbb{E}[f(X)]\right)\right| : f \in F\right) \le C\|\boldsymbol{w}\|_2 \left(\gamma_2(F, d_{L^2(\mu_X)}) + t\,\mathrm{diam}(F, \|\cdot\|_{L^2(\mu_X)})\right),$$

*where $\gamma_2(F, d_{L^2(\mu_X)})$ is the Talagrand's $\gamma_2$-functional of $F$ with respect to the distance generated by $\|\cdot\|_{L^2(\mu_X)}$ while $\mathrm{diam}(F, \|\cdot\|_{L^2(\mu_X)}) = 2\sup(\|f\|_{L^2(\mu_X)} : f \in F)$.*

In Lemma 2, identifying $\boldsymbol{w}$ as $\nabla P_N \ell_{f_{\mathcal{F}}^*}$ yields an upper bound for this multiplier process.

**Corollary 1.** *Let $(\mu_X, f^*, \xi)$ be any real-valued regression problem (Example 1), $f^* \in \mathcal{F} \subset L^2(\mu_X)$ be a subset of a Banach space containing $\mathbf{0}$ and has sub-Gaussian increments with respect to $\|\cdot\|_{L^2(\mu_X)}$. Assume that $\mathbb{E}[f(X)] = 0$ for any $f \in \mathcal{F}$. For any localization subset $\mathcal{G}$, let $\delta_1 = 2\exp(-d_*(\mathcal{G}))$, where $d_*(\mathcal{G}) = (\gamma_2(\mathcal{G}, \|\cdot\|_{L^2(\mu_X)})/\,\mathrm{diam}(\mathcal{G}, \|\cdot\|_{L^2(\mu_X)}))^2$. Suppose $\partial^- P_N \ell_{f_{\mathcal{F}}^*} = \nabla P_N \ell_{f_{\mathcal{F}}^*}$ is independent with $(X_i)_{i=1}^N$, and suppose that there exist some $w > 0$ and $0 < \delta_2 < 1$ such that $\mathbb{P}(\|\nabla P_N \ell_{f_{\mathcal{F}}^*}\|_2 \le w\sqrt{N}) \ge 1 - \delta_2$. Let $\delta_{\mathrm{M}} = \delta_1 + \delta_2$. There then exists an absolute constant $c$ depending on the constant $C$ in Lemma 2 such that*

$$r_{\mathrm{M}}(\mathcal{G}, \delta_{\mathrm{M}}, \kappa, \square) \le \min\left(r > 0 : w\gamma_2\left(\mathcal{G} \cap B_{L^2(\mu_X)}(f_{\mathcal{F}}^*; r), \|\cdot\|_{L^2(\mu_X)}\right) \le c\square r^{\frac{2}{\kappa}}\sqrt{N}\right).$$

The proof of Corollary 1 follows readily and is omitted here.

### 1.3.3 Uniform convergence argument based on two-sided isomorphic fixed point

**The Bernstein's condition**

The Bernstein condition provides an upper bound for $\|\hat{f}_N - f_{\mathcal{F}}^*\|_{L^2(\mu_X)}$ or $P\mathcal{L}_{\hat{f}_N}^{(0,1)} = P\ell_{\hat{f}_N}^{(0,1)} - P\ell_{f_{\mathcal{F}}^*}^{(0,1)}$ in terms of $P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}}$. We define the following condition:

**Definition 6** (Bernstein's condition and local Bernstein's condition)**.** *Denote $\mathcal{L}_f^{(2)}$ by $\|f - f_{\mathcal{F}}^*\|_{L^2(\mu_X)}^2$ and $\mathcal{L}_f^{(0,1)}$ by $\ell_f^{(0,1)} - \ell_{f_{\mathcal{F}}^*}^{(0,1)}$ respectively. If there exist absolute constants $c, \kappa > 0$ such that the following holds:*

1. *for every $f \in \mathcal{F}$, $P\mathcal{L}_f^{(2)} \le c(P\mathcal{L}_f^{\mathcal{F}})^\kappa$ in a regression problem, or $P\mathcal{L}_f^{(0,1)} \le c(P\mathcal{L}_f^{\mathcal{F}})^\kappa$ in a classification problem;*

2. *for every* $f \in \mathcal{F} \cap S_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r)$, $P\mathcal{L}_f^{(2)} \le c(P\mathcal{L}_f^{\mathcal{F}})^\kappa$ *in a regression problem, or* $P\mathcal{L}_f^{(0,1)} \le c(P\mathcal{L}_f^{\mathcal{F}})^\kappa$ *in a classification problem, where* $S_{L^2(\mu_X)}(f^*_{\mathcal{F}}; r) = \{f \in L^2(\mu_X) : \|f - f^*_{\mathcal{F}}\|_{L^2(\mu_X)} = r\}$ *and* $r > 0$ *is a parameter.*

*Then we say that* $(\mathcal{F}, \mu, \ell)$ *satisfies the Bernstein condition with parameters* $(c, \kappa)$*; or that it satisfies a local Bernstein condition with parameters* $(c, \kappa)$ *at scale* $r$*, respectively.*

The following is a trivial example, see, for instance, [Men19].

**Example 14.** *Let* $\Omega_Y = \mathbb{R}$*,* $\ell$ *be the squared loss, and* $\mathcal{F}$ *be any statistical model. Suppose* $Y = f^*_{\mathcal{F}}(X) + \xi$ *satisfies* $\mathbb{E}[\xi(f - f^*_{\mathcal{F}})(X)] \ge 0$ *for any* $f \in \mathcal{F}$*, then* $(\mathcal{F}, \mu, \ell)$ *satisfies the Bernstein's condition with parameters* $(1, 1)$*. The condition* $\mathbb{E}[\xi(f - f^*_{\mathcal{F}})(X)] \ge 0$ *is satisfied, for instance, when either* $\xi$ *is independent with* $X$ *and is centered; or when* $\mathcal{F}$ *is convex.*

The Bernstein condition describes the positional relationship between the supervised learning problem $(\Omega, \mu, \ell)$ and the statistical model $\mathcal{F}$, see [Lec11]. To our knowledge, the Bernstein condition was originally introduced by [BJM03] and has been progressively refined and verified in a series of subsequent works, c.f., [BJM06, BM06, ACL19, CCLN21, LN24]. Among these, the Bernstein condition for the 0/1 loss is often verified via Zhang-type inequalities, [Zha04]. The Bernstein condition heavily relies on the definition of $f^*_{\mathcal{F}} \in \arg\min(P\ell_f : f \in \mathcal{F})$. When we replace $f^*_{\mathcal{F}}$ in the Bernstein condition with $f^*$, i.e., the Bayes rule, the Bernstein condition reduces to Tsybakov's margin condition [MT99, Tsy03], which captures the intrinsic difficulty of the supervised learning problem $(\Omega, \mu, \ell)$, independent of the choice of statistical model.

**When $P\ell_\bullet$ and $\mathcal{F}$ are convex.** The proof of Bernstein's condition generally relies on an extra assumption that $P\ell_\bullet$ and $\mathcal{F}$ are convex – though it is only sufficient but not necessary (see the following Proposition 3 for an example of verifying a local Bernstein's condition when both $\mathcal{F}$ and $\ell$ are nonconvex, and the Example 14 above for a case where $\ell$ is convex but $\mathcal{F}$ is not necessarily convex). In fact, it is sufficient to studying a lower bound on the smallest eigenvalue of the Hessian matrix of $P\ell_\bullet$ around $f^*_{\mathcal{F}}$. That is, if $\mathcal{F}$ is convex, $P\ell_\bullet$ is convex and if $f^*_{\mathcal{F}} \in \operatorname{argmin}(P\ell_f : f \in \mathcal{F})$, then for any $\boldsymbol{g} \in \partial^-(P\ell_\bullet)(f^*_{\mathcal{F}})$, there holds $\langle \boldsymbol{g}, f - f^*_{\mathcal{F}} \rangle \ge 0$ for any $f \in \mathcal{F}$, hence the second order Taylor expansion (if exists) gives

$$P\mathcal{L}_f^{\mathcal{F}} \ge \int_0^1 (1-t)\langle f - f^*_{\mathcal{F}}, \nabla^2(P\ell_\bullet)(f^*_{\mathcal{F}} + t(f - f^*_{\mathcal{F}}))(f - f^*_{\mathcal{F}})\rangle_{L^2(\mu_X)} dt.$$

Therefore, if the Hessian $\nabla^2(P\ell_\bullet) : L^2(\mu_X) \to L^2(\mu_X)$ of $P\ell_\bullet : L^2(\mu_X) \to \mathbb{R}$ is positive definite in a small neighborhood of $f^*_{\mathcal{F}}$, say, there exists an interval $T \subset [0, 1]$ such that $\int_T dt > 0$ on which for any $t \in T$, and any $f \in \mathcal{F}$ (or $f \in \mathcal{F} \cap S_{L^2(\mu_X)}(f^*; r)$), there holds $\nabla^2(P\ell_\bullet)(f^*_{\mathcal{F}} + t(f - f^*_{\mathcal{F}})) \succeq cI_{L^2(\mu_X)}$ for some $c > 0$, where $I_{L^2(\mu_X)}$ is the identity on $L^2(\mu_X)$, then there exists some $c'$ depending only on $c$ such that $(\mathcal{F}, \mu, \ell)$ satisfies the (local) Bernstein condition with parameters $(c', 1)$, see, for instance [ACL19].

**The Bernstein condition depends on $\mu$.** We need to emphasize that the Bernstein condition depends on $\mu$. Hence, even for certain special loss functions that possess second derivatives only in a distributional sense, we can still verify the Bernstein condition because we are examining $\nabla^2(P\ell_\bullet)$, and the probability measure $\mu$ smooths out this function. To the best of our knowledge, the following Proposition 2 and Proposition 3 are novel.

**Proposition 2.** *In a real-valued regression problem* $Y = f^*_{\mathcal{F}}(X) + \xi$ *where* $X$ *is independent with* $\xi$*, let* $\boldsymbol{y} = (Y_1, \ldots, Y_N)$*,* $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_N)$*, and* $\mathbb{X} : f \in \mathcal{F} \mapsto (f(X_i))_{i=1}^N \in \mathbb{R}^N$*. Suppose there exists an even real-valued function* $L \in L^1_{\mathrm{loc}}(\mathbb{R}^N)$ *such that* $P_N\ell_f = L(\boldsymbol{y} - \mathbb{X}f)$*, and suppose that* $\boldsymbol{\xi}$ *has a probability density function* $\varphi$ *with respect to the Lebesgue measure on* $\mathbb{R}^N$*, and that* $\varphi \in C_0^\infty(\mathbb{R}^N)$*, then* $P\ell_\bullet : f \in \mathcal{F} \mapsto \mathbb{R}$ *is* $C^\infty$*.*

**Proof.** Let $g : \boldsymbol{x} \in \mathbb{R}^N \mapsto (L * \varphi)(\boldsymbol{x}) = \int_{\mathbb{R}^N} L(\boldsymbol{x} - \boldsymbol{y})\varphi(\boldsymbol{y})d\boldsymbol{y}$, where $\varphi$ is the probability density function of $\boldsymbol{\xi}$ with respect to the Lebesgue measure, and $*$ denotes convolution. Then $g(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\xi}} L(\boldsymbol{x} - \boldsymbol{\xi})$. Since $\phi \in C_0^\infty(\mathbb{R}^N)$ and $L \in L^1_{\mathrm{loc}}(\mathbb{R}^N)$, there holds $g \in C^\infty(\mathbb{R}^N)$. By Fubini's theorem, $P\ell_f = \mathbb{E}L(\boldsymbol{y} - \mathbb{X}f) = \mathbb{E}L(-\boldsymbol{\xi} + \mathbb{X}(f^*_{\mathcal{F}} - f)) = \mathbb{E}g(\mathbb{X}(f^*_{\mathcal{F}} - f))$, indicating that $P\ell_\bullet : f \in \mathcal{F} \mapsto \mathbb{R}$ is $C^\infty$. ∎

A special case of Proposition 2 is when $L(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ or $L(\boldsymbol{x}) = \|\boldsymbol{x}\|_\infty$, where $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are the $\ell_1$ and $\ell_\infty$ norms on $\mathbb{R}^N$, respectively.

**Proposition 3.** *Under the assumptions of Proposition 2, if we further assume that for any* $f \in \mathcal{F}$*,* $\mathbb{X}f$ *has the same distribution as* $\|f\|_{L^2(\mu_X)}G$*, where* $G$ *is a standard Gaussian random vector in* $\mathbb{R}^N$*. Suppose* $\xi$ *is a Gaussian random*

*variable independent of $X$ with variance $\sigma_\xi^2$, and $L$ is $\alpha$-homogeneous, i.e., for any $t \in \mathbb{R}$, $L(t\boldsymbol{x}) = t^\alpha L(\boldsymbol{x})$, then there exists some absolute constant $c_\alpha$ depending only on $\alpha$ such that for any $r < \sigma_\xi$, $(\mathcal{F}, \mu, \ell)$ satisfies the local Bernstein's condition with parameters $(c_\alpha \mathbb{E}[L(\sigma_\xi^{\frac{\alpha-2}{\alpha}} G)], 1)$ at scale $r$.*

**Proof.** For any $f \in \mathcal{F} \cap S_{L^2(\mu_X)}(f_{\mathcal{F}}^*; r)$, we have $\mathbb{X}(f_{\mathcal{F}}^* - f)$ has the same distribution as $rG$, where $G \sim \mathcal{N}(\mathbf{0}, I_N)$ is a standard Gaussian random vector on $\mathbb{R}^N$. Therefore, there exists some absolute constant $c_\alpha$ such that, for any such $f$, there holds

$$P\mathcal{L}_f^{\mathcal{F}} = \mathbb{E}\left[L\left(\sqrt{r^2 + \sigma_\xi^2} G\right)\right] - \mathbb{E}[L(\sigma_\xi G)] = \mathbb{E}[L(\sigma_\xi G)] \left(\left(1 + \frac{r^2}{\sigma_\xi^2}\right)^{\frac{\alpha}{2}} - 1\right)$$
$$\geq c_\alpha r^2 \frac{\mathbb{E}[L(\sigma_\xi G)]}{\sigma_\xi^2} = c_\alpha \|f - f_{\mathcal{F}}^*\|_{L^2(\mu_X)}^2 \mathbb{E}[L(\sigma_\xi^{\frac{\alpha-2}{\alpha}} G)].$$

Therefore, $(\mathcal{F}, \mu, \ell)$ satisfies the local Bernstein's condition with parameters $(c_\alpha \mathbb{E}[L(\sigma_\xi^{\frac{\alpha-2}{\alpha}} G)], 1)$ at scale $r$. ∎

Note that in Proposition 3, we do not assume that $L$ is convex (the properties of $L$ can even be very poor), nor do we assume that $\mathcal{F}$ is a convex set. A special case is $L(\boldsymbol{x}) = \|\boldsymbol{x}\|_q^q$, where $0 < q < 1$ and $\|\cdot\|_q$ is the $\ell_q$ pseudo-norm. Let $K \subset \mathbb{R}^p$ be an arbitrary set (not necessarily convex) such that $\mathcal{F} = \{f_{\boldsymbol{v}}(\cdot) = \langle \cdot, \boldsymbol{v} \rangle : \boldsymbol{v} \in K\}$, and let $X$ be a Gaussian random vector in $\mathbb{R}^p$ with an arbitrary covariance matrix. The validity of Proposition 3 stems from the special nature of $\mu$. Therefore, Proposition 3 emphasizes that the Bernstein condition depends on $\mu$.

Due to space constraints, we do not provide the derivation of the upper bound on the estimation error for RERM via $r_{\mathrm{iso},2}$ here—it is entirely consistent with Theorem 1.

### 1.3.4 General target functions under squared loss

When $\ell$ is the squared loss, for a general real-valued regression problem $Y = f^*(X) + \xi$, we have the freedom to establish oracle inequalities with respect to an arbitrary target function in $\mathcal{F}$. Consider an arbitrarily chosen $f_{\mathcal{F}} \in \mathcal{F}$; in this section, we aim to derive an upper bound for $\|\hat{f}_N - f_{\mathcal{F}}\|_{L^2(\mu_X)}^2$. Note that here $f_{\mathcal{F}}$ is not necessarily $f_{\mathcal{F}}^* \in \arg\min(P\ell_f : f \in \mathcal{F})$. This situation arises, for instance, in mean-field shallow neural networks trained by Wasserstein gradient flow (Example 7 and Example 15). In this example, when $f^*$ can be approximated by a mean-field shallow neural network that is sparse in a certain sense (see [P5]), the complexity term in the residual of the oracle inequality often becomes infinite for such sparse approximations, rendering the obtained oracle inequality meaningless. We can transform the problem using the following lemma.

**Lemma 3.** *Let $(\mu_X, f^*, \xi)$ be A real-valued regression problem (Example 1), and let $(\hat{f}_N, \mathcal{F})$ be a solution to it. Let $f_{\mathcal{F}} \in \mathcal{F}$ be any function.*

1. *Define $\xi' = f^*(X) - f_{\mathcal{F}}(X)$ and $\zeta = \xi + \xi'$. Then the supervised regression problem $(\mu_X, f^*, \xi)$ is equivalent to the scalar-valued regression problem $(\mu_X, f_{\mathcal{F}}, \zeta)$ in the sense that, the probability distribution of the pair of input, response is the same.*
2. *For any $f_{\mathcal{F}} \in \mathcal{F}$ with $\|f_{\mathcal{F}} - f^*\|_{L^2(\mu_X)} < \infty$ and any $f \in \mathcal{F}$, there holds*

$$P_N \ell_f - P_N \ell_{f_{\mathcal{F}}} \geq P_N(f - f_{\mathcal{F}})^2 - 2\left|P_N \xi (f - f_{\mathcal{F}})\right|$$
$$- 2\left|P_N(f^* - f_{\mathcal{F}})(f - f_{\mathcal{F}}) - P(f^* - f_{\mathcal{F}})(f - f_{\mathcal{F}})\right| - 2r\|f_{\mathcal{F}} - f^*\|_{L^2(\mu_X)}.$$

**Proof.**

1. Conditioned on $X$, the response of the real-valued regression problem is $Y = f^*(X) + \xi$; while the response of the scalar-valued regression problem is $Y' = f_{\mathcal{F}}(X) + \zeta = f^*(X) + \xi$. They are identically distributed. Moreover, the input $X$ has the same probability distribution in both cases. This completes the proof.
2. Notice that for any real numbers $a, b$, we have $a^2 - b^2 = (a-b)^2 + 2b(a-b)$. Applying $a, b$ to $\ell_f(X_i, Y_i)$ and $\ell_{f_{\mathcal{F}}}(X_i, Y_i)$, respectively, and summing over $i = 1$ to $N$, we obtain $P_N \ell_f - P_N \ell_{f_{\mathcal{F}}} = P_N(f - f_{\mathcal{F}})^2 + 2P_N \zeta (f - f_{\mathcal{F}})$. From the definition of $\zeta$ and the fact that $\zeta$ is a centered random variable independent of $X$, the conclusion follows directly via the Cauchy–Schwarz inequality.

∎

Combining Lemma 3 with Lemma 1 yields an upper bound for $\|\hat{f}_N - f_{\mathcal{F}}\|_{L^2(\mu_X)}$. We do not repeat here.

## 1.4 The Future of Statistical Learning Theory: When Computation Comes into Play

Research on complex statistical models, exemplified by neural networks, has a long history, c.f. [Ros62]. However, constrained by computational power, they were once difficult to apply in practice. In recent years, leaps in computational capabilities have finally brought such complex statistical models back into the spotlight of the machine learning field, where they have achieved unexpectedly strong advantages in industrial applications [GBC16]. This glimpse illustrates a trend in the development of mathematical statistics and statistical learning theory—computational properties have become a consideration equally important as statistical properties. From the 0/1 loss function to convex surrogates yielding support vector machines [BBL05], from minimum $\ell_0$ interpolant estimators to basis pursuit [FR13], from infinite-dimensional nonparametric statistics to finite-dimensional statistical learning theory [Lec11], from minimax lower bounds to computational lower bounds [Wei25]—this belief has been verified time and again over decades.

Below we introduce some problems in mathematical statistics that emerge when computational feasibility is taken into consideration.

### 1.4.1 Theory associated to Optimization Algorithms: Beyond Mathematical Definitions

Given a supervised learning problem $(\Omega, \mu, \ell)$ and one of its solutions $(\{\hat{f}_N\}_{N\in\mathbb{N}_+}, \mathcal{F})$, the examination of its statistical properties typically relies on establishing oracle inequalities. From the perspective of computational properties, although statisticians intend to construct the estimator $\hat{f}_N$ according to decision rules and to the training samples $(X_i, Y_i)_{i=1}^N$, constrained by computational limitations, the estimator actually obtained in practice via a computer is usually a different one, which we denote as $\tilde{f}_N$. For example, using gradient descent with a constant step size (say, $\eta > 0$) to train a ERM on the set of deep neural networks, initialized with parameters $(\mathbb{W}_1(0), \cdots, \mathbb{W}_L(0))$, see Example 6. Here, $\tilde{f}_N$ is defined as $f_t := f_{\mathbb{W}_1(t),\cdots,\mathbb{W}_L(t)}$, where

$$\mathbb{W}_\ell(t) = \mathbb{W}_\ell(t-1) - \eta \nabla_{\mathbb{W}_\ell} P_N \ell_{f_t}, \quad \forall 1 \le \ell \le L.$$

We refer to $\tilde{f}_N$ as the algorithm. Consequently, there is often a discrepancy between the theoretically defined $\hat{f}_N$ and the practically computed $\tilde{f}_N$, and this gap can sometimes be substantial enough that their statistical properties may differ significantly. Therefore, if we incorporate into the statistical properties the discrepancy between $\hat{f}_N$ and $\tilde{f}_N$ arising from computational considerations, then in modern statistical learning theory, what we are actually interested in is the following "decomposition", c.f., [Bac24, Section 5.1]:

$$P\ell_{\tilde{f}_N} - P\ell_{f^*} = (P\ell_{\tilde{f}_N} - P\ell_{\hat{f}_N}) + (P\ell_{\hat{f}_N} - P\ell_{f^*_{\mathcal{F}}}) + (P\ell_{f^*_{\mathcal{F}}} - P\ell_{f^*}),$$

Here, $P\ell_{f^*_{\mathcal{F}}} - P\ell_{f^*}$ is the approximation error of the statistical model $\mathcal{F}$; $P\ell_{\hat{f}_N} - P\ell_{f^*_{\mathcal{F}}}$ is the estimation error of the "theoretical" estimator; and $P\ell_{\tilde{f}_N} - P\ell_{\hat{f}_N}$ is the difference in population risk between the output of the "practical" algorithm and that of the theoretical estimator.

This discrepancy between $\hat{f}_N$ and $\tilde{f}_N$ caused by the choice of optimization algorithm (which includes the choice of parameterization) is referred to as implicit regularization or implicit bias. It means that although the practitioner does not explicitly include regularization in the definition of $\hat{f}_N$, the algorithm itself implicitly introduces regularization or bias when actually executed.

**Example 15** (Implicit regularization)**.** *This example considers linear parameterization and convex parameterization.*

1. *Let $\mathcal{F}$ be an RKHS $\mathcal{H}$ (see Example 4), let $\hat{f}_N$ be the ERM (see Example 8) with the squared loss, i.e., $\hat{f}_N \in \arg\min\left(\frac{1}{N}\|\boldsymbol{y} - \mathbb{X}f\|_2^2 : f \in \mathcal{H}\right)$, and suppose in practice we compute it using gradient descent or gradient flow starting from $\mathbf{0}$ with step size $\eta$, stopping at time $t$. Then the output $\tilde{f}_N$ is a spectral algorithm (see Example 9).*

2. *Take $\mathcal{F}$ as the mean-field shallow neural network (see Example 7), and take $\hat{f}_N$ as ERM (see Example 8). If $\hat{f}_N$ is computed via a Wasserstein gradient flow with parameter $\lambda$ (where $\lambda \ge 0$) as the algorithm, then*

$$\tilde{f}_N \in \operatorname{argmin}(P_N \ell_\nu + \lambda \mathrm{Ent}^-(\nu) : \nu \in \mathcal{P}(\Theta)),$$

*where we establish a correspondence $\mathcal{F} \ni f \leftrightarrow \nu \in \mathcal{P}(\Theta)$ and where $\mathrm{Ent}^-(\cdot)$ is the negative Shannon entropy with respect to the Lebesgue measure $d\boldsymbol{\theta}$. See, for instance, [NWS22].*

3. *Take $\mathcal{F}$ as the deep neural networks (see Example 6), and take $\hat{f}_N$ to be the ERM. Suppose $\hat{f}_N$ is computed in practice in the Neural Tangent Kernel (NTK) regime (see [JGH18]) by a first order method starting from* $\mathbf{0}$ *([Boy22]) for infinite time. Then there exists a RKHS $\mathcal{H}$ (called the RKHS generated by the NTK kernel) that is independent with $\mu$, such that the output $\tilde{f}_N$ is the minimum $\|\cdot\|_{\mathcal{H}}$-norm interpolant estimator, see Example 10.*

Example 15 considers estimators trained using training algorithms with guaranteed convergence. Here, the convergence guarantee means that, as the training time $t$ go to $\infty$, $t \mapsto P_N \ell_{f_t}$ converges to 0 when initialized properly—i.e., the global convergence property of the empirical risk/training error. A more challenging problem is that for some solutions, finding a computationally efficient algorithm with guaranteed convergence is often extremely difficult. For example, how to design an optimization algorithm that, with theoretically guaranteed convergence, can efficiently compute the ERM or RERM for deep neural networks has long been a focal research topic in optimization, that is, how to compute the minimizer of the following problem

$$\hat{f}_N \in \operatorname{argmin} \left( L_f((X_i, Y_i)_{i=1}^N) + \lambda \Psi(f) : f \in \mathcal{F} \right),$$

where $\mathcal{F}$ is the class of deep neural networks (Example 6). See [BB21, LRJ23] and the references therein.

A developing trend in statistical learning theory is to establish guarantees on the statistical properties of estimators in nonconvex models and that depend on training algorithms, particularly by taking implicit regularization into account, for instance, [CL19, MU25, HI25].

### 1.4.2 Training-test gap

The problem arising from non-convexity itself appears daunting. Yet even if it were solved—meaning that such an algorithm with global convergence guarantees existed and could compute the ERM or RERM for statistical models like deep neural networks efficiently—it would still pose additional challenges for characterizing its statistical properties: the high complexity of neural networks implies that the loss landscape contains many local minima and global optima, but the test errors of these points may differ drastically. This leads to the following issue: the training–test gap, i.e., the empirical (excess) risk may not reflect the population (excess) risk. This is particularly evident for overfitting estimators, see Example 10.

**Definition 7** (Benign, tempered and catastrophic overfitting, [MSA+22])**.** *In regression and classification problems, let $\mathcal{L}$ be the excess risk $\mathcal{L}^{(2)}$ of the squared loss $\ell^{(2)}$ and $\mathcal{L}^{(0,1)}$ of the 0/1 loss $\ell^{(0,1)}$, respectively. Suppose* $\dim(\mathcal{F}) = \infty$. *We say an estimator is overfitting if $P_N \ell_{\hat{f}_N} = 0$. Furthermore, given a definition of the limit of $\{(\Omega, \mu, \hat{f}_N, \mathcal{F})\}$, we say an overfitting estimator $\hat{f}_N$ exhibits*

1. *benign overfitting, if* $\lim P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} = 0$*;*
2. *tempered overfitting, if* $0 < \lim P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} < \infty$ *in regression, and* $0 < \lim P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} < \frac{1}{2}$ *in (binary) classification;*
3. *catastrophic overfitting, if* $\lim P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} \to \infty$ *in regression, and* $\lim P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} = \frac{1}{2}$ *in (binary) classification.*

For overfitting estimators, the classical proof technique for oracle inequalities introduced in Section 1.3 fails. That is, the uniform convergence argument can only yield the trivial residual $P\ell_{f^*}$ in the oracle inequality for such estimators, that is, tempered overfitting, while the actual residual may be much smaller than this trivial bound, for instance, benign overfitting.

In the case of real-valued regression problem, there exists an empirical–population excess risk gap: that is, $P_N\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} = -(1+o(1))\sigma_\xi^2$, while we wish to show $P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} = o(1)\sigma_\xi^2$, wherein the absolute difference between $P_N\mathcal{L}_{\hat{f}_N}^{\mathcal{F}}$ and $P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}}$ is non-negligible. This requires that the upper bound on the empirical process $\sup((P - P_N)\mathcal{L}_f^{\mathcal{F}} : f \in \mathcal{F})$ be of the order of $\sigma_\xi^2$ for the uniform convergence method to go through. In fact, the upper bound must be accurate up to a $(1 + o(1))\sigma_\xi^2$ factor, so that it cancels with the $-\sigma_\xi^2$ in $P_N\mathcal{L}_{\hat{f}_N}^{\mathcal{F}}$, leaving only a $o(1)\sigma_\xi^2$ term; this is necessary for benign overfitting to result. This extremely fine precision in the constant is exactly why, at present, benign overfitting via a uniform convergence method is known only in the Gaussian setting in [KZSS21, WDY22, DRSY22, ZKS+22]. If one applies the uniform convergence method to general probability measures, to the best of our knowledge, it is currently impossible to establish benign overfitting; only tempered overfitting can be obtained (that is, $P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}} \sim \sigma_\xi^2$), for example, [CLvdG22].

Describing necessary and sufficient conditions for the consistency of ERM is one of the most important tasks in learning theory, [Vap00, Section 1.6]. The emergence of benign overfitting phenomenon challenges the most fundamental methodological approach in statistical learning theory, especially since such a gap frequently appears in the practice of neural networks, see, for instance, [HI25]. Consequently, a developing trend in statistical learning theory is to devise a set of analytical methods compatible with interpolant estimators (in particular, it can characterize the phenomenon of benign overfitting), thereby providing a refinement of the uniform convergence argument that can be used beyond interpolant estimators.

### 1.4.3 Feature learning of neural networks

In recent years, as neural networks have extensively outperformed classical statistical methods in various engineering problems, while its statistical theory has long remained underdeveloped. Among these, comprehending the feature engineering capability of neural networks is a central concern. In the study of neural network theory, the map $\varphi : \boldsymbol{x} \in \mathbb{R}^d \mapsto \varphi(\boldsymbol{x}) := \sigma_{L-1}(\mathbb{W}_{L-1}\sigma_{L-2}(\mathbb{W}_{L-2}\sigma_{L-3}(\cdots \mathbb{W}_2\sigma_1(\mathbb{W}_1\boldsymbol{x})))) \in \mathbb{R}^{W_{L-1}}$ defined in Example 6 is usually called the feature or representation learned by the neural network, or referred to as feature engineering ; see [PHD20, RBPB22, YH21]. The academic community largely believes that the unique feature engineering capability of neural networks is responsible for their remarkable success in industrial practice. However, the definition of neural networks' feature learning ability—both empirically and mathematically—remains rather vague. For instance, [YH21] refers to situations where $\varphi$, after training, differs significantly from its initialized state as feature learning, i.e., the feature evolves during training. This conclusion only describes the change of features from the perspective of training dynamics and does not involve test error. [BES+22] defines feature learning as a scenario where, using the learned $\varphi$ as the feature map of an RKHS (known as data-dependent conjugate kernel) for linear regression on a specific supervised learning problem, the resulting test error is smaller than that obtained by linear regression with a generic RKHS. That is, through training, the neural network learns features relevant to the supervised learning problem, enabling good performance when these features are used for linear regression. Such a definition only focuses on the comparison between deep neural networks and classical kernel methods, yet still fails to touch the essence of "how to understand the learned features." Besides neural networks, [RBPB22] also demonstrates that several other estimators exhibit, experimentally, feature-engineering phenomena similar to those of neural networks.

Providing a sound mathematical definition of the feature learning property is the first step toward understanding the unique and mysterious statistical nature of neural networks. Here, by a sound mathematical definition we mean one that is not only well-defined mathematically, but also capable of characterizing the estimation error of a neural network. Moreover, it should simultaneously be able to explain the various empirical phenomena regarding feature learning in neural networks observed in the deep-learning-theory community, such as neural collapse [PHD20], multiple descent [ZLRB24], grokking, [BMA24].

How to establish a characterization of the population risk for deep neural networks that incorporates the training method, so that it can reflect the feature learning capability of neural networks, is an unavoidable question in statistical learning theory.

## 1.5 Feature Space Decomposition

*"Profound study of nature is the most fertile source of mathematical discoveries."*

— *Joseph Fourier,* The Analytical Theory of Heat*, Ch. 1, p. 7 (1822; English transl. 1878)*

In this section we introduce the main methodological contribution of this thesis: the Feature Space Decomposition (FSD) method. The FSD method was developed in a series of works by [P4, P2, P3, P1]. The Feature Space Decomposition method is first of all a tool to help theorists analyze the population excess risk; at the same time, it could also serve as a potential new theoretical framework for statistical learning theory and mathematical statistics, offering theorists a fresh perspective for understanding the statistical properties of a estimator.

In Section 1.5.1, we present the basic framework of the FSD method for real-valued supervised regression problems and binary classification problems. In Section 1.5.2 and Section 1.5.3, we discuss the roles of the two subspaces produced by the FSD method, respectively, and illustrate them with examples from various supervised learning problems. Finally, in Section 1.5.4, we show how the FSD method can serve as a potential new theoretical framework. Throughout this section, we always assume that $\mathcal{F}$ is a linear space, or at least can be embedded in a linear space. Following the tradition of statistical learning theory, we then refer to $\mathcal{F}$ as the feature space, [VC68].

### 1.5.1 The Feature Space Decomposition method

In this section, we present the FSD method tailored for supervised regression and classification problems. We begin with real-valued supervised regression problems.

**Real-valued supervised regression problem.** We recall from Section 1.2 that the goal of a theorist is: given a real-valued supervised regression problem $(\mu_X, f^*, \xi)$ and one of its solutions $(\mathcal{F}, \hat{f}_N)$, to characterize the estimation error $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$.

For the estimation error, there are two fundamentally different ways to bound it from above:

1. Obtain an upper bound for $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$ via cancellation between $\hat{f}_N$ and $f^*_{\mathcal{F}}$, i.e., by showing that $\hat{f}_N$ and $f^*_{\mathcal{F}}$ are close under the $L^2(\mu_X)$ metric;

2. Use the smallness of $\|\hat{f}_N\|^2_{L^2(\mu_X)}$ and $\|f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$, i.e., apply the triangle inequality to get $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)} \leq 2(\|\hat{f}_N\|^2_{L^2(\mu_X)} + \|f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)})$.

For real-valued supervised regression problems, FSD method can be viewed formally as an interpolation between these two approaches. To see this, we first define FSD.

**Definition 8.** *Any direct-sum decomposition $\mathcal{F} = V_J \oplus V_{J^c}$ of $\mathcal{F}$ is called a Feature Space Decomposition (FSD) of $\mathcal{F}$. Denote by $P_{V_J}$ the projection operator onto the linear subspace $V_J$, and by $P_{V_{J^c}}$ the projection onto $V_{J^c}$; equivalently, the identity operator $I_{\mathcal{F}} = P_{V_J} + P_{V_{J^c}}$ on the feature space $\mathcal{F}$ is decomposed. In particular, if an FSD satisfies that $V_J$ and $V_{J^c}$ are orthogonal with respect to the $L^2(\mu_X)$ inner product, we call it an orthogonal FSD, and denote it by $\mathcal{F} = V_J \oplus^{\perp} V_{J^c}$.*

For any $f \in \mathcal{F}$, write $f_J = P_{V_J} f$ and $f_{J^c} = P_{V_{J^c}} f$. We abbreviate $P_{V_J}\hat{f}_N$ as $\hat{f}_J$, $P_{V_{J^c}}\hat{f}_N$ as $\hat{f}_{J^c}$, $P_{V_J} f^*_{\mathcal{F}}$ as $f^*_J$, and $P_{V_{J^c}} f^*_{\mathcal{F}}$ as $f^*_{J^c}$. Note that we will not confuse $f^*_{\mathcal{F}}$ with $f^*$, because we can always incorporate the approximation error into the noise; see Lemma 3. Given any FSD $\mathcal{F} = V_J \oplus V_{J^c}$, the estimation error admits the decomposition

$$\left\|\hat{f}_N - f^*_{\mathcal{F}}\right\|^2_{L^2(\mu_X)} \begin{cases} = \left\|\hat{f}_J - f^*_J\right\|^2_{L^2(\mu_X)} + \left\|\hat{f}_{J^c} - f^*_{J^c}\right\|^2_{L^2(\mu_X)}, & \text{if } V_J \perp V_{J^c} \text{ in } L^2(\mu_X), \\ \leq 2\left\|\hat{f}_J - f^*_J\right\|^2_{L^2(\mu_X)} + 2\left\|\hat{f}_{J^c} - f^*_{J^c}\right\|^2_{L^2(\mu_X)}, & \text{otherwise.} \end{cases} \tag{1.10}$$

The interpolation between item *1* and item *2* can be expressed as the following inequality.

$$\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)} \leq \min\left(2\|\hat{f}_J - f^*_J\|^2_{L^2(\mu_X)} + 4\|\hat{f}_{J^c}\|^2_{L^2(\mu_X)} + 4\|f^*_{J^c}\|^2_{L^2(\mu_X)} : \mathcal{F} = V_J \oplus V_{J^c}\right). \tag{1.11}$$

The FSD method consists in seeking real-valued functions $r : (V_J, V_{J^c}) \mapsto r(V_J, V_{J^c}) \in \mathbb{R}_+$ and $\delta : (V_J, V_{J^c}) \mapsto \delta(V_J, V_{J^c}) \in [0, 1]$, such that for every (or at least some) FSD, the following inequality holds with probability at least $1 - \delta(V_J, V_{J^c})$ (or in expectation, if one desires an upper bound on the expected estimation error),

$$2\|\hat{f}_J - f^*_J\|^2_{L^2(\mu_X)} + 4\|\hat{f}_{J^c}\|^2_{L^2(\mu_X)} + 4\|f^*_{J^c}\|^2_{L^2(\mu_X)} \leq r^2(V_J, V_{J^c}). \tag{1.12}$$

We call such $r$ the rate function of $(\mu_X, f^*, \xi)$ and $(\mathcal{F}, \hat{f}_N)$. Here, saying that we seek a rate function means seeking a function that is as small as possible; otherwise one could trivially take $r(V_J, V_{J^c}) = \infty$.

As a mathematical proof strategy, the core idea of the FSD method is based on the following belief.

1. On the subspace $V_J$, called the estimation subspace, classical statistics takes place, i.e., $\hat{f}_N$ estimates $f^*_{\mathcal{F}}$ on $V_J$; hence the distance between $\hat{f}_J$ and $f^*_J$ under the $L^2(\mu_X)$ metric is small, contributing to the estimation error via cancellation $\|\hat{f}_J - f^*_J\|^2_{L^2(\mu_X)}$.

2. On the other hand, we believe that $\hat{f}_N$ on $V_{J^c}$ does not estimate $f^*_{\mathcal{F}}$. Therefore we call $V_{J^c}$ the free subspace. On this subspace, $\hat{f}_{J^c}$ fulfills certain tasks determined by the definition of $\hat{f}_N$, but in general not estimation; consequently, we expect the distance between $\hat{f}_{J^c}$ and $f^*_{J^c}$ under the $L^2(\mu_X)$ metric not necessarily to be small compared to the sum of their $L^2(\mu_X)$ norms, so applying the triangle inequality does not necessarily lead to an overestimation of $\|\hat{f}_{J^c} - f^*_{J^c}\|^2_{L^2(\mu_X)}$. In this case, the estimation error receives contributions in the form of the smallness of $\|\hat{f}_{J^c}\|^2_{L^2(\mu_X)}$ and $\|f^*_{J^c}\|^2_{L^2(\mu_X)}$.

Inspecting (1.12), we see that a FSD splits the upper bound on $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$ into three components. Each component carries its own statistical meaning: $\|\hat{f}_J - f^*_J\|_{L^2(\mu_X)}$ is the estimation error incurred because $\hat{f}_J$ estimates $f^*_J$; $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$ is the "energy" of the free part $\hat{f}_{J^c}$; and $\|f^*_{J^c}\|_{L^2(\mu_X)}$ is the approximation error resulting from the fact that $\hat{f}_J$ does not estimate $f^*_{J^c}$.

**Proposition 4.** *For any FSD $\mathcal{F} = V_J \oplus V_{J^c}$ and any rate function $r$, we have*

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

*Define*

$$(V^*_J, V^*_{J^c}) \in \operatorname{argmin}\left(r(V_J, V_{J^c}) : \mathcal{F} = V_J \oplus V_{J^c}\right). \tag{1.13}$$

*We call $(V^*_J, V^*_{J^c})$ the optimal FSD for the solution $(\mathcal{F}, \hat{f}_N)$ of the real-valued supervised regression problem $(\mu_X, f^*, \xi)$. Then in particular,*

$$\mathbb{P}\left(\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)} \leq r^2(V^*_J, V^*_{J^c})\right) \geq 1 - \delta(V^*_J, V^*_{J^c}). \tag{1.14}$$

In the following, we write $P_{V^*_J} f$ as $f_{J_*}$, $P_{V^*_{J^c}} f$ as $f_{J^c_*}$; write $P_{V^*_J}\hat{f}_N$ as $\hat{f}_{J_*}$, $P_{V^*_{J^c}}\hat{f}_N$ as $\hat{f}_{J^c_*}$; and write $P_{V^*_J} f^*_{\mathcal{F}}$ as $f^*_{J_*}$, $P_{V^*_{J^c}} f^*_{\mathcal{F}}$ as $f^*_{J^c_*}$.

The classical statistical learning theory introduced in Section 1.3 corresponds to choosing the trivial FSD $V_J = \mathcal{F}$. In this case, classical statistical learning theory expects classical statistics to perform estimation over the entire feature space, thereby obtaining an upper bound for the estimation error. This approach is intuitive given that when an estimator $\hat{f}_N$ of $f^*_{\mathcal{F}}$ is consistent, we expect $\hat{f}_N$ to estimate $f^*_{\mathcal{F}}$ and not only a part of it. One key idea exposed by the FSD method is that it may not be the case, that is, this trivial FSD is not necessarily optimal; consequently, the upper bound it provides for the estimation error is not always sharp. In fact, for a large class of spectral algorithms—such as ridge regression, gradient descent, gradient flow, etc., see Example 9, and for every real-valued supervised regression problem and for any feature space given by some RKHS, with high probability, we can reverse (1.14), i.e., for those supervised learning problems and solutions, there exist some absolute constant $0 < c < 1$ and some real number $0 < \delta < 1$, the following inequality holds

$$\mathbb{P}\left(\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)} \geq c r^2(V^*_J, V^*_{J^c})\right) \geq 1 - \delta. \tag{1.15}$$

This implies the following remarkable phenomenon: for this class of $(\mu_X, f^*, \xi)$ and $(\mathcal{F}, \hat{f}_N)$, the estimation error $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$ is "characterized" by an interpolation between these two distinct approaches. Here, because $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$ is with high probability equivalent to $r(V^*_J, V^*_{J^c})$, we use the term "characterized". Moreover, there exists no other way to control the estimation error beyond the two avenues described in Proposition 4.

**Binary supervised classification problems.** In this paragraph we consider the population excess risk for binary classification problem $(\mu_X, \eta)$, which we recall is defined as

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

where $\eta : \boldsymbol{x} \in \Omega_X \mapsto \mathbb{P}(Y = 1 | X = \boldsymbol{x})$. As in regression problems, $P\mathcal{L}^{\mathcal{F}}_{\hat{f}_N}$ or $P\mathcal{L}_{\hat{f}_N}$ consists of three contributions. Namely, given an arbitrary decomposition $\mathcal{F} = V_J \oplus V_{J^c}$, let $f^*_J$ be some function in $V_J$ — we will define it later. We decompose the 0-1 risk of $\hat{f}_N$ as follows:

$$\begin{aligned} P\mathcal{L}^{(0,1)}_{\hat{f}_N} &= \mathbb{P}\left(Y\hat{f}_N(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) && (1.16) \\ &\quad + \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y f^*_J(X) < 0\right) && (1.17) \\ &\quad + \mathbb{P}\left(Y f^*_J(X) < 0\right) - \mathbb{P}\left(Y\left(\eta(X) - \frac{1}{2}\right) < 0\right), && (1.18) \end{aligned}$$

where (1.16) is the error caused by free part $\hat{f}_{J^c}$; (1.17) is the prediction error caused by $\hat{f}_J$ compared to the one of $f_J^*$ ; and (1.18) is the prediction error caused by $f_J^*$ compared with the one of the Bayes rule (or, when we replace $\eta(X) - 1/2$ with $f_{\mathcal{F}}^*(X)$, it becomes the approximation error of $f_J^*$ to $f_{\mathcal{F}}^*$). These three terms are precisely the 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 (1.11).

Analogously to the regression case, the FSD method aims to find a non-trivial rate function $r : (V_J, V_{J^c}) \mapsto r(V_J, V_{J^c}) \in \mathbb{R}_+$ and a confidence function $\delta : (V_J, V_{J^c}) \mapsto \delta(V_J, V_{J^c}) \in [0, 1]$ such that for every FSD, the following inequality holds with probability at least $1 - \delta(V_J, V_{J^c})$ (or in expectation):

$$(1.16) + (1.17) + (1.18) \leq r(V_J, V_{J^c}).$$

Similarly, the following proposition holds.

**Proposition 5.** *For any FSD $\mathcal{F} = V_J \oplus V_{J^c}$ and any rate function $r$, we have*

$$\mathbb{P}\left(P\mathcal{L}_{\hat{f}_N}^{(0,1)} \leq r^2(V_J, V_{J^c})\right) \geq 1 - \delta(V_J, V_{J^c}).$$

*Define*

$$(V_J^*, V_{J^c}^*) \in \operatorname{argmin}\left(r(V_J, V_{J^c}) : \mathcal{F} = V_J \oplus V_{J^c}\right). \tag{1.19}$$

*We call $(V_J^*, V_{J^c}^*)$ the optimal FSD for the solution $(\mathcal{F}, \hat{f}_N)$ of the binary supervised classification problem $(\mu_X, \eta)$. Then in particular,*

$$\mathbb{P}\left(P\mathcal{L}_{\hat{f}_N}^{(0,1)} \leq r^2(V_J^*, V_{J^c}^*)\right) \geq 1 - \delta(V_J^*, V_{J^c}^*). \tag{1.20}$$

**FSD as an analytical method.** We emphasize that FSD method serves as a tool to help theorists analyze the excess risk of any estimator as well as to understand its behavior. That is to say, in the construction of estimators $\hat{f}_N$, the practitioners has no control over the choice of $V_J$ and $V_{J^c}$—because the estimator itself does not take $V_J$ or $V_{J^c}$ as input parameters. For instance, the minimum norm interpolating estimator in Example 10 has no tunable parameters whatsoever. Therefore, we assert that the decomposition of $\mathcal{F}$ into two subspaces is performed implicitly by the estimator, not by the practitioners. Consequently, when practitioners execute this statistical algorithm, this decomposition occurs as a black-box operation. For estimators with tunable parameters, given a parameter set by the practitioners, the estimator automatically determines the optimal FSD $(V_J^*, V_{J^c}^*)$ based on both this parameter and the regression problem itself. Certainly, we emphasize that theorists can leverage the new theoretical insights provided by the FSD method to help design practical methods. For example, using the precise characterization of the estimation error offered by the FSD method to design an adaptive estimator via Lepski's method, [Lep91], see also the survey [Lep23] for other adaptive methods.

Below, in Section 1.5.2 and Section 1.5.3, we separately explain the roles of these two subspaces and how they specifically assist theorists in their analysis.

### 1.5.2 $V_J$ defines a morphism in the category of supervised learning problems

For convenience, throughout this section we always assume $f^* \in \mathcal{F}$, and therefore do not distinguish between $f^*$ and $f_{\mathcal{F}}^*$. Before starting this section, we recall that to obtain an upper bound for $\|\hat{f}_N - f^*\|_{L^2(\mu_X)}^2$ or for $P\mathcal{L}_{\hat{f}_N}^{(0,1)}$ in binary classification problems via the FSD method, on $V_J$, we need an upper bound for $\|\hat{f}_J - f_J^*\|_{L^2(\mu_X)}$ or for $\mathbb{P}(Y\hat{f}_J(X) < 0 | (X_i, Y_i)_{i=1}^N) - \mathbb{P}(Yf_J^*(X) < 0)$. This is precisely the task of classical statistical learning theory and mathematical statistics. What, then, is the role of the FSD method on $V_J$?

For any given quintuple $(\mu_X, f^*, \xi, \mathcal{F}, \hat{f}_N)$ consisting of a real-valued supervised regression problem and a solution, the FSD provides, via $V_J$, the following arrow:

$$\bullet_J : (\mu_X, f^*, \xi, \mathcal{F}, \hat{f}_N) \longmapsto (\mu_X, f_J^*, \zeta, V_J, \hat{f}_J), \text{ where } \zeta = \xi + f_{J^c}^*,$$

through the following relation:

$$Y = f^*(X) + \xi = f_J^*(X) + \zeta.$$

In other words, the FSD method endows the theorist with the power to pass from handling a supervised regression problem and its solution $(\mu_X, f^*, \xi, \mathcal{F}, \hat{f}_N)$ to another supervised regression problem and its solution $(\mu_X, f_J^*, \zeta, V_J, \hat{f}_J)$.

Furthermore, if one only wishes to obtain an upper bound for $\|\hat{f}_N - f^*\|^2_{L^2(\mu_X)}$, then the theorist possesses the freedom to choose the arrow, i.e., by selecting an FSD, thereby freely selecting the target supervised regression problem and its solution $(\mu_X, f_J^*, \zeta, V_J, \hat{f}_J)$. This can often grant the theorist extra analytical power beyond the classical statistical learning theory introduced in Section 1.3 — because then it suffices to apply the classical statistical learning theory on the new model $V_J$, and the new signal $f_J^*$ may be easier to analyze. Of course, if one aims to obtain an upper bound for $\|\hat{f}_N - f^*\|^2_{L^2(\mu_X)}$ that is as sharp as possible, or even a precise characterization of $\|\hat{f}_N - f^*\|^2_{L^2(\mu_X)}$ in the sense of (1.15), then it is necessary to choose a good FSD $(V_J, V_{J^c})$, such that the rate function $r(V_J, V_{J^c})$ as small as possible—or even the optimal FSD $(V_J^*, V_{J^c}^*)$.

Let us now illustrate this point with some examples.

**$\bullet_J$ defines the new $\hat{f}_J$.**

Although $\hat{f}_J$ is by definition $P_{V_J}\hat{f}_N$, if $V_J$ is chosen appropriately, $\hat{f}_J$ may admit an equivalent characterization other than $P_{V_J}\hat{f}_N$, which the theorist can then exploit to facilitate the analysis. Three examples follow. Their proofs are readily thus omitted, see also Proposition 20 later in Chapter 2 for the proof of Proposition 8 below.

**Proposition 6** (self-regularization of the minimum $\|\cdot\|_q$-norm interpolant estimator)**.** *Let $p \in \mathbb{N}_+$, $\mathcal{F} = \{\langle \cdot, \boldsymbol{\beta}\rangle : \boldsymbol{\beta} \in \mathbb{R}^p\}$. Let $\boldsymbol{e}_1, \dots, \boldsymbol{e}_p$ be a basis of $\mathbb{R}^p$. Let $1 \le q < \infty$ be a real number, and $\|\cdot\|_q$ be the $\ell_q$ norm on $\mathbb{R}^p$ with respect to this basis. Consider the minimum $\|\cdot\|_q$-norm interpolant estimator defined in Example 10, that is,*

$$\hat{\boldsymbol{\beta}} \in \operatorname{argmin}(\|\boldsymbol{\beta}\|_q : \mathbb{X}\boldsymbol{\beta} = \boldsymbol{y}), \text{ where } \mathbb{X} = [X_1|\cdots|X_N]^\top, \ \boldsymbol{y} = (Y_1, \cdots, Y_N).$$

*Take any FSD $\mathbb{R}^p = V_J \oplus V_{J^c}$, where $V_J = \operatorname{span}(\boldsymbol{e}_j : j \in J)$ for some $J \subset \{1, \cdots, p\}$. Define $\mathcal{A} : \boldsymbol{\mu} \in \mathbb{R}^N \mapsto \mathcal{A}[\boldsymbol{\mu}] \in \arg\min(\|\boldsymbol{\nu}\|_q : \mathbb{X}\boldsymbol{\nu} = \boldsymbol{\mu},\ \boldsymbol{\nu} \in V_{J^c})$. Then $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{A}[\boldsymbol{y} - \mathbb{X}\hat{\boldsymbol{\beta}}_J]$, and*

$$\hat{\boldsymbol{\beta}}_J \in \operatorname*{argmin}_{\boldsymbol{\beta}_J \in V_J} \left(L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) + \|\boldsymbol{\beta}_J\|_q^q\right), \text{ where } L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q.$$

Proposition 6 tells us that although $\hat{\boldsymbol{\beta}}_J$ is by definition equal to $P_{V_J}\hat{\boldsymbol{\beta}}$, as theorists, when we choose a suitable FSD, we can endow it with a new statistical meaning—a RERM whose loss function $L_{\boldsymbol{\beta}_J}$ is in fact a stochastic loss function and $\|\cdot\|_q^q$ is the self-regularization functional. Because this regularization is imposed by $\hat{\boldsymbol{\beta}}$ upon itself, rather than being explicitly defined by the practitioner, we call it self-regularization. This regularization does not depend on the specific training algorithm, and therefore differs from implicit regularization introduced in Section 1.4, see [BMR21, pp. 92].

**Proposition 7** (self-regularization of the minimum $\|\cdot\|_2$-norm interpolant classifier)**.** *If $\mathcal{F}$ is identified with $\mathbb{R}^p$, and $\hat{\boldsymbol{\beta}}$ is the minimum $\|\cdot\|_2$-norm interpolant classifier (Example 10). Take an arbitrary FSD $\mathbb{R}^p = V_J \oplus V_{J^c}$, denote $\mathbb{1} = (1, \dots, 1) \in \mathbb{R}^N$, and let $\mathbb{X}_{\boldsymbol{y},J^c} = [Y_1 P_{V_{J^c}} X_1|\cdots|Y_N P_{V_{J^c}} X_N]^\top$. Define $\mathcal{B} : \boldsymbol{\mu} \in \mathbb{R}^N \mapsto \mathcal{B}[\boldsymbol{\mu}] \in \arg\min(\|\boldsymbol{\nu}\|_{\mathcal{H}} : \mathbb{X}_{\boldsymbol{y},J^c}\boldsymbol{\nu} \succeq \boldsymbol{\mu})$. Then $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\hat{f}_J]$, and*

$$\hat{\boldsymbol{\beta}}_J \in \operatorname{argmin}\left(L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) + \|\boldsymbol{\beta}_J\|_2^2 : f_J \in V_J\right), \text{ where } L_{f_J}((X_i, Y_i)_{i=1}^N) = \|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]\|_2^2.$$

*Here, for any $\boldsymbol{a} = (a_i)_{i=1}^N$ and $\boldsymbol{b} = (b_i)_{i=1}^N$, we write $\boldsymbol{a} \succeq \boldsymbol{b}$, if $a_i \ge b_i$ for any $1 \le i \le N$.*

Similarly, here $\hat{f}_J$ is identified as a RERM whose loss function is a stochastic loss function.

For these two new loss functions, because they incorporate regularization, they do not suffer from overfitting. Consequently, applying classical statistical learning theory on $V_J$ yields an oracle inequality whose residual term can tend to zero. This is precisely the advantage brought by the new estimator $\hat{f}_J$ via FSD.

**Proposition 8** (effective regularization)**.** *If $\mathcal{F}$ is identified with an RKHS $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ with feature map $\phi$, and $\hat{f}_N$ is the ridge regression on $\mathcal{F}$ with parameter $t^{-1}$, i.e., $\hat{f}_N = \frac{1}{N}\mathbb{X}^\top(\frac{1}{N}\mathbb{X}\mathbb{X}^\top + \frac{1}{t}I_N)^{-1}\boldsymbol{y}$, where $\boldsymbol{y} = (Y_1, \dots, Y_N)$ and $\mathbb{X} : f \in \mathcal{H} \mapsto (\langle\phi(X_i), f\rangle_{\mathcal{H}})_{i=1}^N$. Take an arbitrary FSD $\mathcal{H} = V_J \oplus V_{J^c}$, and denote $\mathbb{X}_{J^c} = \mathbb{X}P_{J^c}$. Then*

$$\hat{f}_J \in \operatorname{argmin}\left(L_{f_J}((X_i, Y_i)_{i=1}^N) + \|f_J\|_{\mathcal{H}}^2\right), \text{ where } L_{f_J}((X_i, Y_i)_{i=1}^N) = \|Q(\boldsymbol{y} - \mathbb{X}f_J)\|_{\mathcal{H}}^2,$$

*and $Q : \mathbb{R}^N \to V_{J^c}$ is a bounded linear operator such that $Q^\top Q = (\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + t^{-1}I_N)^{-1}$.*

In other words, $\hat{f}_J$ is identified as a RERM whose loss function $L_{f_J}$ is also a stochastic loss function. Here, ridge regression has a tuning parameter $t^{-1}$; thus, for any tuning parameter $t^{-1}$ given by the practitioner, $\hat{f}_N$ itself selects an FSD, generating a new regularization $(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + t^{-1}I_N)$, which is referred to as effective regularization.

**$\bullet_J$ defines the new signal $f_J^*$.**

By choosing an FSD, the theorist can also select a suitable new signal to work with. In this paragraph we present two examples: latent factor regression and the minimum $\|\cdot\|_{\mathcal{H}}$-norm interpolant classifier.

**Latent factor regression.** Latent factor regression is a special class of real-valued regression problems where the dependence between $(X, Y)$ is governed by a latent random vector $Z$, an unknown embedding matrix $A$, and two types of noise.

**Definition 9** (Latent factor regression problem)**.** *Let $k < p$ be two positive integers, let $\Omega_X = \mathbb{R}^p$, and let $A \in \mathbb{R}^{p\times k}$ be a fixed but unknown matrix. Let $Z \in \mathbb{R}^k$ be a random vector, called latent factor. Let $W \in \mathbb{R}^p$ be a zero-mean random vector, independent of $Z$, with covariance matrix $\Sigma_W = \mathbb{E}[W \otimes W]$. Let $\xi \in \mathbb{R}$ be a zero-mean random variable with variance $\sigma_\xi^2$, independent of $(Z, W)$. The design vector is defined by $X = AZ + W$. Thus, in this model, the observable design vector $X$ arises from a latent factor $Z$ through an unobserved linear transformation $A$, together with an unobserved noise perturbation $W$, so that $X = AZ + W$.*

*Let $\Omega_Y = \mathbb{R}$, and let $Y$ be defined as follows. Let $\boldsymbol{\alpha}^* \in \mathbb{R}^k$ be a location vector, and the response variable by $Y = \langle \boldsymbol{\alpha}^*, Z\rangle + \xi$. The response variable $Y$ depends only on the latent factor $Z$, the unknown signal $\boldsymbol{\alpha}^* \in \mathbb{R}^k$, and an unobserved noise perturbation $\xi$. In latent factor regression, the most common loss function is the squared loss $\ell : (y_1, y_2) \in \mathbb{R} \times \mathbb{R} \mapsto (y_1 - y_2)^2$. See, for instance, [BBSMW21].*

Let $\mathcal{F} = \{f_{\boldsymbol{\beta}}(\cdot) = \langle \boldsymbol{\beta}, \cdot\rangle : \boldsymbol{\beta} \in \mathbb{R}^p\}$. The latent factor regression problem is mis-specified unless $(Z, X)$ is jointly Gaussian. In fact, the Bayes rule is given by $f^* : \boldsymbol{x} \mapsto \langle \boldsymbol{\alpha}^*, \mathbb{E}[Z \mid X = \boldsymbol{x}]\rangle$. However, the statistical model $\mathcal{F}$ is the class of linear functionals. The oracle in $\mathcal{F}$ is given by $f_{\mathcal{F}}^*$, identified by a vector $\boldsymbol{\beta}^*$ through $f_{\mathcal{F}}^*(\cdot) = \langle \cdot, \boldsymbol{\beta}^*\rangle$, defined as $\boldsymbol{\beta}^* \in \operatorname{argmin}(P\ell_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathbb{R}^p) = \operatorname{argmin}(\mathbb{E}[(\langle \boldsymbol{\beta}, X\rangle - Y)^2] : \boldsymbol{\beta} \in \mathbb{R}^p)$ . Let $\Sigma = \mathbb{E}[X \otimes X]$ be the covariance operator of $X$. A direct computation yields $\Sigma = A\Sigma_Z A^\top + \Sigma_W$, where $\Sigma_Z = \mathbb{E}[Z \otimes Z] : \mathbb{R}^k \to \mathbb{R}^k$. Since $\Sigma_W$ is positive definite, $\Sigma$ is also positive definite, and $\Sigma$ can be viewed as the rank-$k$ informative component $A\Sigma_Z A^\top$ perturbed by $\Sigma_W$. It is computed in [BBSMW21, Equation 6] that $\boldsymbol{\beta}^* = \Sigma^{-1}A\Sigma_Z\boldsymbol{\alpha}^*$. Let $\mathbb{Z} : \boldsymbol{\alpha} \in \mathbb{R}^k \mapsto (\langle Z_i, \boldsymbol{\alpha}\rangle)_{i=1}^N \in \mathbb{R}^N$. In the latent factor regression problem, the response vector $\boldsymbol{y} = \mathbb{Z}\boldsymbol{\alpha}^* + \boldsymbol{\xi}$, but we need to solve the problem in $\mathbb{R}^p$, and the oracle in $\mathbb{R}^p$ is $\boldsymbol{\beta}^*$.

Below we show how, by choosing a good FSD—i.e., a good $V_J$—we can explore the rank-$k$ informative component $A\Sigma_Z A^\top$ hidden in $\mathbb{R}^p$, which is exactly the aim of the latent factor regression problem. Take $V_J = \operatorname{Range}(A\Sigma_Z A^\top) = \operatorname{Range}(A)$. In this case, $\boldsymbol{\beta}_J^* = P_{V_J}\boldsymbol{\beta}^* = \boldsymbol{\beta}^*$. Consequently, we have the following supervised regression problem $(\mu_X, \boldsymbol{\beta}_J^*, \zeta)$, where $\zeta = \xi + (\langle Z, \boldsymbol{\alpha}^*\rangle - \langle X, \boldsymbol{\beta}_J^*\rangle)$. Here, the new noise consists of two parts: $\xi$ is the original noise, while $\langle Z, \boldsymbol{\alpha}^*\rangle - \langle X, \boldsymbol{\beta}_J^*\rangle = \langle Z, \boldsymbol{\alpha}^*\rangle - \langle X, \boldsymbol{\beta}^*\rangle$ corresponds to the approximation error of $\boldsymbol{\alpha}^*$ on $\mathbb{R}^p$. In [BBSMW21] it is proved that this term is an irreducible component of the estimation error. Therefore, for the latent factor regression problem, by choosing a suitable $V_J$, we reduce the dimension of the problem to $k$, while guaranteeing that the signal in this space satisfies $\boldsymbol{\beta}_J^* = \boldsymbol{\beta}^*$.

**Minimum $\|\cdot\|_2$ norm interpolant classifier.** In this paragraph we consider the minimum $\|\cdot\|_2$-norm interpolant classifier defined in Example 10, i.e., we assume $\mathcal{F}$ is identified with $\mathbb{R}^p$. We now illustrate that by choosing an FSD appropriately, the approximation error resulting from restricting estimation to $V_J$—namely, (1.18)—can be eliminated. We examine the following standard model for binary supervised classification problems:

**Definition 10** (Logistic classification problem)**.** *Let $\boldsymbol{\mu} \in \mathbb{R}^p$ be called the signal, and $\Lambda \in \mathbb{R}^{p\times p}$ be a positive definite bounded linear operator. Let $X \sim \mathcal{N}(\mathbf{0}, \Lambda)$ be a Gaussian random vector with mean $\mathbf{0}$ and covariance operator $\Lambda$. By defining $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid X = \boldsymbol{x}) = 1/(1 + \exp(-2\langle \Lambda^{-1}\boldsymbol{\mu}, \boldsymbol{x}\rangle))$ and $\mathbb{P}(Y = -1 \mid X = \boldsymbol{x}) = 1 - \eta(\boldsymbol{x})$, we specify the distribution of $Y$. This problem is called the logistic model, [Gir14, Section 11.1.3].*

A straightforward calculation shows that the Bayes classifier for the logistic classification problem is $f^*(\cdot) = \operatorname{sign}(\langle \cdot, \Lambda^{-1}\boldsymbol{\mu}\rangle)$. Hence, the Bayes classifier can be identified with $\Lambda^{-1}\boldsymbol{\mu}$. Therefore, as long as the FSD is chosen so that $f_J^*$ and $\Lambda^{-1}\boldsymbol{\mu}$ are well aligned, (1.18) becomes zero. Later, in Proposition 11, we prove that if $\Lambda^{-1}\boldsymbol{\mu} \in V_J$, then this indeed holds.

The logistic model represents a class of binary supervised classification models; both the Gaussian mixture classification model [WT21] and the latent factor classification model [BW23] share the same characteristic—namely, there exists an optimal linear classifier that corresponds to $f^*$.

**$\bullet_J$ reduces the fixed points.**

Because we believe that estimation occurs only on $V_J$, as a consequence the theorist should apply classical statistical learning theory—i.e., the methods from Section 1.3—only on $V_J$. One outcome of doing so is that, since both the supervised learning problem and its solution have changed, applying classical statistical learning theory on $V_J$ may yield a smaller fixed point, and thus a smaller bound on $\|\hat{f}_J - f_J^*\|^2_{L^2(\mu_X)}$. The main reason why the fixed points on $V_J$ are expected to be smaller than the one over $\mathcal{F}$ is because we usually have $\dim(V_J) \ll \dim(\mathcal{F})$, not because $f_J^*$ and $\hat{f}_J$ have changed. In this section, we illustrate how FSD reduces the fixed points defined in Section 1.3.1, by using the example of benign overfitting for the minimum $\|\cdot\|_q$-norm interpolant estimator. For ridge regression, FSD can also reduce the multiplier and quadratic fixed points, but the proof is more involved and will not be presented here (see [P2]).

**FSD reduces multiplier fixed point.** The formal version and the proof of the following Proposition 9 can be found in [P1], see also Section 4.3.4; we do not repeat it here. Proving these properties requires the geometric tools on $V_{J^c}$ introduced in Section 1.5.3.

**Proposition 9** (informal)**.** *Using the notation of Proposition 6.*

*Under some assumptions, there exist some absolute constants* $0 < \delta_{\mathrm{M}} < \frac{1}{100}$, $c, c' < 1$, $\ell_* > 0$ *and* $c'' = c''(c, c', \delta_{\mathrm{M}}) > 1$ *such that for any localization subset* $\mathcal{G} \subset V_J$, $r_{\mathrm{M}}(\mathcal{G}, \delta_{\mathrm{M}}, \frac{2}{q}, 4c\frac{N^{\frac{q}{2}}}{\ell_*^q}) \le c''\sigma_\xi(\frac{|J|}{N})^{\frac{1}{2(q-1)}}$ *when* $q \ge 2$*; and* $r_{\mathrm{M}}(\mathcal{G}, \delta_{\mathrm{M}}, 1, 4c'\sigma_\xi^{q-2}\frac{N^{\frac{q}{2}}}{\ell_*^q}) \le c''\sigma_\xi^{q-1}(\frac{|J|}{N})^{\frac{1}{2}}$ *when* $1 \le q < 2$.

**FSD reduces quadratic fixed point.** For the minimum $\|\cdot\|_q$-norm interpolant estimator, the FSD provided by Proposition 6 can also reduce the quadratic fixed point. The formal version and the proof of the following proposition can be found in [P1], see also Section 4.3.4 later.

**Proposition 10** (informal)**.** *Under the assumptions of Proposition 9, there exist some absolute constant* $0 < \delta_{\mathrm{Q}} < \frac{1}{100}$, $c = c(q)$, *and* $c' = c'(q)$, *such that the following hold.*

1. *When* $q \ge 2$. *Then for any* $r > 0$, *and any localization subset* $\mathcal{G}$, *with probability at least* $1 - \delta_{\mathrm{Q}}$, *for any* $\boldsymbol{\beta}_J \in \mathcal{G} \cap S_{L^2(P_{V_J}\mu_X)}(\boldsymbol{\beta}_J^*; r)$,

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

   *and* $\boldsymbol{g} = \nabla L_{\boldsymbol{\beta}_J^*}$.

2. *When* $1 \le q < 2$. *Suppose* $X_J$ *is a centered Gaussian random vector, then for any localization subset* $\mathcal{G}$ *and any* $0 < r < \sigma_\xi$, *with probability at least* $1 - \delta_{\mathrm{Q}}$, $(\mathcal{G}, X_J, L_\bullet)$ *satisfies the local Bernstein's condition at scale* $r$, *with parameters* $(\diamondsuit, 1)$, *where*

$$\diamondsuit = c'\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q}.$$

Note: the local Bernstein's condition that holds with high probability in item *2* is due to the fact that in Proposition 6, the loss function $L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q$ is a stochastic loss function that depends on $\mathbb{X}P_{V_{J^c}}$; hence the population excess risk $P\mathcal{L}_{\boldsymbol{\beta}_J}^{V_J}$ is a conditional expectation $\mathbb{E}_{\mathbb{X}_J,\boldsymbol{\xi}}P_N\mathcal{L}_{\boldsymbol{\beta}_J}^{V_J}$, and the local Bernstein condition holds with high probability. One can prove that $\boldsymbol{\beta}_J^* \in \arg\min(P\ell_{\boldsymbol{\beta}_J} : \boldsymbol{\beta}_J \in V_J)$ holds almost surely, where $P\ell_{\boldsymbol{\beta}_J} = \mathbb{E}_{\mathbb{X}_J,\boldsymbol{\xi}}L_{\boldsymbol{\beta}_J}$, see Lemma 21 later.

Proposition 10 tells us that when $q \ge 2$, if the FSD is suitably chosen, there exists $\delta_{\mathrm{Q}} < \frac{1}{100}$ such that for every localization subset $\mathcal{G} \subset V_J$, the quadratic fixed point $r_{\mathrm{Q}}(\mathcal{G}, \delta_{\mathrm{Q}}, \frac{2}{q}) = 0$ when $q \ge 2$. Consequently, in this situation FSD completely eliminates the quadratic fixed point. This implies that in Theorem 1, for any $\rho > 0$, $r_{\mathrm{iso}}(\rho) = r_M(\mathcal{G}, \delta_{\mathrm{M}}, \frac{2}{q}, 4c\frac{N^{q/2}}{\ell_*})$. Hence the system of inequalities in Theorem 1 with $\lambda = 1$ can decouple $\rho$ and $r$—the system can be reduced to the simpler scheme of setting $\rho = \triangle r^{\frac{2}{\kappa}}$ and then solving for the smallest $r > r_{\mathrm{iso}}(\rho)$ such that the inequality $\triangle r^{\frac{2}{\kappa}} > \square r^{\frac{2}{\kappa}} + \lambda\|\nabla\Psi(f_{\mathcal{F}}^*)\|_{(\rho)}$ holds; this $r$ becomes $r_*$. We do not repeat the detailed conclusion here; see [P1], see also Section 4.7.2 later. When $1 < q < 2$, because $P_N\ell_\bullet = L_\bullet$ lacks strong convexity, the quadratic fixed point is not completely removed. The situation here is more complicated and will not be discussed immediately in this chapter; see [P1] and Section 4.7.1 for all details.

**$\bullet_J$ as a shell wrapping classical mathematical-statistical analysis**

The ridge regression and minimum-norm interpolant estimators studied earlier can both be written as RERM (or their limits). Estimators of this form generally fall within the scope of statistical learning theory, [VC68]. In this section, we show that the FSD method is not only applicable to estimators defined by ERM and RERM, which are common in statistical learning theory, but also to classical estimators that belong more broadly to the domain of mathematical statistics: spectral methods (Example 9). Applying the FSD method to the analysis of the estimation error of such estimators amounts to wrapping a shell around the original mathematical-statistical analysis—i.e., confining the analysis of estimation error, which originally covered the whole feature space, to the subspace $V_J$. Even if this may not necessarily create a new estimator or shrink the fixed points as it does for minimum-norm interpolant estimators or ridge regression, it still yields a "correct" signal $f_J^*$ to work with.

Classical statistical theory for spectral methods provides ways to obtain an upper bound on $\|\hat{f}_N - f^*\|_{L^2(\mu_X)}$, for instance, [SZ07, YRC07, BPR07, LGRO+08, BM16, BM18, BMM19, ZLL23, LGSL24]. However, if we first perform an FSD, then we only need to apply the classical theory to obtain an upper bound on $\|\hat{f}_J - f_J^*\|_{L^2(\mu_X)}$. This means we have switched from estimating the full signal $f^*$ to estimating the "effective signal" $f_J^*$, and consequently we can obtain a characterization of the estimation error (in the sense of (1.15))—something that the classical approach cannot achieve. In Chapter 3, we specifically demonstrate how this analytical approach enables us to characterize the sharp convergence rate of the population excess risk for nearly arbitrary spectral methods in linear regression in $\mathbb{R}^p$. This implies that for any estimator studied in mathematical statistics for supervised regression problems, we always have a painless way to handle $\|\hat{f}_J - f_J^*\|_{L^2(\mu_X)}$—simply transplant the classical analysis to the subspace $V_J$. However, one needs to handle the "free part" $V_{J^c}$ that is now what we show how to proceed.

### 1.5.3 $V_{J^c}$: new tools from Geometric Aspects of Functional Analysis

Since no estimation of $f_{J^c}^*$ by $\hat{f}_{J^c}$ takes place in the free subspace, we say that no statistics occur on that subspace. Consequently, the tools required for this subspace do not belong to classical mathematical statistics, and for this reason we still know relatively little about it. Our work therefore constitute the first examples of the analysis of some estimators in the free space. Of course we used existing tools from the Geometric Aspects of Functional Analysis (GAFA) that were not previously used in statistics and we had to extend them to fit our statistical framework. However, we may anticipate that new tools (potentially from GAFA) may be required to fully understand the statistical properties of estimators in the free space. In this section we offer partial insights into the free subspace for some special cases.

Regarding the free subspace and the estimator $\hat{f}_{J^c}$ on it, we focus primarily on the following two issues:

1. the stochastic properties that the free subspace provides for $\hat{f}_J$;
2. the energy $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$ of $\hat{f}_{J^c}$.

**$V_{J^c}$ provides stochastic properties of $\hat{f}_J$**

In this section, we consider the minimum $\|\cdot\|_q$-norm interpolant estimator (Example 10) and the ridge regression.

1. Minimum $\|\cdot\|_q$-norm interpolant estimator.

   In Proposition 6, Proposition 9, and Proposition 10 we have already seen that FSD identifies $\hat{\boldsymbol{\beta}}_J$ equivalently as a RERM whose loss function is given by $L_{\boldsymbol{\beta}_J} : (X_i, Y_i)_{i=1}^N \in \Omega^N \mapsto \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q$. Here we recall its definition: let $\mathbb{X}_J = \mathbb{X}P_{V_J}$ and $\mathbb{X}_{J^c} = \mathbb{X}P_{V_{J^c}}$; then $\mathcal{A} : \boldsymbol{\mu} \in \mathbb{R}^N \mapsto \mathcal{A}[\boldsymbol{\mu}] \in \arg\min(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{\mu})$. Thus $\mathcal{A} : (\mathbb{R}^N, \|\cdot\|_2) \to (V_{J^c}, \|\cdot\|_q)$ is a random embedding operator, and consequently $L_\bullet$ is a stochastic loss function.

2. Ridge regression.

   Similarly, Proposition 8 tells us that for a ridge regression with parameter $t^{-1}$, its $\hat{f}_J$ is also a RERM whose loss function is $L_{f_J}((X_i, Y_i)_{i=1}^N) = \|Q(\boldsymbol{y} - \mathbb{X}f_J)\|_{\mathcal{H}}^2$, where $Q^\top Q = (\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + t^{-1}I_N)^{-1}$.

Following the FSD credo—apply classical mathematical statistics and statistical learning theory (see Section 1.3) on $V_J$—we need to study the properties of these stochastic loss functions in order to complete the proofs of Proposition 9 and Proposition 10, as well as to compute the multiplier and quadratic fixed points for ridge regression. The properties of these stochastic loss functions therefore require analysis using specialized geometric tools. This tool is provided by the celebrated Dvoretzky–Milman theorem, [Dvo59, Dvo61, Mil71].

**The Dvoretzky–Milman theorem and its role in benign overfitting for the minimum $\|\cdot\|_q$-norm interpolant estimator.** For any compact subset $K \subset \mathbb{R}^p$, we define $\ell_*(K) = \mathbb{E}(\sup\langle \boldsymbol{v}, G\rangle : \boldsymbol{v} \in K)$ as the Gaussian mean width of $K$, where $G \in \mathbb{R}^p$ is a standard Gaussian random vector. We let $\operatorname{diam}(K) = \max(\|\boldsymbol{v}\|_2 : \boldsymbol{v} \in K)$ be the $\ell_2$ diameter of $K$. We denote $K^\circ = \{\boldsymbol{v} \in \mathbb{R}^p : \langle \boldsymbol{v}, \boldsymbol{u}\rangle \leq 1, \forall \boldsymbol{u} \in K\}$ as the polar body of $K$. Denote $d_*(K) = (\ell_*(K^\circ)/\operatorname{diam}(K^\circ))^2$ to be the Dvoretzky dimension of $K$. We denote $q'$ by $\frac{q}{q-1}$. Below is Milman's version of Dvoretzky's theorem; see [Pis89].

**Theorem 2** (Dvoretzky-Milman)**.** *There are absolute constants $\kappa_{DM} \leq 1$ and $c_1$ such that the following holds. Let $|||\cdot|||$ be some norm on $\mathbb{R}^p$ and denote by $B$ its unit ball. Denote by $\mathbb{G} := \mathbb{G}^{(N\times p)}$, the $N \times p$ standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ Gaussian entries. Given any $0 < \varepsilon_1 \leq 1$. Assume that $N \leq \kappa_{DM}\varepsilon_1^2 d_*(B)$. Then with probability at least $1 - \exp(-c_1\varepsilon_1^2 d_*(B))$, for every $\boldsymbol{\lambda} \in \mathbb{R}^N$,*

$$(1-\varepsilon_1)\,\|\boldsymbol{\lambda}\|_2\,\ell_*(B^*) \leq |||\mathbb{G}^\top \boldsymbol{\lambda}||| \leq (1+\varepsilon_1)\,\|\boldsymbol{\lambda}\|_2\,\ell_*(B^*). \tag{1.21}$$

For all $0 < \varepsilon_1 < 1$, we define the event

$$\Omega_{\mathrm{DM,reg}}(\varepsilon_1) := \left\{\forall \boldsymbol{\lambda} \in \mathbb{R}^N : \|\boldsymbol{\lambda}\|_2\,(1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_q^p) \leq \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \leq \|\boldsymbol{\lambda}\|_2\,(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_q^p)\right\} \tag{1.22}$$

$$\subset \left\{\forall \boldsymbol{\mu} \in \mathbb{R}^N : \frac{\|\boldsymbol{\mu}\|_2}{(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_q^p)} \leq \|\mathcal{A}[\boldsymbol{\mu}]\|_q \leq \frac{\|\boldsymbol{\mu}\|_2}{(1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_q^p)}\right\}. \tag{1.23}$$

It follows from Theorem 2 applied to the norm $|||\cdot||| = \|\Sigma_{J^c}^{1/2}\cdot\|_{q'}$ that, if $X_{J^c}$ is a Gaussian random vector and $\kappa_{DM}\varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2}B_{q'}^p) \geq N$, then $\mathbb{P}(\Omega_{\mathrm{DM,reg}}(\varepsilon_1)) \geq 1 - \exp(-c_1\varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2}B_{q'}^p))$. The inclusion from (1.23) follows from strong duality: for all $\boldsymbol{\mu} \in \mathbb{R}^N$,

$$\|\mathcal{A}[\boldsymbol{\mu}]\|_q = \min\left(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}^\top \boldsymbol{\nu} = \boldsymbol{\mu}\right) = \max\left(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle : \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \leq 1\right). \tag{1.24}$$

Even though $\mathcal{A} : (\mathbb{R}^N, \ell_2) \to (V_{J^c}, \ell_q)$ is a non-linear metric embedding (except when $q = 2$), it satisfies a DM theorem inherited from $\mathbb{X}_{J^c}^\top$. Since our loss functions in the estimation part of the features space depend on $\mathcal{A}$ in the regression problem, working on the event $\Omega_{\mathrm{DM,reg}}(\varepsilon_1)$ will allow us to greatly simplify its expression because now it is isomorphic to the $\ell_2^N$-norm and so we will work with the classical squared loss function. That is the reason why DM theorem plays a crucial role in our analysis: we use this isomorphic property from DM to greatly simplify the loss function appearing in $V_J$ and then go back to the classical analysis of regularized ERM with respect to the squared loss on $V_J$.

Below, we demonstrate how to use the Dvoretzky–Milman theorem to prove Proposition 10, item *1*.

**Proof.** (of Proposition 10, item *1*) By Example 12, we have

$$\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q - \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q^q \geq \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\rangle + \frac{q-1}{q2^q}\,\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q^q,$$

where $\boldsymbol{g}$ is defined in Proposition 9. From the definition of $\mathcal{A}$, we have $\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q \geq \|\mathcal{A}[\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)]\|_q$. Then using (1.23), we obtain

$$P_N\mathcal{L}_{\boldsymbol{\beta}_J} \geq \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\rangle + \frac{q-1}{q2^q}\,\frac{\|\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)\|_2^q}{(1+\varepsilon_1)^q\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}.$$

Finally, from the assumption $\dim(V_J) \lesssim N$ and the fact that for any $\mathcal{G} \subset V_J$, we have $r_{\mathrm{RIP},-}(\mathcal{G}) = 0$ (see Example 11), the proof of Proposition 10, item *1* is completed. ∎

Of course, if one wants to go beyond the Gaussian design case, one needs to extend DM theorem beyond that case.

**The Dvoretzky–Milman theorem for $\|\cdot\|_{q'}$-norms under general probability measures.** Theorem 2 provides the Dvoretzky–Milman theorem for Gaussian measures. Because we need to study the case where $X_{J^c}$ is distributed according to a general probability measure, we require an extension of the Dvoretzky–Milman theorem for $\|\cdot\|_{q'}$-norms. Extensions of the Dvoretzky–Milman theorem to general probability measures already exist in a substantial body of literature, e.g., [GLPTJ07, MTJ08, BM22a, BM22b, Men22, BM24]. In these works the random

embedding operator is usually induced by row-independent random matrices or by more complex random-matrix models; however, in $\Omega_{\mathrm{DM,reg}}(\varepsilon)$ we need a column-independent random-matrix model. Hence, an entirely new Dvoretzky–Milman theorem for such random matrices is required. The following theorem, taken from [P1], is a contribution to GAFA that was motivated precisely by the FSD method. Its proof may be found in Section 5.1. Denote $\mathrm{Log}(x) = \max\{1, \ln(x)\}$.

**Assumption 1.** *$\boldsymbol{\zeta} = (\zeta_j)_{j=1}^p$ is a centered, isotropic random vector in $\mathbb{R}^p$ with i.i.d. coordinates, satisfying $\mathbb{E}[\zeta_1^2] = 1$, and there exist absolute constants $0 < \kappa \leq 1$ and $\varepsilon > 0$ such that $\mathbb{E}|\zeta_1|^{\max\{4,2q+\varepsilon\}} \leq \kappa^{\max\{4,2q+\varepsilon\}}$.*

**Theorem 3** ([P1])**.** *Let $\boldsymbol{\zeta}$ be a random vector satisfying Assumption 1, and let $\Sigma$ be a positive definite diagonal matrix on $\mathbb{R}^p$, $\Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_p)$. Let $X = \Sigma^{1/2}\boldsymbol{\zeta}$, and let $X_1, \cdots, X_N$ be independent copies of $X$, forming the random matrix $\mathbb{X} = [X_1|\cdots|X_N]^\top = [Z_1|\cdots|Z_p]$, where $(Z_j)_{j=1}^p$ are the column vectors of $\mathbb{X}$. Denote $\ell_* = \ell_*(\Sigma^{1/2}B_q^p)$ and $d_* = d_*(\Sigma^{-1/2}B_{q'}^p)$. Without loss of generality, assume that $d_* \geq 1$. There then exists an absolute constant $0 < \theta < 1$ such that for any $\boldsymbol{\lambda} \in S_2^{N-1}$, $\mathbb{P}(|\langle Z_j, \boldsymbol{\lambda}\rangle| \geq \theta) \geq \kappa$. Moreover, there exist absolute constants $c, c', C, C', C'', \kappa_{DM}, \varepsilon_0 > 0$ such that the following facts hold.*

1. *When $q \geq 2$. If $N \leq \kappa_{DM} d_* \mathrm{Log}^{-2}(p^{\frac{1}{q'}}/d_*)$, then with probability at least*

$$1 - C'\mathrm{Log}\left(\frac{p^{\frac{1}{q'}}}{d_*}\right)\exp\left(-C''\kappa_{DM}\frac{d_*^\theta}{\mathrm{Log}^{2\theta}\left(\frac{p^{\frac{1}{q'}}}{d_*}\right)}\right) - 2\exp\left(-C'd_*\right) - C'd_*^{-c\min\{\varepsilon,\varepsilon_0\}} =: 1 - \bar{p}_{DM},$$

*there holds for any $\boldsymbol{\lambda} \in S_2^{N-1}$,*

$$c\ell_* \leq \left\|\mathbb{X}^\top\boldsymbol{\lambda}\right\|_{q'} \leq C\mathrm{Log}(p)\ell_*.$$

2. *When $q < 2$. If $N \leq \kappa_{DM} d_*(\Sigma^{-1/2}B_{q'}^p)$, then*

$$\mathbb{P}\left(\forall\boldsymbol{\lambda} \in S_2^{N-1},\, c\ell_* \leq \|\mathbb{X}^\top\boldsymbol{\lambda}\|_{q'} \leq C\ell_*\right) \geq 1 - 3\exp(-c'd_*) - C'd_*^{-\frac{q'-2}{4}} =: 1 - \bar{p}_{DM}.$$

Theorem 3 establishes that, under Assumption 1, the linear span of $N$ independent copies of $\Sigma^{1/2}\boldsymbol{\zeta}$ provides a generalization (up to a logarithmic factor when $q \geq 2$) of the Dvoretzky-Milman theorem for the convex body $B_{q'}^p$ under a general probability measure. We emphasize here that if one focuses solely on the sub-Gaussian case, then for $q \geq 2$, the $\mathrm{Log}(p)$ factor in the uniform upper bound for $\|\mathbb{X}^\top\boldsymbol{\lambda}\|_{q'}$ can be removed. To the best of our knowledge, this theorem is the first generalization of the Dvoretzky-Milman theorem for the $\|\Sigma^{1/2}\cdot\|_{q'}$ norm under such broad (almost the most general) conditions.

**The Dvoretzky–Milman theorem for $\|\cdot\|_{\mathcal{H}}$-norms under general probability measures.** When $q = 2$, the Dvoretzky–Milman theorem can hold for more general probability measures, e.g., for feature map generated by RKHS whose kernel functions are polynomials of finite degree. Recall the definition of RKHS $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}}) \subset L^2(\mu_X)$ in Example 4. Let $\Sigma = \mathbb{E}[\phi(X)\otimes\phi(X)] : f \in \mathcal{H} \mapsto \mathbb{E}[\phi(X)\langle\phi(X), f\rangle] \in \mathcal{H}$ be its integral operator. Let $\mathcal{H} = V_J \oplus^\perp V_{J^c}$ be a FSD. Recall that $\phi_{J^c} = P_{V_{J^c}}\phi$, $\Sigma_{J^c} = \Sigma P_{V_{J^c}}$. Let $\mathrm{Tr}(\cdot)$ be the trace, and $\|\cdot\|_{\mathrm{op}}$ be the $\mathcal{H} \to \mathcal{H}$ operator norm. Let $\mathbb{X}_{J^c} : f \in V_{J^c} \mapsto (\langle\phi_{J^c}(X_i), f\rangle_{\mathcal{H}})_{i=1}^N \in \mathbb{R}^N$ and $\mathbb{X}_{J^c}^\top : \boldsymbol{\lambda} \in \mathbb{R}^N \mapsto \sum_{i=1}^N \lambda_i\phi_{J^c}(X_i) \in V_{J^c}$ be its transpose. For any $\lambda \geq 0$, define

$$d_\lambda^*(\Sigma_{J^c}^{-1/2}B_{\mathcal{H}}) := \frac{\mathrm{Tr}(\Sigma_{J^c}) + \lambda}{\|\Sigma_{J^c}\|_{\mathrm{op}}}. \tag{1.25}$$

**Assumption 2.** *There are absolute constants $C_1 > 1$, $C_2 > 1$, $0 \leq \gamma < 1/16$, $0 \leq \delta < 1/(100\sqrt{C_2})$, $\bar{\delta} < C_1^{-1}$, $\epsilon > 0$ and $\kappa > 1$ such that*

- *With probability at least $1 - \gamma$,*

$$\max_{1\leq i\leq N}\left|\frac{\|\phi_{J^c}(X_i)\|_{\mathcal{H}}^2}{(\ell^*)^2} - 1\right| \leq \delta, \tag{1.26}$$

*where we define $\ell^* = \sqrt{\mathbb{E}\,\|\phi_{J^c}(X)\|_{\mathcal{H}}^2} = \sqrt{\mathrm{Tr}\,(\Sigma_{J^c})}$.*

- *For any $f \in V_{J^c}$, we have*
$$\|f\|_{L^{2+\epsilon}(\mu_X)} \leq \kappa \|f\|_{L^2(\mu_X)}. \tag{1.27}$$
- *Depending on the choice of $\epsilon$, there are two cases:*
  1. *if $\epsilon > 2$, then no extra assumption is required.*
  2. *if $0 < \epsilon \leq 2$, then*
$$\kappa N^{\frac{2-\epsilon}{2\epsilon+\epsilon^2}} \log(N) \left( \sqrt{\frac{N\|\Sigma_{J^c}\|_{\mathrm{op}}}{\mathrm{Tr}(\Sigma_{J^c})}} \right) < \bar{\delta}. \tag{1.28}$$

A typical example satisfying Assumption 2 is when $\phi_{J^c}$ is the identity mapping and $X_i$ is a sub-Gaussian random vector; in this case, the result follows from the Hanson–Wright inequality. A more involved example arises when $\phi_{J^c}$ is the feature map of a finite-degree polynomial kernel; see the results in [P2]. The following theorem is proved in [P2], see also Section 5.2.

**Theorem 4.** *Let $X$ be a random vector distributed as $\mu_X$ in a compact set $\Omega_X \subset \mathbb{R}^d$, and let $X_1, \cdots, X_N$ be i.i.d. copies of $X$. Let $\phi : \boldsymbol{x} \in \Omega_X \mapsto K(\boldsymbol{x}, \cdot) \in \mathcal{H}$ be the feature map of the RKHS $\mathcal{H}$. Let $C_3$, $C_4$, $C_5$ and $C_6$ be absolute constants.*

1. *Suppose that $\lambda \leq C_3 \mathrm{Tr}(\Sigma_{J^c})$. Consider $0 < \delta, \bar{\delta} < 1$ from Assumption 2, define*
$$\tilde{\delta} = C_2\delta^2 + C_4\bar{\delta}^2 + 4\sqrt{(3\delta + C_5\bar{\delta})(1 + \delta + C_6\bar{\delta})} \tag{1.29}$$
*Suppose that for some $\lambda \geq 0$, we have $N \leq \kappa_{DM}\bar{\delta}^2 d^*_\lambda\left(\Sigma_{J^c}^{-1/2} B_{\mathcal{H}}\right)$ for a sufficiently small constant $\kappa_{DM} < 1$ which depends only on $\kappa$ (see [P2, Equation 96] for a precise description). We assume that $\phi_{J^c}$ satisfies Assumption 2. Then with probability at least*
$$1 - \gamma - \frac{1}{N^2} - \left(\frac{\kappa}{\bar{\delta}}\right)^{2+\epsilon} \left( \sqrt{\frac{N\|\Sigma_{J^c}\|_{\mathrm{op}}}{\mathrm{Tr}(\Sigma_{J^c})}} \right)^{2+\epsilon} \frac{\log^{2+\epsilon}(N)}{N^{\frac{\epsilon}{2}-1}} =: 1 - \bar{p}_{DM},$$
*for all $\boldsymbol{\lambda} \in \mathbb{R}^N$, it holds that*
$$\left(1 - \tilde{\delta}\right)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}\,\|\boldsymbol{\lambda}\|_2 \leq \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{\mathcal{H}} \leq \left(1 + \tilde{\delta}\right)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}\,\|\boldsymbol{\lambda}\|_2. \tag{1.30}$$
2. *Suppose that $\lambda > C_3 \mathrm{Tr}(\Sigma_{J^c})$. Suppose that $\phi_{J^c}$ satisfies the first two points of Assumption 2. Suppose that for some $\lambda \geq 0$, we have $N \leq (\kappa_{DM}/4) d^*_\lambda(\Sigma_{J^c}^{-1/2} B_{\mathcal{H}})$. Then there exist absolute constants $C_7$ depending on $\epsilon, \kappa, \kappa_{DM}$, and $0 < c_2 < 1$ such that with probability at least*
$$1 - \gamma - N\left(\left(\frac{\kappa_{DM}\kappa^2 \log^2(N)}{N}\right)^{1+\epsilon/2} N\right)^{\lceil (12+2\epsilon)/\epsilon \rceil - 1} - \frac{1}{N^2} =: 1 - \bar{p}_{DM},$$
*we have $\left\|\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + \lambda I_N\right\|_{\mathrm{op}} \leq C_7\lambda + \mathrm{Tr}(\Sigma_{J^c})$ and*
$$\sigma_N(\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + \lambda I_N) \geq c_2\lambda + (1 - c_2)C_3 \mathrm{Tr}(\Sigma_{J^c}).$$

Theorem 4 provides a Dvoretzky–Milman theorem for $q = 2$ (or, more generally, in RKHS settings), and it does not require any special structure among the coordinates of $\phi_{J^c}(X)$. Since $q = 2$, the theorem can be regarded as a statement about the spectrum of random matrices. This theorem can be used to study the properties of the loss function for $\hat{f}_J$ in ridge regression and for $\hat{\boldsymbol{\beta}}_J$ in the minimum $\|\cdot\|_2$-norm interpolant classifier. Formal statement of the next proposition and its proof can be found in [P1] or Section 4.3.3.

**Proposition 11** (informal). *Assume $\mathcal{F}$ is identified with linear functionals on $\mathbb{R}^p$. If $\mathbb{R}^p = V_J \oplus V_{J^c}$ is an FSD satisfying the following properties: 1. $YX_{J^c}$ is a centered sub-Gaussian random vector; 2. $N \le \kappa_{\mathrm{DM}} \tilde{\delta}^2 \frac{\mathrm{Tr}(\Sigma_{J^c})}{\|\Sigma_{J^c}\|_{op}}$, where $\Sigma_{J^c} = \mathbb{E}[X_{J^c} \otimes X_{J^c}]$. Then with high probability the loss function $L_{\boldsymbol{\beta}_J}$ defined in Proposition 7 possesses the following property: for any $\boldsymbol{\beta}_J \in V_J$,*

$$\frac{N}{(1+\tilde{\delta})^2 \,\mathrm{Tr}(\Sigma_{J^c})} P_N \ell^{(\mathrm{sh})}_{\boldsymbol{\beta}_J} \le L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) \le \frac{N}{(1-\tilde{\delta})^2 \,\mathrm{Tr}(\Sigma_{J^c})} P_N \ell^{(\mathrm{sh})}_{\boldsymbol{\beta}_J} \le L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N),$$

*where $\tilde{\delta}$ is from Theorem 4, and $P_N \ell^{(\mathrm{sh})}_{\boldsymbol{\beta}_J} = \frac{1}{N}\sum_{i=1}^N \ell^{(\mathrm{sh})}_{\boldsymbol{\beta}_J}(X_i, Y_i)$, and $\ell^{(\mathrm{sh})}_{\boldsymbol{\beta}_J}(\boldsymbol{x}, y) = (1 - y\langle \boldsymbol{\beta}_J, \boldsymbol{x}\rangle)_+^2$ is the squared hinge loss. Morevoer, for any binary classification problem $(\mu_X, \eta)$, if $f^{**} = \arg\min(P\ell^{(\mathrm{sq})}_f : f$ is measurable), then $f^{**} = f^*$, that is, the Bayes rule.*

In proposition, since $YX_{J^c}$ is a sub-Gaussian random vector, Assumption 2 holds naturally, and Theorem 4 therefore applies. Proposition 11 states that for the minimum $\|\cdot\|_2$-norm interpolant classifier, its loss function is almost "isometrically" equivalent to a squared hinge loss—a loss function that has been extensively studied in classical statistical learning theory.

**Proof.** Applying Theorem 4 to $\phi_{J^c}(X) = YX_{J^c}$ and to $\mathcal{H} = \mathbb{R}^p$, we only need to prove the following inclusion holds

$$\Omega_{\mathrm{DM,class}}(\tilde{\delta}) := \left\{ \forall \boldsymbol{\lambda} \in \mathbb{R}^N : \|\boldsymbol{\lambda}\|_2 (1-\tilde{\delta}) \sqrt{\mathrm{Tr}(\Sigma_{J^c})} \le \left\| \mathbb{X}_{\boldsymbol{y}, J^c}^\top \boldsymbol{\lambda} \right\|_2 \le \|\boldsymbol{\lambda}\|_2 (1+\tilde{\delta}) \sqrt{\mathrm{Tr}(\Sigma_{J^c})} \right\} \tag{1.31}$$

$$\subseteq \left\{ \forall \boldsymbol{\mu} \in \mathbb{R}^N : \frac{\|[\boldsymbol{\mu}]_+\|_2}{(1+\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}} \le \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \le \frac{\|[\boldsymbol{\mu}]_+\|_2}{(1-\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}} \right\}, \tag{1.32}$$

where $[\boldsymbol{\mu}]_+ = (\max(\mu_i, 0))_{i=1}^N$. By standard duality argument, see, for instance, [BV14, Equation 5.11], we obtain that

$$\|\mathcal{B}[\boldsymbol{\mu}]\|_2 = \max\left( \langle \boldsymbol{\mu}, \boldsymbol{\lambda} \rangle : \boldsymbol{\lambda} \succeq \mathbf{0},\ \left\| \mathbb{X}_{\boldsymbol{y}, J^c}^\top \boldsymbol{\lambda} \right\|_2 \le 1 \right). \tag{1.33}$$

Condition on $\Omega_{\mathrm{DM,class}}(\tilde{\delta})$, see (1.31), we have

$$\max_{\boldsymbol{\lambda} \succeq \mathbf{0}} \left( \langle \boldsymbol{\mu}, \boldsymbol{\lambda} \rangle : \|\boldsymbol{\lambda}\|_2 \le \frac{1}{(1+\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}} \right) \le \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \le \max_{\boldsymbol{\lambda} \succeq \mathbf{0}} \left( \langle \boldsymbol{\mu}, \boldsymbol{\lambda} \rangle : \|\boldsymbol{\lambda}\|_2 \le \frac{1}{(1-\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}} \right).$$

Let $H(\boldsymbol{\mu}) := \{i \in [N] : \mu_i < 0\}$ and let $\boldsymbol{\lambda}^-$ be the maximizer of the left-hand-side maximization problem and $\boldsymbol{\lambda}^+$ be the maximizer of the right-hand-side maximization problem. We prove that if $i \in H(\boldsymbol{\mu})$, then $\lambda_i^- = 0$. We prove this by contradiction. Suppose $i \in H(\boldsymbol{\mu})$ but $\lambda_i^- > 0$, then by letting $\tilde{\boldsymbol{\lambda}}^- = (\lambda_1^-, \cdots, \lambda_{i-1}^-, 0, \lambda_{i+1}^-, \cdots, \lambda_N^-)$, we know that $\|\tilde{\boldsymbol{\lambda}}^-\|_2 < \|\boldsymbol{\lambda}^-\|_2 \le \frac{1}{(1+\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}}$. Moreover, $\langle \boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}^- \rangle = \sum_{i' \neq i} \mu_{i'} \lambda_{i'}^- > \sum_{i'=1}^N \mu_{i'} \lambda_{i'}^- = \langle \boldsymbol{\mu}, \boldsymbol{\lambda}^- \rangle$ since $\mu_i \lambda_i^- < 0$. This implies that $\tilde{\boldsymbol{\lambda}}^-$ is a feasible solution but with larger objective function value, hence contradicting the assumption that $\boldsymbol{\lambda}^-$ is the maximizer. Recalling the constraint that $\boldsymbol{\lambda} \succeq \mathbf{0}$, we have: for any $i \in H(\boldsymbol{\mu})$, we necessarily have $\lambda_i^- = 0$. The same also holds for $\boldsymbol{\lambda}^+$. Now, by Cauchy-Schwartz, we have $\boldsymbol{\lambda}^- = (\boldsymbol{\mu}/(\|\boldsymbol{\mu}\|_2(1+\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}))_+$, and $\boldsymbol{\lambda}^+ = (\boldsymbol{\mu}/(\|\boldsymbol{\mu}\|_2(1-\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}))_+$. Therefore, condition on $\Omega_{\mathrm{DM,class}}(\tilde{\delta})$, (1.32) follows. For proofs of properties of the squared hinge loss, see Section 4.9.3. ∎

**Energy of the $\hat{f}_{J^c}$.**

Apart from supplying favorable stochastic properties for the possible new loss function that defines $\hat{\boldsymbol{\beta}}_J$, the energy $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$ of the free part on $V_{J^c}$ is the most difficult component to understand. At present our knowledge of this term remains very limited even though we were able to obtain sharp results of the minimum $\|\cdot\|_2$-norm interpolant estimators, ridge regression and spectral methods. In this paragraph we consider the following cases: 1. $\hat{f}_N$ is a regularized empirical risk minimization; 2. $\hat{f}_N$ is a spectral method; 3. the minimum $\|\cdot\|_q$-norm interpolant estimator; 4. LASSO with support recovery.

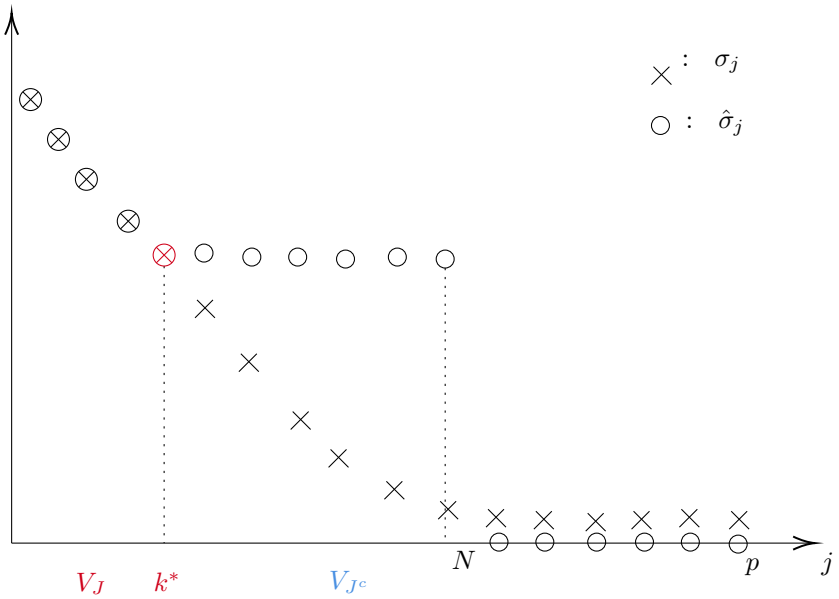


Figure 1.1: By Theorem 4, the spectrum of $\frac{1}{N}\mathbb{X}^\top\mathbb{X} + t^{-1}I$ has a plateau of height $\frac{1}{N}\operatorname{Tr}(\Sigma_{J^c}) + t^{-1}$ that is different from $\operatorname{Spec}(\Sigma_{J^c}) = \{\sigma_j : j \in J^c\}$. This is the reason why there is no possible estimation over $V_{J^c}$.

**Identification of $\hat{f}_{J^c}$.** First, for RERM and the minimum-norm interpolant estimator, we identify that $\hat{f}_{J^c}$ itself is also a RERM and a minimum-norm interpolant estimator. For the minimum-norm interpolant estimator, this has already been proved in Proposition 6; therefore we only address RERM here.

We use the notation of Example 8. The following proposition shows that $\hat{f}_{J^c}$ is a RERM of the residual of $\hat{f}_J$.

**Proposition 12.** *Assume there exists differentiable function $L : \mathbb{R}^N \to \mathbb{R}$ such that for any $f \in \mathcal{F}$, $L_f : (X_i, Y_i)_{i=1}^N \in \Omega^N \mapsto L(\boldsymbol{y} - \mathbb{X}f)$, where $\mathbb{X} : f \in \mathcal{F} \mapsto (f(X_i))_{i=1}^N$ and $\boldsymbol{y} = (Y_1, \cdots, Y_N)$. Let $\lambda \in \mathbb{R}$ be any real number, let $\Psi$ be a differentiable function and assume that $\mathcal{F}$ is a linear space. Define*

$$\hat{f}_N \in \operatorname*{argmin}_{f \in \mathcal{F}} \left(L(\boldsymbol{y} - \mathbb{X}f) + \lambda\Psi(f)\right)$$

*be a RERM. Let $\mathcal{F} = V_J \oplus V_{J^c}$ be a FSD. Suppose $\Psi : \mathcal{F} \to \mathbb{R}$ is decomposable with respect to $V_J \oplus V_{J^c}$, in the sense that for any $f \in \mathcal{F}$, $\Psi(f) = \Psi(f_J) + \Psi(f_{J^c})$. Then*

$$\hat{f}_{J^c} \in \operatorname*{argmin}_{f_{J^c} \in V_{J^c}} \left(L(\boldsymbol{y}' - \mathbb{X}f_{J^c}) + \lambda\Psi(f_{J^c})\right), \text{ where } \boldsymbol{y}' = \boldsymbol{y} - \mathbb{X}\hat{f}_J. \tag{1.34}$$

Here, $\hat{f}_{J^c}$ is learning the residual of $\hat{f}_J$, using $\mathbb{X}_{J^c}$ that are not necessarily isomorphic to $\Sigma_{J^c}^{1/2}$, see Figure 1.1.

**Proof.** By assumption, we can expand the map $(f_J, f_{J^c}) \in V_J \times V_{J^c} \mapsto L(\boldsymbol{y} - \mathbb{X}f) + \lambda\Psi(f)$ as $(f_J, f_{J^c}) \in V_J \times V_{J^c} \mapsto L((\boldsymbol{y} - \mathbb{X}f_J) - \mathbb{X}f_{J^c}) + \lambda\Psi(f_J) + \lambda\Psi(f_{J^c})$. Now fix $f_J = \hat{f}_J = P_{V_J}\hat{f}_N$. We know that $\hat{f}$ must satisfy the first-order optimality condition, i.e, the gradient of this map equals $\mathbf{0}$,

$$\mathbf{0} = -2P_{V_{J^c}}\mathbb{X}^\top\nabla L((\boldsymbol{y} - \mathbb{X}\hat{f}_J) - \hat{f}_{J^c}) + \lambda(\nabla\Psi)(\hat{f}_{J^c}).$$

This is precisely the first-order necessary condition of the optimization problem defined in (1.34). ∎

In the real-valued regression problem (Example 1), $\boldsymbol{y} - \mathbb{X}_J\hat{f}_J = \boldsymbol{\xi} + \mathbb{X}_{J^c}f^*_{J^c} + \mathbb{X}_J(\hat{f}_J - f^*_J)$. Yet, unlike classical theory on RERM, we do not need the estimation error of $\hat{f}_{J^c}$; we only require an upper bound on its $L^2(\mu_X)$ norm. This task is unprecedented in classical mathematical statistics and statistical learning theory.

**Linear estimators and minimum-norm interpolant estimators.** Next, we consider an upper bound for $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$—an aspect of the FSD method that we were able to study in 4 cases but that requires a deeper understanding. In fact, at present we have a relatively complete understanding only when $\hat{f}_{J^c}$ is a linear operator in $\boldsymbol{y}$ in linear regression problem. Below, we illustrate this using two classes of linear operators—ridge regression

and spectral methods—and conclude by discussing the case where $\hat{f}_{J^c}$ is a minimum-norm interpolant estimator, a comparatively simple nonlinear operator.

For linear operators in linear regression, i.e., when there exists a random linear operator $A : \mathbb{R}^N \to V_{J^c}$ independent of $(Y_i)_{i=1}^N$ such that $\hat{f}_{J^c} : (X_i, Y_i)_{i=1}^N \in \Omega^N \mapsto \langle \cdot, A\boldsymbol{y}\rangle \in V_{J^c}$, we identify $\hat{f}_{J^c}$ as $\hat{\boldsymbol{\beta}}_{J^c}(\boldsymbol{y}) = A\boldsymbol{y}$. Then, in the additive regression model, $\boldsymbol{y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$, and consequently $\|\hat{f}_{J^c}\|_{L^2(\mu_X)} \le \|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\mathbb{X}\boldsymbol{\beta}^*_J)\rangle\|_{L^2(\mu_X)} + \|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\mathbb{X}\boldsymbol{\beta}^*_{J^c})\rangle\|_{L^2(\mu_X)} + \|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\boldsymbol{\xi})\rangle\|_{L^2(\mu_X)}$. For the term $\|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\boldsymbol{\xi})\rangle\|_{L^2(\mu_X)}$, we have $\mathbb{E}_{\boldsymbol{\xi}}\|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\boldsymbol{\xi})\rangle\|^2_{L^2(\mu_X)} = \sigma_\xi^2 \operatorname{Tr}(A^\top \Sigma_{J^c} A)$, where $\Sigma_{J^c} = P_{J^c}\Sigma P_{J^c}$. This is precisely the strategy we adopt when analyzing spectral methods.

**Spectral methods.** Recall the spectral methods defined in Example 9. We know that spectral methods are linear estimators, and in this case $A = \frac{1}{N}\varphi(\hat{\Sigma})\mathbb{X}^\top$. Therefore,

$$\|\hat{f}_{J^c}\|_{L^2(\mu_X)} \le \|P_{J^c}\varphi_t(\hat{\Sigma})\hat{\Sigma} f^*_J\|_{L^2(\mu_X)} + \|P_{J^c}\varphi_t(\hat{\Sigma})\hat{\Sigma} f^*_{J^c}\|_{L^2(\mu_X)} + \|P_{J^c}\varphi_t(\hat{\Sigma})[N^{-1}\mathbb{X}^\top]\boldsymbol{\xi}\|_{L^2(\mu_X)}.$$

We do not continue to show the subsequent handling of these terms here; see [P5].

**Ridge regression.** An even more special case occurs when $\hat{f}_N$ is ridge regression, i.e., when $\mathcal{F}$ is identified with some RKHS $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$, $L : \boldsymbol{u} \in \mathbb{R}^N \mapsto \frac{1}{N}\|\boldsymbol{u}\|_2^2$, and $\Psi : f \in \mathcal{H} \mapsto t^{-1}\|f\|^2_{\mathcal{H}}$. In this case we not only know that $\hat{f}_{J^c}$ is a linear operator, but also that $\hat{f}_{J^c}$ is a ridge regression with parameter $t^{-1}$ applied to $\boldsymbol{y} - \mathbb{X}_J\hat{f}_J$. In fact, Proposition 12 shows that for any FSD, the ridge regression $\hat{f}_N$ with tuning parameter $t^{-1}$ satisfies

$$\hat{f}_{J^c} = \frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}(\boldsymbol{y} - \mathbb{X}_J\hat{f}_J). \tag{1.35}$$

From (1.35) we know that $\hat{f}_{J^c}$ is a linear operator on $\boldsymbol{y} - \mathbb{X}_J\hat{f}_J$, and since $\boldsymbol{y} = \mathbb{X}_J\boldsymbol{\beta}^*_J + \mathbb{X}_{J^c}\boldsymbol{\beta}^*_{J^c} + \boldsymbol{\xi}$, the triangle inequality gives

$$\begin{aligned}\|\hat{f}_{J^c}\|_{L^2(\mu_X)} \le{}& \left\|\frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}\mathbb{X}_J(f^*_J - \hat{f}_J)\right\|_{L^2(\mu_X)}\\ &+ \left\|\frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}\mathbb{X}_{J^c}f^*_{J^c}\right\|_{L^2(\mu_X)} + \left\|\frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}\boldsymbol{\xi}\right\|_{L^2(\mu_X)}.\end{aligned} \tag{1.36}$$

Among these three terms, the first two only require operator norms, whereas the last term $\|\frac{1}{N}\mathbb{X}^\top_{J^c}(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N)^{-1}\boldsymbol{\xi}\|_{L^2(\mu_X)}$ is the most peculiar; we must exploit the randomness of $\boldsymbol{\xi}$. The following proposition is called the "upper side" of the Dvoretzky–Milman theorem; its proof can be found in [P2].

**Assumption 3.** *There exist absolute constants $\gamma_1 \in \left(0, \frac{1}{16}\right)$, $\delta_1 \ge 0$, $\epsilon > 0$ and $\kappa' > 1$ such that*

- $$\mathbb{P}\left(\max_{1\le i\le N}\frac{\left\|\Sigma^{1/2}_{J^c}\phi_{J^c}(X_i)\right\|^2_{\mathcal{H}}}{\operatorname{Tr}(\Sigma^2_{J^c})} \le 1 + \delta_1\right) \ge 1 - \gamma_1, \tag{1.37}$$
- *for any $f \in V_{J^c}$, $\|f\|_{L^{4+\epsilon}(\mu_X)} \le \kappa' \|f\|_{L^2(\mu_X)}$.*

**Proposition 13.** *Suppose Assumption 3 holds. There exist some absolute constants $c_3$ and $C_8 > 0$ such that with a probability of at least $1 - \bar{p}_{DMU}$, where $\bar{p}_{DMU} = \frac{c_3}{N^\epsilon} + \gamma_1$, there holds $\left\|\Sigma^{1/2}_{J^c}\mathbb{X}^\top_{J^c}\right\|_{op} \le C\sqrt{\operatorname{Tr}(\Sigma^2_{J^c})} + C\sqrt{N}\|\Sigma_{J^c}\|_{op}$.*

Therefore, if Proposition 13 and Theorem 4 hold, then from (1.35) we know that there exists an absolute constant $C$ such that on the intersection of the two random events the following holds:

$$\begin{aligned}\left\|\langle X, \frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}\mathbb{X}_J(f^*_J - \hat{f}_J)\rangle\right\|_{L^2(\mu_X)} &\le C\frac{\sqrt{\operatorname{Tr}(\Sigma^2_{J^c})} + \sqrt{N}\|\Sigma_{J^c}\|_{\mathrm{op}}}{Nt^{-1} + \operatorname{Tr}(\Sigma_{J^c})}\|\mathbb{X}_J(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J)\|_2 \text{ and}\\ \left\|\langle X, \frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}\mathbb{X}_{J^c}f^*_{J^c}\rangle\right\|_{L^2(\mu_X)} &\le C\frac{\sqrt{\operatorname{Tr}(\Sigma^2_{J^c})} + \sqrt{N}\|\Sigma_{J^c}\|_{\mathrm{op}}}{Nt^{-1} + \operatorname{Tr}(\Sigma_{J^c})}\|\mathbb{X}_{J^c}f^*_{J^c}\|_2.\end{aligned} \tag{1.38}$$

To bound $\|\frac{1}{N}\mathbb{X}^\top_{J^c}(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N)^{-1}\boldsymbol{\xi}\|_{L^2(\mu_X)}$, we first take expectation over $\boldsymbol{\xi}$, yielding

$$\mathbb{E}_{\boldsymbol{\xi}}\left\|\langle X, \frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}\boldsymbol{\xi}\rangle\right\|^2_{L^2(\mu_X)} \le C\sigma_\xi^2\frac{\sum_{i=1}^N\|\Sigma^{1/2}_{J^c}\phi_{J^c}(X_i)\|^2_{\mathcal{H}}}{(Nt^{-1} + \operatorname{Tr}(\Sigma_{J^c}))^2} \le C'\sigma_\xi^2\frac{N\operatorname{Tr}(\Sigma^2_{J^c})}{(Nt^{-1} + \operatorname{Tr}(\Sigma_{J^c}))^2},$$

holds with high probability.

**Minimum $\|\cdot\|_q$-norm interpolant estimator.** Recall that when $\hat{\boldsymbol{\beta}}$ is the minimum $\|\cdot\|_q$-norm interpolant estimator (Example 10), Proposition 6 states: if $(V_J, V_{J^c})$ is an FSD of the form defined in Proposition 6, then $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]$. This endows $\hat{\boldsymbol{\beta}}_{J^c}$ with a statistical meaning: $\hat{\boldsymbol{\beta}}_{J^c}$ is the minimum $\|\cdot\|_q$-norm interpolant estimator of the residual of $\hat{\boldsymbol{\beta}}_J$. Combined with the Dvoretzky–Milman theorem, this statistical interpretation allows us to obtain the following control on the upper bound for $\|\langle X, \hat{\boldsymbol{\beta}}_{J^c}\rangle\|_{L^2(\mu_X)}$.

**Proposition 14.** *Using the notation of Proposition 6. If $N \le \kappa_{\mathrm{DM}}\varepsilon^2 d_*(\Sigma_{J^c}^{1/2}B_{q'}^p)$, then there exist an absolute constant $C$ such that on the random event $\Omega_{DM,reg}(\varepsilon)$, we have*

$$\|\langle X, \hat{\boldsymbol{\beta}}_{J^c}\rangle\|_{L^2(\mu_X)} \le C\frac{\operatorname{diam}(\Sigma_{J^c}^{1/2}B_q^p)}{\ell_*(\Sigma_{J^c}^{1/2}B_{q'}^p)}\|\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J\|_2.$$

**Proof.** From Proposition 6, we have $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]$, and hence under the Dvoretzky-Milman condition $N \le \kappa_{DM}\varepsilon^2 d_*(\Sigma_{J^c}^{1/2}B_{q'}^p)$ holds, since $\|\Sigma_{J^c}^{1/2}\|_{\ell_q\to\ell_2} = \operatorname{diam}(\Sigma_{J^c}^{1/2}B_q^p)$, on $\Omega_{\mathrm{DM,reg}}(\varepsilon)$ by (1.23), there holds $\|\Sigma_{J^c}^{1/2}\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]\|_2 \le \operatorname{diam}(\Sigma_{J^c}^{1/2}B_q^p)\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]\|_q \le C\frac{\operatorname{diam}(\Sigma_{J^c}^{1/2}B_q^p)}{\ell_*(\Sigma_{J^c}^{1/2}B_{q'}^p)}\|\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J\|_2$. ∎

Proposition 14 states that the energy $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$ of the free part $\hat{f}_{J^c}$ is at most the Euclidean norm of the residual of $\hat{\boldsymbol{\beta}}_J$, scaled by the factor $\frac{\operatorname{diam}(\Sigma_{J^c}^{1/2}B_q^p)}{\ell_*(\Sigma_{J^c}^{1/2}B_{q'}^p)}$. The Euclidean norm of the residual of $\hat{\boldsymbol{\beta}}_J$ can be obtained readily; we do not repeat it here.

**Minimum $\|\cdot\|_2$-norm interpolant classifier.** For binary classification problems, we have a more unified understanding of the energy of $\hat{f}_{J^c}$. The following proposition is proved in [P1], see also Section 4.8.3.

**Proposition 15.** *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.,*

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

**LASSO with support recovery.** As another example of a nonlinear estimator, we consider in this paragraph the case where $\hat{\boldsymbol{\beta}}$ is the LASSO estimator, i.e., $\hat{\boldsymbol{\beta}} \in \operatorname{argmin}(\frac{1}{2N}\|\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1)$. Let $\boldsymbol{\beta}^* \in \mathbb{R}^p$ be an unknown vector and denote its support by $S = \operatorname{supp}(\boldsymbol{\beta}^*)$, i.e., $S = \{j \in [p] : \langle \boldsymbol{\beta}^*, \boldsymbol{e}_j\rangle \neq 0\}$, where we recall that $\boldsymbol{e}_1, \dots, \boldsymbol{e}_p$ is the canonical basis of $\mathbb{R}^p$. Let $s = |S|$ and assume $s \le \lfloor c\frac{N}{\log(p/N)}\rfloor$ for some constant $c < 1$. To highlight the FSD and avoid being distracted by stochastic arguments, we work on the following stochastic event

$$\Omega_{\mathrm{LASSO}} = \left\{\operatorname{supp}(\hat{\boldsymbol{\beta}}) = S,\ \left\|\frac{1}{N}\mathbb{X}_S\mathbb{X}_S^\top\right\|_{\mathrm{op}} \le 10,\ \left\|\frac{1}{N}\mathbb{X}_S^\top\boldsymbol{\xi}\right\|_2 \le \sigma_\xi\sqrt{\frac{2s}{N}}\right\},$$

where $\mathbb{X}_S = [X_{1,S}|\cdots|X_{N,S}]^\top \in \mathbb{R}^{N\times s}$ be the restriction of $\mathbb{X}$ to $S$, and $X_{i,S}$ is the restriction of $X_i$ to $S$. When the event $\operatorname{supp}(\hat{\boldsymbol{\beta}}) = S$ occurs, we say that $\hat{\boldsymbol{\beta}}$ achieves support recovery. Sufficient conditions for this event have been extensively studied in the theory of LASSO, e.g., in [Gir14, Section 5.5.2]. The advantage of working on this stochastic event is that, if we let $J = S$ and $V_J = \operatorname{span}(\{\boldsymbol{e}_j : j \in S\})$, we have $\hat{\boldsymbol{\beta}}_{J^c} = \boldsymbol{\beta}^*_{J^c} = \boldsymbol{0}$, which eliminates the need to consider the energy of $\hat{\boldsymbol{\beta}}_{J^c}$. The case where support recovery does not necessarily occur remains a particularly interesting direction for future research. In this case, we have the following proposition.

**Proposition 16.** *Assume that $\lambda > \sigma_\xi\sqrt{\frac{\log(ep/s)}{N}}$ and $p > e^7 s$. Then on $\Omega_{\mathrm{LASSO}}$, we have $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \ge \frac{1}{20}\sigma_\xi\sqrt{\frac{s\log(ep/s)}{N}}$.*

**Proof.** By the definition of $\hat{\boldsymbol{\beta}}$ and the KKT conditions, we have $\frac{1}{N}\mathbb{X}_S^\top(\mathbb{X}_S\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda\operatorname{sign}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$. Substituting $\boldsymbol{y} = \mathbb{X}_S\boldsymbol{\beta}^*_S + \boldsymbol{\xi}$ and using the fact that $\operatorname{sign}(\hat{\boldsymbol{\beta}}) = \operatorname{sign}(\boldsymbol{\beta}^*_S)$, we obtain $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_S = (\frac{1}{N}\mathbb{X}_S^\top\mathbb{X}_S)^{-1}[\lambda\operatorname{sign}(\boldsymbol{\beta}^*_S) - \frac{1}{N}\mathbb{X}_S^\top\boldsymbol{\xi}]$. Taking the $\ell_2$ norm on both sides and applying the triangle inequality, we have

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 \ge \frac{1}{10}\left(\lambda\left\|\operatorname{sign}(\boldsymbol{\beta}^*_S)\right\|_2 - \sigma_\xi\sqrt{\frac{2s}{N}}\right). \tag{1.39}$$

Given that $\|\operatorname{sign}(\boldsymbol{\beta}^*_S)\|_2 = \sqrt{s}$ and the assumption $\lambda > \sigma_\xi\sqrt{\frac{\log(ep/s)}{N}}$, the proof is complete. ∎

Proposition 16 indicates that when support recovery occurs, the estimation error bound of LASSO is instance optimal for *this* specific $\boldsymbol{\beta}^*$. Specifically, if $X$ is an isotropic random vector, then $(V_J^*, V_{J^c}^*)$ is given by $V_J^* = \mathrm{span}(\{\boldsymbol{e}_j : j \in S\})$, and in this case $\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle\|_{L^2(\mu_X)} \sim r(V_J^*, V_{J^c}^*)$, where $r(V_J^*, V_{J^c}^*) = \sigma_\xi \sqrt{\frac{s \log(ep/s)}{N}}$ with high probability. This result is stronger than minimax optimality, as it does not merely assert the existence of some $\boldsymbol{\beta}^*$ for which the estimation error is no less than this lower bound; instead, the lower bound holds for every $\boldsymbol{\beta}^*$ that satisfies the support recovery condition. In fact, from (1.39), we can observe that $\lambda\|\operatorname{sign}(\boldsymbol{\beta}_S^*)\|_2$ and $\sigma_\xi\sqrt{\frac{2s}{N}}$ correspond respectively to the bias and variance terms (recall that $s = \dim(V_J^*)$) of the estimation subspace in the FSD rate function (cf. (1.41) below). Here, since the projections of both $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}^*$ onto the free subspace are $\mathbf{0}$, there are no bias or variance terms from the free subspace in the rate function.

**Research direction.** When $\hat{f}_{J^c}$ is a nonlinear estimator—as in the case where $\hat{f}_{J^c}$ is the minimum $\|\cdot\|_q$-norm interpolant estimator, LASSO, or a general RERM—how can we develop a systematic mathematical toolkit that allows us to obtain a possibly sharp high-probability upper bound for $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$? Here, a systematic method means one that works for general $\mu_X$ and general $\hat{f}_{J^c}$, and when specialized to ridge regression and spectral methods, it can recover the sharp upper bounds that were obtained using the knowledge that $\hat{f}_{J^c}$ is a linear estimator, [P4]. For (R)ERM, this constitutes a class of problems that have never been explored; it may spur the development of new geometric tools.

### 1.5.4 FSD as a theoretical framework

Through the discussion in Section 1.5.1, Section 1.5.2, and Section 1.5.3, we have explained the role of FSD as an analytic method. In this section we further argue that FSD method also serves as a theoretical framework for understanding the properties of how a solution tackles a specific supervised learning problem, offering theorists a potential new perspective and way of thinking.

In this section we focus on situations where an optimal FSD $(V_{J_*}, V_{J_*^c})$ can be constructed, such that (1.15) can be proved. In such cases, the FSD method reveals how an estimator actually employs the feature space for estimation—which features the estimator uses when solving the problem - and how it uses $V_{J_*^c}$ to handle signal and noise. This differs from the classical statistical-learning-theory goal of "establishing an oracle inequality that matches the minimax lower bound" as in Section 1.1; instead, it focuses more on understanding, from a mathematical perspective, the fitting relationship between the solution and the problem itself. Thus we say that the FSD method introduces a potential new research paradigm into statistical learning theory.

The FSD method, as a theoretical framework, can—like all successful theoretical frameworks—supply the "right" definitions. We illustrate this point by defining the following three concepts, namely:

1. a definition of a pre-order on spectral methods for a given supervised regression problem,
2. a definition of the generalized saturation effect and its necessary and sufficient conditions, and
3. a mathematical definition of the feature learning property as well as the signal-features alignment property.

Definitions of these concepts need to be built upon the study of spectral methods via the FSD method. In [P5] we apply the FSD method to study spectral methods (Example 9) for solving linear regression problems in $\mathbb{R}^p$, that is, we assume $\mathcal{H}$ is identified with $\mathbb{R}^p$ via $f(\cdot) = \langle \cdot, \boldsymbol{\beta}\rangle$; then the spectral method $\hat{f}_N$ is identified with $\hat{\boldsymbol{\beta}}$, and the signal $f^*$ is identified with $\boldsymbol{\beta}^*$.

**Definition 11.** *Recall that $\sigma_1 \geq \cdots$ are the eigenvalues of $\Sigma$. Let $b > 0$ and $t \geq 1$. The estimation dimension of the spectral method $\hat{\boldsymbol{\beta}}$ with filter function $\varphi_t$ is defined as*

$$k^* = k^*_{t^{-1},b} = \min\Big\{k \in [p] : \sigma_{k+1} \leq bt^{-1}\Big\}. \tag{1.40}$$

Let $V_{J_*} = \mathrm{span}(\boldsymbol{e}_j : j \in J_*)$, $J_* = \{1, \cdots, k^*\}$, $(\boldsymbol{e}_j)_j$ are the eigenvectors of $\Sigma$ and $\psi_t$ is the residual function defined by $\psi_t(x) = 1 - x\varphi_t(x)$. Define

$$r(V_{J_*}, V_{J_*^c}) = \left\|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\right\|_2 + \sigma_\xi\sqrt{\frac{|J_*|}{N}} + \left\|\Sigma_{J_*^c}^{1/2}\boldsymbol{\beta}^*_{J_*^c}\right\|_2 + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma^2_{J_*^c})}{N}}, \tag{1.41}$$

where we recall that we denote $\Sigma_{J_*} = P_{V_{J_*}}\Sigma P_{V_{J_*}}$ and $\Sigma_{J_*^c} = P_{V_{J_*^c}}\Sigma P_{V_{J_*^c}}$. The main conclusion of [P5], see also Chapter 3, is that, under general assumptions, the spectral methods $\hat{\boldsymbol{\beta}} = \frac{1}{N}\varphi_t(\hat{\Sigma})\mathbb{X}^\top \boldsymbol{y}$ satisfies the following property with high probability:

$$\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle\|_{L^2(\mu_X)} \sim r(V_{J_*}, V_{J_*^c}). \tag{1.42}$$

From Section 1.5.1, we know that estimation of $\boldsymbol{\beta}^*$ occurs only on $V_{J_*}$, while "absorption of noise" occurs on the free space $V_{J_*^c}$. Equation (1.42) shows that, for any given linear regression problem $(\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$ and tuning parameter $t$, the space $V_{J_*}$ where estimation takes place is determined solely by the spectrum of $\Sigma$ and the tuning parameter, and is independent of the signal $\boldsymbol{\beta}^*$ to be estimated, the eigenvectors of $\Sigma$, and the family of filter functions $(\varphi_t)_{t\geq 1}$. This observation indicates the following facts:

1. Since $V_{J_*}$ is independent of $(\varphi_t)_{t\geq 1}$, we know that for a given linear regression problem, all algorithms in the class of spectral methods (such as ridge regression, gradient descent and gradient flow) decompose the feature space in the same way to estimate the signal. By examining the definition of $r(V_{J_*}, V_{J_*^c})$ in (1.41), we find that only the term $\|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\|_2$ depends on the specific choice of the filter / residual functions. In other words— the only difference in the statistical properties of different spectral methods for a given linear regression problem lies in how close the residual function $\psi_t$ is to 0 on $\{x > 0 : tx > b\}$—the closer it is to 0, the better the statistical properties (i.e., the faster the convergence rate).

2. The FSD approach yields some understanding on the behavior of spectral methods. For instance, the estimation dimension $k^*$ tells us that gradient descent at step $t$ is estimating only the first $k^*$ coordinates of the signal in the basis of eigenvectors of $\Sigma$, no more no less. This means that along the path of gradient descent, there are more and more coordinates (in the eigenbasis of $\Sigma$) that are estimated; the estimation dimension quantifying this phenomenon and the estimation space $V_{J^*}$ localizing the space where this estimation is happening. We may suspect that Newton's method behaves similarly but with an estimation dimension growing faster than the one of gradient descent.

3. Since $k^*$ and $V_{J_*}$ does not depend on the signal, spectral methods do not have the feature learning property that is defined later in Definition 16, meaning that they are not able to design features in order to improve their prediction ability.

### Pre-order of Spectral Methods

Thanks to the FSD method, (1.42) provides matching upper and lower bounds for arbitrary linear regression problem $\mathcal{R} = (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, rather than being restricted to a specific spectrum decay or a particular class of $\boldsymbol{\beta}^*$. Consequently, for any $\mathcal{R}$, comparing the population excess risk of two spectral methods is reduced to comparing two real numbers given by their optimal FSD.

Since a spectral algorithm is uniquely determined by its filter function, we consider two spectral methods $\hat{\boldsymbol{\beta}}^{(A)}_{t_A}$ and $\hat{\boldsymbol{\beta}}^{(B)}_{t_B}$ with parameters $t_A, t_B$ , and with filter functions $\varphi^{(A)}_{t_A}$ and $\varphi^{(B)}_{t_B}$, respectively. By (1.42), there exist $r^{(A)}_{t_A}(V^{(A)}_{J_*}, V^{(A)}_{J_*^c})$ and $r^{(B)}_{t_B}(V^{(B)}_{J_*}, V^{(B)}_{J_*^c})$ characterizing the squared loss population excess risk $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}^{(A)}_{t_A} - \boldsymbol{\beta}^*)\|_2$ and $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}^{(B)}_{t_B} - \boldsymbol{\beta}^*)\|_2$ for these two spectral methods in this linear regression problem. Given any $\mathcal{R} = (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi) \in \mathbb{R}^{p\times p} \times \mathbb{R}^p \times \mathbb{R}$, we define the following pre-order "$\preceq_{\mathcal{R}}$" on the set of all spectral methods.

**Definition 12** (Pre-order of Spectral Algorithms in Linear Regression Problems)**.** *For the linear regression problem $\mathcal{R} := (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, we write*

$$\hat{\boldsymbol{\beta}}^{(A)}_{t_A} \preceq_{\mathcal{R}} \hat{\boldsymbol{\beta}}^{(B)}_{t_B} \quad \text{if} \quad r^{(A)}_{t_A}(V^{(A)}_{J_*}, V^{(A)}_{J_*^c}) = O\Big(r^{(B)}_{t_B}(V^{(B)}_{J_*}, V^{(B)}_{J_*^c})\Big)$$

*as $N$ and $p$ go to infinity. In particular, if $r^{(A)}_{t_A}(V^{(A)}_{J_*}, V^{(A)}_{J_*^c}) = \Theta\Big(r^{(B)}_{t_B}(V^{(B)}_{J_*}, V^{(B)}_{J_*^c})\Big)$, we write $\hat{\boldsymbol{\beta}}^{(A)}_{t_A} \asymp_{\mathcal{R}} \hat{\boldsymbol{\beta}}^{(B)}_{t_B}$. It is straightforward to verify that "$\asymp_{\mathcal{R}}$" defines an equivalence relation on the set of all spectral methods, while $\preceq_{\mathcal{R}}$ defines a pre-order.*

Definition 12 describes, for a specific linear regression problem $\mathcal{R} = (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, the relative speed of convergence of the population excess risk for any two given spectral methods $\hat{\boldsymbol{\beta}}^{(A)}_{t_A}$ and $\hat{\boldsymbol{\beta}}^{(B)}_{t_B}$, thereby characterizing the relative performance of different spectral methods for that problem.

In the following, we consider the case when $t_A = t_B$. Since the choice of $V_{J_*}$ for a given $t \geq 1$ in the optimal decomposition of the feature space given by (1.42) is universal for any spectral algorithm, it follows that, for any fixed $(\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, (1.42) can be applied to any spectral algorithm to obtain the corresponding $r(V_{J_*}, V_{J_*^c})$. In the sense of equality up to a multiplicative constant, the squared loss population excess risk of each spectral algorithm differs only in the bias term $\|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\|_2$ of $\hat{\boldsymbol{\beta}}_{J_*}$. This means that, for any spectral algorithm $\hat{\boldsymbol{\beta}}$, the variance of $\hat{\boldsymbol{\beta}}_{J_*}$ and both the bias and variance of $\hat{\boldsymbol{\beta}}_{J_*^c}$ are identical - the only difference lies in the convergence rate of $\hat{\boldsymbol{\beta}}_{J_*}$ used to estimate $\boldsymbol{\beta}^*_{J_*}$. Therefore, we have the following corollary.

**Corollary 2.** *Given any linear regression problem $\mathcal{R} = (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$. For any $t \geq 1$, $\hat{\boldsymbol{\beta}}_t^{(A)} \preceq_{\mathcal{R}} \hat{\boldsymbol{\beta}}_t^{(B)}$ if and only if as $N$ and $p$ go to infinity*

$$\left\|\Sigma_{J_*}^{\frac{1}{2}}\psi_t^{(A)}(\Sigma)\boldsymbol{\beta}^*_{J_*}\right\|_2 = O\left(\left\|\Sigma_{J_*}^{\frac{1}{2}}\psi_t^{(B)}(\Sigma)\boldsymbol{\beta}^*_{J_*}\right\|_2\right).$$

The following corollary is a direct consequence of the elementary inequality $\exp(-tx) \leq 1/(1+xt)$.

**Corollary 3** (GF outperforms Ridge)**.** *For any linear regression problem, $\varphi_t^{(\mathrm{GF})} \preceq_{\mathcal{R}} \varphi_t^{(\mathrm{Ridge})}$, where $\varphi_t^{(\mathrm{Ridge})}(x) = \frac{1}{x+t^{-1}}$ is the filter function of ridge regression,; while $\varphi_t^{(\mathrm{GF})}(x) = \frac{1-\exp(-tx)}{x}$ is the filter function of gradient flow.*

### Generalized saturation effect

For a fixed parameter $t$, the difference in population excess risk between different spectral methods arises from the structure of their residual function $\psi_t$, and this naturally leads to the saturation effect – the cause of the saturation effect also lies in the properties of the residual function. We first introduce the following generalized definition.

**Definition 13** (Generalized Saturation Effect)**.** *For any linear regression problem $\mathcal{R}$, any interval $I \subset [1, +\infty)$ and families of filter functions $\{\varphi_t^{(A)}\}_{t\geq 1}$ and $\{\varphi_t^{(B)}\}_{t\geq 1}$, we write $\{\varphi_t^{(A)}\}_{t\in I} \preceq_{\mathcal{R}} \{\varphi_t^{(B)}\}_{t\in I}$ if as $N$ and $p$ go to infinity*

$$\inf\left(r_{t_A}^{(A)}(V_{J_*}, V_{J_*^c}) : t_A \in I\right) = O\left(\inf\left(r_{t_B}^{(B)}(V_{J_*}, V_{J_*^c}) : t_B \in I\right)\right).$$

*If $\{\varphi_t^{(A)}\}_{t\in I} \preceq_{\mathcal{R}} \{\varphi_t^{(B)}\}_{t\in I}$, we say that the spectral algorithm $\hat{\boldsymbol{\beta}}^{(B)}$ defined by the filter function family $\{\varphi_t^{(B)}\}_{t\geq 1}$ is saturated compared to the filter function family $\{\varphi_t^{(A)}\}_{t\geq 1}$ in $I$. In particular, if $I = \mathbb{R}_+$, we write $\{\varphi_t^{(A)}\}_{t\geq 1} \preceq_{\mathcal{R}} \{\varphi_t^{(B)}\}_{t\geq 1}$ and say that the spectral algorithm $\hat{\boldsymbol{\beta}}^{(B)}$ defined by the filter function family $\{\varphi_t^{(B)}\}_{t\geq 1}$ is saturated compared to the filter function family $\{\varphi_t^{(A)}\}_{t\geq 1}$. It is straightforward to verify that $\preceq_{\mathcal{R}}$ is a pre-order. Similarly, we can define an equivalence relation $\asymp_{\mathcal{R}}$ on families of filter functions. When big-O is replaced by small-o, we denote by $\prec_{\mathcal{R}}$ .*

Definition 12 describes the relative performance of two spectral methods for given parameters $t_A$ and $t_B$, whereas Definition 13 concerns their relative performance under their respective optimal parameters within interval $I$. It is easy to see that the classical saturation effect defined in [BPR07] corresponds to the pre-order on the following set of linear regression problems.

$$\mathcal{R} \in \mathfrak{R}_{\mathrm{Sob}}(s, \alpha) := \Bigg\{(\Sigma, \boldsymbol{\beta}^*, \sigma_\xi) : \Sigma = \sum_{j=1}^{p} \sigma_j \boldsymbol{e}_j \otimes \boldsymbol{e}_j,\ \sigma_j \sim j^{-\alpha},\ \|\Sigma^{\frac{1-s}{2}}\boldsymbol{\beta}^*\|_2 < \infty, \sigma_\xi \text{ is constant}\Bigg\}.$$

Moreover, in [BPR07], $\{r_t^{(B)}\}_{t\geq 1}$ is the family of ridge regression. In addition, on $\mathfrak{R}_{\mathrm{Sob}}(s, \alpha)$, the optimal tuning parameter is $t^{-1} \sim N^{-\frac{\alpha}{1+\tilde{s}\alpha}}$, where $\tilde{s} = s \wedge \tau$ and $\tau = 2$ for ridge regression, $\tau = \infty$ for gradient flow. We say this choice is optimal, because it achieves the minimax rate on $\mathfrak{R}_{\mathrm{Sob}}$, [LZL23]. Applying to $\varphi_t^{(A)} : x \mapsto (1 - \exp(-tx))/x$, i.e., gradient flow, and to $\varphi_t^{(B)} : x \mapsto (x + t^{-1})^{-1}$, i.e., ridge regression, we have the following. For the same $t \sim N^{\frac{\alpha}{1+\tilde{s}\alpha}}$, [P2] computed that $\|\Sigma_{J_*}^{1/2}\psi_t^{(B)}(\Sigma)\boldsymbol{\beta}^*_{J_*}\|_2 \sim N^{-\frac{\alpha(s\wedge 2)}{1+\alpha(s\wedge 2)}}$, while the following corollary yields $\|\Sigma_{J_*}^{1/2}\psi_t^{(A)}(\Sigma)\boldsymbol{\beta}^*_{J_*}\|_2 \sim N^{-\frac{\alpha s}{1+\alpha s}}$. Combined with Corollary 2, this recovers the classical saturation effect in the sense of [BPR07]. The proof of Corollary 4 may be found in Section 3.5.1, see also [P5].

**Corollary 4** (Saturation Effect in Sobolev Space). *Let* $\varphi_t^{(\mathrm{GF})} : x \mapsto (1-\exp(-tx))/x$ *and* $\varphi_t^{(\mathrm{Ridge})} : x \mapsto (x+t^{-1})^{-1}$. *Let* $\mathcal{R} \in \mathfrak{R}_{\mathrm{Sob}}(s,\alpha)$. *We have* $\{\varphi^{(\mathrm{GF})}\}_{t\geq 1} \preceq_{\mathcal{R}} \{\varphi^{(\mathrm{Ridge})}\}_{t\geq 1}$. *Moreover, when* $t^{-1} \sim N^{-\frac{\alpha}{1+\tilde{s}\alpha}}$, *where* $\tilde{s} = s \wedge 2$ *for ridge regression, and* $\tilde{s} = s$ *for gradient flow, we have* $(r_t^{(\mathrm{GF})}(V_{J_*}, V_{J_*^c}))^2 \sim N^{-\frac{\alpha s}{1+s\alpha}}$ *and* $(r_t^{(\mathrm{Ridge})}(V_{J_*}, V_{J_*^c}))^2 \sim N^{-\frac{\alpha \tilde{s}}{1+\tilde{s}\alpha}}$.

Here, however, we offer a geometric perspective on the classical saturation effect: its occurrence is due to the fact that, on $V_{J_*}$, the residual function of ridge regression decays too slowly in the eigen-basis with power decay, compared to the residual function of gradient flow. We emphasize that Corollary 2 provides not only this most classical example of the saturation effect in Sobolev spaces, but also necessary and sufficient conditions for the occurrence of more general saturation effects.

**Corollary 5** (Saturation effect in the plateau covariance model). *Suppose there exists some* $k \lesssim N \lesssim p-k$, $\sigma > \varepsilon > 0$ *such that* $\sigma_1 = \cdots = \sigma_k = \sigma$, *and* $\sigma_{k+1} = \cdots = \sigma_p = \varepsilon$. *Let* $J = \{1, \cdots, k\}$ *and suppose there exists a real number* $\alpha_* > 0$ *such that* $|\langle \boldsymbol{\beta}^*, \boldsymbol{e}_j \rangle| = \alpha_*$ *for any* $j \in J$ *while* $\langle \boldsymbol{\beta}^*, \boldsymbol{e}_j \rangle = 0$ *otherwise. Let*

$$\mathrm{SNR} = \frac{\|\Sigma^{1/2}\boldsymbol{\beta}^*\|_2}{\sigma_\xi} \frac{\sigma\sqrt{N}}{\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)}}.$$

*Suppose* $4 < \mathrm{SNR} \leq b\frac{\sigma}{\varepsilon}$, *where* $b$ *is from* (1.40). *Let* $I = \{t > 1 : b^{-1}\varepsilon \leq t^{-1} < \sigma\}$. *Then*

$$\min_{t\in I} r^{(\mathrm{GF})}(V_{J_*}, V_{J_*^c}) \leq \min_{t\in I} r^{(\mathrm{Ridge})}(V_{J_*}, V_{J_*^c}).$$

*Moreover, when* $\mathrm{SNR} \to \infty$ *and* $\sigma = \Omega(\varepsilon)$, $\{\varphi_t^{(\mathrm{Ridge})}\}_{t\in I} \prec_{\mathcal{R}} \{\varphi_t^{(\mathrm{GF})}\}_{t\in I}$.

The proof of Corollary 5 may be found in Section 3.5.2, see also [P5]. The quantity SNR in Corollary 5 can be interpreted as a signal-to-noise ratio, but it is rescaled according to the sample size and the spectrum of $\Sigma$. The lower bound in the condition $4 < \mathrm{SNR} \leq b\frac{\sigma}{\varepsilon}$ is intended to ensure that the signal-to-noise ratio is not too small, while the upper bound is rather mild. For example, if we take $\sigma = 1$, $\varepsilon = (p-k)^{-1}$, and $\|\Sigma_{J_*}^{1/2}\boldsymbol{\beta}^*\|_2/\sigma_\xi$ to be a constant, then this condition is satisfied. This corollary considers the case where the signal $\boldsymbol{\beta}^*$ is well aligned with the covariance structure, and shows that the saturation effect occurs over a rather broad range of tuning parameters (which is reasonable, since the tuning parameter is neither too large, causing overfitting, nor too small, leading to underfitting). This illustrates our claim that the saturation effect is a fairly general phenomenon in linear regression problems.

### FSD for defining feature learning property

The following definition provides a mathematical definition, in the realm of deep-learning theory, of what is called feature learning property (see Section 1.4.3). In the following definition, the top-$k$ eigenvectors involved in the alignment property are specifically related to the understanding gained from studying spectral methods via the FSD method—namely, that spectral methods learn using the eigenvectors (features) corresponding to the $k$ largest eigenvalues, see Definition 11. Roughly speaking, FSD itself provides how an estimator utilizes features in the feature space, while the feature learning property focuses on how features that are beneficial for solving a specific supervised learning problem are implicitly constructed and used by estimators, particularly neural networks. Therefore, the role of FSD in the definition of the feature learning property is to help us understand how these features constructed by neural networks are related to the specific problem.

**Definition 14** (Alignment Property). *Let* $(\mu_X, f^*, \xi)$ *be a supervised regression problem, let* $\mathcal{H}$ *be an RKHS, and let* $\hat{g}_N$ *be an estimator taking values in* $\mathcal{H}$. *Let* $g_{\mathcal{H}}^* \in \mathrm{argmin}\{\|g - f^*\|_{L^2(\mu_X)} : g \in \mathcal{H}\}$ *denote the oracle in* $\mathcal{H}$. *Given a monotonically increasing function* $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, *a sequence of non-negative real numbers* $\{\gamma_j\}_{j=k+1}^\infty$ *and a real number* $0 < \delta < 1$, *we say that* $\hat{g}_N$ *satisfies the* $(\Phi, k, \delta)$ *alignment property with respect to* $\{\gamma_j\}_{j=k+1}^\infty$ *if with probability at least* $1-\delta$, *there holds* $\|\hat{g}_N - g_{\mathcal{H}}^*\|_{L^2(\mu_X)}^2 \leq \Phi(\sum_{j>k}^\infty \gamma_j \langle g_{\mathcal{H}}^*, \boldsymbol{e}_j \rangle^2)$, *where* $\langle g_{\mathcal{H}}^*, \boldsymbol{e}_j \rangle$ *is the inner product in* $\mathcal{H}$ *between* $g_{\mathcal{H}}^*$ *and* $\boldsymbol{e}_j$.

An estimator $\hat{g}_N$ satisfying the alignment property has the following characteristic: when the oracle $g_{\mathcal{H}}^*$ is well aligned with the eigenvectors (functions) corresponding to the largest $k$ eigenvalues of the covariance operator $\Sigma$, the estimator can exploit this structure and thereby achieve a smaller estimation error $\|\hat{g}_N - g_{\mathcal{H}}^*\|_{L^2(\mu_X)}^2$. Many estimators are known to satisfy this property, including ridge regression, gradient flow, gradient descent, principal components regression, and RERM with regularization terms whose Bregman divergence is non-trivial; see Theorem 1. For

general RERM, the weights $\gamma_j$ may be chosen as 1; for ridge regression gradient flow, gradient descent, and principal components regression, $\gamma_j$ may be chosen as $\sigma_j$, see (1.41).

The alignment property characterizes the ability of an estimator to exploit an existing favorable geometric alignment; however, such an alignment is not always inherently present. When this alignment relationship is suboptimal, certain estimators yield large prediction errors. To investigate the impact of the unfavorable alignment between the oracle $g^*_{\mathcal{H}}$ and the eigenfunctions of the covariance operator of $\mathcal{H}$ on the estimation error of a broad class of estimators, we introduce the following concept of the efficiency of alignment.

**Definition 15.** *Let $(\mu_X, f^*, \xi)$ be a supervised regression problem, $\sigma_\xi^2$ be the variance of $\xi$, $\mathcal{H}$ an RKHS, and $N$ be the sample size. Let $0 < \alpha < 1$ be a pre-specified threshold. We define the balance dimension $k^\circ(N) = \min\{k \in \mathbb{N} : \|P_{k+1:\infty} g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)} \le \sigma_\xi^2 \frac{k}{N}\}$, where $P_{k+1:\infty} = \sum_{j>k} \boldsymbol{e}_j \otimes \boldsymbol{e}_j$ is the projection onto $\operatorname{span}(\boldsymbol{e}_j : j > k)$. If $k^\circ(N) \le \alpha N$, we say that alignment is efficient; if $k^\circ(N) > \alpha N$, we say that alignment is deficient.*

Definition 15 is intended to characterize the relationship between the balance dimension and the sample size $N$. Based on the FSD, the projection $P_{k+1:\infty} g^*_{\mathcal{H}}$ is not estimated and thus enters the estimation error as the price for no estimation. Meanwhile, $\sigma_\xi^2 \frac{k}{N}$ represents the variance term associated with the estimation subspace. Since both terms are independent of the specific choice of the estimator, the balance dimension—by describing the equilibrium between these two components—reveals the impact on the estimation error that depends exclusively on the alignment of $g^*_{\mathcal{H}}$ with the eigenvectors of the covariance operator of $\mathcal{H}$, regardless of the specific algorithm.

In the rate function of FSD, the terms $\|P_{k+1:\infty} g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)}$ and $\sigma_\xi^2 \frac{k}{N}$ are universal, meaning that they are independent of the specific choice of the estimator and always appear in the rate function. Consequently, the balance between these two terms characterizes, in an estimator-free sense, the impact of the alignment between $g^*_{\mathcal{H}}$ and the eigenvectors of the covariance operator on the estimation error of any estimator. In particular, by the definition of $k^\circ(N)$, for any $k \in \mathbb{N}$ (especially for the estimation dimension $k^*$), we always have $\|P_{k^\circ+1:\infty} g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)} + \sigma_\xi^2 \frac{k^\circ}{N} \le 2(\|P_{k+1:\infty} g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)} + \sigma_\xi^2 \frac{k}{N})$. This implies that the optimal rate function of FSD is always subject to a lower bound provided by $\sigma_\xi^2 \frac{k^\circ}{N}$, regardless of which estimator is selected for the FSD. When the balance dimension is excessively large, this lower bound can be quite substantial, leading to a suboptimal estimation error. A key element of the feature learning property introduced in the sequel is that feature learning can automatically construct favorable geometric alignments through the autonomous learning of features.

**Definition 16** (Alignment property and Feature Learning property)**.** *Consider a real-valued supervised regression problem $(\mu_X, f^*, \xi)$. Let $\hat{f}_N$ be an estimator. We say that $\hat{f}_N$ performs feature learning property in solving $(\mu_X, f^*, \xi)$, if there exist a RKHS $\mathcal{H}_{\text{fea}}$ with its canonical feature map denoted by $\phi_{\text{fea}} : \Omega_X \to \mathcal{H}_{\text{fea}}$, and an element $\hat{g}_N \in \mathcal{H}_{\text{fea}}$ such that the following conditions hold under the limit when $N \to \infty$:*

1. *$\hat{g}_N$ satisfies the alignment property with estimation dimension $k$ for some $k \in \mathbb{N}_+$;*
2. *$\hat{f}_N(\cdot) = \hat{g}_N(\phi_{\text{fea}}(\cdot))$ where $\hat{g}_N(\phi_{\text{fea}}(\cdot)) = \langle \hat{g}_N, \phi_{\text{fea}}(\cdot)\rangle_{\mathcal{H}_{\text{fea}}}$;*
3. *the oracle $g^*_{\mathcal{H}_{\text{fea}}} \in \operatorname{argmin}\{\|g - f^*\|_{L^2(\mu_X)} : g \in \mathcal{H}_{\text{fea}}\}$ satisfies that $\|f^* - g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)} = o_{\mathbb{P}}(1)$;*
4. *$k = O(d)$, and $\|P_{k+1:\infty} g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)} = o_{\mathbb{P}}(1)$ where $P_{k+1:\infty} = \sum_{j>k} \boldsymbol{e}_j \otimes \boldsymbol{e}_j$.*

*We refer to such $\mathcal{H}_{\text{fea}}$ as the learned feature subspace.*

The meaning of Definition 16 is the following: if $\hat{f}_N$ constructs from the training samples a feature subspace $\mathcal{H}_{\text{fea}}$ (which we assume to be an RKHS) such that, for the real-valued regression problem $(\mu_X, f^*, \xi)$, the approximation error of the constructed feature space $\mathcal{H}_{\text{fea}}$, $\|f^* - g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)}$, is small, and within this feature subspace $\mathcal{H}_{\text{fea}}$, those features that are helpful for estimating $f^*$ are indeed used to estimate $f^*$. In other words: the feature engineering capability of $\hat{f}_N$ on this regression problem is manifested by the data-constructed feature subspace $\mathcal{H}_{\text{fea}}$, which possesses small approximation error, and the constructed top $k$ features are beneficial for estimating $f^*$ in the sense that on $\mathcal{H}_{\text{fea}}$ there exists a latent estimator $\hat{g}_N$ that is able to utilize the top-$k$ eigenvectors most important for achieving a small estimation error in order to estimate the oracle $g^*_{\mathcal{H}_{\text{fea}}}$. Finally, $\hat{f}_N$ is close to this latent estimator $\hat{g}_N$ in $L^2(\mu_X)$ distance.

Consequently, the phenomenon observed by both practitioners and theorists when $\hat{f}_N$ solves the problem $(\mu_X, f^*, \xi)$ is the following: based on the training data $(X_i, Y_i)_{i=1}^N$, $\hat{f}_N$ appears to automatically construct, within $L^2(\mu_X)$, a feature space $\mathcal{H}_{\text{fea}}$ that possesses favorable statistical properties for the problem, learns "within" this space, and achieves a small estimation error — as if $\hat{f}_N$ itself were an estimator defined on $\mathcal{H}_{\text{fea}}$. This phenomenon is precisely what is often referred to as the feature learning property in the theory of deep learning.

Therefore, if $\hat{f}_N$ possesses the feature learning property when solving $(\mu_X, f^*, \xi)$, then its estimation error (up to square) can be bounded as follows:

$$\|\hat{f}_N - f^*\|_{L^2(\mu_X)} \leq \|\hat{f}_N - \hat{g}_N\|_{L^2(\mu_X)} + \|\hat{g}_N - g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)} + \|g^*_{\mathcal{H}_{\text{fea}}} - f^*\|_{L^2(\mu_X)}.$$

The requirement in Definition 16 that $\|\hat{f}_N - \hat{g}_N\|_{L^2(\mu_X)}$ be small cannot be dropped, because under very broad conditions, for any $f^* \in L^2(\mu_X)$, there always exists a deterministic RKHS $\mathcal{H}$ (with feature map denoted by $\phi$) such that there is a $\hat{g}_N$ possessing the alignment property for which $\|f^* - g^*_{\mathcal{H}}\|_{L^2(\mu_X)}$ is small, where $g^*_{\mathcal{H}} : \boldsymbol{x} \in \Omega_X \mapsto \mathbb{E}[Y \mid \phi(\boldsymbol{x})]$ is the regression function in $\mathcal{H}$. In fact, spectral methods with analytic filter functions always possess the alignment property, and many RKHS $\mathcal{H}$ are dense in $L^2(\mu_X)$. Therefore, what Definition 16 emphasizes is that such an RKHS must be dependent with $(X_i, Y_i)_{i=1}^N$ as well as to $(\mathcal{F}, \hat{f}_N)$ (which is why we denote it by $\mathcal{H}_{\text{fea}}$), and such that $\hat{g}_N$ is close to $\hat{f}_N$ in the $L^2(\mu_X)$ metric. Only then does this latent feature subspace and the latent estimator $\hat{g}_N$ on it become capable of explaining the feature learning property of $\hat{f}_N$ for the given problem.

# Chapter 2

# A Geometrical Analysis of Kernel Ridge Regression

## 2.1 Introduction

We focus on regression problems in the context of kernel learning. Let $\lambda \geq 0$ be a tuning parameter and $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be some Reproducing Kernel Hilbert Space (RKHS) containing functions from the probability space $(\Omega, \mu)$ to $\mathbb{R}$. Given $N$ independent design vectors $(X_i)_{i=1}^N$ distributed as the unknown probability measure $\mu$ and associated responses $(Y_i)_{i=1}^N \subset \mathbb{R}$, let the Kernel Ridge Regression (KRR) estimator be

$$\hat{f}_\lambda \in \operatorname*{argmin}_{f \in \mathcal{H}} \left( \sum_{i=1}^N \left(f(X_i) - Y_i\right)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right). \tag{2.1}$$

KRR is a highly effective and adaptable method utilized in various domains, including finance, biology, natural language processing, image analysis and partial differential equations, [STC04, RW05, SC08, SS16]. Additionally, it serves as a tool for developing mathematical foundations for deep neural networks, [JGH18, BMM18]. This paper presents general bounds for the estimation error of KRR, aiming to investigate various phenomena in KRR and provide insights in the field of statistical deep learning.

### 2.1.1 Notation

We use $c, c_0, c_1, \cdots, C, C_0, C_1, \cdots$ to denote absolute constants. Usually, $C$ stands for large but finite constants and $c$ stands for small but non-zero constants. Such constants are always assumed to be positive, but may change from one instance to another. Given two quantities $A, B$, we write $A \lesssim B$ (or $A \gtrsim B$) if there exists an absolute constant $C$ such that $A \leq CB$ (or $A \geq CB$). If a constant is assumed to depend on some parameter (say $K$), we use the expression $C_K$, and write $A \lesssim_K B$ (or $A \gtrsim_K B$) if there exists a constant $C_K$ such that $A \leq C_K B$ (or $A \geq C_K B$). We write $A \sim B$ if $B \lesssim A \lesssim B$. We use $O_d(\cdot)$ (respectively $o_d(\cdot)$) for the big-O (respectively small-O) notation, where $d$ emphasizes the asymptotic variable. Further, if $g(d) = O_d(f(d))$, we write $f = \Omega_d(g)$ and if $g(d) = o_d(f(d))$, we write $f = \omega_d(g)$. Given a sequence of random variables $(Z_d)_d$ and a deterministic sequence $(s_d)_d$, we say $Z_d = o_{d,\mathbb{P}}(s_d)$ if $Z_d/s_d \to 0$ in probability.

Given $r \in \mathbb{N}_+$, we define $[r] := \{1, 2, \cdots, r\}$. Let $(\Omega, \mu)$ be a probability space and, for $q \in \mathbb{N}_+$, let $L_q(\Omega, \mu)$ be the $L_q$ space with the norm $\|f\|_{L_q} = \left(\int_\Omega |f(x)|^q\right)^{1/q} d\mu(x)$. When there is no ambiguity, we abbreviate $L_q(\Omega, \mu)$ to $L_q(\mu)$ or $L_q$. Given a random variable $X_1$, we write $\mathbb{E}_{X_1}$ for the conditional expectation with respect to $X_1$ conditionally on all other random variables. Given probability measures $\mu_1, \mu_2$, we denote by $\mu_1 \times \mu_2$ the product probability measure. We say a real-valued random variable $X$ is sub-Gaussian, if $\|X\|_{\psi_2} := \inf\left(t > 0 : \mathbb{E}\exp\left(X^2/t^2\right) \leq 2\right) < \infty$. We write $\mathcal{N}(0,1)$ for the standard Gaussian random variable, and $\mathcal{N}(0, I_d)$ for the standard Gaussian random vector in $\mathbb{R}^d$.

Given two Hilbert spaces $\mathcal{H}, \mathcal{G}$, we characterize $\mathcal{H} \otimes \mathcal{G}$ by defining $f \otimes g \in \mathcal{H} \otimes \mathcal{G}$ as the mapping $(f \otimes g) : h \in \mathcal{G} \mapsto \langle g, h\rangle_{\mathcal{G}} f \in \mathcal{H}$. We use angle brackets $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ to denote an inner product in some Hilbert space $\mathcal{H}$, and omit it when $\mathcal{H}$ is a Euclidean space. We use $\|\cdot\|_{\mathcal{H}}$ to denote the Hilbert norm, and denote the Euclidean norm by $\|\cdot\|_2$. Given a

bounded linear operator $T : \mathcal{H} \to \mathcal{G}$, we denote by $\|T\|_{\mathrm{op},\mathcal{H}\to\mathcal{G}}$ the operator norm of $T$, that is,

$$\|T\|_{\mathrm{op},\mathcal{H}\to\mathcal{G}} = \sup\left(\|Tx\|_{\mathcal{G}} : \|x\|_{\mathcal{H}} \le 1\right).$$

When there is no ambiguity, we abbreviate the operator norm as $\|T\|_{\mathrm{op}}$. For any ONB $(\varphi_j)_{j\in\mathbb{N}}$ of $\mathcal{H}$, if $\sum_{j\in\mathbb{N}} \|T\varphi_j\|_{\mathcal{G}}^2 < \infty$, we say $T$ is a Hilbert-Schmidt operator from $\mathcal{H}$ to $\mathcal{G}$, and denote

$$\|T\|_{HS,\mathcal{H}\to\mathcal{G}} = \sqrt{\sum_{j\in\mathbb{N}} \|T\varphi_j\|_{\mathcal{G}}^2}.$$

One can prove that the HS norm of $T$ is independent of the choice of ONB, for instance, [Pis89, pp.7]. When there is no ambiguity, we abbreviate the Hilbert-Schmidt norm of $T$ to $\|T\|_{HS}$. For any $j \in \mathbb{N}$, let $\mathrm{He}_j(x)$ be the $j$-th (probabilist) Hermite polynomial, [Pis89, pp.16]. For a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$, we set $B_{\mathcal{H}} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1\}$ and $S_{\mathcal{H}} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} = 1\}$.

### 2.1.2 Reproducing Kernel Hilbert Spaces

In this section, essential RKHS background knowledge is presented. Readers who are not familiar with RKHSs can think of $\phi(X)$ in the upcoming discussion as a Gaussian random vector (which is precisely the essence of the Gaussian Equivalence Property).

**Structural aspect** Let $\Omega \subset \mathbb{R}^d$ be a compact Hausdorff space[1] and $\mu$ a probability measure on $\Omega$. Let $L_2(\Omega, \mu)$ be the space of real-valued, square-integrable functions with respect to $\mu$. Suppose $K : \Omega \times \Omega \to \mathbb{R}$ is a positive definite continuous function, and without loss of generality, we assume that $\|K\|_\infty \le 1$[2]. In [P2, Section 5.3], we present a crucial example of an RKHS that does not satisfy $\|K\|_{L_\infty(\mu\times\mu)} < \infty$. Nevertheless, our analysis remains valid. We say that a Hilbert space $\mathcal{H} \subset L_2(\mu)$ of functions is an RKHS (with RKHS inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and associated RKHS norm $\|\cdot\|_{\mathcal{H}}$) over $\Omega$ if for every $x \in \Omega$, there exists a constant $C_x > 0$ (depending on $x$), such that $|f(x)| \le C_x \|f\|_{\mathcal{H}}$ for every $f \in \mathcal{H}$, that is, the evaluation functional $\mathrm{ev}_x : f \mapsto f(x)$, $\mathrm{ev}_x : \mathcal{H} \to \mathbb{R}$ is a bounded linear functional. By Riesz's representation theorem for Hilbert spaces, this is equivalent to saying that the inner product of $\mathcal{H}$ can be characterized as follows: for every $x \in \Omega$, there exists $K(x,\cdot) \in \mathcal{H}$ such that $\langle f, K(x,\cdot)\rangle_{\mathcal{H}} = f(x) = \langle f, \phi(x)\rangle_{\mathcal{H}}$. This is called the *Reproducing Property*. Given a kernel $K$, the (canonical) feature map is defined as $\phi : x \mapsto K(x,\cdot) \in \mathcal{H}$. The RKHS can be considered as a linear model on $\mathcal{H}$, where the design vector $X \in \Omega \subset \mathbb{R}^d$ is embedded into $\mathcal{H}$ via the feature map $\phi$; hence $\phi(X)$ plays the role of a design vector and may therefore be called the RKHS design vector.

By mapping $X$ from $\mathbb{R}^d$ to $\mathcal{H}$, we clarify the prediction function, covariance structure and the design matrix. Given that $\phi(X)$ now serves as the design vector, its integral operator $\Gamma = \mathbb{E}\left[\phi(X) \otimes \phi(X)\right]$ (i.e. $\Gamma : f \in \mathcal{H} \to \mathbb{E}\left[\phi(X)\langle\phi(X), f\rangle_{\mathcal{H}}\right] \in \mathcal{H}$) will play the role of the covariance matrix. We denote the operator norm of $\Gamma : \mathcal{H} \to \mathcal{H}$ by $\|\Gamma\|_{\mathrm{op}}$. Since $\|K\|_\infty \le 1$, $\Gamma$ is compact and positive semi-definite, so it has a discrete spectrum of non-negative eigenvalues. By Mercer's theorem, see, for instance [Wai19, Theorem 12.20] or [RW05, section 4.3], for any $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, it holds that $K(\boldsymbol{x},\boldsymbol{y}) = \langle\phi(\boldsymbol{x}), \phi(\boldsymbol{y})\rangle_{\mathcal{H}} = \sum_{j\ge1} \varphi_j(\boldsymbol{x})\varphi_j(\boldsymbol{y}) = \sum_{j\ge1} \sigma_j h_j(\boldsymbol{x}) h_j(\boldsymbol{y})$ where $(\varphi_j)_{j\in\mathbb{N}^*}$ are eigenfunctions of $\Gamma$, and $(\sigma_j)_{j\ge1}$ are the eigenvalues of $\Gamma$ associated to $(\varphi_j)_{j\ge1}$, and where $\varphi_j = \sqrt{\sigma_j} h_j$. It is possible to check that this decomposition is unique. By [Lax02, section 30.5], $\Gamma$ is a trace-class operator, that is, $\mathrm{Tr}\,(\Gamma) < \infty$ where $\mathrm{Tr}\,(\Gamma) = \mathbb{E}\,\|\phi(X)\|_{\mathcal{H}}^2 = \mathbb{E}K(X,X) = \sum_{j\in\mathbb{N}^*} \sigma_j$.

It is also widely-used to embed $\mathcal{H}$ into $\ell_2$ by $\phi : \boldsymbol{x} \in \Omega \mapsto \sum_{j=1}^\infty \sqrt{\sigma_j} h_j(\boldsymbol{x}) \boldsymbol{e}_j$ (we use the same notation as for the feature map $\phi : x \to K(x,\cdot)$ introduced above; however, which definition of $\phi(x)$ we use is clear from the context). Therefore for every $\boldsymbol{v} \in \ell_2$, the corresponding element of $\mathcal{H}$ is $f_{\boldsymbol{v}}(\boldsymbol{x}) = \langle\phi(\boldsymbol{x}), \boldsymbol{v}\rangle_{\ell_2}$. Therefore, $\mathcal{H}$ can be defined equivalently as the image of $\ell_2$ under the map $\boldsymbol{v} \mapsto f_{\boldsymbol{v}}$, with the inner product $\langle f_{\boldsymbol{v}}, f_{\boldsymbol{u}}\rangle_{\mathcal{H}} = \langle\boldsymbol{v},\boldsymbol{u}\rangle_{\ell_2}$ when $\sigma_j > 0$ for all $j$ (in case $\Gamma$ is of rank $r$, $\mathcal{H}$ is in a one-to-one correspondence with $\ell_2^r$).

We will frequently use a decomposition of $\mathcal{H}$ so it is necessary to introduce the following notation. Given $p < q \in \mathbb{N}\cup\{\infty\}$ and $k \in \mathbb{N}\cup\{\infty\}$, we set $\Gamma_{p:q} = \sum_{p\le j\le q} \sigma_j \varphi_j \otimes \varphi_j$. Then, $\Gamma$ can be decomposed as $\Gamma = \Gamma_{1:k} + \Gamma_{k+1:\infty}$. We set $\mathcal{H}_{p:q} = \mathrm{span}\,(\varphi_j : p \le j \le q)$ and let $P_{p:q}$ be the orthogonal projection onto $\mathcal{H}_{p:q}$. For any $f \in \mathcal{H}$, denote $P_{p:q} f$ as $f_{p:q}$, for example, $\phi_{p:q}(X) = \sum_{p\le j\le q} \langle\phi(X), \varphi_j\rangle_{\mathcal{H}} \varphi_j$. Consequently, we decompose $\mathcal{H} = \mathcal{H}_{1:k} \oplus^\perp \mathcal{H}_{k+1:\infty}$. Given $\iota \in \mathbb{N}_+$, we denote $f^*_{\le\iota} := f^*_{1:\sum_{l\le\iota} d^l}$, $f^*_{>\iota} := f^* - f^*_{\le\iota}$, $\Gamma_{\le\iota} := \Gamma_{1:\sum_{l\le\iota} d^l}$ and $\Gamma_{>\iota} := \Gamma - \Gamma_{\le\iota}$, where we recall

[1] We emphasize that we do not actually require $\Omega$ to be compact; it suffices for $K$ to have a spectral decomposition.

[2] We remark that we do not really need $\|K\|_\infty \le 1$, but need only the integral operator $\Gamma$ defined below to be a positive, compact, symmetric trace-class operator.

that $d$ is the dimension of the design vector $X$. Also, let $P_{\leq l}$ be the projection onto the eigen-space of $\Gamma_{\leq l}$, and $P_{>l}$ its complement.

By the Reproducing Property, for any $f \in \mathcal{H}$, it holds that

$$\|f\|_{L_2(\mu)}^2 = \mathbb{E}\,\langle \phi(X), f\rangle_{\mathcal{H}}^2 = \left\|\Gamma^{1/2} f\right\|_{\mathcal{H}}^2 \leq \|\Gamma\|_{\mathrm{op}}\, \|f\|_{\mathcal{H}}^2 . \tag{2.2}$$

Define the RKHS design matrix $\mathbb{X}_\phi : \mathcal{H} \to \mathbb{R}^N$ as

$$\mathbb{X}_\phi = \begin{pmatrix} \phi(X_1)^\top \\ \vdots \\ \phi(X_N)^\top \end{pmatrix}, \text{ so that } \mathbb{X}_\phi f = \begin{pmatrix} \langle \phi(X_1), f\rangle_{\mathcal{H}} \\ \vdots \\ \langle \phi(X_N), f\rangle_{\mathcal{H}} \end{pmatrix} = \begin{pmatrix} f(X_1) \\ \vdots \\ f(X_N) \end{pmatrix},$$

where for all $x \in \Omega$, we use $\phi^\top(x) : \mathcal{H} \to \mathbb{R}$, that is, the operator $\phi^\top(x) : f \mapsto \langle \phi(x), f\rangle_{\mathcal{H}} = f(x)$.

**Statistical model and closed-form solution to (2.1)** In this paper, we always assume $f^* \in \mathcal{H}$ (except in Remark 4). Let $\boldsymbol{\xi} = (\xi_i)_{i\in[N]}$ be the noise vector with i.i.d. zero-mean coordinates with variance $\sigma_\xi^2$, which are independent from the design vectors $(X_i)_{i\in[N]}$. The kernel ridge estimator $\hat{f}_\lambda$ for $(X_i, Y_i)_{i\in[N]} \subset (\Omega \times \mathbb{R})^N$ and the tuning parameter $\lambda \geq 0$ is defined as

$$\hat{f}_\lambda \in \operatorname*{argmin}_{f\in\mathcal{H}} \left( \|\mathbb{X}_\phi f - \boldsymbol{y}\|_2^2 + \lambda \|f\|_{\mathcal{H}}^2 \right),$$

where $Y_i = f^*(X_i) + \xi_i$ and $\boldsymbol{y} = (Y_1, \cdots, Y_N)^\top$. This coincides with (2.1). By [Wai19, Proposition 12.33], $\hat{f}_\lambda$ has an explicit form:

$$\hat{f}_\lambda = \mathbb{X}_\phi^\top \left( \mathbb{X}_\phi \mathbb{X}_\phi^\top + \lambda I_N \right)^{-1} \boldsymbol{y}.$$

As $(\phi(X_i))_{i\in[N]} \subset \mathcal{H}$, we write $\mathbb{X}_{\phi,p:q} = ((P_{p:q}\phi(X_1))| \cdots |(P_{p:q}\phi(X_N)))^\top$, thus we can decompose $\mathbb{X}_\phi$ into two parts for any $k \in \mathbb{N}_+$:

$$\mathbb{X}_\phi = \begin{pmatrix} (P_{1:k}\phi(X_1))^\top \\ \vdots \\ (P_{1:k}\phi(X_N))^\top \end{pmatrix} + \begin{pmatrix} (P_{k+1:\infty}\phi(X_1))^\top \\ \vdots \\ (P_{k+1:\infty}\phi(X_N))^\top \end{pmatrix} =: \mathbb{X}_{\phi,1:k} + \mathbb{X}_{\phi,k+1:\infty}.$$

Recall that $f^\top : g \in \mathcal{H} \mapsto \langle f, g\rangle_{\mathcal{H}} \in \mathbb{R}$ for all $f \in \mathcal{H}$, that is, $f^\top g = \langle f, g\rangle_{\mathcal{H}}$ for all $g \in \mathcal{H}$.

**Key quantities driving the rate of convergence of KRR** We first define some key quantities related to the convergence rate of KRR, and then provide an informal version of the main conclusions of this paper. For the formal version, please refer to Theorem 5 in Section 3 and [P2, Theorem 6]. Let $k \in \mathbb{N}_+$, and let $\kappa_{DM}$ be some absolute constant. We define

$$J_1 := \left\{ j \in [k] : \sigma_j \geq \frac{\kappa_{DM}\left(4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)}{N} \right\}, \; J_2 = [k]\backslash J_1, \tag{2.3}$$

and

$$\tilde{\Gamma}_{1,\mathrm{thre}}^{-1/2} = \sum_{j=1}^{k} \left( \sigma_j \vee \frac{\kappa_{DM}\left(4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)}{N} \right)^{-1/2} \varphi_j \otimes \varphi_j.$$

For the optimal choice of $k$, we will have $J_1 = [k]$ and $\tilde{\Gamma}_{1,\mathrm{thre}}^{-1/2} = \Gamma_{1,k}^{-1/2}$.

The main conclusion of this paper is the following:

**Theorem 5 (informal).** *Under certain conditions on the kernel, the design vector and the noise, for any $k \leq N$ and $\lambda \geq 0$ that satisfies certain conditions, the following fact holds with high probability:*

$$\left\| \hat{f}_\lambda - f^* \right\|_{L_2(\mu)} \lesssim r_{\lambda,k}^*,$$

*where the convergence rate $r^*_{\lambda,k}$ is defined as*

$$r^*_{\lambda,k} = \sigma_\xi \sqrt{\frac{|J_1|}{N}} + \sigma_\xi \sqrt{\frac{\sum_{j\in J_2}\sigma_j}{4\lambda \operatorname{Tr}(\Gamma_{k+1:\infty})}} + \left\|\tilde{\Gamma}^{-1/2}_{1,\mathrm{thre}} f^*_{1:k}\right\|_{\mathcal{H}} \frac{2\lambda + 3\operatorname{Tr}(\Gamma_{k+1:\infty})}{N} + \left\|\Gamma^{1/2}_{k+1:\infty} f^*_{k+1:\infty}\right\|_{\mathcal{H}} + \sigma_\xi \frac{\sqrt{N \operatorname{Tr}(\Gamma^2_{k+1:\infty})}}{\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}. \tag{2.4}$$

Examples concerning $r^*_{\lambda,k}$ can be found in [P2]. For instance, suppose there exist $\alpha > 1$ and $s \geq 0$ such that $\sigma_j \sim j^{-\alpha}$ and $\|\Gamma^{\frac{1-s}{2}} f^*\|_{\mathcal{H}} < \infty$, that is, the Sobolev regression setting. In this case, one can show that if $\lambda \sim N^{1-\frac{\alpha}{1+\alpha(s\wedge 2)}}$, then $r^*_{\lambda,k} \sim N^{-\frac{\alpha(s\wedge 2)}{2(1+\alpha(s\wedge 2))}}$, which is the known minimax rate for this problem, [LZL23]. We provide some comments on Theorem 5. In general, based on the choice of $\lambda$ and the spectrum of $\Gamma$, KRR automatically decomposes the feature space $\mathcal{H}$ into a direct sum of two subspaces: $\mathcal{H} = \mathcal{H}_{1:k} \oplus^{\perp} \mathcal{H}_{k+1:\infty}$. In $\mathcal{H}_{1:k}$, KRR estimates the target function $f^*$, while in $\mathcal{H}_{k+1:\infty}$, KRR absorbs noise. Here, $k$ is a key quantity that determines this space decomposition and thus the estimation error of KRR. Therefore, there exists an optimal $k$ that minimizes the convergence rate $r^*_{\lambda,k}$. We emphasize that $k$ is a free parameter; that is, for KRR, statisticians do not select a priori a specific $k$; rather, this $k$ is needed for analysis. Thus, this $k$ is determined by KRR itself; somehow KRR learns by itself the best feature space decomposition.

Let us now provide some explanation on the five terms appearing in the definition of $r^*_{\lambda,k}$. In the space $\mathcal{H}_{1:k}$ (where estimation happens), $[k]$ is further divided into $J_1$ and $J_2$, which correspond to different parts in the final bound: the first row of (2.4) is made of three terms, the first two are variance terms (restricted on $J_1$ and $J_2$) and the last one can be interpreted as a bias term on the entire space $\mathcal{H}_{1:k}$. In most applications, $J_1 = [k]$ and $J_2 = \emptyset$. In the classical case (for instance, when $\lambda \sim N\sigma_k$), the dominating term in the definition of $\tilde{\Gamma}^{-1/2}_{1,\mathrm{thre}}$ is $\sigma_j$, which means that KRR does not filter out the largest $k$ eigenvalues. However, theoretically, there may exist choices of $k$ and $\lambda$ such that KRR filters out only certain eigenvalues, with the number of these eigenvalues being less than $k$. In $\mathcal{H}_{k+1:\infty}$, $f^*_{k+1:\infty}$ is treated as noise, which leads to the term $\|\Gamma^{1/2}_{k+1:\infty} f^*_{k+1:\infty}\|_{\mathcal{H}}$ in (2.4). The space $\mathcal{H}_{k+1:\infty}$ is responsible for noise absorption, corresponding to the term $\sigma_\xi \frac{\sqrt{N \operatorname{Tr}(\Gamma^2_{k+1:\infty})}}{\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}$ in (2.4). Here, it is necessary for $\operatorname{Tr}(\Gamma^2_{k+1:\infty})$ to be relatively small compared to $\lambda^2$ and $(\operatorname{Tr}(\Gamma_{k+1:\infty}))^2$ in order for the estimator to effectively absorb noise. When $\lambda$ dominates $\operatorname{Tr}(\Gamma_{k+1:\infty})$ (which corresponds to the case of sufficiently strong regularization), then $\lambda$ needs to be large enough to absorb noise. Conversely, if $\operatorname{Tr}(\Gamma_{k+1:\infty})$ dominates (which corresponds to insufficient regularization), we require that the spectrum of $\Gamma_{k+1:\infty}$ is well spread, meaning that the ratio of its $\ell_2$ norm to its $\ell_1$ norm must be sufficiently small.

### 2.1.3 Restricted Isomorphy Property

The Restricted Isomorphy Property characterizes the geometric properties of the RKHS design matrix restricted to $\mathcal{H}_{1:k}$, that is, $\mathbb{X}_{\phi,1:k}$. We will see later that this is the part of the space where estimation happens. It identifies the set on which, with high probability, the operator $\mathbb{X}_{\phi,1:k} : \mathcal{H}_{1:k} \to \ell^N_2$ forms an isomorphism when $k \lesssim N$ or a restricted isomorphism when $k \gtrsim N$. This property was used for linear functionals of sub-Gaussian random vectors in [LS24] in the context of benign overfitting in linear regression. In this paper, since we need to study RKHSs, we must establish a corresponding version of this restricted isomorphy property.

**When $k \lesssim N$** When $k \lesssim N$, the RKHS design operator $\mathbb{X}_{\phi,1:k}$ behaves like an isomorphy over the entire space $\mathcal{H}_{1:k}$ under the following assumption.

**Assumption 4.** *There exist absolute constants $\kappa'' \geq 1$, $c_{19}$ depending on $\kappa''$ ($c_{19} \leq \frac{9}{1568(\kappa'')^4}$ is sufficient), $0 \leq \gamma_2 < 1/16$, $\epsilon > 0$, $\delta_2 \geq 0$ such that*

- *$k \leq c_{19} N$.*
- *with probability at least $1 - \gamma_2$,*

$$\max_{1\leq i\leq N} \left\|\Gamma^{-1/2}_{1:k} \phi_{1:k}(X_i)\right\|^2_{\mathcal{H}} \leq \delta_2 k, \tag{2.5}$$

- *for any $f \in \mathcal{H}_{1:k}$, $\|f\|_{L_{4+\epsilon}} \leq \kappa'' \|f\|_{L_2}$;*

The next result shows the isomorphy property of $\mathbb{X}_{\phi,1:k}$ on $\mathcal{H}_{1:k}$ under Assumption 4.

**Proposition 17.** *Under Assumption 4, there exist absolute constants $c_4$, $c_5$ and $C_9$ such that with probability at least $1-\gamma_2-\frac{c_4}{N^\epsilon}-2\exp(-k)$, for all $f_{1:k}\in\mathcal{H}_{1:k}$,*

$$c_5\left\|\Gamma_{1:k}^{1/2}f_{1:k}\right\|_{\mathcal{H}}\leq\frac{1}{\sqrt{N}}\left\|\mathbb{X}_{\phi,1:k}f_{1:k}\right\|_2\leq C_9\left\|\Gamma_{1:k}^{1/2}f_{1:k}\right\|_{\mathcal{H}},$$

*where $c_5$ can be taken equal to $\frac{1}{2}$ and $C_9$ equal to $\sqrt{2C_8^2(1+c_{19})}$. We denote $\gamma_2+\frac{c_4}{N^\epsilon}+2\exp(-k)$ by $\bar{p}_{RIP}$.*

**When $k$ is not necessarily smaller than $N$** When $k$ is not necessarily smaller than $N$, the design matrix $\mathbb{X}_{\phi,1:k}$ cannot behave like an isomorphy over the entire space $\mathcal{H}_{1:k}$ because it has a non-trivial kernel, but it can be an isomorphy restricted to a subset of $\mathcal{H}_{1:k}$. This set can be taken to be a cone defined below in (2.7). We refer to this property as the Restricted Isomorphy Property (RIP) as in [LS24], in reminiscence to the RIP used in Compressed sensing [FR13]. Please note that we can still use the following proposition to replace Proposition 17 when $k\lesssim N$, albeit at the cost of incurring a logarithmic factor.

**Proposition 18.** *Let $C_{10}$ and $C_{11}$ be absolute constants. For any $R>0$, let $\bar{\Gamma}_{1:k}^{-1/2}=\sum_{j\leq k}\min\left(\frac{1}{R},\frac{1}{\sqrt{\sigma_j}}\right)\varphi_j\otimes\varphi_j$. For some $c_6$ sufficiently small ($c_6<\frac{1}{100C_{10}^2C_{11}^2}$ is sufficient), let*

$$R_N(c_6)=\inf\left\{R>0:\left\|\max_{i\in[N]}\left\|\bar{\Gamma}_{1:k}^{-1/2}\phi(X_i)\right\|_{\mathcal{H}}\right\|_{L_\infty}\leq c_6\frac{\sqrt{N}}{\log N}\right\}. \tag{2.6}$$

*For any $0<\delta_3<1$, there exist absolute constants $c_5$, and $C_9$ depending on $\delta_3$ such that when $R\geq R_N(c_6)$, then with probability at least $1-\delta_3$, for all $f\in\operatorname{cone}(\mathcal{C}(R))$, where*

$$\mathcal{C}(R)=R^{-1}B_{\mathcal{H}_{1:k}}\cap\Gamma_{1:k}^{-1/2}S_{\mathcal{H}_{1:k}}\text{ and }\operatorname{cone}(\mathcal{C}(R))=\left\{f\in\mathcal{H}_{1:k}:R\left\|f\right\|_{\mathcal{H}}\leq\left\|f\right\|_{L_2}\right\}, \tag{2.7}$$

*we have*

$$c_5\left\|\Gamma_{1:k}^{1/2}f_{1:k}\right\|_{\mathcal{H}}\leq\frac{1}{\sqrt{N}}\left\|\mathbb{X}_{\phi,1:k}f_{1:k}\right\|_2\leq C_9\left\|\Gamma_{1:k}^{1/2}f_{1:k}\right\|_{\mathcal{H}}.$$

In contrast to Proposition 17, we only have constant probability deviation in Proposition 18. We refer to Proposition 18 as the "Restricted Isomorphic Property under the embedding index condition" because the estimation of the fixed point $R_N(c_6)$ requires the embedding index condition. An example of the estimate of $R_N(c_6)$ may be found in [P2, Section 6.5.4]. The proofs of Proposition 17 and Proposition 18 are postponed to [P2, Section 6.5.2].

## 2.2 Main Results

In this section, we present the upper bounds on the estimation error of KRR.

### 2.2.1 Review of the Assumptions

**Assumption 2 (Recall).** *There are absolute constants $C_1>1$, $C_2>1$, $0\leq\gamma<1/16$, $0\leq\delta<1/(100\sqrt{C_2})$, $\bar{\delta}<C_1^{-1}$, $\epsilon>0$ and $\kappa>1$ such that*

- *With probability at least $1-\gamma$,*

$$\max_{1\leq i\leq N}\left|\frac{\left\|\phi_{J^c}(X_i)\right\|_{\mathcal{H}}^2}{(\ell^*)^2}-1\right|\leq\delta, \tag{1.26}$$

  *where we define $\ell^*=\sqrt{\mathbb{E}\left\|\phi_{J^c}(X)\right\|_{\mathcal{H}}^2}=\sqrt{\operatorname{Tr}\left(\Sigma_{J^c}\right)}$.*

- *For any $f\in V_{J^c}$, we have*

$$\left\|f\right\|_{L^{2+\epsilon}(\mu_X)}\leq\kappa\left\|f\right\|_{L^2(\mu_X)}. \tag{1.27}$$

- *Depending on the choice of $\epsilon$, there are two cases:*

1. *if* $\epsilon > 2$, *then no extra assumption is required.*
2. *if* $0 < \epsilon \leq 2$, *then*

$$\kappa N^{\frac{2-\epsilon}{2\epsilon+\epsilon^2}} \log(N) \left( \sqrt{\frac{N\|\Sigma_{J^c}\|_{\mathrm{op}}}{\mathrm{Tr}(\Sigma_{J^c})}} \right) < \bar{\delta}. \tag{1.28}$$

**Assumption 3 (Recall).** *There exist absolute constants* $\gamma_1 \in \left(0, \frac{1}{16}\right)$, $\delta_1 \geq 0$, $\epsilon > 0$ *and* $\kappa' > 1$ *such that*

- $$\mathbb{P}\left( \max_{1\leq i\leq N} \frac{\left\|\Sigma_{J^c}^{1/2}\phi_{J^c}(X_i)\right\|_{\mathcal{H}}^2}{\mathrm{Tr}(\Sigma_{J^c}^2)} \leq 1+\delta_1 \right) \geq 1-\gamma_1, \tag{1.37}$$
- *for any* $f \in V_{J^c}$, $\|f\|_{L^{4+\epsilon}(\mu_X)} \leq \kappa' \|f\|_{L^2(\mu_X)}$.

**Assumption 4 (Recall).** *There exist absolute constants* $\kappa'' \geq 1$, $c_{19}$ *depending on* $\kappa''$ *(*$c_{19} \leq \frac{9}{1568(\kappa'')^4}$ *is sufficient),* $0 \leq \gamma_2 < 1/16$, $\epsilon > 0$, $\delta_2 \geq 0$ *such that*

- $k \leq c_{19}N$.
- *with probability at least* $1-\gamma_2$,

$$\max_{1\leq i\leq N} \left\|\Gamma_{1:k}^{-1/2}\phi_{1:k}(X_i)\right\|_{\mathcal{H}}^2 \leq \delta_2 k, \tag{2.5}$$

- *for any* $f \in \mathcal{H}_{1:k}$, $\|f\|_{L_{4+\epsilon}} \leq \kappa'' \|f\|_{L_2}$;

In this chapter, we also assume that the noise $\xi$ is independent of $X$ with mean zero and variance $\sigma_\xi^2$. We assume that for some $\kappa_1 > 0$ and $r > 4$, $\|\xi\|_{L_r} \leq \kappa_1 \sigma_\xi$.

### 2.2.2 Our results

As remarked in several works [BMR21, TB23, LS24], there is a fundamental parameter $k$ which is the dimension of the space endowed by the top $k$ eigenvectors of $\Gamma$ where 'estimation' happens whereas, on the orthogonal space, 'absorption of the noise (and even overfitting of the noise when $\lambda = 0$)' happens. Our analysis depends on the case where this parameter is smaller or larger than the number of data $N$.

**When** $k \leq c_{19}N$ In this paragraph, we provide conclusions for the case of $k \lesssim N$. This scenario is precisely what linear regression and Multiple Descents problems are most concerned with. The definition of $d_\lambda^*(\Gamma_{k+1:\infty}^{-1/2}B_{\mathcal{H}})$ is given in Equation (1.25) and the ones of $\bar{p}_{RIP}$, $\bar{p}_{DM}$, and $\bar{p}_{DMU}$ in Proposition 17, Theorem 4, and Proposition 13, respectively.

**Theorem 5.** *Suppose Assumptions 2, 3 and 4 hold. There then exist absolute constants* $C_{12}$, $c_{19}$, $c_7$, $c_8$, $C_{13}$ *(*$C_{13}$ *depends on* $\kappa_1$*) and* $C_{14}$, *such that the following holds. Suppose the noise* $\xi$ *is independent of* $X$ *with mean zero and variance* $\sigma_\xi^2$. *We assume that for some* $\kappa_1 > 0$ *and* $r > 4$, $\|\xi\|_{L_r} \leq \kappa_1\sigma_\xi$. *Let* $\lambda \geq 0$. *We assume that there exists* $k \in \mathbb{N}$ *so that* $C_{12} \leq N \leq c_7\kappa_{DM} d_\lambda^*\left(\Gamma_{k+1:\infty}^{-1/2}B_{\mathcal{H}}\right)$, $\sigma_1 N > \kappa_{DM}(4\lambda + \mathrm{Tr}(\Gamma_{k+1:\infty}))$ *and* $k \leq c_{19}N$. *Let* $\bar{p}_\xi$ *be some probability deviation strictly less than* 1 *(defined later in [P2, Equation 68]). Then with probability at least*

$$1 - \bar{p}_{RIP} - \bar{p}_{DM} - \bar{p}_{DMU} - \frac{c_8}{N} - \bar{p}_\xi - \left( \frac{C_{13}\,\mathrm{Tr}(\Gamma_{k+1:\infty})}{|J_1|\,\mathrm{Tr}(\Gamma_{k+1:\infty}) + N\left(\sum_{j\in J_2}\sigma_j\right)} \right)^{\frac{r}{4}},$$

*we have, for* $r_{\lambda,k}^*$ *defined in (2.4), that* $\left\|\hat{f}_\lambda - f^*\right\|_{L_2} \leq C_{14} r_{\lambda,k}^*$.

**When $k$ is not necessarily smaller than $c_{19}N$** In this paragraph, we provide the informal version of our conclusion for the case of $k$ is not necessarily smaller than $c_{19}N$ (the formal version may be found in [P2, Section 5.6]). In fact, for the context of the minimum $\|\cdot\|_2$ norm interpolant estimator in linear regression, the authors of [LS24] have already conducted research on this scenario. They have demonstrated that when the design vector is symmetric, the optimal value of $k$ — the one that minimizes the estimation error among all possible $k$ — falls precisely within the range of $k \lesssim N$ if one wants benign overfitting to happen. However, in the case of KRR, there is no lower bound indicating that the optimal $k$ must satisfy $k \lesssim N$. We therefore present the following theorem.

**Theorem 6** (informal, [P2])**.** *Under certain conditions, for any $\lambda \geq 0$ and $k$ satisfying the Dvoretzky-Milman condition, with constant probability, we have $\|\hat{f}_\lambda - f^*\|_{L_2} \lesssim r^*_{\lambda,k}$ where $r^*_{\lambda,k}$ is defined in (2.4).*

**Remark 4** (Misspecified model)**.** *In practical applications, we often encounter cases where $f^* \notin \mathcal{H}$, as exemplified by our [P2, Proposition 11]. When $f^* \notin \mathcal{H}$, we define $f^{**} = \arg\min(\|f^* - f\|_{L_2(\mu)} : f \in \mathcal{H})$, that is, the orthogonal projection of $f^*$ onto $\mathcal{H}$ in the $L_2(\mu)$ inner product (we could also choose other oracles as we have the liberty to choose the oracle). We replace the target function in Theorem 5 and [P2, Theorem 6] with $f^{**}$ and replace the noise $\boldsymbol{\xi}$ with $\boldsymbol{\epsilon} = \boldsymbol{r} + \boldsymbol{\xi}$, where $\boldsymbol{r} = (f^*(X_i) - f^{**}(X_i))_{i=1}^N$. This new noise $\boldsymbol{\epsilon}$ is dependent on the kernel design matrix $\mathbb{X}_\phi$. See [P2, Proposition 30, Proposition 31] and the subsequent discussion of this property in supplementary material and [P2, Proposition 11] for an example when $f^* \notin \mathcal{H}$. This conclusion is not contradictory to the counterexamples provided in [CLvdG22] and [Sha22], as indicated in [P2, Remark 9]. For the estimation error in the misspecified setting, see Proposition 25 in page 58.*

**Remark 5** (Uniform results in $\lambda$)**.** *It is possible to have results equivalent to Theorem 5 and [P2, Theorem 6] uniform in the regularization parameter $\lambda$. This type of results is particularly useful when one wants to use a data-dependent regularization parameter in order to achieve optimal and adaptive results. It is for instance the case, when $\lambda$ is chosen according to the Lepski's method [BMM19]. In that case, our results hold with the same probability and convergence rates. However, for instance in Theorem 5, we just need to assume that $N \lesssim d^*_{\lambda_0}\left(\Gamma^{-1/2}_{k+1:\infty} B_{\mathcal{H}}\right)$ for some $\lambda_0 \geq 0$ and then the result of Theorem 5 holds uniformly for all $\lambda \geq \lambda_0$. This may be particularly useful when $\lambda_0 = 0$.*

## 2.3 Proof of Theorem 5 (the $k \lesssim N$ case)

We first provide some definitions. Define

$$\square = \max\left\{\sigma_\xi \sqrt{\frac{\operatorname{Tr}(\Gamma_{1:k})}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}}, \sqrt{\frac{\sigma_1 N}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}} \left\|\Gamma^{1/2}_{k+1:\infty} f^*_{k+1:\infty}\right\|_{\mathcal{H}}, \right. \\ \left. \|f^*_{1:k}\|_{\mathcal{H}} \sqrt{\frac{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}{N}}\right\}, \tag{2.8}$$

if $\sigma_1 N \leq \kappa_{DM}(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))$; and

$$\square = \max\left\{\sigma_\xi \sqrt{\frac{|J_1|}{N}}, \sigma_\xi \sqrt{\frac{\sum_{j\in J_2}\sigma_j}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}}, \left\|\Gamma^{1/2}_{k+1:\infty} f^*_{k+1:\infty}\right\|_{\mathcal{H}}, \right. \\ \left. \left\|\tilde{\Gamma}^{-1/2}_{1,\text{thre}} f^*_{1:k}\right\|_{\mathcal{H}} \frac{2\lambda + 3\operatorname{Tr}(\Gamma_{k+1:\infty})}{N}\right\}, \tag{2.9}$$

if $\sigma_1 N > \kappa_{DM}(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))$. Let $\triangle = \square\sqrt{N}/\sqrt{\kappa_{DM}(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))}$.

For any $\lambda \geq 0$ and $k \in \mathbb{N}$, define

$$r^*_{\lambda,k} := \max\left\{\square, \sigma_\xi \frac{\sqrt{N \operatorname{Tr}(\Gamma^2_{k+1:\infty})}}{\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}\right\}. \tag{2.10}$$

Define

$$\tilde{\Gamma}^{1/2}_{1:k} = \sum_{j=1}^{k} \max\left(\frac{\sqrt{\sigma_j}}{\square}, \frac{1}{\triangle}\right) \varphi_j \otimes \varphi_j. \tag{2.11}$$

For the sake of simplicity, we denote by $A : \mathbb{R}^N \to \mathcal{H}$, a random matrix with i.i.d. column vectors denoted by $\phi(X_1), \cdots, \phi(X_N)$: $A = [\phi(X_1)|\cdots|\phi(X_N)]$ such that for any $\boldsymbol{\lambda} \in \mathbb{R}^N$, $A\boldsymbol{\lambda} = \sum_{i=1}^N \lambda_i \phi(X_i)$. In other words, $A$ is the adjoint of $\mathbb{X}_\phi$, i.e., $A = \mathbb{X}_\phi^\top$. We denote $\ell^* = \sqrt{\mathbb{E}\,\|\phi(X)\|_{\mathcal{H}}^2} = \sqrt{\mathbb{E}K(X,X)} = \sqrt{\mathrm{Tr}\,(\Gamma)}$.

In this section, we establish the proof of Theorem 5. The proof is generally divided into two main parts: the Stochastic Argument and the Deterministic Argument. In the following subsection, we commence with the Stochastic Argument.

**Stochastic event behind Theorem 5.**

Let $C_8, C_{15}, C_{16}, c_5$, and $C_9$ be absolute constants. We denote by $\Omega_0$ the event which we have:

- for all $\boldsymbol{\lambda} \in \mathbb{R}^N$,

$$\begin{aligned}\left(\frac{1}{2}\,\mathrm{Tr}(\Gamma_{k+1:\infty}) + \lambda\right)\|\boldsymbol{\lambda}\|_2 &\le \left\|\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I\right)\boldsymbol{\lambda}\right\|_2 \\ &\le \left(\frac{3}{2}\,\mathrm{Tr}(\Gamma_{k+1:\infty}) + \lambda\right)\|\boldsymbol{\lambda}\|_2\end{aligned} \tag{2.12}$$

- for all $f_{1:k} \in \mathcal{H}_{1:k}$,

$$c_5\left\|\Gamma_{1:k}^{1/2} f_{1:k}\right\|_{\mathcal{H}} \le (1/\sqrt{N})\,\|\mathbb{X}_{\phi,1:k} f_{1:k}\|_2 \le C_9\left\|\Gamma_{1:k}^{1/2} f_{1:k}\right\|_{\mathcal{H}} \tag{2.13}$$

- for all $\boldsymbol{\lambda} \in \mathbb{R}^N$,

$$\left\|\Gamma_{k+1:\infty}^{1/2}\mathbb{X}_{\phi,k+1:\infty}^\top\boldsymbol{\lambda}\right\|_{\mathcal{H}} \le C_8\left(\sqrt{\mathrm{Tr}\left(\Gamma_{k+1:\infty}^2\right)} + \sqrt{N}\,\|\Gamma_{k+1:\infty}\|_{\mathrm{op}}\right)\|\boldsymbol{\lambda}\|_2 \tag{2.14}$$

- $$\left\|\mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^*\right\|_2 \le C_{15}\kappa\sqrt{N}\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}} \tag{2.15}$$

- $$\sum_{i=1}^N \left\|\left(\Gamma_{k+1:\infty}^{1/2}\phi_{k+1:\infty}\right)(X_i)\right\|_{\mathcal{H}}^2 \le C_{16} N\,\mathrm{Tr}\left(\Gamma_{k+1:\infty}^2\right). \tag{2.16}$$

By the definition of $\tilde{\delta}$, see (1.29), there exists an absolute constant $c_7$ such that if $\delta^2, \bar{\delta}^2 < c_7$, we have $\tilde{\delta} < 1/2$. When $\lambda > C_3\,\mathrm{Tr}\,(\Gamma_{k+1:\infty})$, one may replace the absolute constants $\frac{1}{2}$ and $\frac{3}{2}$ in (2.12) by those in Theorem 4. It follows from Theorem 4, Proposition 13 and Proposition 17 that if $N \le c_7\kappa_{DM} d_\lambda^*(\Gamma_{k+1:\infty}^{-1/2} B_{\mathcal{H}})$, then with probability larger than $1 - \bar{p}_{RIP} - \bar{p}_{DM} - \bar{p}_{DMU}$, (2.12), (2.13) and (2.14) hold.

For (2.15), we use [Men16, Lemma 3.2] on the $L_r$-norm of a sum of i.i.d. random variables to deduce the following result.

**Lemma 4.** *There exist absolute constants $c_9, c_{10}, c_{11}$ such that the following holds. Let $1 \le r < q$, set $Z \in L_q$ and put $Z_1, \cdots, Z_N$ to be independent copies of $Z$. Fix $1 \le p \le N$, let $j_0 = \lceil (c_9 p)/\left(((q/r) - 1)\log\left(4 + eN/p\right)\right)\rceil$ and $t > 2$. If $j_0 = 1$ and $0 < \beta < (q/r) - 1$ then with probability at least $1 - c_{11}t^{-q}N^{-\beta}$,*

$$\left(\sum_{j=1}^N |Z_i|^r\right)^{1/r} \le c_{10}\left(\frac{q}{q - (\beta+1)r}\right)^{1/r} t\,\|Z\|_{L_q} N^{1/r}.$$

Without loss of generality, we take $c_{10} > 1$. Let $Z_i = f_{k+1:\infty}^*(X_i)$, $r = 2$, $p = 1$, $q = 4 + \epsilon$ (where $\epsilon$ is from Assumption 2, Assumption 3 and Assumption 4) and $\beta = 1$ in Lemma 4, and by the fact that $\left\|f_{k+1:\infty}^*\right\|_{L_{4+\epsilon}} \le$

$\kappa \left\| f^*_{k+1:\infty} \right\|_{L_2}$, Lemma 4 indicates that if $N \geq (e^{c_9} - 4)/e \vee c_{11}^2$, then there exists absolute constant $c_8$ with probability at least $1 - c_8/N$,

$$\begin{aligned} \left\| \mathbb{X}_{\phi,k+1:\infty} f^*_{k+1:\infty} \right\|_2 &= \left( \sum_{i=1}^N \left( f^*_{k+1:\infty}(X_i) \right)^2 \right)^{1/2} \leq C_{15} \sqrt{N} \left\| f^*_{k+1:\infty} \right\|_{L_{4+\epsilon}} \\ &\leq C_{15} \kappa \sqrt{N} \left\| \Gamma^{1/2}_{k+1:\infty} f^*_{k+1:\infty} \right\|_{\mathcal{H}}, \end{aligned}$$

where $c_8 = c_{11}$, $C_{15} = c_{10}\sqrt{(4+\epsilon)/\epsilon}$.

We are left with checking (2.16). However, this is a simple consequence of Assumption 3. By Assumption 3, (2.16) holds with probability at least $1 - \gamma_1$ with constant $C_{16} = 1 + \delta_1$.

Combining the above probabilistic estimates, we have the following Proposition

**Proposition 19.** *Suppose Assumption 2, Assumption 3 and Assumption 4 hold. There exist absolute constants $c_9$, $c_{11}$, $c_{19}$, $c_7$ and $c_8$, such that if $N \geq (e^{c_9} - 4)/e \vee c_{11}^2$, and if there exists $k \leq c_{19} N$, such that $N \leq c_7 \kappa_{DM} d^*_\lambda \left( \Gamma^{-1/2}_{k+1:\infty} B_{\mathcal{H}} \right)$, then*

$$\mathbb{P}(\Omega_0) \geq 1 - \bar{p}_{RIP} - \bar{p}_{DM} - \bar{p}_{DMU} - \frac{c_8}{N} - \gamma_1.$$

We now place ourselves on the event $\Omega_0$ up to the end of the proof of Theorem 5. All the remaining material does not rely on any stochastic arguments since they all have been collectd in $\Omega_0$.

### Decomposition of $\hat{f}_\lambda$

As in the linear situation, the KRR estimator $\hat{f}_\lambda$ is decomposed into two components: $\hat{f}_{1:k} \in \mathcal{H}_{1:k} = \operatorname{span}(\varphi_j : 1 \leq j \leq k)$ and $\hat{f}_{k+1:\infty} \in \mathcal{H}_{k+1:\infty} = \operatorname{span}(\varphi_j : j > k)$. The two components have their own role in estimating $f^*$: $\hat{f}_{1:k}$ is used as a ridge estimator of $P_{1:k} f^*$ whereas $\hat{f}_{k+1:\infty}$ is used to absorb noise, thus is not expected to be a good estimator of $P_{k+1:\infty} f^*$.

**Proposition 20.** *For any $k \in \mathbb{N}_+$, the KRR defined by (2.1) can be written as $\hat{f}_\lambda = \hat{f}_{1:k} + \hat{f}_{k+1:\infty}$, where*

$$\hat{f}_{1:k} \in \operatorname*{argmin}_{f_{1:k} \in \mathcal{H}_{1:k}} \left( \left\| Q \left( \boldsymbol{y} - \mathbb{X}_{\phi,1:k} f_{1:k} \right) \right\|^2_{\mathcal{H}} + \left\| f_{1:k} \right\|^2_{\mathcal{H}} \right), \tag{2.17}$$

*and*

$$\hat{f}_{k+1:\infty} = \mathbb{X}^\top_{\phi,k+1:\infty} \left( \mathbb{X}_{\phi,k+1:\infty} \mathbb{X}^\top_{\phi,k+1:\infty} + \lambda I_N \right)^{-1} \left( \boldsymbol{y} - \mathbb{X}_{\phi,1:k} \hat{f}_{1:k} \right), \tag{2.18}$$

*where $Q : \mathbb{R}^N \to \mathcal{H}_{k+1:\infty}$ is a bounded linear operator such that*

$$Q^\top Q = \left( \mathbb{X}_{\phi,k+1:\infty} \mathbb{X}^\top_{\phi,k+1:\infty} + \lambda I_N \right)^{-1}.$$

*Such an operator exists because $Q^\top Q$ is semi positive-definite.*

**Proof.** The empirical regularized loss functional is defined as $L : f \in \mathcal{H} \mapsto \|\boldsymbol{y} - \mathbb{X}_\phi f\|_2^2 + \lambda \|f\|^2_{\mathcal{H}}$. As $\hat{f}_\lambda$ is a minimizer of $L(f)$, we decompose $\hat{f}_\lambda = \hat{f}_{1:k} + \hat{f}_{k+1:\infty}$ and take the derivative of $L$ with respect to the canonical inner product on $\mathcal{H}$ at $\hat{f}_{1:k}$ and $\hat{f}_{k+1:\infty}$ and set them to 0, as a result, we obtain

$$\left( \mathbb{X}^\top_{\phi,1:k} \mathbb{X}_{\phi,1:k} + \lambda I \right) \hat{f}_{1:k} + \mathbb{X}^\top_{\phi,1:k} \mathbb{X}_{\phi,k+1:\infty} \hat{f}_{k+1:\infty} = \mathbb{X}^\top_{\phi,1:k} \boldsymbol{y} \tag{2.19}$$

$$\mathbb{X}^\top_{\phi,k+1:\infty} \mathbb{X}_{\phi,1:k} \hat{f}_{1:k} + \left( \mathbb{X}^\top_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty} + \lambda I \right) \hat{f}_{k+1:\infty} = \mathbb{X}^\top_{\phi,k+1:\infty} \boldsymbol{y}, \tag{2.20}$$

where $I : \mathcal{H} \to \mathcal{H}$ is identity operator. Solving (2.20) gives

$$\hat{f}_{k+1:\infty} = \left( \mathbb{X}^\top_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty} + \lambda I \right)^{-1} \mathbb{X}^\top_{\phi,k+1:\infty} \left( \boldsymbol{y} - \mathbb{X}_{\phi,1:k} \hat{f}_{1:k} \right),$$

which coincides with (2.18) because of the Woodbury formula

$$\mathbb{X}^\top_{\phi,k+1:\infty} \left( \mathbb{X}_{\phi,k+1:\infty} \mathbb{X}^\top_{\phi,k+1:\infty} + \lambda I_N \right)^{-1} = \left( \mathbb{X}^\top_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty} + \lambda I \right)^{-1} \mathbb{X}^\top_{\phi,k+1:\infty}.$$

For (2.17), we plug (2.18) into (2.19) to obtain

$$\left(\mathbb{X}_{\phi,1:k}^\top \mathbb{X}_{\phi,1:k} + \lambda I - \mathbb{X}_{\phi,1:k}^\top \mathbb{X}_{\phi,k+1:\infty} \left(\mathbb{X}_{\phi,k+1:\infty}^\top \mathbb{X}_{\phi,k+1:\infty} + \lambda I\right)^{-1} \mathbb{X}_{\phi,k+1:\infty}^\top \mathbb{X}_{\phi,1:k}\right) \hat{f}_{1:k}$$
$$= \mathbb{X}_{\phi,1:k}^\top \left(I - \mathbb{X}_{\phi,k+1:\infty} \left(\mathbb{X}_{\phi,k+1:\infty}^\top \mathbb{X}_{\phi,k+1:\infty} + \lambda I\right)^{-1} \mathbb{X}_{\phi,k+1:\infty}^\top\right) \boldsymbol{y}.$$

Let $F = I - \mathbb{X}_{\phi,k+1:\infty} \left(\mathbb{X}_{\phi,k+1:\infty}^\top \mathbb{X}_{\phi,k+1:\infty} + \lambda I\right)^{-1} \mathbb{X}_{\phi,k+1:\infty}^\top$. The above equation is then equivalent to

$$\left(\mathbb{X}_{\phi,1:k}^\top F \mathbb{X}_{\phi,1:k} + \lambda I\right) \hat{f}_{1:k} = \mathbb{X}_{\phi,1:k}^\top F \boldsymbol{y}.$$

Applying the Woodbury formula to $F$ gives $F = \lambda \left(\lambda I + \mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top\right)^{-1}$. As $\mathbb{X}_{\phi,1:k}^\top F \mathbb{X}_{\phi,1:k} + \lambda I$ is invertible (because $F \succeq 0$), we have

$$\begin{aligned} \hat{f}_{1:k} = &\left(\mathbb{X}_{\phi,1:k}^\top \left(\lambda I + \mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top\right)^{-1} \mathbb{X}_{\phi,1:k} + I\right)^{-1} \\ &\mathbb{X}_{\phi,1:k}^\top \left(\lambda I + \mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top\right)^{-1} \boldsymbol{y}. \end{aligned} \tag{2.21}$$

To check that (2.21) is equivalent to (2.17), we take the gradient of the convex objective function $f_{1:k} \mapsto \|Q\left(\boldsymbol{y} - \mathbb{X}_{\phi,1:k} f_{1:k}\right)\|_{\mathcal{H}}^2 + \|f_{1:k}\|_{\mathcal{H}}^2$ from (2.17) and set it to 0. This gives $\hat{f}_{1:k} = \mathbb{X}_{\phi,1:k}^\top Q^\top Q(\boldsymbol{y} - \mathbb{X}_{\phi,1:k} \hat{f}_{1:k})$. Recall that $Q^\top Q = \left(\mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}$ and $\mathbb{X}_{\phi,1:k}^\top Q^\top Q \mathbb{X}_{\phi,1:k} + I$ is invertible. Hence $\hat{f}_{1:k}$ from (2.21) is the unique solution to the optimization problem from (2.17) and so (2.17) holds. ■

**Estimation properties of the "ridge estimator" $\hat{f}_{1:k}$**

For any $f_{1:k} \in \mathcal{H}_{1:k}$, we define its (empirical) excess risk as follows (note that KRR's empirical excess risk includes a regularization term):

$$\begin{aligned} \mathcal{L}_{f_{1:k}} &= \|Q\left(\boldsymbol{y} - \mathbb{X}_{\phi,1:k} f_{1:k}\right)\|_{\mathcal{H}}^2 + \|f_{1:k}\|_{\mathcal{H}}^2 - \left(\|Q\left(\boldsymbol{y} - \mathbb{X}_{\phi,1:k} f_{1:k}^*\right)\|_{\mathcal{H}}^2 + \|f_{1:k}^*\|_{\mathcal{H}}^2\right) \\ &= \left\|(\mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1/2} \mathbb{X}_{\phi,1:k}(f_{1:k} - f_{1:k}^*)\right\|_2^2 \\ &+ 2\left\langle \mathbb{X}_{\phi,1:k}^\top (\mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1} (\mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^* + \boldsymbol{\xi}) - f_{1:k}^*, f_{1:k} - f_{1:k}^* \right\rangle_{\mathcal{H}} \\ &+ \|f_{1:k} - f_{1:k}^*\|_{\mathcal{H}}^2, \end{aligned} \tag{2.22}$$

where we have used the fact that $Q^\top Q = \left(\mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}$, $\|Q\boldsymbol{\lambda}\|_{\mathcal{H}} = \left\|\left(\mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1/2} \boldsymbol{\lambda}\right\|_2$ from Proposition 20 and $\|f_{1:k}\|_{\mathcal{H}}^2 - \|f_{1:k}^*\|_{\mathcal{H}}^2 = \|f_{1:k} - f_{1:k}^*\|_{\mathcal{H}}^2 - 2\langle f_{1:k}^*, f_{1:k}^* - f_{1:k}\rangle_{\mathcal{H}}$.

We denote the three terms of the decomposition (2.22) by $\mathcal{Q}_{f_{1:k}}$, $\mathcal{M}_{f_{1:k}}$ and $\mathcal{R}_{f_{1:k}}$ respectively:

$$\mathcal{Q}_{f_{1:k}} = \left\|(\mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1/2} \mathbb{X}_{\phi,1:k}(f_{1:k} - f_{1:k}^*)\right\|_2^2, \tag{2.23}$$

$$\mathcal{M}_{f_{1:k}} = 2\left\langle \mathbb{X}_{\phi,1:k}^\top (\mathbb{X}_{\phi,k+1:\infty} \mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1} (\mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^* + \boldsymbol{\xi}) - f_{1:k}^*, f_{1:k} - f_{1:k}^* \right\rangle_{\mathcal{H}}, \tag{2.24}$$

$$\mathcal{R}_{f_{1:k}} = \|f_{1:k} - f_{1:k}^*\|_{\mathcal{H}}^2. \tag{2.25}$$

We notice that of these three terms, only the multiplier term $\mathcal{M}_{f_{1:k}}$ can take negative values, whereas the quadratic term $\mathcal{Q}_{f_{1:k}}$ and the regularization term $\mathcal{R}_{f_{1:k}}$ are always positive.

We will show that with high probability, $\left\|\Gamma_{1:k}^{1/2}(\hat{f}_{1:k} - f_{1:k}^*)\right\|_{\mathcal{H}} \leq \square$ and $\left\|\hat{f}_{1:k} - f_{1:k}^*\right\|_{\mathcal{H}} \leq \triangle$, where $\square, \triangle > 0$ will be defined later. In other words, we want to show that $\hat{f}_{1:k} \in f_{1:k}^* + B$ where $B$ is the unit ball of the norm $|||\cdot|||$

defined as

$$|||f||| := \max\left\{\frac{\left\|\Gamma_{1:k}^{1/2} f\right\|_{\mathcal{H}}}{\square}, \frac{\|f\|_{\mathcal{H}}}{\triangle}\right\}.$$

From the definition of $\hat{f}_{1:k}$ in (2.17), we know that $\mathcal{L}_{\hat{f}_{1:k}} \leq 0$ so it suffices to show that for all $f_{1:k} \notin f_{1:k}^* + B$ we have $\mathcal{L}_{f_{1:k}} > 0$. We denote the border of $B$ in $V_{1:k}$ by $\partial B$. Let $f_{1:k} \in V_{1:k}$ be such that $f_{1:k} \notin f_{1:k}^* + B$. There exists $f_0 \in \partial B$ and $\theta > 1$ such that $f_{1:k} - f_{1:k}^* = \theta(f_0 - f_{1:k}^*)$. Using (2.22), it follows from the convexity that $\mathcal{L}_{f_{1:k}} \geq \theta \mathcal{L}_{f_0}$. As a consequence, if we prove that $\mathcal{L}_{f_{1:k}} > 0$ for all $f_{1:k} \in f_{1:k}^* + \partial B$, this will imply that $\mathcal{L}_{f_{1:k}} > 0$ for all $f_{1:k} \notin f_{1:k}^* + B$. Hence, we only need to show the positivity of the excess regularized risk $\mathcal{L}_{f_{1:k}}$ on the border $f_{1:k}^* + \partial B$.

For $f_{1:k} \in \partial B$, there are two cases:

1. $\left\|\Gamma_{1:k}^{1/2}(f_{1:k} - f_{1:k}^*)\right\|_{\mathcal{H}} = \square$ and $\|f_{1:k} - f_{1:k}^*\|_{\mathcal{H}} \leq \triangle$, or
2. $\left\|\Gamma_{1:k}^{1/2}(f_{1:k} - f_{1:k}^*)\right\|_{\mathcal{H}} \leq \square$ and $\|f_{1:k} - f_{1:k}^*\|_{\mathcal{H}} = \triangle$.

We will prove that either we have $\mathcal{Q}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$ (this occurs in case [1]), or $\mathcal{R}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$ (this occurs in case [2]). Combined with (2.22), this will show that $\mathcal{L}_{f_{1:k}} > 0$. To achieve this goal, we need to obtain a lower bound for $\mathcal{Q}_{f_{1:k}}$ in case [1] and an upper bound for $\mathcal{M}_{f_{1:k}}$ in case [1] and [2]. The lower bound for $\mathcal{R}_{f_{1:k}}$ is straightforward because this term is not random and is positive.

**Bound of the multiplier term** We show an upper bound on $\mathcal{M}_{f_{1:k}}$ when $f_{1:k} \in f_{1:k}^* + \partial B$.

Observe that

$$\begin{aligned}
&|\mathcal{M}_{f_{1:k}}| \\
&\leq 2\sup_{f \in B} \left|\left\langle \mathbb{X}_{\phi,1:k}^\top (\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}(\mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^* + \boldsymbol{\xi}) - f_{1:k}^*, f\right\rangle_{\mathcal{H}}\right|.
\end{aligned}$$

Also, for $f \in \mathcal{H}_{1:k}$, we have $|||f||| \leq \left\|\tilde{\Gamma}_{1:k}^{1/2} f\right\|_{\mathcal{H}} \leq \sqrt{2}|||f|||$ where $\tilde{\Gamma}_{1:k}$ is defined in (2.11). Therefore, $|||\cdot|||$'s dual norm $|||\cdot|||_*$ is also equivalent to $\left\|\tilde{\Gamma}_{1:k}^{1/2}\cdot\right\|_{\mathcal{H}}$'s dual norm which is given by $\left\|\tilde{\Gamma}_{1:k}^{-1/2}\cdot\right\|_{\mathcal{H}}$: for all $f \in \mathcal{H}_{1:k}$, $(1/\sqrt{2})|||f|||_* \leq \left\|\tilde{\Gamma}_{1:k}^{-1/2} f\right\|_{\mathcal{H}} \leq |||f|||_*$. Hence, for all $f_{1:k} \in f_{1:k}^* + \partial B$, we have

$$\begin{aligned}
|\mathcal{M}_{f_{1:k}}| &\leq 2\sqrt{2}\left\|\tilde{\Gamma}_{1:k}^{-1/2}\left(\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\right.\right. \\
&\quad \left.\left.(\mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^* + \boldsymbol{\xi}) - f_{1:k}^*\right)\right\|_{\mathcal{H}} \\
&\leq 2\sqrt{2}\left(\left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^*\right\|_{\mathcal{H}}\right. \\
&\quad \left. + \left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\boldsymbol{\xi}\right\|_{\mathcal{H}} + \left\|\tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^*\right\|_{\mathcal{H}}\right)
\end{aligned} \tag{2.26}$$

We next handle the first two terms in (2.26) in the next two lemmas.

**Lemma 5.** *Under the assumptions of Proposition 19,*

$$\begin{aligned}
&\left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}\mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^*\right\|_{\mathcal{H}} \\
&\leq \frac{4C_{15}C_9\kappa N\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}\sigma(\square, \triangle),
\end{aligned} \tag{2.27}$$

*where*

$$\sigma(\square, \triangle) := \begin{cases} \square, & \text{if } \triangle\sqrt{\sigma_1} \geq \square \\ \triangle\sqrt{\sigma_1}, & \text{otherwise.} \end{cases} \tag{2.28}$$

**Proof.** On the event $\Omega_0$ we have

$$\left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top\right\|_{\mathrm{op}} = \left\|\mathbb{X}_{\phi,1:k}\tilde{\Gamma}_{1:k}^{-1/2}\right\|_{\mathrm{op}} \le C_9\sqrt{N}\left\|\Gamma_{1:k}^{1/2}\tilde{\Gamma}_{1:k}^{-1/2}\right\|_{\mathrm{op}},$$

because of the isomorphic property of $\mathbb{X}_{\phi,1:k}$ and

$$\left\|\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}\right\|_{\mathrm{op}} \le \frac{4}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}.$$

Hence, on $\Omega_0$,

$$\begin{aligned}
&\left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}\mathbb{X}_{\phi,k+1:\infty}f_{k+1:\infty}^*\right\|_{\mathcal{H}}\\
&\le \left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top\right\|_{\mathrm{op}}\left\|\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}\right\|_{\mathrm{op}}\left\|\mathbb{X}_{\phi,k+1:\infty}f_{k+1:\infty}^*\right\|_2\\
&\le \frac{4C_{15}C_9\kappa N}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}\left\|\Gamma_{1:k}^{1/2}\tilde{\Gamma}_{1:k}^{-1/2}\right\|_{\mathrm{op}}\left\|\Gamma_{k+1:\infty}^{1/2}f_{k+1:\infty}^*\right\|_{\mathcal{H}}\\
&\le \frac{4C_{15}C_9\kappa N\left\|\Gamma_{k+1:\infty}^{1/2}f_{k+1:\infty}^*\right\|_{\mathcal{H}}}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}\sigma\left(\square,\triangle\right),
\end{aligned}$$

where the last inequality follows from the definition of $\Gamma_{1:k}$ and $\tilde{\Gamma}_{1:k}$. ∎

We define the sets

$$J_1 := \left\{j \in [k] :\, \sigma_j \ge \left(\frac{\square}{\triangle}\right)^2\right\}, \quad J_2 := [k]\backslash J_1.$$

They will match the definition of $J_1$ and $J_2$ in (2.3) once $\triangle$ and $\square$ have been chosen.

We prove the following lemma:

**Lemma 6.** *Under the assumptions of Proposition 19. We define*

$$t(\square,\triangle) := \frac{1}{\sigma^2\left(\square,\triangle\right)}\left(|J_1|\,\square^2 + \triangle^2\sum_{j\in J_2}\sigma_j\right). \tag{2.29}$$

*Recall $r$ and $\kappa_1$ from Theorem 5. There then exists an absolute constant $C_{13}$ depending only on $\kappa_1$ and there exists an absolute constant $C_{17}$ such that with probability at least $1-(C_{13}/\lfloor t(\square,\triangle)\rfloor)^{r/4} - \mathbb{P}(\Omega_0^c)$,*

$$\begin{aligned}
&\left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}\boldsymbol{\xi}\right\|_{\mathcal{H}}\\
&\le \frac{8C_{17}\sigma_\xi\sqrt{N}}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}\sqrt{|J_1|\,\square^2 + \triangle^2\sum_{j\in J_2}\sigma_j}.
\end{aligned} \tag{2.30}$$

**Proof.** Let $D = \tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}$. We calculate separately the upper bounds for $\sqrt{\mathrm{Tr}\left(DD^\top\right)}$ and $\|D\|_{\mathrm{op}}$. For the basis $(\varphi_j)_{j\in\mathbb{N}}$ of eigenfunctions $\Gamma$, and since Tr is independent with the choice of the basis, on $\Omega_0$ we have

$$\begin{aligned}
\mathrm{Tr}\left(DD^\top\right) &= \sum_{j\in\mathbb{N}}\left\langle DD^\top\varphi_j,\varphi_j\right\rangle_{\mathcal{H}} = \sum_{j\in\mathbb{N}}\left\|D^\top\varphi_j\right\|_2^2 = \sum_{j\in[k]}\left\|D^\top\varphi_j\right\|_2^2\\
&= \sum_{j=1}^k\left(\frac{\sqrt{\sigma_j}}{\square}\vee\frac{1}{\triangle}\right)^{-2}\left\|\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}\mathbb{X}_{\phi,1:k}\varphi_j\right\|_2^2\\
&\le \sum_{j=1}^k\left(\frac{\sqrt{\sigma_j}}{\square}\vee\frac{1}{\triangle}\right)^{-2}\left\|\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}\right\|_{\mathrm{op}}^2\left\|\mathbb{X}_{\phi,1:k}\varphi_j\right\|_2^2\\
&\le \sum_{j=1}^k\left(\frac{\sqrt{\sigma_j}}{\square}\vee\frac{1}{\triangle}\right)^{-2}\left(\lambda + \frac{\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}{4}\right)^{-2}C_9^2N\sigma_j,
\end{aligned}$$

where the last inequality follows from (2.12). Hence

$$\sqrt{\operatorname{Tr}\left(DD^\top\right)} \leq \frac{4C_9\sqrt{N}}{4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)} \sqrt{|J_1|\,\square^2 + \triangle^2 \sum_{j\in J_2} \sigma_j}. \tag{2.31}$$

This implies that $D$ is a Hilbert-Schmidt operator. Using the inequality $\left\|\Gamma_{1:k}^{1/2}\tilde{\Gamma}_{1:k}^{-1/2}\right\|_{\text{op}} \leq \sigma(\square, \triangle)$ we obtain

$$\begin{aligned}
\|D\|_{\text{op}} = \left\|D^\top\right\|_{\text{op}} &\leq \left\|\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1}\right\|_{\text{op}} \left\|\mathbb{X}_{\phi,1:k}\tilde{\Gamma}_{1:k}^{-1/2}\right\|_{\text{op}} \\
&\leq C_9\sqrt{N}\left\|\Gamma_{1:k}^{1/2}\tilde{\Gamma}_{1:k}^{-1/2}\right\|_{\text{op}} \cdot \frac{4}{4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)} \\
&\leq \frac{4C_9\sqrt{N}\sigma\left(\square, \triangle\right)}{4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}.
\end{aligned} \tag{2.32}$$

We finish the proof by Proposition 23 with the $k$ from Proposition 23 set as

$$k = \left\lfloor \frac{|J_1|\,\square^2 + \triangle^2 \sum_{j\in J_2}\sigma_j}{\sigma^2\left(\square, \triangle\right)} \right\rfloor$$

and with $C_{17} = \frac{3}{2}C_9$. ∎

**Bound of the quadratic term and choice of $\square$ and $\triangle$** In the previous section, we obtained an upper bound on $\mathcal{M}_{f_{1:k}}$. Our main approach, as outlined in the previous section, is to separately prove $\mathcal{Q}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$ in case [1], and $\mathcal{R}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$ in case [2]. Now that we have the upper bound for $\mathcal{M}_{f_{1:k}}$, it only remains to bound $\mathcal{Q}_{f_{1:k}}$ in case [1].

Before we begin, we need to make another classification. This time, the classification is based on the values of $\sigma(\square, \triangle)$. In the upcoming proof, we will *firstly* start by classifying based on $\sigma(\square, \triangle)$, and *then* proceed to prove the desired propositions separately in cases [1] and [2]. This parameter is crucial in the analysis as it determines whether the regularization is too strong, potentially completely submerging the signal. One can revisit the classification discussion regarding $\sigma_1 N$ and $4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)$ in Theorem 5. Doing so will reveal that this corresponds to the classification based on the values of $\sigma(\square, \triangle)$. When $\sigma_1 N$ is too small, it signifies excessive regularization that drowns out the signal.

**If $\sigma\left(\square, \triangle\right) = \square$** Let us first study case [1]. Consider $f_{1:k} \in \mathcal{H}_{1:k}$ such that $\left\|\Gamma_{1:k}^{1/2}(f_{1:k} - f_{1:k}^*)\right\|_{\mathcal{H}} = \square$ and $\|f_{1:k} - f_{1:k}^*\|_{\mathcal{H}} \leq \triangle$. In this case, we show that $\mathcal{Q}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$. Notice that on $\Omega_0$ we have

$$\begin{aligned}
\mathcal{Q}_{f_{1:k}} &= \left\|\left(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right)^{-1/2}\mathbb{X}_{\phi,1:k}(f_{1:k} - f_{1:k}^*)\right\|_2^2 \\
&\geq \left(\lambda + \frac{3\operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}{2}\right)^{-1} c_5^2 N \left\|\Gamma_{1:k}^{1/2}(f_{1:k} - f_{1:k}^*)\right\|_{\mathcal{H}}^2 \\
&= \frac{2c_5^2 N\square^2}{2\lambda + 3\operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}.
\end{aligned} \tag{2.33}$$

To prove that $\mathcal{Q}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$, it suffices to show that

$$\begin{aligned}
&\frac{c_5^2 N\square^2}{\sqrt{2}\left(2\lambda + 3\operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} \\
&> \left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\mathbb{X}_{\phi,k+1:\infty}f_{k+1:\infty}^*\right\|_{\mathcal{H}} \\
&+ \left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\boldsymbol{\xi}\right\|_{\mathcal{H}} + \left\|\tilde{\Gamma}_{1:k}^{-1/2}f_{1:k}^*\right\|_{\mathcal{H}}
\end{aligned}$$

Lemma 5 and Lemma 6 then make clear that is suffices to prove that the following conditions hold for well-chosen $\square$ and $\triangle$.

- $$\frac{c_5^2 N \square^2}{\sqrt{2}\left(2\lambda + 3\,\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} > \frac{4C_{15}C_9\kappa N \left\|\Gamma_{k+1:\infty}^{1/2} f^*_{k+1:\infty}\right\|_{\mathcal{H}}}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)} \sigma\left(\square, \triangle\right).$$

  This is equivalent to $\square > \frac{12\sqrt{2}C_{15}C_9}{c_5^2}\kappa \left\|\Gamma_{k+1:\infty}^{1/2} f^*_{k+1:\infty}\right\|_{\mathcal{H}}$.

- $$\frac{c_5^2 N \square^2}{\sqrt{2}\left(2\lambda + 3\,\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} > \frac{8C_{17}\sqrt{N}\sigma_\xi}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)} \sqrt{|J_1|\,\square^2},$$

  which holds if $\square > \frac{24\sqrt{2}C_{17}}{c_5^2}\sigma_\xi\sqrt{\frac{|J_1|}{N}}$.

- $$\frac{c_5^2 N \square^2}{\sqrt{2}\left(2\lambda + 3\,\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} > \frac{8C_{17}\sqrt{N}\sigma_\xi}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)} \sqrt{\triangle^2 \sum_{j\in J_2}\sigma_j},$$

  which holds if $\square > \left(\frac{24C_{17}}{c_5^2}\triangle\sigma_\xi\sqrt{\frac{2}{N}\sum_{j\in J_2}\sigma_j}\right)^{1/2}$.

- $$\frac{c_5^2 N \square^2}{\sqrt{2}\left(2\lambda + 3\,\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} > \left\|\tilde{\Gamma}_{1:k}^{-1/2} f^*_{1:k}\right\|_{\mathcal{H}},$$

  which is equivalent to $\square > \sqrt{\left\|\tilde{\Gamma}_{1:k}^{-1/2} f^*_{1:k}\right\|_{\mathcal{H}} \frac{\sqrt{2}(2\lambda + 3\,\mathrm{Tr}(\Gamma_{k+1:\infty}))}{c_5^2 N}}$.

In conclusion, there exists an absolute constant $C_{18}$ (for example, $C_{18} = \frac{24\sqrt{2}C_{15}C_{17}}{c_5^2}$) so that if we have

$$\begin{aligned}\square > C_{18}\kappa \max\Bigg\{&\sigma_\xi\sqrt{\frac{|J_1|}{N}}, \left(\triangle\sigma_\xi\sqrt{\frac{1}{N}\sum_{j\in J_2}\sigma_j}\right)^{1/2}, \left\|\Gamma_{k+1:\infty}^{1/2} f^*_{k+1:\infty}\right\|_{\mathcal{H}}, \\ &\sqrt{\left\|\tilde{\Gamma}_{1:k}^{-1/2} f^*_{1:k}\right\|_{\mathcal{H}} \frac{2\lambda + 3\,\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}{N}}\Bigg\},\end{aligned} \tag{2.34}$$

then $\mathcal{Q}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$.

In case [2]. We consider a function $f_{1:k} \in \mathcal{H}_{1:k}$ such that $\left\|\Gamma_{1:k}^{1/2}(f_{1:k} - f^*_{1:k})\right\|_{\mathcal{H}} \le \square$ and $\|f_{1:k} - f^*_{1:k}\|_{\mathcal{H}} = \triangle$. In this case, we show that $\mathcal{R}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$. Since $\mathcal{R}_{f_{1:k}} = \triangle^2$, this amounts to showing that

$$\begin{aligned}\triangle^2 > 2\sqrt{2}&\left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\mathbb{X}_{\phi,k+1:\infty} f^*_{k+1:\infty}\right\|_{\mathcal{H}} \\ &+ 2\sqrt{2}\left(\left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\boldsymbol{\xi}\right\|_{\mathcal{H}} + \left\|\tilde{\Gamma}_{1:k}^{-1/2} f^*_{1:k}\right\|_{\mathcal{H}}\right).\end{aligned}$$

By Lemma 5 and Lemma 6, $\triangle$ must satisfy the following conditions:

- $$\triangle^2 > \frac{4C_{15}C_9\kappa N \left\|\Gamma_{k+1:\infty}^{1/2} f^*_{k+1:\infty}\right\|_{\mathcal{H}}}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)} \sigma\left(\square, \triangle\right).$$

- $$\triangle^2 > \frac{8C_{17}\sqrt{N}\sigma_\xi}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)} \sqrt{|J_1|\,\square^2}.$$

- $$\triangle^2 > \frac{8C_{17}\sqrt{N}\sigma_\xi}{4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)} \sqrt{\triangle^2 \sum_{j\in J_2}\sigma_j},$$

  which is equivalent to $\triangle^2 > \frac{64C_{17}^2 N\sigma_\xi^2}{(4\lambda + \mathrm{Tr}(\Gamma_{k+1:\infty}))^2}\left(\sum_{j\in J_2}\sigma_j\right)$.

- $$\triangle^2 > \left\| \tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^* \right\|_{\mathcal{H}}.$$

Hence, there exists an absolute constant $C_{19}$ (for example, $C_{19} = 64 C_{15} C_{17}^2$). We need to choose

$$\triangle^2 > C_{19}\kappa \max\left\{ \frac{\sigma_\xi \square \sqrt{|J_1|\, N}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}, \frac{\sigma_\xi^2 N \sum_{j\in J_2} \sigma_j}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2}, \frac{N\square \left\| \Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^* \right\|_{\mathcal{H}}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}, \left\| \tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^* \right\|_{\mathcal{H}} \right\}. \tag{2.35}$$

Then $\mathcal{R}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$.

While we have shown that $\mathcal{Q}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$ in case [1] and $\mathcal{R}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$ in case [2] if (2.34) and (2.35) hold, our task is not yet complete because (2.35) and (2.34) do not explicitly define for $\triangle$ and $\square$. We next derive explicit definition for $\square$ and $\triangle$ through these two equations.

We fix $\triangle$ so that

$$\frac{\square}{\triangle} = \sqrt{\frac{\kappa_{DM}\,(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))}{N}} \tag{2.36}$$

and take $C_{20} = C_{18}^2 \kappa^2 \kappa_{DM}^{-1/2} \vee 2C_{19}\kappa$ and

$$\square > C_{20} \max\left\{ \sigma_\xi \sqrt{\frac{|J_1|}{N}}, \sigma_\xi \left( \frac{\sum_{j\in J_2} \sigma_j}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})} \right)^{1/2}, \left\| \Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^* \right\|_{\mathcal{H}}, \sqrt{\left\| \tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^* \right\|_{\mathcal{H}} \frac{2\lambda + 3\operatorname{Tr}(\Gamma_{k+1:\infty})}{N}} \right\}. \tag{2.37}$$

One may observe that the second term inside the max is different from the corresponding term in (2.34). However, if $\square > C_{20}\sigma_\xi \left( \frac{\sum_{j\in J_2} \sigma_j}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})} \right)^{1/2}$, and $\square, \triangle$ satisfy (2.36), it follows that

$$\begin{aligned} \frac{\square}{\sqrt{\triangle \sigma_\xi \sqrt{\frac{1}{N}\sum_{j\in J_2}\sigma_j}}} &= \sqrt{\frac{\square}{\triangle} \frac{\square}{\sigma_\xi \sqrt{\frac{1}{N}\sum_{j\in J_2}\sigma_j}}} \\ &> \left( \sqrt{\frac{\kappa_{DM}\,(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))}{N}} \cdot \frac{C_{20}\sigma_\xi \left( \frac{\sum_{j\in J_2}\sigma_j}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})} \right)^{1/2}}{\sigma_\xi \sqrt{\frac{1}{N}\sum_{j\in J_2}\sigma_j}} \right)^{1/2} \\ &= \kappa_{DM}^{1/4} \sqrt{C_{20}} \geq C_{18}\kappa. \end{aligned}$$

Hence the new choice of $\square$ in (2.37) satisfies (2.34).

We now need to check that for this choice of $\square$, (2.35) is also satisfied.

- $$\frac{\triangle^2}{\frac{\sigma_\xi \square \sqrt{|J_1| N}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}} = \square^2 \frac{N}{\kappa_{DM}\,(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))} \cdot \frac{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}{\sigma_\xi \square \sqrt{|J_1| N}} > \frac{C_{20}}{\kappa_{DM}} > C_{19}\kappa.$$

- $$\begin{aligned} \frac{\triangle^2}{\frac{\sigma_\xi^2 N \sum_{j\in J_2}\sigma_j}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2}} &= \square^2 \frac{N}{\kappa_{DM}\,(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))} \cdot \frac{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2}{\sigma_\xi^2 N \sum_{j\in J_2}\sigma_j} \\ &> C_{20}^2 \frac{\sigma_\xi^2 \sum_{j\in J_2}\sigma_j}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})} \frac{N}{\kappa_{DM}\,(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))} \cdot \frac{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2}{\sigma_\xi^2 N \sum_{j\in J_2}\sigma_j} > C_{19}\kappa. \end{aligned}$$

- $$\frac{\triangle^2}{\frac{N\square\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}}{4\lambda+\mathrm{Tr}(\Gamma_{k+1:\infty})}} = \square^2 \frac{N}{\kappa_{DM}\left(4\lambda+\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} \cdot \frac{4\lambda+\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}{N\square\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}}$$
$$> \frac{C_{20}}{\kappa_{DM}} > C_{19}\kappa.$$

- $$\frac{\triangle^2}{\left\|\tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^*\right\|_{\mathcal{H}}} = \square^2 \frac{N}{\kappa_{DM}\left(4\lambda+\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} \cdot \frac{1}{\left\|\tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^*\right\|_{\mathcal{H}}} > \frac{C_{20}}{2\kappa_{DM}} > C_{19}\kappa.$$

We deduce that with the right choice of the absolute constants, such a choice of $\square, \triangle$ satisfies (2.34) and (2.35).

We have established that by selecting appropriate values for $\square$ and $\triangle$, we can conclude the following: if $f_{1:k} - f_{1:k}^* \in \partial B$, then we necessarily have $\mathcal{L}_{f_{1:k}} > 0$, and thanks to a homogeneity argument, it follows that for all $f_{1:k} \notin f_{1:k}^* + B$, we have $\mathcal{L}_{f_{1:k}} \leq 0$ hence $\hat{f}_{1:k} \in f_{1:k}^* + B$.

In the beginning of the analysis, we assumed that $\sigma(\square, \triangle) = \square$, which is true if and only if $\sigma_1 \geq \kappa_{DM} \frac{4\lambda+\mathrm{Tr}(\Gamma_{k+1:\infty})}{N}$. Hence if this inequality is satisfied, $\square$ is an upper bound on the estimation error $\left\|\Gamma_{1:k}^{1/2}(\hat{f}_{1:k} - f_{1:k}^*)\right\|_{\mathcal{H}}$ and $\triangle$ is an upper bound on $\left\|\hat{f}_{1:k} - f_{1:k}^*\right\|_{\mathcal{H}}$. Notice also that $\tilde{\Gamma}_{1:k}^{-1/2} = U\tilde{D}_1^{-1/2}U^\top$ where $\tilde{D}_1^{-1/2} =: \square D_{1,\mathrm{thre}}^{-1/2}$. Hence we can express $\square$ as in Equation (2.9).

$$\square = C_{20} \max\left\{\sigma_\xi \sqrt{\frac{|J_1|}{N}}, \sigma_\xi \sqrt{\frac{\sum_{j\in J_2}\sigma_j}{4\lambda+\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}}, \left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}},\right.$$
$$\left.\left\|\tilde{\Gamma}_{1,\mathrm{thre}}^{-1/2} f_{1:k}^*\right\|_{\mathcal{H}} \frac{2\lambda+3\,\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}{N}\right\}.$$

**If $\sigma(\square, \triangle) = \triangle\sqrt{\sigma_1}$** In this case, it follows by definition that $J_1 = \emptyset$, $J_2 = [k]$,

$$t(\square, \triangle) = \frac{\mathrm{Tr}\left(\Gamma_{1:k}\right)}{\sigma_1}, \text{ and } \tilde{D}_1^{1/2} = \frac{1}{\triangle}\,\mathrm{diag}\left(1, \cdots, 1, 0, \cdots\right),$$

where there are $k$ ones in the definition of $\tilde{D}_1^{1/2}$. Since we have completed a similar proof in the previous paragraph, we will expedite the presentation in this paragraph.

Suppose that $\left\|\Gamma_{1:k}^{1/2}(f_{1:k} - f_{1:k}^*)\right\|_{\mathcal{H}} = \square$ and $\|f_{1:k} - f_{1:k}^*\|_{\mathcal{H}} \leq \triangle$. As we discussed in the previous subsections, on $\Omega_0$, $\mathcal{Q}_{f_{1:k}} \geq \frac{N\square^2}{4\lambda+6\,\mathrm{Tr}(\Gamma_{k+1:\infty})}$. To show that $\mathcal{Q}_{f_{1:k}} > \mathcal{M}_{f_{1:k}}$, it suffices to show that

$$\frac{c_5^2 N\square^2}{\sqrt{2}\left(2\lambda+3\,\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)}$$
$$> \left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^*\right\|_{\mathcal{H}}$$
$$+ \left\|\tilde{\Gamma}_{1:k}^{-1/2}\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\boldsymbol{\xi}\right\|_{\mathcal{H}} + \left\|\tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^*\right\|_{\mathcal{H}}.$$

Recall that Lemma 5 and Lemma 6 hold true for all possible values of $\square, \triangle$, so we can still use them in the current setting. Hence:

- $$\frac{c_5^2 N\square^2}{\sqrt{2}\left(2\lambda+3\,\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} > \frac{4C_{15}C_9\kappa N\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}}{4\lambda+\mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)}\sigma(\square, \triangle),$$

  which is equivalent to $\square^2 > \frac{12\sqrt{2}C_{15}C_9}{c_5^2}\kappa\triangle\sqrt{\sigma_1}\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}$.

- $$\frac{c_5^2 N \square^2}{\sqrt{2}\left(2\lambda + 3\operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} > \frac{8C_{17}\sqrt{N}\sigma_\xi}{4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}\sqrt{|J_1|\,\square^2},$$

  which is true since $|J_1| = 0$.

- $$\frac{c_5^2 N \square^2}{\sqrt{2}\left(2\lambda + 3\operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} > \frac{8C_{17}\sqrt{N}\sigma_\xi}{4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}\sqrt{\triangle^2 \sum_{j\in J_2}\sigma_j},$$

  which is true if $\square^2 > \frac{24\sqrt{2}C_{17}}{c_5^2}\sigma_\xi \triangle \sqrt{\frac{\operatorname{Tr}(\Gamma_{1:k})}{N}}$.

- $$\frac{c_5^2 N \square^2}{\sqrt{2}\left(2\lambda + 3\operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)\right)} > \left\|\tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^*\right\|_{\mathcal{H}},$$

  which is true if $\square^2 > \frac{\sqrt{2}}{c_5^2}\left\|\tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^*\right\|_{\mathcal{H}} \frac{2\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}{N}$.

We conclude that there exists an absolute constant $C_{21} = \frac{24\sqrt{2}C_{15}C_{17}}{c_5^2}\kappa$, such that we can take

$$\square^2 > C_{21}\max\left\{\triangle\sqrt{\sigma_1}\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}, \sigma_\xi\triangle\sqrt{\frac{\operatorname{Tr}\left(\Gamma_{1:k}\right)}{N}}, \frac{2\lambda + 3\operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}{N}\left\|\tilde{\Gamma}_{1:k}^{-1/2} f_{1:k}^*\right\|_{\mathcal{H}}\right\}.$$

Since $\left\|\tilde{\Gamma}_{1:k}^{-1/2}\right\|_{\mathrm{op}} = \triangle$, it suffices to choose

$$\square^2 > C_{21}\max\left\{\triangle\sqrt{\sigma_1}\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}, \sigma_\xi\triangle\sqrt{\frac{\operatorname{Tr}\left(\Gamma_{1:k}\right)}{N}}, \frac{2\lambda + 3\operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}{N}\triangle\left\|f_{1:k}^*\right\|_{\mathcal{H}}\right\}.$$

In the case where $\left\|\Gamma_{1:k}^{1/2}(\hat{f}_{1:k} - f_{1:k}^*)\right\|_{\mathcal{H}} \le \square$ and $\left\|\hat{f}_{1:k} - f_{1:k}^*\right\|_{\mathcal{H}} = \triangle$, a similar analysis gives us that there exists an absolute constant $C_{22}$ depending on $\kappa$ such that we can take

$$\triangle^2 > C_{22}\max\left\{\left\|f_{1:k}^*\right\|_{\mathcal{H}}^2, \frac{N\sigma_\xi^2\operatorname{Tr}\left(\Gamma_{1:k}\right)}{\left(4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)\right)^2}, \frac{\sigma_1 N^2\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}^2}{\left(4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)\right)^2}\right\}. \tag{2.38}$$

Again, we choose that $\triangle = \square\sqrt{N/\left(\kappa_{DM}(4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right))\right)}$. There exists an absolute constant $C_{23}$ such that we can express $\square$ as in (2.8).

$$\square = C_{23}\max\left\{\sigma_\xi\sqrt{\frac{\operatorname{Tr}\left(\Gamma_{1:k}\right)}{4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}}, \sqrt{\frac{\sigma_1 N}{4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}}\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}, \left\|f_{1:k}^*\right\|_{\mathcal{H}}\sqrt{\frac{4\lambda + \operatorname{Tr}\left(\Gamma_{k+1:\infty}\right)}{N}}\right\}.$$

In particular, we check that for this choice, $\triangle$ satisfies (2.38).

Summarizing the above arguments, we have established the following proposition.

**Proposition 21.** *Under the assumption of Theorem 5, there exist absolute constants $C_{20}$, $C_{13}$ and $C_{23}$ such that the following holds for all such $k$'s and all $\lambda \geq 0$. Recall the definition of $t(\square, \triangle)$ from (2.29), with probability at least $1 - (C_{13}/\lfloor t(\square, \triangle)\rfloor)^{r/4} - \mathbb{P}(\Omega_0^c)$ we have*

$$\left\|\Gamma_{1:k}^{1/2}\left(\hat{f}_{1:k} - f^*_{1:k}\right)\right\|_{\mathcal{H}} \leq \square, \quad \left\|\hat{f}_{1:k} - f^*_{1:k}\right\|_{\mathcal{H}} \leq \square\sqrt{\frac{N}{\kappa_{DM}(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))}},$$

*where*

1. *If $\sigma_1 N \leq \kappa_{DM}(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))$,*

$$\square = C_{20}\max\left\{\sigma_\xi\sqrt{\frac{\operatorname{Tr}(\Gamma_{1:k})}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}}, \sqrt{\frac{\sigma_1 N}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}}\left\|\Gamma_{k+1:\infty}^{1/2} f^*_{k+1:\infty}\right\|_{\mathcal{H}}, \right.$$
$$\left. \|f^*_{1:k}\|_{\mathcal{H}}\sqrt{\frac{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}{N}}\right\}.$$

2. *If $\sigma_1 N > \kappa_{DM}(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))$,*

$$\square = C_{23}\max\left\{\sigma_\xi\sqrt{\frac{|J_1|}{N}}, \sigma_\xi\sqrt{\frac{\sum_{j\in J_2}\sigma_j}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}}, \left\|\Gamma_{k+1:\infty}^{1/2} f^*_{k+1:\infty}\right\|_{\mathcal{H}}, \right.$$
$$\left. \left\|\tilde{\Gamma}_{1,\mathrm{thre}}^{-1/2} f^*_{1:k}\right\|_{\mathcal{H}}\frac{2\lambda + 3\operatorname{Tr}(\Gamma_{k+1:\infty})}{N}\right\}.$$

**Upper bound on $\left\|\Gamma_{k+1:\infty}^{1/2}(\hat{f}_{k+1:\infty} - f^*_{k+1:\infty})\right\|_{\mathcal{H}}$**

We do not expect $\hat{f}_{k+1:\infty}$ to be a good estimator of $f^*_{k+1:\infty}$ because the minimum $\|\cdot\|_{\mathcal{H}}$-norm estimator $\hat{f}$ is using the 'remaining part' of $\mathcal{H}$ endowed by the eigenfunctions $(\varphi_j)_{j\geq k+1}$ of $\Gamma$ (we denoted this space by $\mathcal{H}_{k+1:\infty}$) to absorb the influence of noise introduced by $\boldsymbol{\xi}$ and not to estimate $f^*_{k+1:\infty}$ which is why we call the error term $\left\|\Gamma_{k+1:\infty}^{1/2}(\hat{f}_{k+1:\infty} - f^*_{k+1:\infty})\right\|_{\mathcal{H}}$ a price for noise absorption instead of an estimation error. A consequence is that we can only upper bound this term by

$$\left\|\Gamma_{k+1:\infty}^{1/2}(\hat{f}_{k+1:\infty} - f^*_{k+1:\infty})\right\|_{\mathcal{H}} \leq \left\|\Gamma_{k+1:\infty}^{1/2}\hat{f}_{k+1:\infty}\right\|_{\mathcal{H}} + \left\|\Gamma_{k+1:\infty}^{1/2} f^*_{k+1:\infty}\right\|_{\mathcal{H}}.$$

We then just need to find a high probability upper bound on $\left\|\Gamma_{k+1:\infty}^{1/2}\hat{f}_{k+1:\infty}\right\|_{\mathcal{H}}$.

We have for $A := \mathbb{X}_{\phi,k+1:\infty}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}$,

$$\begin{aligned}\left\|\Gamma_{k+1:\infty}^{1/2}\hat{f}_{k+1:\infty}\right\|_{\mathcal{H}} &= \left\|\Gamma_{k+1:\infty}^{1/2} A(\boldsymbol{y} - \mathbb{X}_{\phi,1:k}\hat{f}_{1:k})\right\|_{\mathcal{H}} \\ &\leq \left\|\Gamma_{k+1:\infty}^{1/2} A\mathbb{X}_{\phi,1:k}(f^*_{1:k} - \hat{f}_{1:k})\right\|_{\mathcal{H}} + \left\|\Gamma_{k+1:\infty}^{1/2} A\mathbb{X}_{\phi,k+1:\infty} f^*_{k+1:\infty}\right\|_{\mathcal{H}} + \left\|\Gamma_{k+1:\infty}^{1/2} A\boldsymbol{\xi}\right\|_{\mathcal{H}} \qquad (2.39)\end{aligned}$$

and now we obtain high probability upper bounds on the three terms in (2.39).

On $\Omega_0$, for all $\boldsymbol{\lambda} \in \mathbb{R}^N$,

$$\left\|\Gamma_{k+1:\infty}^{1/2}\mathbb{X}_{\phi,k+1:\infty}^\top\boldsymbol{\lambda}\right\|_{\mathcal{H}} \leq C_8\left(\sqrt{\operatorname{Tr}(\Gamma^2_{k+1:\infty})} + \sqrt{N}\,\|\Gamma_{k+1:\infty}\|_{\mathrm{op}}\right)\|\boldsymbol{\lambda}\|_2. \qquad (2.40)$$

Notice that this result holds without any extra assumption on $N$. We have

$$\begin{aligned}&\left\|\Gamma_{k+1:\infty}^{1/2} A\mathbb{X}_{\phi,1:k}(f^*_{1:k} - \hat{f}_{1:k})\right\|_{\mathcal{H}} \\ &= \left\|\Gamma_{k+1:\infty}^{1/2}\mathbb{X}_{\phi,k+1:\infty}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\mathbb{X}_{\phi,1:k}(f^*_{1:k} - \hat{f}_{1:k})\right\|_{\mathcal{H}} \\ &\leq \left\|\Gamma_{k+1:\infty}^{1/2}\mathbb{X}_{\phi,k+1:\infty}^\top\right\|_{\mathrm{op}}\left\|(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\right\|_{\mathrm{op}}\left\|\mathbb{X}_{\phi,1:k}(f^*_{1:k} - \hat{f}_{1:k})\right\|_2 \qquad (2.41) \\ &\leq 4C_8C_9\frac{\left(\sqrt{N\operatorname{Tr}(\Gamma^2_{k+1:\infty})} + N\,\|\Gamma_{k+1:\infty}\|_{\mathrm{op}}\right)}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}\left\|\Gamma_{1:k}^{1/2}(f^*_{1:k} - \hat{f}_{1:k})\right\|_{\mathcal{H}}.\end{aligned}$$

On $\Omega_0$,

$$\begin{aligned}
&\left\|\Gamma_{k+1:\infty}^{1/2} A \mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^*\right\|_{\mathcal{H}} \\
&\le \left\|\Gamma_{k+1:\infty}^{1/2} \mathbb{X}_{\phi,k+1:\infty}^\top\right\|_{\text{op}} \left\|(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\right\|_{\text{op}} \left\|\mathbb{X}_{\phi,k+1:\infty} f_{k+1:\infty}^*\right\|_2 \\
&\le 4C_8C_{15}\kappa \frac{\sqrt{N \operatorname{Tr}(\Gamma_{k+1:\infty}^2)} + N \left\|\Gamma_{k+1:\infty}\right\|_{\text{op}}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})} \left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}}.
\end{aligned} \tag{2.42}$$

Finally, let $D = \Gamma_{k+1:\infty}^{1/2} A$. As $\mathbb{X}_{\phi,k+1:\infty}\Gamma_{k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top : \mathbb{R}^N \to \mathbb{R}^N$,

$$\begin{aligned}
\operatorname{Tr}\left(\mathbb{X}_{\phi,k+1:\infty}\Gamma_{k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top\right) &= \left\|\mathbb{X}_{\phi,k+1:\infty}\Gamma_{k+1:\infty}^{1/2}\right\|_{HS}^2 \\
&= \sum_{i=1}^N \left\|\left(\Gamma_{k+1:\infty}^{1/2}\phi_{k+1:\infty}\right)(X_i)\right\|_{\mathcal{H}}^2
\end{aligned}$$

is the sum of $N$ i.i.d. random variables appearing in (2.16). On $\Omega_0$,

$$\sum_{i=1}^N \left\|\left(\Gamma_{k+1:\infty}^{1/2}\phi_{k+1:\infty}\right)(X_i)\right\|_{\mathcal{H}}^2 \le C_{16} N \operatorname{Tr}\left(\Gamma_{k+1:\infty}^2\right).$$

So

$$\operatorname{Tr}(DD^\top) = \operatorname{Tr}\left(D^\top D\right) \le \frac{\operatorname{Tr}\left(\mathbb{X}_{\phi,k+1:\infty}\Gamma_{k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top\right)}{\left\|\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N\right\|_{\text{op}}^2} \le \frac{16C_{16}N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)}{\left(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})\right)^2} \tag{2.43}$$

and

$$\begin{aligned}
\|D\|_{\text{op}} &= \left\|\Gamma_{k+1:\infty}^{1/2}\mathbb{X}_{\phi,k+1:\infty}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\right\|_{\text{op}} \\
&\le \left\|\Gamma_{k+1:\infty}^{1/2}\mathbb{X}_{\phi,k+1:\infty}^\top\right\|_{\text{op}} \left\|(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\right\|_{\text{op}} \\
&\le \frac{4C_8}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}\left(\sqrt{\operatorname{Tr}(\Gamma_{k+1:\infty}^2)} + \sqrt{N}\left\|\Gamma_{k+1:\infty}\right\|_{\text{op}}\right).
\end{aligned} \tag{2.44}$$

Set $k$ from Proposition 23 as

$$k = \left\lfloor \frac{\sqrt{C_{16}}}{C_8} \frac{\sqrt{N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)}}{\sqrt{\operatorname{Tr}(\Gamma_{k+1:\infty}^2)} + \sqrt{N}\left\|\Gamma_{k+1:\infty}\right\|_{\text{op}}} \right\rfloor^2 =: C_{13}(\bar{p}_\xi)^{-4/r}. \tag{2.45}$$

Then by Proposition 23, with probability at least $1 - \bar{p}_\xi - \mathbb{P}(\Omega_0^c)$, we have

$$\left\|\Gamma_{k+1:\infty}^{1/2} A\boldsymbol{\xi}\right\|_{\mathcal{H}} \le \frac{3}{2}\sigma_\xi \frac{\sqrt{16C_{16}N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}. \tag{2.46}$$

Let us summarize the above discussion into the following Proposition:

**Proposition 22.** *Under the assumption of Theorem 5. The following then holds for all such $k$'s and all $\lambda \ge 0$. With probability at least $1 - \bar{p}_\xi - \mathbb{P}(\Omega_0^c)$ we have*

$$\begin{aligned}
&\left\|\hat{f}_{k+1:\infty} - f_{k+1:\infty}^*\right\|_{L_2} \le C_8 \frac{\left(\sqrt{N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)} + N\left\|\Gamma_{k+1:\infty}\right\|_{op}\right)}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})} \left\|\Gamma_{1:k}^{1/2}(f_{1:k}^* - \hat{f}_{1:k})\right\|_{\mathcal{H}} \\
&+ 4C_8C_{16}\kappa \frac{\sqrt{N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)} + N\left\|\Gamma_{k+1:\infty}\right\|_{op}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})} \left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}} + \left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^*\right\|_{\mathcal{H}} \\
&+ \frac{3}{2}\sigma_\xi \frac{\sqrt{16C_{16}N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}.
\end{aligned} \tag{2.47}$$

**Concentration of noise**

The following lemma is taken from [P4, Lemma 10].

**Lemma 7.** *Let $\boldsymbol{\xi} = (\xi_i)_{i=1}^N$ be a random vector with independent mean zero and variance $\sigma_\xi$ real-valued coordinates. We assume that for all $i$'s, $\|\xi_i\|_{L_r} \leq \kappa_1 \sigma_\xi$ for some $\kappa_1 > 0$ and $r > 4$. There then exists some absolute constant $C_{\kappa_1}$ (depending only on $\kappa_1$) such that for any matrix $D \in \mathbb{R}^{p\times N}$ the following holds: if for some integer $k$ for which $\sqrt{k}\|D\|_{op} \leq \sqrt{\mathrm{Tr}(DD^\top)}$ then with probability at least $1-(C_{\kappa_1}/k)^{r/4}$,*

$$\|D\boldsymbol{\xi}\|_2 \leq (3/2)\sigma_\xi \sqrt{\mathrm{Tr}(DD^\top)}.$$

We emphasize that Lemma 7 does not depend on $p$, so we can set $p = \infty$. Note that there exists an isometric embedding from $\mathcal{H}$ to $\ell_2$, given by $f \in \mathcal{H} \mapsto \sum_{j=1}^\infty \langle f, \varphi_j\rangle_{\mathcal{H}} \boldsymbol{e}_j$, where we recall that $(\boldsymbol{e}_j)_{j=1}^\infty$ is ONB of $\ell_2$. Therefore, we extend $D : \mathbb{R}^N \to \ell_2$ to $D : \mathbb{R}^N \to \mathcal{H}$. As a result, we have the following proposition:

**Proposition 23.** *Let $\boldsymbol{\xi} = (\xi_i)_{i=1}^N$ be a random vector with independent mean zero and variance $\sigma_\xi$ real-valued coordinates. We assume that for all $i$'s, $\|\xi_i\|_{L_r} \leq \kappa_1 \sigma_\xi$ for some $\kappa_1 > 0$ and $r > 4$. There then exists some absolute constant $C_{13}$ (depending only on $\kappa_1$) such that for any Hilbert-Schmidt operator $D : \mathbb{R}^N \to \mathcal{H}$ the following holds: if for some integer $k$ for which $\sqrt{k}\|D\|_{op} \leq \sqrt{\mathrm{Tr}(DD^\top)}$ then with probability at least $1-(C_{13}/k)^{r/4}$,*

$$\|D\boldsymbol{\xi}\|_{\mathcal{H}} \leq (3/2)\sigma_\xi \sqrt{\mathrm{Tr}(DD^\top)}.$$

Proposition 23 is used to establish Lemma 6, which concerns the concentration property of the noise. To extend the control of the noise to the model-misspecified setting, we establish the following proposition.

**Proposition 24.** *Suppose $f^* - f^{**} \in L_{2+\varepsilon}(\mu)$ for some $\varepsilon \geq 0$, where we recall that $\mu$ is the probability distribution of design vector $X$. Suppose $\xi, \xi_1, \cdots, \xi_N$ are i.i.d. mean zero sub-Gaussian random variables with variance $\sigma_\xi^2$ and suppose $\xi$ is independent with $X$. Let $\boldsymbol{\epsilon} = (f^*(X_i) - f^{**}(X_i) + \xi_i)_{i=1}^N$.*

1. *When $A = (\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\mathbb{X}_{\phi,1:k}\tilde{\Gamma}_{1:k}^{-1}\mathbb{X}_{\phi,1:k}^\top(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}$.*

   (a) *When $\varepsilon > 0$. Suppose $N \geq e^{-1}(\exp(2c_9/\varepsilon) - 4)$. Recall the definition of $\sigma(\square, \triangle)$ from (2.28) and the definition of $t(\square, \triangle)$ from (2.29). There exist absolute constants $C_{25}$ depending on $\varepsilon$, $C_9$ and $c_{11}$ such that for any $t_1 \geq 0$ and $t_3 > 2$,*

$$\begin{aligned}
&\mathbb{P}\Bigg(\boldsymbol{\epsilon}^\top A\boldsymbol{\epsilon} \leq 2(1+t_1)\sigma_\xi^2 \frac{16C_9^2 N\left(|J_1|\,\square^2 + \triangle^2 \sum_{j\in J_2}\sigma_j\right)}{\left(4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)^2} \\
&+ t_3^2 C_{25}^2 \frac{\left(\sqrt{N}\sigma\left(\square, \triangle\right)\right)^2}{\left(4\lambda + \mathrm{Tr}\left(\Gamma_{k+1:\infty}\right)\right)^2}\|f^* - f^{**}\|_{L_{2+\varepsilon}}^2 N\Bigg) \\
&\geq 1 - \mathbb{P}(\Omega_0^c) - \exp\left(-c(t_1^2 \wedge t_1)\frac{|J_1|\,\square^2 + \triangle^2\sum_{j\in J_2}\sigma_j}{\sigma^2\left(\square, \triangle\right)}\right) - c_{11}t_3^{-(2+\varepsilon)}N^{-\frac{\varepsilon}{4}}.
\end{aligned} \tag{2.48}$$

   (b) *When $\varepsilon = 0$, (2.48) is still valid with not necessarily $N \geq e^{-1}(\exp(2c_9/\varepsilon) - 4)$, but with $C_{25}$ replaced by $C_{24}$, where $C_{24} = 16C_9^2$; and $c_{11}$ replaced by 1.*

2. *When*

$$\begin{aligned}
A =&(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}\mathbb{X}_{\phi,k+1:\infty}\Gamma_{k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top \\
&(\mathbb{X}_{\phi,k+1:\infty}\mathbb{X}_{\phi,k+1:\infty}^\top + \lambda I_N)^{-1}.
\end{aligned}$$

   (a) *When $\varepsilon > 0$. Suppose $N \geq e^{-1}(\exp(2c_9/\varepsilon) - 4)$. There exist absolute constants $C_{26}$ depending on $\varepsilon$, $C_{16}$,*

*$C_8$ and $c_{11}$, such that for any $t_1, t_2 \geq 0$ and $t_3 > 2$,*

$$\begin{aligned}
&\mathbb{P}\Bigg(\boldsymbol{\epsilon}^\top A\boldsymbol{\epsilon} \leq (1+t_1)\sigma_\xi^2 \frac{16C_{16}^2 N \operatorname{Tr}(\Gamma_{k+1:\infty}^2)}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2} \\
&+ t_2 t_3 \sigma_\xi \frac{C_{26}\left(\operatorname{Tr}(\Gamma_{k+1:\infty}^2) + N\left\|\Gamma_{k+1:\infty}\right\|_{op}^2\right)}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2}\sqrt{N}\left\|f^* - f^{**}\right\|_{L_2} \\
&+ t_3^2 \frac{C_{26}\left(\operatorname{Tr}(\Gamma_{k+1:\infty}^2) + N\left\|\Gamma_{k+1:\infty}\right\|_{op}^2\right)}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2} N\left\|f^* - f^{**}\right\|_{L_{2+\varepsilon}}^2\Bigg) \\
&\geq 1 - \mathbb{P}(\Omega_0^c) - \exp(-ct_2^2) - c_{11}t_3^{-(2+\varepsilon)}N^{-\frac{\varepsilon}{4}} \\
&- \exp\left(-c(t_1^2 \wedge t_1)\frac{8C_{16}N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)}{C_8^2\left(\operatorname{Tr}(\Gamma_{k+1:\infty}^2) + N\left\|\Gamma_{k+1:\infty}\right\|_{op}\right)}\right)
\end{aligned} \tag{2.49}$$

*(b) When $\varepsilon = 0$, (2.49) is still valid with not necessarily $N \geq e^{-1}(\exp(2c_9/\varepsilon) - 4)$, but with $C_{26}$ replaced by $32C_8c_{10}$ and with $c_{11}$ replaced by 1.*

**Proof.** Recall that $\boldsymbol{\epsilon} = \boldsymbol{r} + \boldsymbol{\xi}$. We have:

$$\boldsymbol{\epsilon}^\top A\boldsymbol{\epsilon} = \boldsymbol{r}^\top A\boldsymbol{r} + \boldsymbol{\xi}^\top A\boldsymbol{\xi} + \boldsymbol{r}^\top A\boldsymbol{\xi} + \boldsymbol{\xi}^\top A\boldsymbol{r} \leq 2\|A\|_{\mathrm{op}}\|\boldsymbol{r}\|_2^2 + 2\boldsymbol{\xi}^\top A\boldsymbol{\xi}.$$

By Hanson-Wright inequality, see, for example [Ver18, Theorem 6.2.1], there exists some absolute constant $c > 0$ such that for any $t_1 \geq 0$,

$$\mathbb{P}\left(\boldsymbol{\xi}^\top A\boldsymbol{\xi} - \sigma_\xi^2\operatorname{Tr}(A) \leq t_1\sigma_\xi^2\operatorname{Tr}(A)\right) \geq 1 - \exp\left(-c(t_1^2 \wedge t_1)\frac{\operatorname{Tr}(A)}{\|A\|_{\mathrm{op}}}\right).$$

Let $\Omega_{\mathrm{noise},1}$ as the random event on which

$$\sqrt{\operatorname{Tr}(A)} \leq \frac{4C_9\sqrt{N}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}\sqrt{|J_1|\square^2 + \triangle^2\sum_{j\in J_2}\sigma_j}, \text{ and } \|A\|_{\mathrm{op}}^{1/2} \leq \frac{4C_9\sqrt{N}\sigma(\square,\triangle)}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}.$$

By (2.31) and (2.32), $\mathbb{P}(\Omega_{\mathrm{noise},1}) \geq 1 - \mathbb{P}(\Omega_0^c)$. In Lemma 4, let $Z = f^*(X) - f^{**}(X)$, $q = 2+\varepsilon$ for some $\varepsilon > 0$, $r = 2$, $p = 1$. When $N \geq e^{-1}(\exp(2c_9/\varepsilon) - 4)$, by Lemma 4, we have $j_0 = 1$ and thus for $\beta = \varepsilon/4$, for any $t_3 > 2$, with probability at least $1 - c_{11}t_3^{-(2+\varepsilon)}N^{-\frac{\varepsilon}{4}}$,

$$\|\boldsymbol{r}\|_2 \leq c_{10}\left(\frac{2+\varepsilon}{2+\varepsilon-2(\varepsilon/4+1)}\right)^{1/2} t_3\left\|f^* - f^{**}\right\|_{L_{2+\varepsilon}}\sqrt{N}.$$

Combining the above together, we obtain that for any $t_1 \geq 0$ and $t_3 > 2$,

$$\begin{aligned}
&\mathbb{P}\Bigg(\boldsymbol{\epsilon}^\top A\boldsymbol{\epsilon} \leq 2(1+t_1)\sigma_\xi^2\frac{16C_9^2N\left(|J_1|\square^2 + \triangle^2\sum_{j\in J_2}\sigma_j\right)}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2} \\
&\quad + t_3^2C_{25}^2\frac{\left(\sqrt{N}\sigma(\square,\triangle)\right)^2}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2}\left\|f^* - f^{**}\right\|_{L_{2+\varepsilon}}^2 N\Bigg) \\
&\geq 1 - \mathbb{P}(\Omega_0^c) - \exp\left(-c(t_1^2 \wedge t_1)\frac{|J_1|\square^2 + \triangle^2\sum_{j\in J_2}\sigma_j}{\sigma^2(\square,\triangle)}\right) - c_{11}t_3^{-(2+\varepsilon)}N^{-\frac{\varepsilon}{4}},
\end{aligned}$$

where $C_{25} = 16C_9^2c_{10}\left(\frac{2+\varepsilon}{2+\varepsilon-2(\varepsilon/4+1)}\right)^{1/2}$. In particular, when we use Markov's inequality to replace Lemma 4, with the probability deviation $t_3^{-2}$ instead of $c_{11}t_3^{-(2+\varepsilon)}N^{-\frac{\varepsilon}{4}}$, the term $\sqrt{N}\left\|f^* - f^{**}\right\|_{L_{2+\varepsilon}}$ can be improved to $\sqrt{N}\left\|f^* - f^{**}\right\|_{L_2}$.

Similarly, let $\Omega_{\mathrm{noise},2}$ as the random event on which

$$\begin{aligned}
&\operatorname{Tr}(A) \leq \frac{16C_{16}N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)}{\left(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})\right)^2} \text{ and} \\
&\|A\|_{\mathrm{op}}^{1/2} \leq \frac{4C_8}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}\left(\sqrt{\operatorname{Tr}(\Gamma_{k+1:\infty}^2)} + \sqrt{N}\left\|\Gamma_{k+1:\infty}\right\|_{\mathrm{op}}\right).
\end{aligned}$$

By (2.43) and (2.44), $\mathbb{P}(\Omega_{\text{noise},2}) \geq 1 - \mathbb{P}(\Omega_0^c)$. Repeat the above arguments, we obtain that for any $t_1, t_2 \geq 0$ and $t_3 > 2$,

$$\begin{aligned}
&\mathbb{P}\Bigg(\boldsymbol{\epsilon}^\top A\boldsymbol{\epsilon} \leq (1+t_1)\sigma_\xi^2 \frac{16C_{16}^2 N \operatorname{Tr}(\Gamma_{k+1:\infty}^2)}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2} \\
&+ t_2 t_3 \sigma_\xi \frac{C_{26}\left(\operatorname{Tr}(\Gamma_{k+1:\infty}^2) + N\left\|\Gamma_{k+1:\infty}\right\|_{\text{op}}^2\right)}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2} \sqrt{N}\left\|f^* - f^{**}\right\|_{L_2} \\
&+ t_3^2 \frac{C_{26}\left(\operatorname{Tr}(\Gamma_{k+1:\infty}^2) + N\left\|\Gamma_{k+1:\infty}\right\|_{\text{op}}^2\right)}{(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))^2} N\left\|f^* - f^{**}\right\|_{L_{2+\varepsilon}}^2\Bigg) \\
&\geq 1 - \mathbb{P}(\Omega_0^c) - \exp(-ct_2^2) - c_{11} t_3^{-(2+\varepsilon)} N^{-\frac{\varepsilon}{4}} \\
&- \exp\left(-c(t_1^2 \wedge t_1) \frac{8C_{16} N \operatorname{Tr}(\Gamma_{k+1:\infty}^2)}{C_8^2\left(\operatorname{Tr}(\Gamma_{k+1:\infty}^2) + N\left\|\Gamma_{k+1:\infty}\right\|_{\text{op}}\right)}\right)
\end{aligned}$$

where $C_{26} = 32C_8c_{10}\left(\frac{2+\varepsilon}{2+\varepsilon-2(\varepsilon/4+1)}\right)^{1/2}$. ■

For simplicity, we only consider the case where $\sigma_1 N > \kappa_{DM}(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))$ and $k \lesssim N$. Replace the usage of Proposition 23 in Section 2.3 with Proposition 24 and repeat the proof. We can conclude as follows:

**Proposition 25.** *Grant the assumptions of Theorem 5 except that $f^* \in \mathcal{H}$ and Proposition 24. Assume that $\sigma_1 N > \kappa_{DM}(4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty}))$. Then with the same probability deviation as in Theorem 5 but with $\bar{p}_\xi$ replaced by (2.48) and*

$$\left(\frac{C_{13}\operatorname{Tr}(\Gamma_{k+1:\infty})}{|J_1|\operatorname{Tr}(\Gamma_{k+1:\infty}) + N\left(\sum_{j\in J_2}\sigma_j\right)}\right)^{\frac{r}{4}}$$

*replaced by (2.49), we have:*

$$\begin{aligned}
\left\|\hat{f}_\lambda - f^*\right\|_{L_2} \lesssim{}& \left\|f^* - f^{**}\right\|_{L_2} + \sigma_\xi\sqrt{\frac{|J_1|}{N}} + \sigma_\xi\sqrt{\frac{\sum_{j\in J_2}\sigma_j}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}} + \left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^{**}\right\|_{\mathcal{H}} \\
&+ \left\|\tilde{\Gamma}_{1,\text{thre}}^{-1/2} f_{1:k}^{**}\right\|_{\mathcal{H}} \frac{2\lambda + 3\operatorname{Tr}(\Gamma_{k+1:\infty})}{N} + 4C_{16}\sigma_\xi \frac{\sqrt{N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})} \\
&+ \frac{C_{26}^{1/2}\sqrt{\operatorname{Tr}(\Gamma_{k+1:\infty}^2) + N\left\|\Gamma_{k+1:\infty}\right\|_{op}^2}}{4\lambda + \operatorname{Tr}(\Gamma_{k+1:\infty})}\sqrt{N}\left\|f^* - f^{**}\right\|_{L_{2+\varepsilon}}.
\end{aligned} \tag{2.50}$$

[Bac24, Section 7.5.2] uses $\inf\{\|f^* - f\|_{L_2}^2 + \lambda\|f\|_{\mathcal{H}}^2 : f \in \mathcal{H}\}$ to characterize the approximation error of $\mathcal{H}$, that is, the trade-off between the approximation error and $\|f\|_{\mathcal{H}}$. Our Proposition 25 shows that the approximation error is actually traded off against $\left\|\tilde{\Gamma}_{1,\text{thre}}^{-1/2} f_{1:k}^{**}\right\|_{\mathcal{H}} \frac{2\lambda+3\operatorname{Tr}(\Gamma_{k+1:\infty})}{N}$ and $\left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^{**}\right\|_{\mathcal{H}}$ instead of $\|f^{**}\|_{\mathcal{H}}$.

We observe that when the following holds, we still have benign overfitting, even though there is model-misspecification, when $\sigma_\xi \sim 1$: for $k \lesssim N$ such that $N\|\Gamma_{k+1:\infty}\|_{\text{op}} \lesssim \operatorname{Tr}(\Gamma_{k+1:\infty})$,

$$\begin{aligned}
&|J_1| = o(N),\ \sum_{j\in J_2}\sigma_j = o(\operatorname{Tr}(\Gamma_{k+1:\infty})),\ \left\|\Gamma_{k+1:\infty}^{1/2} f_{k+1:\infty}^{**}\right\|_{\mathcal{H}} = o(1), \\
&\left\|\tilde{\Gamma}_{1,\text{thre}}^{-1/2} f_{1:k}^{**}\right\|_{\mathcal{H}} \frac{\operatorname{Tr}(\Gamma_{k+1:\infty})}{N} = o(1), \\
&\frac{\sqrt{N\operatorname{Tr}(\Gamma_{k+1:\infty}^2)}}{\operatorname{Tr}(\Gamma_{k+1:\infty})} = o(1),\ \text{and} \\
&\sqrt{N}\left\|f^* - f^{**}\right\|_{L_{2+\varepsilon}} = o\left(\frac{\operatorname{Tr}(\Gamma_{k+1:\infty})}{\sqrt{\operatorname{Tr}(\Gamma_{k+1:\infty}^2)}} \wedge \frac{\operatorname{Tr}(\Gamma_{k+1:\infty})}{\sqrt{N}\left\|\Gamma_{k+1:\infty}\right\|_{\text{op}}} \wedge \sqrt{N}\right).
\end{aligned} \tag{2.51}$$

**Remark 6.** *We emphasize that this conclusion does not contradict the counterexamples presented in [CLvdG22] and [Sha22] in the setting of adversarial noise and model-misspecification. This is because in [CLvdG22], $\|f^* - f^{**}\|_{L_2} = \Theta(\sigma_\xi)$ (see [CLvdG22, Appendix D]); in [Sha22], $f^* : \boldsymbol{x} \in \mathbb{R}^d \mapsto \exp(x_1)$, where $x_1$ is the first coordinate of $\boldsymbol{x}$, and $f^{**} = \langle \boldsymbol{x}, \boldsymbol{\beta}^* \rangle$ for some $\boldsymbol{\beta}^* \in \mathbb{R}^d$. Under the probability measure $\mu$ assumed in [Sha22, Example 1], $\|f^* - f^{**}\|_{L_2}$ is $\inf_{\mathbf{w} \in \mathbb{R}^d} R_d(\mathbf{w})$ in the language of [Sha22, Example 1], and it is grater than a constant depending only on $\mu$. Therefore, when $\|f^* - f^{**}\|_{L_2} = \Theta(1)$, our (2.51) does not necessarily yield benign overfitting, thus not conflicting with [CLvdG22, Sha22].*

# Chapter 3

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

This paper is the third one in the series on the Feature Space Decomposition following [P4], [P2] and the up-coming one [P1]. The position of this chapter within the FSD series is as follows: by studying spectral methods and the saturation effect, it illustrates how the FSD method improves the analysis of the population excess risk for these classical estimators as it did previously for minimum norm interpolant estimators as well as for ridge regression.

## 3.1 Introduction

We are concerned with a supervised regression problem where we observe a vector of output $\boldsymbol{y} \in \mathbb{R}^N$ and a design matrix $\mathbb{X} \in \mathbb{R}^{N\times p}$ such that

$$\boldsymbol{y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$$

where $\mathbb{X} = [X_1|\cdots|X_N]^\top \in \mathbb{R}^{N\times p}$, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ and $\boldsymbol{\xi} = (\xi_i)_{i=1}^N$. We assume that $X_1, \ldots, X_N$ are $N$ i.i.d. vectors in $\mathbb{R}^p$ with probability distribution denoted by $\mu$ and $\xi_1, \ldots, \xi_N$ are $N$ i.i.d. centered Gaussian random variable with variance $\sigma_\xi^2$ independent of the $X_i$'s. Let $\Sigma = \mathbb{E}[X \otimes X] : \boldsymbol{v} \in \mathbb{R}^p \mapsto \mathbb{E}[\langle \boldsymbol{v}, X\rangle X] \in \mathbb{R}^p$ and $\Sigma = \sum_{j=1}^p \sigma_j \boldsymbol{e}_j \otimes \boldsymbol{e}_j$ be the spectral decomposition of $\Sigma$ such that $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$. Given a linear regression problem characterized by a triple $(\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, our goal is to obtain sharp convergence rates for the estimation error $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2$ of estimators $\hat{\boldsymbol{\beta}}$ in a large class of spectral methods.

### 3.1.1 Spectral Methods

We now introduce the family of estimators of interest in this chapter, namely, the spectral methods. We denote $\hat{\Sigma} = \frac{1}{N}\mathbb{X}^\top\mathbb{X} = \frac{1}{N}\sum_{i=1}^N X_i \otimes X_i$ the empirical version of $\Sigma$.

**Definition 17** (Spectral method)**.** *Let $(\varphi_t)_{t\geq 1}$ be a family of real-valued functions defined on $\mathbb{R}^+$ called the filter functions. For all $t \geq 1$, we define the spectral method associated with $\varphi_t$ by:*

$$\hat{\boldsymbol{\beta}} : \boldsymbol{y} \in \mathbb{R}^N \mapsto \hat{\boldsymbol{\beta}}(\boldsymbol{y}) = \frac{1}{N}\varphi_t(\hat{\Sigma})\mathbb{X}^\top\boldsymbol{y} = \frac{1}{N}\mathbb{X}^\top\varphi_t(\frac{1}{N}\mathbb{X}\mathbb{X}^\top)\boldsymbol{y} \tag{3.1}$$

*where $\varphi_t(\hat{\Sigma})$ and $\varphi_t(\frac{1}{N}\mathbb{X}\mathbb{X}^\top)$ are defined via the spectral calculus. When there is no ambiguity, we abbreviate $\hat{\boldsymbol{\beta}}(\boldsymbol{y})$ as $\hat{\boldsymbol{\beta}}$.*

A spectral method is uniquely characterized by its filter function. There is also a compagnion function to a given filter function that plays an important role regarding the statistical properties of the associated spectral method: it is called the *residual function* defined for all $t \geq 1$ as $\psi_t : x \in \mathbb{R}^+ \to 1 - x\varphi_t(x)$. Spectral methods encapsulte several important estimators and algorithms. We are now listing several of them.

**Example 16.**

- **Gradient flow** *with respect to the square loss and linear parameterization initialized at* $\mathbf{0}$*: that is, the solution of the ODE* $\dot{\boldsymbol{\beta}}_t = -(\nabla \frac{1}{2N}\|\boldsymbol{y} - \mathbb{X}\cdot\|_2^2)(\boldsymbol{\beta}_t)$ *for any* $t \geq 1$*, starting from* $\boldsymbol{\beta}_1 = \mathbf{0}$*. Then* $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_t$ *is the spectral method associated with the filter and residual functions*

$$\varphi_t : x \in \mathbb{R}^+ \mapsto \begin{cases} \frac{1-\exp(-tx)}{x} & \text{if } x > 0 \\ t & \text{if } x = 0 \end{cases} \quad \text{and } \psi_t : x \in \mathbb{R}^+ \mapsto \exp(-tx). \tag{3.2}$$

- **Ridge regression** *with regularization 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 the spectral method for the choice of filter and associated residual functions*

$$\varphi_t(x) = (t^{-1} + x)^{-1} \text{ and } \psi_t(x) = \frac{1}{xt+1}. \tag{3.3}$$

- **Gradient descent** *starting at* $\boldsymbol{\beta}_1 = \mathbf{0}$ *with step-size* $0 < \eta < 1/8$ *and at step* $t \in \mathbb{N}^*$ *for minimizing* $\boldsymbol{\beta} \mapsto \frac{1}{2N}\|\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}\|_2^2$*, i.e.* $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} - \eta\nabla(\frac{1}{2N}\|\boldsymbol{y} - \mathbb{X}\cdot\|_2^2)(\boldsymbol{\beta}_{t-1})$*, is the spectral method for the filter function* $\varphi_t(x) = (1-(1-\eta x)^t)/x$ *and its associated residual function* $\psi_t(x) = (1-\eta x)^t$*.*

- **The heavy-ball method, [Pol87, Section 3.2.1] and Nesterov's acceleration, [Nes83]** *with variable parameters are also examples of spectral algorithms (see [PR19]). Their residual functions admit recursive definitions with no known closed-form expressions.*

- **Principle Components Regression (PCR)** *estimator is* $\hat{\boldsymbol{\beta}} \in \operatorname{argmin}(\|\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}\|_2^2 : \boldsymbol{\beta} \in \hat{V}_{\leq k})$ *where* $\hat{V}_{\leq k}$ *is the subspace spanned by the first* $k$ *eigenvectors of* $\hat{\Sigma}$*. PCR equals to the spectral method with tuning parameter* $\hat{\sigma}_{k+1} \leq bt^{-1} < \hat{\sigma}_k$ *- where* $\hat{\sigma}_k$ *and* $\hat{\sigma}_{k+1}$ *are the* $k$*-th and* $k+1$*-th largest eigenvalue of* $\hat{\Sigma}$ *- for the filter function and its associated residual function given for some constant* $b > 0$ *by*

$$\varphi_t : x \in \mathbb{R}^+ \mapsto \frac{1}{x}\mathbb{1}(bt^{-1} \leq x) \text{ and } \psi_t(x) = \mathbb{1}(bt^{-1} > x).$$

*Here,* $b$ *is an absolute constant. By rescaling the tuning parameter* $t$*, it can be removed. We keep* $b$ *here in order to maintain the formal consistency with* $k^*$ *defined in (3.5).*

We are now describing the class of spectral methods considered in this work.

**Assumption 5.** *The family of filter functions* $(\varphi_t)_{t\geq 1}$ *is such that for all* $t \geq 1$*,* $\varphi_t$ *has an holomorphic extension to an open subset of* $\mathbb{C}$ *containing the contour* $\mathcal{C}_t$ *defined in Section 3.5.3. Furthermore, there are two absolute constants* $0 \leq c_{12} \leq C_{27}$ *such that for all* $t \geq 1$ *and all* $x \in [0, 8]$*:*

$$\frac{c_{12}}{x + t^{-1}} \leq \varphi_t(x) \leq \frac{C_{27}}{x + t^{-1}}. \tag{3.4}$$

Filter functions of gradient flow, ridge regression and gradient descent all satisfy Assumption 5. Indeed, for gradient flow, (3.4) holds for all $x \geq 0$ if one take $c_1 = 1$ and $C_1 = 2$ and the same does for ridge regression with $c_1 = C_1 = 1$. For gradient descent, (3.4) holds only for $x \in [0, 8]$ and for $c_1 = \eta/2$ and $C_1 = 2$. In Assumption 5, we only ask (3.4) to be true for $x \in [0, 8]$ because later we will apply this inequality only on an event where both spectra of $\Sigma$ and $\hat{\Sigma}$ are in $[0, 8]$.

We assume the existence of an holomorphic extension for technical reason related to the residual theorem, it however discards the PCR estimator for which we develop a special analysis. Regarding the assumption on the shape of the residual functions in (3.4): we ask for the filter function to be equivalent to the one of the ridge estimator with regularization parameter $t^{-1}$ in (3.4). However, the family of spectral methods satisfying this assumption is pretty wide. We also note that (3.4) is weaker than the classical assumptions used in the field of spectral methods that we recall below in Remark 7.

**Remark 7** (Classical assumptions)**.** *In several works (see for instance, [BMM19]), the filter function is assumed to satisfy the following: there exist absolute constants* $\tau \in \mathbb{N}_+ \cup \{\infty\}$*,* $C_{28} = C_{28}(\tau) \geq 1$ *such that*

1. *for any* $0 \leq \alpha \leq 1$ *and any* $t \geq 1$*,* $\sup(x^\alpha\varphi_t(x) : 0 \leq x \leq 1) \leq C_{27}t^{1-\alpha}$*;*
2. *for any* $t \geq 1$*,* $\sup(|\psi_t(x)|(x + t^{-1})^\tau : 0 \leq x \leq 1) \leq C_{28}t^{-\tau}$*;*
3. *for any* $0 \leq x \leq 1$ *and* $1 < t < \infty$*,* $c_{12} \leq (x + t^{-1})\varphi_t(x)$*.*

*It is straightforward to see that item 1. for $\alpha = 0$ and $\alpha = 1$ together with item 3. implies (3.4).*

The study of spectral methods, as far as we know, originated with Tikhonov regularization [EHN00] (ridge regression) and Landweber regularization (gradient descent) for (ill-posed) statistical inverse problems. The classical analysis of the statistical properties of spectral methods is generally based on regression problems in Sobolev spaces i.e. under regularity assumptions. Specifically, one assumes that $\Sigma$ exhibits power decay, i.e., there exists $\alpha > 1$ such that $\sigma_j \sim j^{-\alpha}$ for all $j$, and that there exists $s \geq 1$ such that $\|\Sigma^{\frac{1-s}{2}}\boldsymbol{\beta}^*\|_2$ is bounded, known as the Hölder source condition. Under this framework, the properties of spectral methods are well understood; to name a few, [SZ07, YRC07, BPR07, LGRO+08, BM16, PVRB18, PR19, BMM19, CW21, ZLL23, LGSL24, VPY24].

However, beyond this setting, the statistical properties of spectral methods are not yet fully understood-even though such algorithms have existed for almost three decades [EHN00]. We emphasize that in modern mathematical statistics, particularly in problems motivated by machine learning, a linear regression setup often does not satisfy the above Hölder source condition. In fact, in such problems, $\Sigma$ and $\boldsymbol{\beta}^*$ may follow arbitrary patterns. Thus, it is genuinely necessary to understand the statistical properties of spectral methods for arbitrary linear regression problems.

Our first objective is, for a given linear regression problem $(\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, to obtain matching high-probability upper and lower bounds for $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2$ where $\hat{\boldsymbol{\beta}}$ is a spectral method whose filter function satisfy Assumption 5. Our second objective is to show how the the Feature Space Decomposition method can be used on spectral methods to achieve this goal.

### 3.1.2 Notation

We use $a \lesssim b$ (respectively $a \gtrsim b$) to represent the fact that there exists an absolute constant $C$ such that $a \leq Cb$ ($a > Cb$). We use $a \sim b$ if $a \lesssim b$ and $b \lesssim a$. We say $a \lesssim_K b$ if $C = C(K)$. For a probability measure $\mu$, we write $\mu^{\otimes N}$ as its $N$-fold tensor product. We denote the $\ell_2 \to \ell_2$ operator norm of a matrix by $\|\cdot\|_{\text{op}}$ and by $\|\cdot\|_{HS}$ its Hilbert-Schmidt norm.

## 3.2 Main Results

In this section, we present the main results of this chapter. We first gather all the model assumptions.

**Assumption 6.** *We assume that $\|\Sigma\|_{op} \leq 1$. The noise $\xi$ satisfies $\xi \sim \mathcal{N}(0, \sigma_\xi^2)$ and it is independent with $X$. Assume $X$ is sub-Gaussian: there exists an absolute constant $C > 0$ such that for any $\boldsymbol{v} \in \mathbb{R}^p$ and $q \geq 2$, $\|\langle X, \boldsymbol{v}\rangle\|_{L^q(\mu)} \leq C\sqrt{q}\|\langle X, \boldsymbol{v}\rangle\|_{L^2(\mu)}$.*

Next, we introduce the optimal dimension used to split the feature space in the case of spectral methods.

**Definition 18.** *Let $b > 0$ and $t \geq 1$. The **estimation dimension** of the spectral method $\hat{\boldsymbol{\beta}}$ with filter function $\varphi_t$ is defined as*

$$k^* = k^*_{t^{-1},b} = \min\Big\{k \in [p] : \sigma_{k+1} \leq bt^{-1}\Big\}. \tag{3.5}$$

Here the absolute constant $b$ does not carry any particular significance; it arises artificially in the course of proving the lower bound. At present, we do not know how to obtain the result with $b = 1$, as in the ridge regression case as in [CM22, Proposition 2.2].

The estimation dimension $k^*$ is the dimension of the space $V_{J_*}$ where estimation of the spectral method $\hat{\boldsymbol{\beta}}$ with filter function $\varphi_t$ happens. It coincides with the optimal one for ridge regression defined in [P2] when $\text{Tr}[\Sigma_{J_*^c}] \leq Nt^{-1}$. In particular, we see that this dimension does not depend on the shape of the filter function but just on its parameter $t$. However, the optimal convergence rate of a spectral method depends on its filter function via its residual function since we will show that it is given by

$$r(V_{J_*}, V_{J_*^c}) = \left\|\Sigma^{1/2}_{J_*}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\right\|_2 + \sigma_\xi\sqrt{\frac{|J_*|}{N}} + \left\|\Sigma^{1/2}_{J_*^c}\boldsymbol{\beta}^*_{J_*^c}\right\|_2 + \sigma_\xi t\sqrt{\frac{\text{Tr}(\Sigma^2_{J_*^c})}{N}}, \tag{3.6}$$

where $V_{J_*} = \mathrm{span}(\boldsymbol{e}_j : j \in J_*)$, $J_* = [k^*]$, $(\boldsymbol{e}_j)_j$ are the eigenvectors of $\Sigma$ and $\psi_t$ is the residual function defined in Definition 17. For concrete examples of $r(V_{J_*}, V_{J_*^c})$, the reader is referred to Corollary 4 and Corollary 5 in page 34.

We are now in a position to state our main results: two upper and lower bounds for the excess risk of spectral methods and a corollary identifying the conditions where the two bounds match, giving the optimal rate from (3.6). The proof of the following results may be found in Section 3.3 for the upper bound and in Section 3.4 for the lower bound.

**Theorem 7** (Main Result - upper bound)**.** *We consider a linear regression model with parameter $(\boldsymbol{\beta}^*, \Sigma, \sigma_\xi)$ satisfying Assumption 6. Let $(\varphi_t)_{t\geq 1}$ be a family of filter functions satisfying Assumption 5 for $c_1 = 0$. Let $t \geq 1$. Then, there exists an absolute constant $c > 0$ such that for all $0 < \square < 1/9$, if $\square^2 N \gtrsim \mathrm{Tr}\left(\Sigma(\Sigma + t^{-1}I_p)^{-1}\right) \vee 1$ and $\square \lesssim \log^{-1}(et)$ then with probability at least $1 - 2\exp(-c|J_*|) - \exp(-c\square^2 N)$,*

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \lesssim r(V_{J_*}, V_{J_*^c}) + \frac{\square}{t}\left\|\Sigma_{J_*}^{-\frac{1}{2}}\boldsymbol{\beta}^*_{J_*}\right\|_2.$$

**Theorem 8** (Main result - lower bound)**.** *There are absolute positive constants $c_0, c, c_2$ and $c_3$ such that the following holds. Let $(\boldsymbol{\beta}^*, \Sigma, \sigma_\xi)$ be the parameters of a linear regression model under Assumption 6 where $X$ is assumed to have independent and centered coordinates with respect to $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_p\}$. Let $\hat{\boldsymbol{\beta}}$ be a spectral method with filter function satisfying Assumption 5 for $0 < c_1 \leq C_1$. Let $0 < \square < 1/9$ be such that $\square \lesssim \log^{-1}(et)$ and $\square^2 N \gtrsim \mathrm{Tr}\left(\Sigma(\Sigma + t^{-1}I_p)^{-1}\right) \vee 1$. Let $k^*$ be the estimation dimension introduced in Definition 18 for some $0 < b \leq c_0$ and $J_* = [k^*]$. Then, with probability at least $1 - c\exp(-k^*/c) - \exp(-\square^2 N/c)$,*

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \geq c_2 r(V_{J_*}, V_{J_*^c}) - \frac{c_3\square}{t}\left\|\Sigma_{J_*}^{-\frac{1}{2}}\boldsymbol{\beta}^*_{J_*}\right\|_2. \tag{3.7}$$

The next result is a high probability upper and lower bound for spectral methods showing that $r(V_{J_*}, V_{J_*^c})$ is the right quantity describing the statistical properties of these estimators for a given linear regression model. It follows from Theorem 7 and Theorem 8.

**Corollary 6.** *There are absolute positive constants $c_0, c, (c_k)_{k=2,3,4,5}$ such that the following holds. Under the same assumptions as in Theorem 8. Let $t \geq 1$ and $0 < \square < 1/9$ be such that $\square \leq c_0 \log^{-1}(et)$, $\square^2 N \geq c(\mathrm{Tr}\left(\Sigma(\Sigma + t^{-1}I_p)^{-1}\right) \vee 1)$, $k^* \geq c$ and*

$$\frac{\square}{t}\left\|\Sigma_{J_*}^{-\frac{1}{2}}\boldsymbol{\beta}^*_{J_*}\right\|_2 \leq c_2 r(V_{J_*}, V_{J_*^c}). \tag{3.8}$$

*Then, with probability at least $1 - c_3\exp(-k^*/c_3) - \exp(-\square^2 N/c_3)$,*

$$c_4 r(V_{J_*}, V_{J_*^c}) \leq \left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \leq c_5 r(V_{J_*}, V_{J_*^c}).$$

Condition (3.8) holds when $(\square/t)\left\|\Sigma_{J_*}^{-\frac{1}{2}}\boldsymbol{\beta}^*_{J_*}\right\|_2$ is smaller than one of the four terms in $r(V_{J_*}, V_{J_*^c})$; for instance, it holds when

1. $\frac{1}{t\sigma_\xi}\left\|\Sigma_{J_*}^{-\frac{1}{2}}\boldsymbol{\beta}^*_{J_*}\right\|_2 \lesssim \frac{1}{\square}\sqrt{\frac{|J_*|}{N}}$, where we recall that $t^{-1}\|\Sigma_{J_*}^{-\frac{1}{2}}\boldsymbol{\beta}^*_{J_*}\|_2$ is the bias of $\hat{\boldsymbol{\beta}}_J^{(\mathrm{Ridge})}$ when $\hat{\boldsymbol{\beta}}_J^{(\mathrm{Ridge})}$ is the ridge regression with tuning parameter $t^{-1}$, and $\frac{1}{\square}$ may be taken to be $\sqrt{\frac{N}{k^* + t\,\mathrm{Tr}(\Sigma_{J_*^c})}}$;
2. or when $\frac{\square}{t}\left\|\Sigma_{J_*}^{-\frac{1}{2}}\boldsymbol{\beta}^*_{J_*}\right\|_2 \lesssim \left\|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\right\|_2$, which is the case when $\square/t$ is small enough so that $\psi_t(x) \geq (\square/t)x$ for all $x \in [0,1]$ (recall that we assumed that $\|\Sigma\|_{\mathrm{op}} \leq 1$ in Assumption 6) which is equivalent to assume that $\varphi_t(x) \leq (t - \square x)/(xt)$.

As mentioned earlier the case of PCR is special since it requires a property on the $k^*$-th spectral gap of $\Sigma$. We therefore state a result devoted to PCR. The proof of the following result is different from the one of Theorem 7 and may be found in Section 3.6.

**Theorem 9** (Upper bound for PCR)**.** *We consider a linear regression model with parameter $(\boldsymbol{\beta}^*, \Sigma, \sigma_\xi)$ satisfying Assumption 6. Let $t \geq 1$ and $0 < b < 1$. Denote by $\hat{\boldsymbol{\beta}}$ the PCR estimator with filter function $\varphi_t : x > 0 \mapsto x^{-1}\mathbb{1}(x \geq bt^{-1})$. Let $0 < \square < 1/9$ and assume that $\square^2 N \gtrsim \mathrm{Tr}\left(\Sigma(\Sigma + t^{-1}I_p)^{-1}\right) \vee 1$ and that $\theta > 0$ where*

$$\theta := \min\left(bt^{-1} - \left(\sigma_{k^*+1} + \square(\sigma_{k^*+1} + t^{-1})\right), \left(\sigma_{k^*} - \square(\sigma_{k^*} + t^{-1})\right) - bt^{-1}\right). \tag{3.9}$$

*Then, there exists an absolute constant $c > 0$ such that with probability at least $1 - 2\exp(-c|J_*|) - \exp(-c\square^2 N)$,*

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \lesssim r(V_{J_*}, V_{J_*^c}) + \frac{\square}{\theta^2}\left\|\Sigma_{J_*}^{-\frac{1}{2}}\boldsymbol{\beta}^*_{J_*}\right\|_2.$$

In the case of PCR, the convergence rate $r(V_{J_*}, V_{J_*^c})$ contains only the three terms

$$\left\|\Sigma_{J_*^c}^{1/2}\boldsymbol{\beta}^*_{J_*^c}\right\|_2 + \sigma_\xi\sqrt{\frac{|J_*|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma^2_{J_*^c})}{N}}$$

since $\left\|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\right\|_2 = 0$ because $\psi_t(\Sigma) = P_{J_*^c}$. Note also that compare with Theorem 7 we don't need to choose $\square$ less than $\log^{-1}(et)$ and so one can choose $\square$ to be of the order of a constant. The choice $\square \sim \sqrt{k^*/N}$ is also legitimate as long as the sample complexity assumption $\square^2 N \gtrsim \mathrm{Tr}\left(\Sigma(\Sigma + t^{-1}I_p)^{-1}\right) \vee 1$ is satisfied that is when $k^* \gtrsim \mathrm{Tr}\left(\Sigma(\Sigma + t^{-1}I_p)^{-1}\right) \vee 1$ which holds (see the discussion below (3.15)) when $k^* \gtrsim t\,\mathrm{Tr}[\Sigma_{J_*^c}]$. This is for instance, the case when $\sigma(\Sigma)$ has a fast decay. However, Theorem 9 requires $\theta > 0$ that holds iff the $k^*$-th spectral gap of $\Sigma$ is large enough:

$$\sigma_{k^*} - \sigma_{k^+1} > \square\left(\sigma_{k^*} + \sigma_{k^*+1} + 2t^{-1}\right)$$

and when $bt^{-1} \in \left[\sigma_{k^*+1} + \square(\sigma_{k^*+1} + t^{-1})\right), \sigma_{k^*} - \square(\sigma_{k^*} + t^{-1})\right]$.

Let us now comment on the consequences of the results above.

### 3.2.1 Contribution to the understanding of the statistical properties of spectral methods

For an arbitrary linear regression problem $(\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, Corollary 6 provides, under fairly general conditions, high probability matching upper and lower bounds (up to a multiplicative constant) for the population excess risk of spectral methods in this specific regression problem.

1. Compared with classical results in the statistical properties of spectral methods, such as [SZ07, YRC07, BPR07, LGRO+08, BM16, BM18, BMM19, CW21, ZLL23, LGSL24, VPY24], we observe that the classical results are typically restricted to Sobolev spaces (which impose a power decay on the eigenvalues of $\Sigma$), or require certain eigenvalue decay conditions. Among them, [BM16] does not rely on power decay, but still requires the eigenvalues to satisfy certain specific decay conditions. In contrast, Theorem 7 imposes no restrictions on the spectrum of $\Sigma$.

2. In addition, the aforementioned classical literature typically assumes that $\boldsymbol{\beta}^*$ satisfies a certain Hölder-type source condition, namely, that there exists $s > 1$ such that $\|\Sigma^{\frac{1-s}{2}}\boldsymbol{\beta}^*\|_2$ is bounded. In contrast, our Theorem 7 requires no assumptions whatsoever on $\boldsymbol{\beta}^*$, yet still yields a precise characterization of its statistical properties.

3. [AKT19] also examines the population excess risk of gradient descent and gradient flow for general linear regression problems when $\boldsymbol{\beta}^*$ is a random vector. However, they upper bound the residual functions of gradient flow and gradient descent by that of ridge regression, thereby failing to obtain the precise characterization presented in this work, and consequently, they do not derive conclusions such as the saturation effect and the generalized saturation effect that are introduced in the next section. After the completion of the main body of this chapter, we became aware of the concurrent works [HW23, WBL+25], which extend the analysis of [AKT19] but do not assume $\boldsymbol{\beta}^*$ is a random vector. Nevertheless, they also upper bound the residual functions of gradient flow and gradient descent by that of ridge regression, hence they share the same limitations as [AKT19]. Furthermore, the lower bound in Theorem 4.1 of [WBL+25] holds only in the overfitting regime, whereas our primary focus is on appropriate regularization. From this perspective, our findings and those of [WBL+25] are complementary. Furthermore, we provide a sharp bound applicable to arbitrary linear regression problems under appropriate regularization, whereas the corresponding result in [WBL+25] for the properly regularized regime is far from sharp.

Precisely because Theorem 7 yields a precise (up to a multiplicative constant) characterization of the population excess risk for any linear regression problem, it allows us to describe the statistical properties of spectral methods in the most general linear regression setting. To the best of our knowledge, this is the first result that establishes an universal statistical property of spectral methods valid for any linear regression problem.

From Section 1.5.1, we know that estimation of $\boldsymbol{\beta}^*$ occurs only on $V_{J_*}$, while absorption of noise occurs on $V_{J_*^c}$. Theorem 7 shows that, for any given linear regression problem $(\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$ and tuning parameter $t$, the space $V_{J_*}$ where estimation takes place is determined solely by the spectrum of $\Sigma$ and the tuning parameter, and is independent of the signal $\boldsymbol{\beta}^*$ to be approximated, the eigenvectors of $\Sigma$, and the family of filter functions $(\varphi_t)_{t\geq 1}$. This observation indicates the following facts:

1. Since $V_{J_*}$ is independent of $(\varphi_t)_{t\geq 1}$, we know that for a given linear regression problem, all algorithms in the class of spectral methods decompose the feature space in the same way to estimate the signal. By examining the definition of $r(V_{J_*}, V_{J_*^c})$ in (3.6), we find that only the term $\|\Sigma_{J_*}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*_{J_*}\|_2$ depends on the specific choice of the filter / residual functions. In other words- the only difference in the statistical properties of different spectral methods for a given linear regression problem lies in how close the residual function $\psi_t$ is to 0 on $\{x > 0 : tx > b\}$-the closer it is to 0, the better the statistical properties (i.e., the faster the convergence rate). For example, when the eigenvalues of $\Sigma$ satisfy power decay, i.e., there exists $\alpha > 1$ such that $\sigma_j \sim j^{-\alpha}$ for all $j$ (corresponding to regression problems in Sobolev spaces with sufficient smoothness), the residual function of ridge regression is $\psi_t(x) = \frac{1}{xt+1}$, that of gradient flow is $\psi_t(x) = \exp(-tx)$, and that of gradient descent is $\psi_t(x) = (1-\eta x)^t$, see Example 16. For the latter two, when $tx > b$, their convergence to 0 as functions of $x$ is much faster than that of ridge regression. This provides an explanation of the saturation effect [BPR07]: on the set $\{x > 0 : tx > b\}$, the residual function of ridge regression decays too slowly. We provide more general situations in [P3].

2. The FSD approach yields some understanding on the behavior of spectral methods. For instance, the estimation dimension $k^*$ tells us that gradient descent at step $t$ is estimating only the first $k^*$ coordinates of the signal in the basis of eigenvectors of $\Sigma$, no more no less. This means that along the path of gradient descent, there are more and more coordinates (in the eigenbasis of $\Sigma$) that are estimated; the estimation dimension quantifying this phenomenon and the estimation space $V_{J^*}$ localizing the space where this estimation is happening. We may suspect that Newton's method behaves similarly but with an estimation dimension growing faster than the one of gradient descent.

3. Corollary 6 may also help us to identify the best possible choice for parameter $t$: for gradient descent, that is the best possible stopping time and for ridge that is the best regularization parameter. Indeed, the optimal choice of $t$ depends on the specific regression problem and is given by the one minimizing the optimal rate $r(V_{J_*}, V_{J_*^c})$ subject to $\square^2 N \gtrsim \mathrm{Tr}(\Sigma(\Sigma + t^{-1}I_p)^{-1})$, keeping in mind that $J_* = [k^*]$ and that $k^*$ depends on $t$. The identification of such an optimal choice for $t$ may help to show that a given data-driven choice for $t$ (see for instance the Lepski's method from [BMM19]) is optimal if the resulting spectral method achieves the optimal rate $r(V_{J_*}, V_{J_*^c})$ for the optimal choice of $t$.

4. Since $V_{J_*}$ is independent of $\boldsymbol{\beta}^*$, it follows from Definition 16 that any spectral algorithm satisfying the conditions of Theorem 7 (such as gradient flow/descent) does not possess the feature learning property. We emphasize that the gradient flow/descent studied in this chapter refers to ODEs for quadratically minimized problems with linear parameterization on linear spaces, which differ from the gradient flow/descent in neural network theory, where Riemannian manifolds [NWS22], [P5] or nonlinear parameterizations [PVRF22] are often used.

5. The lack of feature learning capability has the following drawback: if the alignment between $\boldsymbol{\beta}^*$ and $\Sigma$ is poor, spectral methods exhibit unfavorable statistical properties. For example, when the support of $\boldsymbol{\beta}^*$ satisfies $\mathrm{supp}(\boldsymbol{\beta}^*) = V_{J_*^c}$, the term $\|\Sigma_{J_*^c}^{1/2}\boldsymbol{\beta}^*_{J_*^c}\|_2$ in $r(V_{J_*}, V_{J_*^c})$ reduces to $\|\langle X, \boldsymbol{\beta}^*\rangle\|_{L^2(\mu)}$, which may be big. Of course, one can change $V_{J_*}$ by adjusting the tuning parameter $t$, but we stress that statisticians usually do not know the support of $\boldsymbol{\beta}^*$ in the basis of eigenvectors of $\Sigma$ in advance, and hence cannot preselect an appropriate $t$. Therefore, unlike statistical algorithms with the sparsity inducing property such as basis pursuit or the LASSO, the fact that spectral methods lack the feature learning property implies that, when the signal and the eigenvectors of $\Sigma$ are poorly aligned, spectral methods generally have inferior statistical performance. We discuss further in [P3] on the lack of feature learning of spectral methods.

From Example 16, we know that the residual function of gradient flow is smaller than the one of ridge regression. Therefore, for a given linear regression problem and for the same tuning parameter, we always have $r^{(\mathrm{GF})}(V_{J_*}, V_{J_*^c}) \leq r^{(\mathrm{Ridge})}(V_{J_*}, V_{J_*^c})$. This means that, from the perspective of population excess risk, whenever one can choose between ridge regression and gradient flow, gradient flow should always be preferred, regardless of the specific linear regression problem under consideration. In [P3], we further discuss the notion of partial order on the set of spectral algorithms.

### 3.2.2 Contribution within the FSD series of papers

The high-level idea of the proof of Theorem 7 is to wrap the classical analysis of the statistical properties of spectral methods, such as [LGSL24], with a FSD layer-namely, instead of analyzing the statistical properties over the entire feature space $\mathbb{R}^p$, we restrict the analysis to $V_J$, while on $V_{J^c}$ we perform a signal-free analysis (considering $\langle X, \boldsymbol{\beta}^* \rangle$ as part of the noise). Remarkably, we obtain the precise result of Theorem 7. We therefore believe that the proof of Theorem 7 itself suggests that the FSD method may serve as a systematic tool in mathematical statistics for deriving precise non-asymptotic results on the population excess risk of general supervised learning algorithms.

Theorem 7 can be regarded as an extension of the results of [MMM22, TB23, CM22, BS24, P2] on ridge regression to spectral methods. In this theorem, we apply the FSD method for the first time to estimators beyond ridge regression and the minimum norm interpolant estimator. Unlike the ridge results in [MMM22, TB23, CM22, BS24, P2], in (3.5) we do not observe an "effective regularization" term of the form $Nt^{-1} + \mathrm{Tr}(\Sigma_{J^c})$. This is because we only consider the well-regularized regime, namely, when the spectral algorithm is far from overfitting. The overfitting regime of spectral methods-for example, when the running time $t$ of gradient descent/flow tends to infinity-yields the minimum $\ell_2$ norm interpolant estimator, which has already been studied in [TB23] and [P4].

## 3.3 Proof of the upper bound in Theorem 7

We abbreviate $J_*$ by $J$ in this section, i.e. $J = [k^*]$ where $k^*$ is the estimation dimension from Definition 18. Following the FSD method, we recall the risk decomposition of $\hat{\boldsymbol{\beta}}$ given by

$$\left\|\Sigma^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right\|_2 \le \left\|\Sigma_J^{1/2}\left(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*\right)\right\|_2 + \left\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\right\|_2 + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 \tag{3.10}$$

where $\hat{\boldsymbol{\beta}}_J = P_J\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{J^c} = P_{J^c}\hat{\boldsymbol{\beta}}$. The next two sections are devoted to show high probability upper bounds on the estimation part $\left\|\Sigma_J^{1/2}\left(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*\right)\right\|_2$ and the noise absorption part $\left\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\right\|_2$ appearing in (3.10).

In multiple occasions, we will use the following relations that follows for instance from SVD: we recall that $P_J : \mathbb{R}^p \to \mathbb{R}^p$ is the projection operator onto $V_J$ and $\mathbb{X}_J^\top := [P_J X_1 | \cdots | P_J X_N]$. We have $\mathbb{X}_J = \mathbb{X}P_J$, $\mathbb{X}_J^\top = P_J\mathbb{X}^\top$ and $\hat{\Sigma}_J := \frac{1}{N}\mathbb{X}_J^\top\mathbb{X}_J = P_J\hat{\Sigma}P_J$. Since, $V_J$ is an eigen-space of $\Sigma$, we also have $P_J\varphi_t(\Sigma)\Sigma = \varphi_t(\Sigma_J)\Sigma_J$ where $\Sigma_J := \mathbb{E}(P_JX)(P_JX)^\top = P_J\Sigma P_J$. We also define $\Sigma_t = \Sigma + t^{-1}I_p$ and $\hat{\Sigma}_t = \hat{\Sigma} + t^{-1}I_p$.

It also follows from the definition of $k^*$ that $b^{-1}\sigma_{k^*+1} \le t^{-1} \le b^{-1}\sigma_{k^*}$. Consequently,

$$\left\|\Sigma_J^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\right\|_{\mathrm{op}} \le \left\|\Sigma^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\right\|_{\mathrm{op}} \le 1, \left\|\Sigma_{J^c}^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\right\|_{\mathrm{op}} \le \sqrt{\frac{b}{1+b}} \text{ and } \left\|\Sigma_J^{-\frac{1}{2}}\Sigma_t^{\frac{1}{2}}\right\|_{\mathrm{op}} \le \sqrt{\frac{1+b}{b}}. \tag{3.11}$$

We also have from the definition of $k^*$ that for all $x \in V_J$,

$$\left\|\Sigma_t^{1/2}x\right\|_2^2 = \left\|\Sigma_J^{1/2}x\right\|_2^2 + t^{-1}\|x\|_2^2 \le \frac{1+b}{b}\left\|\Sigma_J^{1/2}x\right\|_2^2 \tag{3.12}$$

because $bt^{-1}\|x\|_2^2 \le \sigma_{k^*}\|x\|_2^2 \le \left\|\Sigma_J^{1/2}x\right\|_2^2$.

### 3.3.1 The main property of $\hat{\Sigma}$ required for the analysis and the event $\Omega_t$

The main uniform property we need $\hat{\Sigma}$ to satisfy for the analysis is the one from the following event: let $0 < \square < 1/9$ (a typical choice of $\square$ could be of the order of $\log^{-1}(et)$ or $\sqrt{\mathrm{Tr}(\Sigma(\Sigma + t^{-1})^{-1})/N)}$, we consider the event

$$\Omega_t := \left\{\left\|\Sigma_t^{-1/2}(\hat{\Sigma} - \Sigma)\Sigma_t^{-1/2}\right\|_{\mathrm{op}} \le \square\right\}. \tag{3.13}$$

We show in the next result that $\Omega_t$ holds with large probability as long as $\square^2 N$ is larger than the effective rank $\mathrm{Tr}\left[\Sigma(\Sigma + t^{-1}I_p)^{-1}\right]$.

**Lemma 8.** *Grant Assumption 6. Let $t \ge 1$ and assume that $\square^2 N \gtrsim \mathrm{Tr}\left[\Sigma(\Sigma + t^{-1}I_p)^{-1}\right]$ and $\square^2 N \gtrsim 1$. There exists an absolute constant $c > 0$ such that $\Omega_t$ happens with probability at least $1 - \exp(-c\square^2 N)$.*

*Proof.* It follows from Theorem 5.5 in [Dir15] on the control of empirical quadratic processes and the sub-gaussian assumption from Assumption 6 that there is an absolute constant $C \geq 1$ such that for all $r \geq 1$, with probability at least $1 - \exp(-r)$,

$$\sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} f^2(X_i) - \mathbb{E} f^2(X) \right| \leq C \left( \frac{D\gamma_2}{\sqrt{N}} + \frac{\gamma_2^2}{N} + D^2 \left( \sqrt{\frac{r}{N}} + \frac{r}{N} \right) \right) \tag{3.14}$$

where $\gamma_2 = \gamma_2(F, \|\cdot\|_{L^2(\mu)})$ is Talagrand's $\gamma_2$-functional of $\mathcal{F}$ with respect the $L^2(\mu)$-norm [Tal14, Definition 2.2.19] and $D = \mathrm{diam}(F, L^2(\mu)) := \sup(\|f\|_{L^2(\mu)} : f \in F)$. Applying (3.14) to $F = \{\langle \cdot, \boldsymbol{v} \rangle : \boldsymbol{v} \in \Sigma_t^{-1/2} S_2^{p-1}\}$ where $S_2^{p-1}$ is the unit $\ell_2^p$-sphere, we have $D = \mathrm{diam}(F, L^2(\mu)) = \left\| \Sigma^{1/2} \Sigma_t^{-1/2} \right\|_{op} \leq 1$ and $\gamma_2(F, \|\cdot\|_{L^2(\mu)}) \sim \mathbb{E} \left\| \Sigma^{1/2} \Sigma_t^{-1/2} G \right\|_2 \sim \sqrt{\mathrm{Tr}(\Sigma(\Sigma + t^{-1} I_p)^{-1})}$ where $G \sim \mathcal{N}(0, I_p)$. As a consequence, it follows from the sample complexity assumption $\square^2 N \gtrsim \mathrm{Tr}\left[\Sigma(\Sigma + t^{-1} I_p)^{-1}\right]$ that for $r = \square^2 N/(16C^2)$ (which is larger than 1 since we assumed that $\square^2 N \gtrsim 1$), with probability at least $1 - \exp(-\square^2 N/(16C^2))$,

$$\begin{aligned}
\left\| \Sigma_t^{-1/2} (\hat{\Sigma} - \Sigma) \Sigma_t^{-1/2} \right\|_{\mathrm{op}} &= \sup_{\boldsymbol{u} \in S_2^{p-1}} \left| \boldsymbol{u}^\top \Sigma_t^{-1/2} (\hat{\Sigma} - \Sigma) \Sigma_t^{-1/2} \boldsymbol{u} \right| \\
&= \sup_{\boldsymbol{u} \in S_2^{p-1}} \left| \|\hat{\Sigma}^{\frac{1}{2}} \Sigma_t^{-\frac{1}{2}} \boldsymbol{u}\|_2^2 - \|\Sigma^{\frac{1}{2}} \Sigma_t^{-\frac{1}{2}} \boldsymbol{u}\|_2^2 \right| \\
&= \sup_{\boldsymbol{u} \in S_2^{p-1}} \left| \frac{1}{N} \sum_{i=1}^{N} \langle \Sigma_t^{-\frac{1}{2}} \boldsymbol{u}, X_i \rangle^2 - \mathbb{E} \langle \Sigma_t^{-\frac{1}{2}} \boldsymbol{u}, X_i \rangle^2 \right| \leq \square.
\end{aligned}$$

■

The sample complexity assumption $\square^2 N \gtrsim \mathrm{Tr}\left[\Sigma(\Sigma + t^{-1} I_p)^{-1}\right]$ is classical in the analysis of spectral methods. It has some consequences on the definition of the estimation dimension $k^*$. Indeed, one has

$$\mathrm{Tr}\left[\Sigma(\Sigma + t^{-1} I_p)^{-1}\right] = \sum_j \frac{\sigma_j}{\sigma_j + t^{-1}} = \sum_{j \in J} \frac{\sigma_j}{\sigma_j + t^{-1}} + \sum_{j \notin J} \frac{\sigma_j}{\sigma_j + t^{-1}}$$

where we recall that $J = \{j : \sigma_j \geq b t^{-1}\}$ is of cardinality $k^*$, by definition of $k^*$ and so

$$\frac{b k^*}{1+b} + \frac{t}{1+b} \mathrm{Tr}[\Sigma_{J^c}] \leq \mathrm{Tr}\left[\Sigma(\Sigma + t^{-1} I_p)^{-1}\right] \leq k^* + t \, \mathrm{Tr}[\Sigma_{J^c}]. \tag{3.15}$$

As a consequence, the sample complexity assumption implies both $\square^2 N \gtrsim b k^*$ - meaning that we require the estimation dimension to be smaller than $N$ - and $\square^2 N \gtrsim t \, \mathrm{Tr}[\Sigma_{J^c}]$ implying that the estimation dimension of ridge obtained in [P2] coincides with the one used here in Definition 18, i.e. $k^{**} = k^*$, for other spectral methods.

In the classical analysis of spectral methods, the property induced by the event $\Omega_t$ is referred as the "Change-of-Norm argument" (see, for example, [CW21]). From a geometrical perspective, the event $\Omega_t$ is the union of two type of events that are part of the FSD method. Indeed, $\Omega_t$ is equivalent to: for all $\boldsymbol{u} \in \mathbb{R}^p$,

$$\left| \left\| \hat{\Sigma}^{1/2} \boldsymbol{u} \right\|_2^2 - \left\| \Sigma^{1/2} \boldsymbol{u} \right\|_2^2 \right| \leq \square \left\| \Sigma_t^{1/2} \boldsymbol{u} \right\|_2^2. \tag{3.16}$$

As a consequence, there are two regimes depending on the relative values of $\left\| \Sigma^{1/2} \boldsymbol{u} \right\|_2$ and $\left\| \Sigma_t^{1/2} \boldsymbol{u} \right\|_2$ that can be described via the following cone

$$\begin{aligned}
C &:= \left\{ \boldsymbol{u} \in \mathbb{R}^p : \square \left\| \Sigma_t^{1/2} \boldsymbol{u} \right\|_2^2 \leq \frac{1}{2} \left\| \Sigma^{1/2} \boldsymbol{u} \right\|_2^2 \right\} \\
&= \left\{ \boldsymbol{u} \in \mathbb{R}^p : \square t^{-1} \|\boldsymbol{u}\|_2^2 \leq \left( \frac{1}{2} - \square \right) \left\| \Sigma^{1/2} \boldsymbol{u} \right\|_2^2 \right\}.
\end{aligned} \tag{3.17}$$

Then, we consider the decomposition of $\mathbb{R}^p$ as the union: $\mathbb{R}^p = C \cup C^c$. This decomposition is closed to the one of the FSD $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ since one can see that $C$ contains all singular vectors of $\Sigma$ with singular values such that

$\sigma_j \gtrsim \square t^{-1}$ which is, up to the $\square$ term, the inequality appearing in the definition of $k^*$. We see that an isomorphic property restricted to this cone follows from (3.16): for all $\boldsymbol{u} \in C$,

$$\frac{1}{\sqrt{2}} \left\| \Sigma^{1/2} \boldsymbol{u} \right\|_2 \leq \left\| \hat{\Sigma}^{1/2} \boldsymbol{u} \right\|_2 \leq \sqrt{\frac{3}{2}} \left\| \Sigma^{1/2} \boldsymbol{u} \right\|_2 .$$

This type of 'RIP' (i.e. restricted isomorphic property) is expected in the FSD method on the estimation part of the feature space i.e. $V_J$ or the slightly bigger cone $C$. On the 'noise absorption part' of the feature space, i.e. $V_{J^c}$ - or the slightly bigger cone $C^c$, when $\square$ is of the order of a constant - we don't need such an isomorphic property but only a control of the largest 'restricted' singular value of $\hat{\Sigma}$: for all $\boldsymbol{u} \notin C$,

$$\begin{aligned} \left\| \hat{\Sigma}^{1/2} \boldsymbol{u} \right\|_2 \leq \sqrt{3\square} \left\| \Sigma_t^{1/2} \boldsymbol{u} \right\|_2 &= \sqrt{3\square} \left( \left\| \Sigma^{1/2} \boldsymbol{u} \right\|_2^2 + t^{-1} \left\| \boldsymbol{u} \right\|_2^2 \right)^{1/2} \\ &\leq 3\sqrt{t^{-1}\square} \left\| \boldsymbol{u} \right\|_2 \leq \sqrt{t^{-1}} \left\| \boldsymbol{u} \right\|_2 . \end{aligned}$$

In particular, we see that, on the event $\Omega_t$, for all $\boldsymbol{u} \in \mathbb{R}^p$, we have

$$\left\| \hat{\Sigma}^{1/2} \boldsymbol{u} \right\|_2 \leq \max \left( \sqrt{3/2} \left\| \Sigma^{1/2} \boldsymbol{u} \right\|_2 , \sqrt{t^{-1}} \left\| \boldsymbol{u} \right\|_2 \right) .$$

In particular, the following Lemma holds.

**Lemma 9.** *On the event $\Omega_t$, $\hat{\sigma}_1 = \left\| \hat{\Sigma} \right\|_{op} \leq 4(\sigma_1 + t^{-1})$.*

For our proof strategy, it is important to localize the spectrum of $\hat{\Sigma}$. Indeed, the spectral method $\hat{\boldsymbol{\beta}}$ depends on the filter function via the term $\varphi_t(\hat{\Sigma})$ in its definition from (3.1)). In particular, we will need to tell how $\varphi_t(\hat{\Sigma})$ is close to $\varphi_t(\Sigma)$ . However, it is well-known that for a general non-linear function $f$ (for which the spectral calculus is well-defined), $\mathbb{E}[f(\hat{\Sigma})] \neq f(\Sigma)$; for example, when $f(x) = x^2$. This illustrates that $f(\hat{\Sigma})$, as a plug-in estimator for $f(\Sigma)$, is a biased estimator (in fact, this is one of the motivations behind [Kol18]). Methods for handling this bias have been developed in [LGSL24], they are based on the residue theorem: for any counterclock-wise contour $\mathcal{C}_t$ surrounding both spectra of $\hat{\Sigma}$ and $\Sigma$, we have

$$\begin{aligned} \varphi_t(\hat{\Sigma}) - \varphi_t(\Sigma) &= -\frac{1}{2\pi i} \oint_{\mathcal{C}_t} \varphi_t(z) \left[ (\hat{\Sigma} - zI_p)^{-1} - (\Sigma - zI_p)^{-1} \right] dz \\ &= \frac{1}{2\pi i} \oint_{\mathcal{C}_t} (\hat{\Sigma} - zI_p)^{-1} (\hat{\Sigma} - \Sigma)(\Sigma - zI_p)^{-1} \varphi_t(z) dz. \end{aligned} \tag{3.18}$$

In particular, for the choice of contour $\mathcal{C}_t$ from Section 3.5.3, we have $\mathcal{C}_t$ surrounding both spectra of $\hat{\Sigma}$ and of $\Sigma$ on the event $\Omega_t$ thanks to Lemma 9. So that the residue theorem applies to both $\varphi_t(\hat{\Sigma})$ and $\varphi_t(\Sigma)$ and the formulae above is valid on $\Omega_t$. Next, to handle the summand in this integral, we use the following lemma taken from [LGSL24].

**Lemma 10** ([LGSL24])**.** *There exists an absolute constant $C > 1$ such that the following holds. Let $t \geq 1$. For the contour $\mathcal{C}_t$ defined in (3.50) and for any $z \in \mathcal{C}_t$, we have*

$$\left\| \Sigma_t^{\frac{1}{2}} (\Sigma - zI_p)^{-1} \Sigma_t^{\frac{1}{2}} \right\|_{op} \leq C, \ \oint_{\mathcal{C}_t} |\varphi_t(z) dz| \leq C \log(t), \ \text{and} \ \oint_{\mathcal{C}_t} |\psi_t(z) dz| \leq C t^{-1}.$$

*Moreover, on $\Omega_t$, for any $z \in \mathcal{C}_t$, we further have*

$$\left\| \Sigma_t^{\frac{1}{2}} \left( \hat{\Sigma} - zI_p \right)^{-1} \Sigma_t^{\frac{1}{2}} \right\|_{op} \leq C.$$

For the sake of completeness, we provide the proof of Lemma 10 in Section 3.5.3. On the event $\Omega_t$, other properties that will be useful in our analysis hold. For instance, to obtain an upper bound for $\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\|_2$, we will further require the following result.

**Lemma 11.** *Let $t \geq 1$ and recall that $\hat{\Sigma}_t = \hat{\Sigma} + t^{-1} I_p$. On the event $\Omega_t$, we have $\|\Sigma_J^{\frac{1}{2}} \hat{\Sigma}_t^{-\frac{1}{2}}\|_{op}^2 \leq \|\Sigma_t^{\frac{1}{2}} \hat{\Sigma}_t^{-\frac{1}{2}}\|_{op}^2 \leq 2$ and $\|\Sigma_t^{-\frac{1}{2}} \hat{\Sigma}_t^{\frac{1}{2}}\|_{op}^2 \leq 2$.*

Lemma 11 provides the following insight: for a suitably chosen $J$, the (modified) population covariance and the (modified) sample covariance can be interchanged. The proof of Lemma 11 may be found in Section 3.5.4

The event $\Omega_t$ contains all the properties on $\hat{\Sigma}$ that are enough for our analysis. The only remaining stochastic argument used in the proof from now are only dealing with the noise. As a consequence, if one wants to extend the conclusion from Theorem 7 beyond Assumption 6, one may only focus on proving that $\Omega_t$ happens with large probability under the new considered setup. Now, that we have dealt with mostly all the stochastic aspect of the proof we can move to the deterministic one, as long as we work on the event $\Omega_t$.

### 3.3.2 The estimation property of $\hat{\boldsymbol{\beta}}_J$

In this subsection, we investigate the estimation properties of $\hat{\boldsymbol{\beta}}_J$, i.e. we obtain a high probability upper bound on $\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\right\|_2$. In the following analysis, we will see that the estimation error analysis for the estimator on $V_J$, namely $\hat{\boldsymbol{\beta}}_J$, is similar to the classical analysis of spectral methods but performed over $V_J$. This is because on this subspace the problem reduces to standard estimation. From this perspective, the FSD method can be viewed as an additional layer around classical analysis only requiring an isomorphic property on the estimation space instead of the entire space, thereby providing better estimation properties under smaller sample complexity.

#### Risk decomposition of the estimation part $\hat{\boldsymbol{\beta}}_J$

We start with a risk decomposition of the estimation part $\hat{\boldsymbol{\beta}}_J$ of the spectral method $\hat{\boldsymbol{\beta}}$. Let the "population" spectral method be defined as $\tilde{\boldsymbol{\beta}} = \varphi_t(\Sigma)\Sigma\boldsymbol{\beta}^*$. It is the 'population version' of $\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) = \varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*$ where $\hat{\Sigma}$ has been replaced by $\Sigma$; we therefore look at $\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*)$ as a plug-in estimator of $\tilde{\boldsymbol{\beta}}$ in the noise free case and in the estimation part of the feature space. Then, by linearity of $\hat{\boldsymbol{\beta}}$, we may decompose $\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*$ as follows:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_J(\boldsymbol{y}) - \boldsymbol{\beta}_J^* &= \hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*) - \boldsymbol{\beta}_J^* + \hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}) \\ &= \left(\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*) - \tilde{\boldsymbol{\beta}}_J\right) + \left(\tilde{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*\right) + \hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}).\end{aligned}$$

Here, $\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*) - \tilde{\boldsymbol{\beta}}_J$ plays the role of a bias term of the plug-in estimator $\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*)$ in the free noise case, while $\tilde{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*$ denotes an approximation error and $\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi})$ is considered as a variance term. The following risk decomposition follows from the decomposition above:

$$\begin{aligned}\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\right\|_2 \leq \left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*) - \tilde{\boldsymbol{\beta}}_J)\right\|_2 + \left\|\Sigma_J^{1/2}(\tilde{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\right\|_2 \\ + \left\|\Sigma_J^{1/2}\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi})\right\|_2.\end{aligned} \tag{3.19}$$

Next, we upper bound the three terms from this sum.

#### Upper bound on the approximation term $\left\|\Sigma_J^{1/2}(\tilde{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\right\|_2$

It follows from the definition of the residual function $\psi_t : x \in \mathbb{R}^+ \to 1 - x\varphi_t(x)$ that $\tilde{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^* = (\varphi_t(\Sigma)\Sigma - I_p)\boldsymbol{\beta}_J^* = -\psi_t(\Sigma)\boldsymbol{\beta}_J^*$ and so

$$\left\|\Sigma_J^{1/2}(\tilde{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\right\|_2 = \left\|\Sigma_J^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2. \tag{3.20}$$

Next, we move to an upper bound on the bias of the plug-in estimator $\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*)$. We will see that the approximation term above is dominating the bias term.

**Upper bound on the bias term** $\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*) - \tilde{\boldsymbol{\beta}}_J)\right\|_2$

The filter and residual functions satisfy the relation $\varphi_t(x)x + \psi_t(x) = 1$, hence, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*) - \tilde{\boldsymbol{\beta}}_J &= P_J\varphi(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}_J^* - P_J(\varphi_t(\hat{\Sigma})\hat{\Sigma} + \psi_t(\hat{\Sigma}))\tilde{\boldsymbol{\beta}}_J \\
&= P_J\varphi_t(\hat{\Sigma})\hat{\Sigma}(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J) - P_J\psi_t(\hat{\Sigma})\tilde{\boldsymbol{\beta}}_J \\
&= P_J\varphi_t(\hat{\Sigma})(\hat{\Sigma} - \Sigma)(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J) + P_J\varphi_t(\hat{\Sigma})\Sigma(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J) - P_J\psi_t(\hat{\Sigma})\tilde{\boldsymbol{\beta}}_J \\
&= P_J\varphi_t(\hat{\Sigma})(\hat{\Sigma} - \Sigma)(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J) + P_J\left(\varphi_t(\hat{\Sigma}) - \varphi_t(\Sigma)\right)\Sigma\psi_t(\Sigma)\boldsymbol{\beta}_J^* \\
&+ P_J\left(\psi_t(\Sigma) - \psi_t(\hat{\Sigma})\right)\Sigma\varphi_t(\Sigma)\boldsymbol{\beta}_J^*
\end{aligned}$$

where we used the fact that $\tilde{\boldsymbol{\beta}}_J := P_J\tilde{\boldsymbol{\beta}} = \varphi_t(\Sigma_J)\Sigma_J\boldsymbol{\beta}_J^* = \varphi_t(\Sigma)\Sigma\boldsymbol{\beta}_J^*$ and so $\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J = \psi_t(\Sigma)\boldsymbol{\beta}_J^*$ because $V_J$ is an eigenspace of $\Sigma$ and the fact that $\Sigma$, $\varphi_t(\Sigma)$ and $\psi_t(\Sigma)$ commute. Now, by taking $\|\Sigma_J^{1/2}\cdot\|_2$ on both sides, we obtain the following decomposition of the bias term:

$$\begin{aligned}
\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*) - \tilde{\boldsymbol{\beta}}_J)\|_2 \leq &\left\|\Sigma_J^{1/2}\varphi_t(\hat{\Sigma})(\hat{\Sigma} - \Sigma)(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J)\right\|_2 \\
&+ \left\|\Sigma_J^{1/2}\left(\varphi_t(\hat{\Sigma}) - \varphi_t(\Sigma)\right)\Sigma\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2 \\
&+ \left\|\Sigma_J^{1/2}\left(\psi_t(\Sigma) - \psi_t(\hat{\Sigma})\right)\Sigma\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2.
\end{aligned} \tag{3.21}$$

Next, we provide upper bounds on the three terms in this sum.

**Upper bound for** $\left\|\Sigma_J^{1/2}\varphi_t(\hat{\Sigma})(\hat{\Sigma} - \Sigma)(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J)\right\|_2$ We recall that $\hat{\Sigma}_t = \hat{\Sigma} + t^{-1}I_p$. We have

$$\begin{aligned}
&\left\|\Sigma_J^{1/2}\varphi_t(\hat{\Sigma})(\hat{\Sigma} - \Sigma)(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J)\right\|_2 \\
&\leq \|\Sigma_J^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\|_{\mathrm{op}}\|\Sigma_t^{\frac{1}{2}}\varphi_t(\hat{\Sigma})\Sigma_t^{\frac{1}{2}}\|_{\mathrm{op}}\|\Sigma_t^{-\frac{1}{2}}(\hat{\Sigma} - \Sigma)\Sigma_t^{-\frac{1}{2}}\|_{\mathrm{op}}\|\Sigma_t^{\frac{1}{2}}(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J)\|_2.
\end{aligned} \tag{3.22}$$

Under Assumption 5, we know that $\varphi_t(x) \leq C_{27}(x + t^{-1})^{-1}$ hence, by Lemma 11, we have, on $\Omega_t$,

$$\|\Sigma_t^{\frac{1}{2}}\varphi_t(\hat{\Sigma})\Sigma_t^{\frac{1}{2}}\|_{\mathrm{op}} \leq \left\|\Sigma_t^{\frac{1}{2}}\hat{\Sigma}_t^{-\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\hat{\Sigma}_t^{\frac{1}{2}}\varphi_t(\hat{\Sigma})\hat{\Sigma}_t^{\frac{1}{2}}\right\|_{\mathrm{op}}\left\|\hat{\Sigma}_t^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\right\|_{\mathrm{op}} \leq 2C_{27}. \tag{3.23}$$

Moreover, by (3.11), $\|\Sigma_J^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\|_{\mathrm{op}} \leq 1$. Plugging (3.23) into (3.22) together with (3.13), on $\Omega_t$, we have

$$\begin{aligned}
&\left\|\Sigma_J^{\frac{1}{2}}\varphi_t(\hat{\Sigma})(\hat{\Sigma} - \Sigma)(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J)\right\|_2 \leq 2\square C_1\|\Sigma_t^{\frac{1}{2}}(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J)\|_2 \\
&\leq 2C_1\left(\frac{1+b}{b}\right)\square\|\Sigma_J^{1/2}(\boldsymbol{\beta}_J^* - \tilde{\boldsymbol{\beta}}_J)\|_2 \leq 2C_1\left(\frac{1+b}{b}\right)\square\left\|\Sigma_J^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2
\end{aligned} \tag{3.24}$$

where we used (3.20) and (3.12) in the last inequality.

**Upper bound for** $\left\|\Sigma_J^{1/2}(\varphi_t(\hat{\Sigma}) - \varphi_t(\Sigma))\Sigma\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2$ To handle this term, we use (3.18) which is valid on $\Omega_t$: on $\Omega_t$, we have

$$\begin{aligned}
&\Sigma_J^{\frac{1}{2}}(\varphi_t(\hat{\Sigma}) - \varphi_t(\Sigma))\Sigma\psi_t(\Sigma)\boldsymbol{\beta}_J^* \\
&= \frac{1}{2\pi i}\oint_{\mathcal{C}_t}\Sigma_J^{\frac{1}{2}}\left(\hat{\Sigma} - zI_p\right)^{-1}\left(\hat{\Sigma} - \Sigma\right)(\Sigma - zI_p)^{-1}\Sigma\psi_t(\Sigma)\boldsymbol{\beta}_J^*\varphi_t(z)dz \\
&= \frac{1}{2\pi i}\oint_{\mathcal{C}_t}\Sigma_J^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\Sigma_t^{\frac{1}{2}}\left(\hat{\Sigma} - zI_p\right)^{-1}\Sigma_t^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\left(\Sigma - \hat{\Sigma}\right)\Sigma_t^{-\frac{1}{2}}\Sigma_t^{\frac{1}{2}}(\Sigma - zI_p)^{-1}\Sigma^{\frac{1}{2}} \\
&\qquad\cdot\Sigma^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\varphi_t(z)dz.
\end{aligned}$$

Taking the $\|\cdot\|_2$ norm on both sides and applying Lemma 10 yields, on $\Omega_t$,

$$\begin{aligned}
&\left\|\Sigma_J^{\frac{1}{2}}\left(\varphi_t(\hat{\Sigma})-\varphi_t(\Sigma)\right)\Sigma\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2\\
\leq&\|\Sigma_J^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\|_{\mathrm{op}}\oint_{\mathcal{C}_t}\left\|\Sigma_t^{\frac{1}{2}}\left(\hat{\Sigma}-zI_p\right)^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{op}\left\|\Sigma_t^{-\frac{1}{2}}\left(\Sigma-\hat{\Sigma}\right)\Sigma_t^{-\frac{1}{2}}\right\|_{op}\\
&\cdot\left\|\Sigma_t^{\frac{1}{2}}(\Sigma-zI_p)^{-1}\Sigma^{\frac{1}{2}}\right\|_{op}\left\|\Sigma^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2|\varphi_t(z)dz|\\
\lesssim&\square\left\|\Sigma^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2\oint_{\mathcal{C}_t}|\varphi_t(z)dz|\lesssim\square\log(t)\left\|\Sigma^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2,
\end{aligned}\tag{3.25}$$

where we have used that $\Sigma\preceq\Sigma_t$ to get $\left\|\Sigma_t^{\frac{1}{2}}(\Sigma-zI_p)^{-1}\Sigma^{\frac{1}{2}}\right\|_{op}\leq\left\|\Sigma_t^{\frac{1}{2}}(\Sigma-zI_p)^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{op}\lesssim 1$ from Lemma 10.

**Upper bound for** $\left\|\Sigma_J^{1/2}\left(\psi_t(\Sigma)-\psi_t(\hat{\Sigma})\right)\Sigma\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2$ We have on $\Omega_t$ and from (3.18) :

$$\begin{aligned}
&\Sigma_J^{\frac{1}{2}}\left(\psi_t(\hat{\Sigma})-\psi_t(\Sigma)\right)\Sigma\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\\
&=\frac{1}{2\pi i}\oint_{\mathcal{C}_t}\Sigma_J^{\frac{1}{2}}\left(\hat{\Sigma}-zI_p\right)^{-1}\left(\hat{\Sigma}-\Sigma\right)(\Sigma-zI_p)^{-1}\Sigma\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\psi_t(z)dz\\
&=\frac{1}{2\pi i}\oint_{\mathcal{C}_t}\Sigma_J^{\frac{1}{2}}\left(\hat{\Sigma}-zI_p\right)^{-1}\Sigma_t^{\frac{1}{2}}\cdot\Sigma_t^{-\frac{1}{2}}\left(\hat{\Sigma}-\Sigma\right)\Sigma_t^{-\frac{1}{2}}\Sigma_t^{\frac{1}{2}}(\Sigma-zI_p)^{-1}\Sigma_J^{\frac{1}{2}}\\
&\qquad\cdot\Sigma_J^{\frac{1}{2}}\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\psi_t(z)dz.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left\|\Sigma_J^{\frac{1}{2}}\left(\psi_t(\Sigma)-\psi_t(\hat{\Sigma})\right)\Sigma\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2\\
&\leq\frac{1}{2\pi}\oint_{\mathcal{C}_t}\left\|\Sigma_t^{\frac{1}{2}}\left(\hat{\Sigma}-zI_p\right)^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{op}\left\|\Sigma_t^{-\frac{1}{2}}\left(\hat{\Sigma}-\Sigma\right)\Sigma_t^{-\frac{1}{2}}\right\|_{op}\\
&\cdot\left\|\Sigma_t^{\frac{1}{2}}(\Sigma-zI_p)^{-1}\Sigma_J^{\frac{1}{2}}\right\|_{op}\cdot\left\|\Sigma_J^{\frac{1}{2}}\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2|\psi_t(z)dz|\\
&\lesssim\square\cdot\left\|\Sigma_J^{\frac{1}{2}}\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2\cdot\oint_{\mathcal{C}_t}|\psi_t(z)dz|\lesssim\square\left\|\Sigma_J^{1/2}\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2t^{-1}.
\end{aligned}\tag{3.26}$$

Collecting (3.24), (3.25) and (3.26) all together in (3.21), we obtain that, on $\Omega_t$, it holds

$$\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*)-\tilde{\boldsymbol{\beta}}_J)\right\|_2\lesssim\square\left(\log(et)\left\|\Sigma_J^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2+t^{-1}\left\|\Sigma_J^{1/2}\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2\right)\tag{3.27}$$

and since $\varphi_t(\Sigma)\preceq C_1\Sigma_t^{-1}$, we obtain $\left\|\Sigma_J^{1/2}\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2\leq C_1\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}_J^*\right\|_2$, we finally get, on $\Omega_t$,

$$\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*)-\tilde{\boldsymbol{\beta}}_J)\right\|_2\lesssim\square\left(\log(et)\left\|\Sigma_J^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2+t^{-1}\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}_J^*\right\|_2\right).\tag{3.28}$$

**Upper bound on the variance term** $\left\|\Sigma_J^{1/2}\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_{J^c}^*+\boldsymbol{\xi})\right\|_2$.

By linearity of the spectral estimator (see (3.1)), we have

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

Now, we prove high probability upper bounds on the two terms from the sum above.

**Upper bound for $\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\hat\Sigma\boldsymbol{\beta}^*_{J^c}\|_2$** We have

$$\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\hat\Sigma\boldsymbol{\beta}^*_{J^c}\|_2 = \frac{1}{N}\left\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\Sigma_t^{\frac12}\Sigma_t^{-\frac12}\mathbb{X}^\top\mathbb{X}\boldsymbol{\beta}^*_{J^c}\right\|_2$$
$$\le \frac{1}{\sqrt{N}}\left\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\Sigma_t^{1/2}\right\|_{\rm op}\left\|\Sigma_t^{-1/2}\mathbb{X}^\top\right\|_{\rm op}\frac{\|\mathbb{X}\boldsymbol{\beta}^*_{J^c}\|_2}{\sqrt{N}}.$$

It follows from (3.23), (3.11) and Lemma 11, that on the event $\Omega_t$,

$$\frac{1}{\sqrt{N}}\left\|\Sigma_t^{-1/2}\mathbb{X}^\top\right\|_{\rm op} = \left\|\Sigma_t^{-1/2}\hat\Sigma^{1/2}\right\|_{\rm op} \le \sqrt{2}$$

and

$$\left\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\Sigma_t^{1/2}\right\|_{\rm op} \le \left\|\Sigma_J^{1/2}\Sigma_t^{-1/2}\right\|_{\rm op}\left\|\Sigma_t^{1/2}\varphi_t(\hat\Sigma)\Sigma_t^{1/2}\right\|_{\rm op} \le 2C_1.$$

Next, it follows from the sub-gaussian property of the design vector $X$ from Assumption 6 and Lemma 13 that, for some absolute constant $c>0$, with probability at least $1-\exp(-cN)$,

$$\frac{1}{N}\|\mathbb{X}\boldsymbol{\beta}^*_{J^c}\|_2^2 = \frac{1}{N}\sum_{i=1}^N\langle X_i,\boldsymbol{\beta}^*_{J^c}\rangle^2 \le 2\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2^2.$$

For the convenience of the reader, we reproduce Lemma 13 here; its proof will be given on page 80.

**Lemma 13 (recall).** *There is some absolute constant $c>0$ such that the following holds. Let $X$ be a sub-gaussian vector in $\mathbb{R}^p$ and denote $\Sigma = \mathbb{E}XX^\top$ ($X$ is not necessarily centered). Let $\boldsymbol{v}\in\mathbb{R}^p$. With probability at least $1-\exp(-cN)$, we have*

$$\frac12\left\|\Sigma^{\frac12}\boldsymbol{v}\right\|_2^2 \le \frac1N\sum_{i=1}^N\langle X_i,\boldsymbol{v}\rangle^2 \le \frac32\left\|\Sigma^{\frac12}\boldsymbol{v}\right\|_2^2. \tag{3.47}$$

As a result, there exist an absolute constants $c>0$ such that with probability at least $1-\exp(-c|J|)-\mathbb{P}[\Omega_t^c]$,

$$\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\hat\Sigma\boldsymbol{\beta}^*_{J^c}\|_2 \le 16C_1\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2. \tag{3.29}$$

**Upper bound for $(1/N)\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\mathbb{X}^\top\boldsymbol{\xi}\|_2^2$** We first work conditionally on $\mathbb{X}$ and consider the randomness coming only from the Gaussian vector $\boldsymbol{\xi}$ so that we can apply the Borel-TIS inequality (see Theorem 7.1 in [Led96] or p.56-57 in [LT91]) in order to get: for almost all $\mathbb{X}$, for all $t\ge1$ with probability at least $1-\exp(-t/2)$, $\|A\boldsymbol{\xi}\|_2 \le \sigma_\xi\sqrt{\mathrm{Tr}[AA^\top]}+\sigma_\xi\|A\|_{op}\sqrt{t}$ where $A=\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\mathbb{X}^\top$. This implies that for almost all $\mathbb{X}$, with probability at least $1-\exp(-|J|/2)$,

$$\frac1N\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\mathbb{X}^\top\boldsymbol{\xi}\|_2^2 \le 2\sigma_\xi^2\,\mathrm{Tr}\left[\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\hat\Sigma\varphi_t(\hat\Sigma)\Sigma_J^{1/2}\right] + \frac{2\sigma_\xi^2}{N}\left\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\mathbb{X}^\top\right\|_{op}^2|J|.$$

For the weak variance term in the inequality above, we have $\hat\Sigma_t^{1/2}\varphi_t(\hat\Sigma)\hat\Sigma\varphi_t(\hat\Sigma)\hat\Sigma_t^{1/2} \preceq C_{27}^2 I_p$ and so by Lemma 11 we get, on $\Omega_t$,

$$\frac1N\left\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\mathbb{X}^\top\right\|_{\rm op}^2 \le \left\|\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\hat\Sigma\varphi_t(\hat\Sigma)\Sigma_J^{1/2}\right\|_{\rm op}$$
$$\le \left\|\Sigma_J^{1/2}\hat\Sigma_t^{-1/2}\right\|_{\rm op}\left\|\hat\Sigma_t^{1/2}\varphi_t(\hat\Sigma)\hat\Sigma\varphi_t(\hat\Sigma)\hat\Sigma_t^{1/2}\right\|_{\rm op}\left\|\hat\Sigma_t^{-1/2}\Sigma_J^{1/2}\right\|_{\rm op} \le 2C_{27}^2.$$

For the strong variance term in the inequality above, we use that $\varphi_t(\hat\Sigma)\hat\Sigma\varphi_t(\hat\Sigma) \preceq C_{27}^2\hat\Sigma_t^{-1}$ and apply Lemma 11 to get, on $\Omega_t$,

$$\mathrm{Tr}\left[\Sigma_J^{1/2}\varphi_t(\hat\Sigma)\hat\Sigma\varphi_t(\hat\Sigma)\Sigma_J^{1/2}\right] \le C_{27}^2\,\mathrm{Tr}\left[\Sigma_J^{1/2}\hat\Sigma_t^{-1}\Sigma_J^{1/2}\right] = C_{27}^2\left(\mathrm{Tr}\left[\hat\Sigma_t^{-1}(\Sigma_J-\hat\Sigma_J)\right]+\mathrm{Tr}\left[\hat\Sigma_t^{-1}\hat\Sigma_J\right]\right)$$
$$\le C_{27}^2\left(\mathrm{Tr}\left[\hat\Sigma_t^{-1/2}(\Sigma_J-\hat\Sigma_J)\hat\Sigma_t^{-1/2}\right]+|J|\right)$$
$$\le C_{27}^2\left(|J|\left\|\hat\Sigma_t^{-1/2}(\Sigma_J-\hat\Sigma_J)\hat\Sigma_t^{-1/2}\right\|_{op}+|J|\right) \le 2C_{27}^2|J|.$$

As a consequence, we obtain that with probability at least $1-2\exp(-c|J|)-\mathbb{P}[\Omega_t^c]$, $(1/N)\|\Sigma_J^{1/2}\varphi_t(\hat{\Sigma})\mathbb{X}^\top\boldsymbol{\xi}\|_2^2 \lesssim \sigma_\xi^2|J|$.

Finally, gathering the last inequality together with (3.29) we obtain that with probability at least $1-2\exp(-c|J|)-\mathbb{P}[\Omega_t^c]$,

$$\|\Sigma_J^{1/2}\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}^*_{J^c}+\boldsymbol{\xi})\|_2 \lesssim \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2 + \sigma_\xi\sqrt{\frac{|J|}{N}}.$$

**Conclusion on the estimation property of $\hat{\boldsymbol{\beta}}_J$**

It follows from the results obtained in the previous sections, that with probability at least $1-2\exp(-c|J|)-\mathbb{P}[\Omega_t^c]$,

$$\begin{aligned}\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}^*_J)\right\|_2 \lesssim \sigma_\xi\sqrt{\frac{|J|}{N}} + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\right\|_2 + (\square\log(t)+1)\left\|\Sigma_J^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*\right\|_2 \\ + \frac{\square}{t}\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}^*_J\right\|_2.\end{aligned} \tag{3.30}$$

This result finishes our analysis of the statistical property of the estimation part $\hat{\boldsymbol{\beta}}_J$ of the spectral method $\hat{\boldsymbol{\beta}}$. The next step of the FSD method is to handle the 'noise absorption part' of $\hat{\boldsymbol{\beta}}$.

### 3.3.3 Control of the noise absorption part $\hat{\boldsymbol{\beta}}_{J^c}$

In this section, we derive an upper bound for $\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$, where $\hat{\boldsymbol{\beta}}_{J^c}=P_{J^c}\hat{\boldsymbol{\beta}}$. We recall that $\hat{\boldsymbol{\beta}}=N^{-1}\varphi_t(\hat{\Sigma})\mathbb{X}^\top\boldsymbol{y}$ and $\boldsymbol{y}=\mathbb{X}\boldsymbol{\beta}^*+\boldsymbol{\xi}=\mathbb{X}\boldsymbol{\beta}^*_J+\mathbb{X}\boldsymbol{\beta}^*_{J^c}+\boldsymbol{\xi}$. Therefore, we have

$$\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2 \le \left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*_J\right\|_2 + \left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*_{J^c}\right\|_2 + \left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})[N^{-1}\mathbb{X}^\top]\boldsymbol{\xi}\right\|_2. \tag{3.31}$$

Next, we prove high probability upper bounds on the three terms in the sum above.

**Upper bound for $\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*_J\right\|_2$**

By definition of the residual function, we have $\varphi_t(\hat{\Sigma})\hat{\Sigma}=I_p-\psi_t(\hat{\Sigma})$ and so $\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*_J=-\Sigma_{J^c}^{1/2}\psi_t(\hat{\Sigma})\boldsymbol{\beta}^*_J$ where we have used the fact that $\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_J=0$. Next, we take the $\ell_2^p$-norm on both sides and use the fact that $\Sigma_{J^c}^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*_J=0$ to get

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

Next, on $\Omega_t$, we can apply the residual theorem to $\psi_t(\hat{\Sigma})$ and $\psi_t(\Sigma)$ and get a result similar to the one of (3.18) where $\varphi_t$ is replaced by $\psi_t$. Thanks to this result we get (on $\Omega_t$)

$$\begin{aligned}&\left\|\Sigma_{J^c}^{1/2}\big(\psi_t(\hat{\Sigma})-\psi_t(\Sigma)\big)\boldsymbol{\beta}^*_J\right\|_2 = \left\|\Sigma_{J^c}^{\frac12}\Sigma_t^{-\frac12}\Sigma_t^{\frac12}\big(\psi_t(\hat{\Sigma})-\psi_t(\Sigma)\big)\boldsymbol{\beta}^*_J\right\|_2 \\ &\le \left\|\Sigma_{J^c}^{\frac12}\Sigma_t^{-\frac12}\right\|_{\mathrm{op}}\left\|\Sigma_t^{\frac12}\big(\psi_t(\hat{\Sigma})-\psi_t(\Sigma)\big)\boldsymbol{\beta}^*_J\right\|_2 \\ &\le \sqrt{\frac{b}{1+b}}\left\|\oint_{\mathcal{C}_t}\Sigma_t^{\frac12}(\hat{\Sigma}-zI_p)^{-1}(\hat{\Sigma}-\Sigma)(\Sigma-zI_p)^{-1}\boldsymbol{\beta}^*_J\psi_t(z)dz\right\|_2 \\ &\le \sqrt{\frac{b}{1+b}}\oint_{\mathcal{C}_t}\left\|\Sigma_t^{\frac12}(\hat{\Sigma}-zI_p)^{-1}\Sigma_t^{\frac12}\right\|_{\mathrm{op}}\left\|\Sigma_t^{-\frac12}(\hat{\Sigma}-\Sigma)\Sigma_t^{-\frac12}\right\|_{\mathrm{op}} \\ &\cdot\left\|\Sigma_t^{\frac12}(\Sigma-zI_p)^{-1}\Sigma_t^{\frac12}\right\|_{\mathrm{op}}\left\|\Sigma_t^{-\frac12}\boldsymbol{\beta}^*_J\right\|_2|\psi_t(z)dz|\end{aligned}$$

and so, on $\Omega_t$, by applying Lemma 10 we obtain

$$\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*_J\right\|_2 = \left\|\Sigma_{J^c}^{1/2}\big(\psi_t(\hat{\Sigma})-\psi_t(\Sigma)\big)\boldsymbol{\beta}^*_J\right\|_2 \lesssim \frac{\square}{t}\left\|\Sigma_J^{-\frac12}\boldsymbol{\beta}^*_J\right\|_2. \tag{3.32}$$

**Upper bound for $\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*_{J^c}\right\|_2$**

For the convenience of the reader, we recall the following property, which can be found on page 29.

**Proposition 13 (Recall).** *Suppose Assumption 3 holds. There exist some absolute constants $c_3$ and $C_8 > 0$ such that with a probability of at least $1 - \bar{p}_{DMU}$, where $\bar{p}_{DMU} = \frac{c_3}{N^\epsilon} + \gamma_1$, there holds $\left\|\Sigma_{J^c}^{1/2}\mathbb{X}_{J^c}^\top\right\|_{op} \leq C_8\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)} + C_8\sqrt{N}\|\Sigma_{J^c}\|_{op}$.*

It is straightforward to verify that Assumption 6 implies Assumption 3. It follows from Proposition 13 that under Assumption 6, there are absolute constants $C, c > 0$ such that with probability at least $1 - \exp(-cN)$,

$$\mathbb{P}\left(\left\|\Sigma_{J^c}^{1/2}\mathbb{X}^\top\right\|_{\mathrm{op}} \leq C\left(\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)} + \sqrt{N}\left\|\Sigma_{J^c}\right\|_{\mathrm{op}}\right)\right) \geq 1 - \exp(-cN). \tag{3.33}$$

Moreover, we have $\|\mathbb{X}\boldsymbol{\beta}^*_{J^c}\|_2 \leq C\sqrt{N}\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2$ with probability at least $1 - \exp(-cN)$. Next, we observe that thanks to Assumption 5, $\varphi_t(x) \leq C_1(x + t^{-1})^{-1} \leq C_1 t$ so that we have $\varphi_t(\hat{\Sigma}) \leq C_1 t I_p$ and (since $\hat{\Sigma}$ and $I_p$ commute) for all $x \in \mathbb{R}^p$, $\left\|\varphi_t(\hat{\Sigma})x\right\|_2 \leq C_1 t\|x\|_2$. It follows that with probability at least $1 - 2\exp(-cN)$,

$$\begin{aligned}\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*_{J^c}\right\|_2 &= \frac{1}{N}\left\|\Sigma_{J^c}^{1/2}\mathbb{X}^\top\varphi_t(\hat{\Sigma})\mathbb{X}\boldsymbol{\beta}^*_{J^c}\right\|_2 \\ &\leq CC_1\frac{\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)} + \sqrt{N}\left\|\Sigma_{J^c}\right\|_{\mathrm{op}}}{\sqrt{N}t^{-1}}\left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\right\|_2.\end{aligned} \tag{3.34}$$

Finally, it follows from the definition of $k^*$ that $\sigma_{k^*+1} = \left\|\Sigma_{J^c}\right\|_{\mathrm{op}} \leq bt^{-1}$ and from the sample complexity assumption (i.e. $\square^2 N \gtrsim \mathrm{Tr}\left[\Sigma(\Sigma + t^{-1}I_p)^{-1}\right]$) - see the discussion below (3.15) - that $\square^2 N \gtrsim t\,\mathrm{Tr}[\Sigma_{J^c}]$ so that

$$\frac{\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)} + \sqrt{N}\left\|\Sigma_{J^c}\right\|_{\mathrm{op}}}{\sqrt{N}t^{-1}} \leq \sqrt{\frac{\left\|\Sigma_{J^c}\right\|_{\mathrm{op}}}{t^{-1}}}\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c})}{Nt^{-1}}} + \frac{\left\|\Sigma_{J^c}\right\|_{\mathrm{op}}}{t^{-1}} \leq \sqrt{b\square} + b \leq 2b \tag{3.35}$$

as long as $\square \leq b$. We conclude that with probability at least $1 - 2\exp(-cN)$,

$$\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}^*_{J^c}\right\|_2 \leq CC_1 b\left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\right\|_2. \tag{3.36}$$

**Upper bound for $\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})[N^{-1}\mathbb{X}^\top]\boldsymbol{\xi}\right\|_2$**

As in the previous section we first condition on $\mathbb{X}$ and apply the Borell-TIS inequality: for almost all $\mathbb{X}$, for all $r > 0$, with probability at least $1 - \exp(-r/2)$, $\|A\boldsymbol{\xi}\|_2 \leq \sigma_\xi\sqrt{\mathrm{Tr}[AA^\top]} + \sigma_\xi\|A\|_{op}\sqrt{r}$ where $A = \Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})[N^{-1}\mathbb{X}^\top]$. Hence, we have with probabiltity at least $1 - \exp(-|J|/2)$,

$$\begin{aligned}\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})[N^{-1}\mathbb{X}^\top]\boldsymbol{\xi}\right\|_2 &\leq \sigma_\xi\sqrt{\frac{\mathrm{Tr}[\Sigma_{J^c}\hat{\Sigma}\varphi_t^2(\hat{\Sigma})]}{N}} + \sigma_\xi\left\|\Sigma_{J^c}^{1/2}\hat{\Sigma}^{1/2}\varphi_t(\hat{\Sigma})\right\|_{op}\sqrt{\frac{|J|}{N}} \\ &\leq \sigma_\xi C_{27}t\sqrt{\frac{\mathrm{Tr}[\Sigma_{J^c}\hat{\Sigma}]}{N}} + \sigma_\xi C_{27}t\left\|\Sigma_{J^c}^{1/2}\hat{\Sigma}^{1/2}\right\|_{op}\sqrt{\frac{|J|}{N}}\end{aligned} \tag{3.37}$$

where in the last inequality we used that $\varphi_t(x) \leq C_{27}(x + t^{-1})^{-1} \leq C_{27}t$. We use the following Lemma 15, whose proof will be given on page 81.

**Lemma 15 (recall).** *There exists an absolute constant $c > 0$ such that the following holds. Let $X$ be a sub-gaussian vector in $\mathbb{R}^p$ and denote $\Sigma = \mathbb{E}XX^\top$ ($X$ is not necessarily centered). Let $A$ be a matrix in $\mathbb{R}^{p\times d}$. With probability at least $1 - \exp(-cN)$,*

$$\frac{1}{2}\left\|\Sigma^{1/2}A^\top\right\|_{HS}^2 \leq \frac{1}{N}\sum_{i=1}^N\|AX_i\|_2^2 \leq \frac{3}{2}\left\|\Sigma^{1/2}A^\top\right\|_{HS}^2.$$

Next, it follows from Lemma 15 that there exists an absolute constant $c > 0$ such that with probability at least $1 - \exp(-cN)$,

$$\mathrm{Tr}[\Sigma_{J^c}\hat{\Sigma}] = \frac{1}{N}\mathrm{Tr}(\mathbb{X}\Sigma_{J^c}\mathbb{X}^\top) = \frac{1}{N}\sum_{i=1}^N \left\|\Sigma_{J^c}^{1/2}X_i\right\|_2^2 \leq 2\,\mathrm{Tr}(\Sigma_{J^c}^2).$$

Then, it follows from (3.33) that there are absolute constants $C, c > 0$ such that with probability at least $1-\exp(-cN)$,

$$\left\|\Sigma_{J^c}^{1/2}\hat{\Sigma}^{1/2}\right\|_{op} = \frac{1}{\sqrt{N}}\left\|\Sigma_{J^c}^{1/2}\mathbb{X}^\top\right\|_{op} \leq C\left(\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}} + \|\Sigma_{J^c}\|_{\mathrm{op}}\right). \tag{3.38}$$

Finally, collecting the last two results together with (3.35) in the Borell-TIS inequality above, we get that with probability at least $1 - 2\exp(-c|J|)$,

$$\begin{aligned}\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})[N^{-1}\mathbb{X}^\top]\boldsymbol{\xi}\right\|_2 &\lesssim \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}} + \sigma_\xi t\left(\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}} + \|\Sigma_{J^c}\|_{\mathrm{op}}\right)\sqrt{\frac{|J|}{N}} \\ &\lesssim \sigma_\xi\sqrt{\frac{|J|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}}.\end{aligned} \tag{3.39}$$

**Concluding on the noise absorption property**

Combining (3.32), (3.36) and (3.39), we obtain that with probability at least $1 - 2\exp(-c|J|) - \mathbb{P}[\Omega_t^c]$,

$$\left\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\right\|_2 \lesssim \frac{\square}{t}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}_J^*\right\|_2 + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 + \sigma_\xi\sqrt{\frac{|J|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}}. \tag{3.40}$$

### 3.3.4 End of the proof of the upper bound from Theorem 7

Going back to the original risk decomposition from the FSD method in (3.10) and collecting both results on the estimation part and the noise absorption part from (3.30) and (3.40), we obtain that with probability at least $1 - \exp(-c|J|) - \mathbb{P}[\Omega_t^c]$,

$$\begin{aligned}&\left\|\Sigma^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right)\right\|_2 \leq \left\|\Sigma_J^{1/2}\left(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*\right)\right\|_2 + \left\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\right\|_2 + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 \\ &\lesssim \left(\sigma_\xi\sqrt{\frac{|J|}{N}} + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 + (\square\log(et)+1)\left\|\Sigma_J^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*\right\|_2 + \frac{\square}{t}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}_J^*\right\|_2\right) \\ &+ \left(\frac{\square}{t}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}_J^*\right\|_2 + \sigma_\xi\sqrt{\frac{|J|}{N}} + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}}\right) + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 \\ &\lesssim \sigma_\xi\sqrt{\frac{|J|}{N}} + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 + (\square\log(et)+1)\left\|\Sigma_J^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}^*\right\|_2 \\ &+ \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}} + \frac{\square}{t}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}_J^*\right\|_2\end{aligned}$$

and the result follows if one takes $\square \lesssim \log^{-1}(et)$.

## 3.4 Proof of the lower bound result from Theorem 8

In this section, we prove the lower bound result from Theorem 8. We first work conditionally to $\mathbb{X}$ so that we can use the concentration inequality of a Lipschitz function of the Gaussian vector $\boldsymbol{\xi}$ (see Eq.(2.35) in [Led05] or Theorem 5.2.2 in [Ver18]): for almost all $\mathbb{X}$, for all $r > 0$, with probability at least $1 - \exp(-r)$, $\phi(\boldsymbol{\xi}) \geq \mathbb{E}_{\boldsymbol{\xi}}\phi(\boldsymbol{\xi}) - \sigma_\xi\|\phi\|_{Lip}\sqrt{2r}$ where $\phi(\boldsymbol{\xi}) = \left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}) - \boldsymbol{\beta}^*)\right\|_2$ and $\|\phi\|_{Lip}$ is the Lipschitz constant of $\phi$ with respect to the Euclidean

norm. Moreover, thanks to the concentration of Lipschitz functions of Gaussian vectors recalled above we have: for almost all $\mathbb{X}$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\phi(\boldsymbol{\xi})^2 - [\mathbb{E}_{\boldsymbol{\xi}}\phi(\boldsymbol{\xi})]^2 &= \mathbb{E}_{\boldsymbol{\xi}}\left[(\phi(\boldsymbol{\xi}) - \mathbb{E}_{\boldsymbol{\xi}}\phi(\boldsymbol{\xi}))^2\right] \\ &= \int_0^\infty \mathbb{P}_{\boldsymbol{\xi}}\left[|\phi(\boldsymbol{\xi}) - \mathbb{E}_{\boldsymbol{\xi}}\phi(\boldsymbol{\xi})| \geq \sqrt{r}\right] dr \leq 2\sigma_\xi^2 \|\phi\|_{Lip}^2 .\end{aligned}$$

As a consequence, $[\mathbb{E}_{\boldsymbol{\xi}}\phi(\boldsymbol{\xi})]^2 \geq \mathbb{E}_{\boldsymbol{\xi}}[\phi(\boldsymbol{\xi})^2] - 2\sigma_\xi^2 \|\phi\|_{Lip}^2$ and so, for almost all $\mathbb{X}$, with $\mathbb{P}_{\boldsymbol{\xi}}$-probability at least $1-\exp(-r)$,

$$\phi(\boldsymbol{\xi}) \geq \mathbb{E}_{\boldsymbol{\xi}}\phi(\boldsymbol{\xi}) - \sigma_\xi \|\phi\|_{Lip}\sqrt{2r} \geq \sqrt{\frac{\mathbb{E}_{\boldsymbol{\xi}}[\phi(\boldsymbol{\xi})^2]}{2}} - \sigma_\xi \|\phi\|_{Lip}\sqrt{2r}, \tag{3.41}$$

when $\mathbb{E}_{\boldsymbol{\xi}}\phi(\boldsymbol{\xi})^2 \geq 4\sigma_\xi^2 \|\phi\|_{Lip}^2$. We note that (3.41) also holds when $\mathbb{E}_{\boldsymbol{\xi}}\phi(\boldsymbol{\xi})^2 \leq 4\sigma_\xi^2 \|\phi\|_{Lip}^2$ as long as $r \geq 4\sqrt{2}$ since $\phi(\boldsymbol{\xi}) \geq 0$ a.s.. As a consequence, we (always) have for all $r \geq 4\sqrt{2}$,

$$\phi(\boldsymbol{\xi}) \geq \sqrt{\frac{\mathbb{E}_{\boldsymbol{\xi}}[\phi(\boldsymbol{\xi})^2]}{2}} - \sigma_\xi \|\phi\|_{Lip}\sqrt{2r}.$$

Next, thanks to the linearity of the estimator $\hat{\boldsymbol{\beta}}$ we have for all $\xi_1, \xi_2 \in \mathbb{R}^p$, $|\phi(\xi_1) - \phi(\xi_2)| \leq \left\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}(\xi_1 - \xi_2)\right\|_2$ and so $\|\phi\|_{Lip} \leq \|A\|_{op}$ where $A = \Sigma^{1/2}\varphi_t(\hat{\Sigma})[N^{-1}\mathbb{X}^\top]$ and

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}[\phi(\boldsymbol{\xi})^2] = \mathbb{E}_{\boldsymbol{\xi}}\left\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2^2 &= \left\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\right\|_2^2 + \mathbb{E}_{\boldsymbol{\xi}}\left\|\Sigma^{\frac{1}{2}}\hat{\boldsymbol{\beta}}(\boldsymbol{\xi})\right\|_2^2 \\ &= \left\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\right\|_2^2 + \sigma_\xi \operatorname{Tr}[AA^\top].\end{aligned}$$

Finally, we have for almost all $\mathbb{X}$ and all $r \geq 4\sqrt{2}$, with probability at least $1 - \exp(-r)$,

$$\begin{aligned}\left\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \geq \frac{1}{\sqrt{2}}\left\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\right\|_2 + \frac{\sigma_\xi}{\sqrt{2}}\sqrt{\frac{\operatorname{Tr}[\Sigma\varphi_t^2(\hat{\Sigma})\hat{\Sigma}]}{N}} \\ - \sigma_\xi\left\|\Sigma^{1/2}\varphi_t(\hat{\Sigma})\frac{\mathbb{X}^\top}{\sqrt{N}}\right\|_{op}\sqrt{\frac{2r}{N}}.\end{aligned} \tag{3.42}$$

In the next two sections, we obtain lower bounds on the three main terms appearing in the right hand side of (3.42).

### 3.4.1 A lower bound for the bias term $\left\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\right\|_2$

As before, we decompose the feature space as $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ where $J = J_*$ is the optimal decomposition, so that the bias term can be decomposed as

$$\left\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\right\|_2^2 = \left\|\Sigma_J^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\right\|_2^2 + \left\|\Sigma_{J^c}^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\right\|_2^2.$$

**A lower bound for the bias term on $V_J$**

In Section 3.3.2, we introduced $\tilde{\boldsymbol{\beta}} = \varphi_t(\Sigma)\Sigma\boldsymbol{\beta}^*$ and proved in (3.20) that

$$\left\|\Sigma_J^{1/2}(\tilde{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\right\|_2 = \left\|\Sigma_J^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2$$

and in (3.27) that, for some absolute constant $C > 0$, on $\Omega_t$,

$$\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}_J^*) - \tilde{\boldsymbol{\beta}}_J)\right\|_2 \leq C\square\left(\log(et)\left\|\Sigma_J^{1/2}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2 + t^{-1}\left\|\Sigma_J^{1/2}\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2\right).$$

Next, it follows from Assumption 5 that $\varphi_t(\Sigma) \preceq C_1\Sigma_t^{-1}$ and so $\left\|\Sigma_J^{1/2}\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2 \le C_1\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}_J^*\right\|_2$. As a consequence, as long as $\square\log(e^2t) \lesssim 1$, the following lower bound holds on $\Omega_t$:

$$\begin{aligned}
&\|\Sigma_J^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\|_2 \ge \|\Sigma_J^{\frac{1}{2}}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2 - \|\Sigma_J^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \tilde{\boldsymbol{\beta}})\|_2 \\
&\ge \|\Sigma_J^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\|_2 - C\square\left(\log(et)\|\Sigma_J^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\|_2 + t^{-1}\left\|\Sigma_J^{1/2}\varphi_t(\Sigma)\boldsymbol{\beta}_J^*\right\|_2\right) \\
&\ge \left(1 - C\square\log(e^2t)\right)\|\Sigma_J^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\|_2 - \frac{CC_1\square}{t}\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}_J^*\right\|_2 \\
&\ge \frac{1}{2}\|\Sigma_J^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}^*\|_2 - \frac{CC_1\square}{t}\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}_J^*\right\|_2.
\end{aligned}$$

**A lower bound for the bias on $V_{J^c}$**

We have

$$\left\|\Sigma_{J^c}^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\right\|_2 \ge \left\|\Sigma_{J^c}^{\frac{1}{2}}\boldsymbol{\beta}_{J^c}^*\right\|_2 - \left\|\Sigma_{J^c}^{\frac{1}{2}}\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*)\right\|_2$$

and using that $\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) = \hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}_J^*) + \hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}_{J^c}^*) = \varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}_J^* + \varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}_{J^c}^*$ we get

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

In (3.32), we proved that on $\Omega_t$,

$$\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}_J^*\right\|_2 \lesssim \frac{\square}{t}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}_J^*\right\|_2.$$

Next, it follows from (3.36) that with probability at least $1 - 2\exp(-cN)$

$$\left\|\Sigma_{J^c}^{1/2}\varphi_t(\hat{\Sigma})\hat{\Sigma}\boldsymbol{\beta}_{J^c}^*\right\|_2 \le Cb\left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2$$

and so when $b \le 1/(2C)$, we obtain

$$\left\|\Sigma^{\frac{1}{2}}(\hat{\boldsymbol{\beta}}(\mathbb{X}\boldsymbol{\beta}^*) - \boldsymbol{\beta}^*)\right\|_2 \ge \frac{1}{2}\|\Sigma_J^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\|_2 + \frac{1}{2}\left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 - \frac{C\square}{t}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}_J^*\right\|_2. \tag{3.43}$$

### 3.4.2 Lower bound for the conditional variance term $\mathbb{E}_{\boldsymbol{\xi}}\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}(\boldsymbol{\xi})\|_2^2$

In this section, we obtain a lower bound on the conditional (with respect to $\mathbb{X}$) variance of $\hat{\boldsymbol{\beta}}$: $\mathbb{E}_{\boldsymbol{\xi}}\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}(\boldsymbol{\xi})\|_2^2$. It follows from Assumption 5 that for all $t \ge 1$ and $x \in [0, 8]$ , we have

$$\varphi_t(x) \ge \frac{c_{12}}{x + t^{-1}} := c_{12}\varphi_t^{(\mathrm{Ridge})}(x) \tag{3.44}$$

where we recall (see (3.3)) that $\varphi_t^{(\mathrm{Ridge})}(x) = (x + t^{-1})^{-1}$ is the filter function of ridge regression with regularization parameter $t^{-1}$.

**Lemma 12.** *Grant Assumption 6 and assume that $X$ has independent and centered coordinates with respect to the basis $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_p\}$. Let $\hat{\boldsymbol{\beta}}$ be a spectral algorithm defined in Definition 17 with filter function $\varphi_t$ satisfying (3.44). Then, there exists absolute constants $c, c_{13} > 0$ such that with probability at least $1 - c\exp(-N/c) - \mathbb{P}[\Omega_t^c]$,*

$$\sigma_\xi^2\frac{\mathrm{Tr}[\Sigma\varphi_t^2(\hat{\Sigma})\hat{\Sigma}]}{N} = \mathbb{E}_{\boldsymbol{\xi}}\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}(\boldsymbol{\xi})\|_2^2 \ge c_{13}c_{12}\sigma_\xi^2\left(\frac{|J|}{N} + t^2\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}\right).$$

*Proof.* Let $\sum_{j=1}^p \hat{\sigma}_i^{\frac{1}{2}}\hat{\boldsymbol{u}}_i \otimes \hat{\boldsymbol{e}}_i$ be the singular value decomposition of $\frac{1}{\sqrt{N}}\mathbb{X}$, where $\hat{\sigma}_j = 0$ if $j > N$, $\{\hat{\boldsymbol{u}}_i\}_{i=1}^N$ is an orthonormal basis of $\mathbb{R}^N$ and $\{\hat{\boldsymbol{e}}_j\}_{j=1}^p$ is an orthonormal basis of $\mathbb{R}^p$. It follows from (3.1) that

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\xi}) = \frac{1}{N}\varphi(\hat{\Sigma})\mathbb{X}^\top\boldsymbol{\xi} = \frac{1}{\sqrt{N}}\varphi_t(\hat{\Sigma})\sum_{i=1}^N \hat{\boldsymbol{e}}_i\sqrt{\hat{\sigma}_i}\langle\hat{\boldsymbol{u}}_i, \boldsymbol{\xi}\rangle = \frac{1}{\sqrt{N}}\sum_{i=1}^N \sqrt{\hat{\sigma}_i}\varphi_t(\hat{\sigma}_i)\langle\hat{\boldsymbol{u}}_i, \boldsymbol{\xi}\rangle\hat{\boldsymbol{e}}_i$$

and by taking $\|\Sigma^{1/2}\cdot\|_2^2$, we obtain

$$\left\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}(\boldsymbol{\xi})\right\|_2^2 = \frac{1}{N}\left\|\sum_{i=1}^N \sqrt{\hat{\sigma}_i}\varphi_t(\hat{\sigma}_i)\langle\hat{\boldsymbol{u}}_i,\boldsymbol{\xi}\rangle\Sigma^{1/2}\hat{\boldsymbol{e}}_i\right\|_2^2$$
$$= \frac{1}{N}\sum_{i,j=1}^N \sqrt{\hat{\sigma}_i\hat{\sigma}_j}\varphi_t(\hat{\sigma}_i)\varphi_t(\hat{\sigma}_j)\langle\hat{\boldsymbol{u}}_i,\boldsymbol{\xi}\rangle\langle\hat{\boldsymbol{u}}_j,\boldsymbol{\xi}\rangle\langle\Sigma^{1/2}\hat{\boldsymbol{e}}_i,\Sigma^{1/2}\hat{\boldsymbol{e}}_j\rangle.$$

Taking expectation with respect to $\boldsymbol{\xi}$ and using that $\mathbb{E}_{\boldsymbol{\xi}}[\langle\hat{\boldsymbol{u}}_i,\boldsymbol{\xi}\rangle\langle\hat{\boldsymbol{u}}_j,\boldsymbol{\xi}\rangle] = \sigma_\xi^2\langle\hat{\boldsymbol{u}}_i,\hat{\boldsymbol{u}}_j\rangle = \sigma_\xi^2\mathbb{1}_{\{i=j\}}$, we obtain that for almost all $\mathbb{X}$,

$$\mathbb{E}_{\boldsymbol{\xi}}\left\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}(\boldsymbol{\xi})\right\|_2^2 = \frac{\sigma_\xi^2}{N}\sum_{i=1}^N \hat{\sigma}_i\varphi_t^2(\hat{\sigma}_i)\left\|\Sigma^{1/2}\hat{\boldsymbol{e}}_i\right\|_2^2.$$

The latter result is actually true for any filter function. By applying it to the filter function from ridge regression and using (3.44), we have on the event $\Omega_t$ (where we know, thanks to Lemma 9, that the spectrum of $\hat{\Sigma}$ is in $[0,8]$ because $\sigma_1, t^{-1}\le 1$) that

$$\mathbb{E}_{\boldsymbol{\xi}}\left\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}(\boldsymbol{\xi})\right\|_2^2 \ge c_{12}^2\frac{\sigma_\xi^2}{N}\sum_{i=1}^N \hat{\sigma}_i\big(\varphi_t^{(\mathrm{Ridge})}(\hat{\sigma}_i)\big)^2\|\Sigma^{1/2}\hat{\boldsymbol{e}}_i\|_2^2 = c_{12}^2\mathbb{E}_{\boldsymbol{\xi}}\left\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}^{(Ridge)}(\boldsymbol{\xi})\right\|_2^2.$$

Finally, by [TB23, Lemma 7 and Theorem 2], there exists an absolute constant $0 < c_{14} < 1$ such that with probability at least $1 - c\exp(-N/c)$,

$$\mathbb{E}_{\boldsymbol{\xi}}\left\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}^{(\mathrm{Ridge})}(\boldsymbol{\xi})\right\|_2^2 \ge c_{14}\sigma_\xi^2\left(\frac{|J|}{N} + \frac{N\,\mathrm{Tr}(\Sigma_{J^c}^2)}{(Nt^{-1}+\mathrm{Tr}(\Sigma_{J^c}))^2}\right).$$

Lemma 12 then follows since $\mathrm{Tr}(\Sigma_{J^c}) \lesssim \square t^{-1}N \lesssim t^{-1}N$ thanks to the sampling complexity assumption (see the discussion below (3.15)). ∎

### 3.4.3 An upper bound for the weak variance term and the conclusion

In this section, we provide a high probability upper bound on the weak variance term coming from Borell's inequality in (3.42) i.e. $\sigma_\xi\left\|\Sigma^{1/2}\varphi_t(\hat{\Sigma})(\mathbb{X}^\top/\sqrt{N})\right\|_{op}$. It follows from (3.11) and Lemma 11 that, on the event $\Omega_t$, we have

$$\left\|\Sigma^{1/2}\varphi_t(\hat{\Sigma})(\mathbb{X}^\top/\sqrt{N})\right\|_{op} \le \left\|\Sigma^{1/2}\Sigma_t^{-1/2}\right\|_{\mathrm{op}}\left\|\Sigma_t^{1/2}\hat{\Sigma}_t^{-1/2}\right\|_{\mathrm{op}}\left\|\hat{\Sigma}_t\varphi_t(\hat{\Sigma})^2\hat{\Sigma}\right\|_{op}^{1/2} \lesssim 1 \tag{3.45}$$

where we used Assumption 5 to get $\hat{\Sigma}_t\varphi_t(\hat{\Sigma})^2\hat{\Sigma} \preceq C_1\hat{\Sigma}_t\hat{\Sigma}_t^{-2}\hat{\Sigma} \preceq C_1 I_p$.

Finally, plugging (3.45) and (3.43) together with Lemma 12 in (3.42), we get that for all $r \ge 4\sqrt{2}$, with probability at least $1 - \exp(-r) - c\exp(-N/c) - \mathbb{P}[\Omega_t^c]$,

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right\|_2 \ge \left(\|\Sigma_J^{\frac{1}{2}}\psi_t(\Sigma)\boldsymbol{\beta}_J^*\|_2 + \frac{1}{2}\left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2 - \frac{C\square}{t}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}_J^*\right\|_2\right)$$
$$+ c_{13}\left(\sigma_\xi\sqrt{\frac{|J|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}}\right) - c_0\sigma_{\boldsymbol{\xi}}\sqrt{\frac{r}{N}}$$
$$\ge cr(V_J, V_{J^c}) - \frac{C\square}{t}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}_J^*\right\|_2 - c_0\sigma_{\boldsymbol{\xi}}\sqrt{\frac{r}{N}}$$

as long as $b \lesssim 1$. Finally, the result follows by taking $r \sim k^*$ in the inequality above.

## 3.5 Auxiliaries results

We start with some results on the concentration of sum of independent sub-exponential variables. We first start with the definition of $\psi$-norm (see for instance, Chapter 1 in [CGLP12]). Let $\psi$ be an Orlicz function. We define the Orlicz norm of a random variable $Z$ as

$$\|Z\|_{\psi} = \inf\left(c : \mathbb{E}\psi(|Z|/c) \leq \psi(1)\right).$$

Orlicz functions that are of particular interest to us are, for all $\alpha \geq 1$, $\psi_\alpha : t \geq 0 \to \exp(t^\alpha) - 1$. It follows from Theorem 1.1.5 in [CGLP12] that for all $\alpha \geq 1$, there is equivalence between:

(a) there is a constant $K_1 > 0$ such that $\|Z\|_{\psi_\alpha} \leq K_1$

(b) there is a constant $K_2 > 0$ such that for all $p \geq \alpha$, $\|Z\|_{L_p} \leq K_2 p^{1/\alpha}$

(c) there exists $K_3, K_3'$ such that for all $t \geq K_3'$,

$$\mathbb{P}(|Z| \leq t) \geq 1 - \exp(-t^\alpha / K_3^\alpha).$$

Moreover, $K_2 \leq 2eK_1$, $K_3 \leq eK_2$, $K_3' \leq e^2 K_2$ and $K_1 \leq 2\max(K_3, K_3')$. It follows from these equivalence that

$$\|Z\|_{\psi_\alpha} \sim \sup_{p \geq \alpha} \frac{\|Z\|_{L_p}}{p^{1/\alpha}}.$$

In particular, if $X$ is a sub-gaussian vector as defined in Assumption 6 then there exists some absolute constant $C > 0$ such that for all $\boldsymbol{v} \in \mathbb{R}^p$, $\left\|\langle X, \boldsymbol{v}\rangle\right\|_{\psi_2} \leq C \left\|\Sigma^{\frac{1}{2}}\boldsymbol{v}\right\|_2$. It is also clear from the definition of the $\psi_1$ and $\psi_2$ norm that for all $\boldsymbol{v}$ we have $\left\|\langle X, \boldsymbol{v}\rangle^2\right\|_{\psi_1} = \left\|\langle X, \boldsymbol{v}\rangle\right\|_{\psi_2}^2 \leq C^2 \left\|\Sigma^{\frac{1}{2}}\boldsymbol{v}\right\|_2^2$. Finally, the last tool we need is Bernstein's inequality for the sum of independent $\psi_1$ variable (see for instance Theorem 1.2.7 in [CGLP12]): if $Z_1, \ldots, Z_N$ are independent $\psi_1$ random variables then for all $t \geq 1$, with probability at least $1 - \exp(-ct)$,

$$\left|\frac{1}{N}\sum_{i=1}^N Z_i - \mathbb{E}Z_i\right| \leq \sigma_1\sqrt{\frac{t}{N}} + M_1\frac{t}{N}$$

where $M_1 = \max_{1\leq i\leq N}\|Z_i - \mathbb{E}Z_i\|_{\psi_1}$ and $\sigma_1^2 = (1/N)\sum_{i=1}^N \|Z_i - \mathbb{E}Z_i\|_{\psi_1}^2$. In particular, if we apply this result for $Z_i = \langle X_i, \boldsymbol{v}\rangle^2$ (which is a $\psi_1$ random variable according to the argument above), we get that with probability at least $1 - \exp(-ct)$,

$$\left|\frac{1}{N}\sum_{i=1}^N \langle X_i, \boldsymbol{v}\rangle^2 - \mathbb{E}\langle X_i, \boldsymbol{v}\rangle^2\right| \leq \left\|\langle X, \boldsymbol{v}\rangle^2 - \mathbb{E}\langle X, \boldsymbol{v}\rangle^2\right\|_{\psi_1}\left(\sqrt{\frac{t}{N}} + \frac{t}{N}\right) \tag{3.46}$$

and

$$\begin{aligned}\left\|\langle X, \boldsymbol{v}\rangle^2 - \mathbb{E}\langle X, \boldsymbol{v}\rangle^2\right\|_{\psi_1} &\leq \left\|\langle X, \boldsymbol{v}\rangle^2\right\|_{\psi_1} + \left\|\mathbb{E}\langle X, \boldsymbol{v}\rangle^2\right\|_{\psi_1} \\ &\leq \left\|\langle X, \boldsymbol{v}\rangle\right\|_{\psi_2}^2 + \|1\|_{\psi_1}\mathbb{E}\langle X, \boldsymbol{v}\rangle^2 \leq C\left\|\Sigma^{\frac{1}{2}}\boldsymbol{v}\right\|_2^2\end{aligned}$$

where we used the subgaussian property of $X$. As a consequence, we proved the following result.

**Lemma 13.** *There is some absolute constant $c > 0$ such that the following holds. Let $X$ be a sub-gaussian vector in $\mathbb{R}^p$ and denote $\Sigma = \mathbb{E}XX^\top$ ($X$ is not necessarily centered). Let $\boldsymbol{v} \in \mathbb{R}^p$. With probability at least $1 - \exp(-cN)$, we have*

$$\frac{1}{2}\left\|\Sigma^{\frac{1}{2}}\boldsymbol{v}\right\|_2^2 \leq \frac{1}{N}\sum_{i=1}^N \langle X_i, \boldsymbol{v}\rangle^2 \leq \frac{3}{2}\left\|\Sigma^{\frac{1}{2}}\boldsymbol{v}\right\|_2^2. \tag{3.47}$$

Next we use the classical generic chaining bound for sub-gaussian processes that follows from Theorem 2.2.27 in [Tal96]. Note that the following result requires less assumptions than the one required in Hanson-Wright inequality from Theorem 6.2.1 in [Ver18].

**Lemma 14.** *There is an absolute constant $c > 0$ such that the following holds. Let $X$ be a sub-gaussian vector in $\mathbb{R}^p$ and denote $\Sigma = \mathbb{E}XX^\top$ ($X$ is not necessarily centered). Let $A$ be a matrix in $\mathbb{R}^{p\times d}$. We have for all $t > 0$, with probability at least $1 - \exp(-t)$,*

$$\|AX\|_2 \leq c\left(\left\|\Sigma^{1/2}A^\top\right\|_{HS} + \left\|\Sigma^{1/2}A^\top\right\|_{op}\sqrt{t}\right).$$

*We also have*

$$\mathbb{E}\,\|AX\|_2^2 = \left\|\Sigma^{1/2}A^\top\right\|_{HS}^2.$$

*Proof.* We first note that

$$\|AX\|_2 \leq \|A(X-\mathbb{E}X)\|_2 + \|A\mathbb{E}X\|_2.$$

Then, we write $\|A(X-\mathbb{E}X)\|_2$ as the supremum of a centered sub-gaussian process: $\|A(X-\mathbb{E}X)\|_2 = \sup(Z_x : x \in A^\top B_2^d)$ where $Z_x = \langle X - \mathbb{E}X, x\rangle$. The canonical metric associated with this process is $(u,v) \to \left(\mathbb{E}(Z_u - Z_v)^2\right)^{1/2} = \left\|\Sigma_0^{1/2}(u-v)\right\|_2$ where $\Sigma_0 = \mathbb{E}\left[(X-\mathbb{E}X)(X-\mathbb{E}X)^\top\right]$. It follows from Theorem 2.2.27 in [Tal96], that for all $t > 0$, with probability at least $1-\exp(-t)$, $\|A(X-\mathbb{E}X)\|_2 \lesssim \gamma_2 + \sqrt{t}D$ where $\gamma_2 = \gamma_2(\Sigma_0^{1/2}A^\top B_2^d, \ell_2^p)$ is Talagrand's $\gamma_2$-functional and $D$ is the diameter of $\Sigma_0^{1/2}A^\top B_2^d$ with respect to $\ell_2^p$. It follows from Talagrand's majorizing measure that

$$\gamma_2(\Sigma_0^{1/2}A^\top B_2^d, \ell_2^p) \lesssim \mathbb{E}\left\|\Sigma_0^{1/2}A^\top G\right\|_2 \lesssim \mathrm{Tr}[A\Sigma_0 A^\top]^{1/2} = \left\|\Sigma_0^{1/2}A^\top\right\|_{HS}$$

and $D = \left\|\Sigma_0^{1/2}A^\top\right\|_{op}$. We conclude the proof of the exponential bound by using that $\|A\mathbb{E}X\|_2 + \left\|\Sigma_0^{1/2}A^\top\right\|_{HS} \lesssim \left\|\Sigma^{1/2}A^\top\right\|_{HS}$. The result in expectation follows from

$$\mathbb{E}\,\|AX\|_2^2 = \mathrm{Tr}[\mathbb{E}[AXX^\top A^\top]] = \left\|\Sigma^{1/2}A^\top\right\|_{HS}^2.$$

■

Under the same assumptions as in Lemma 14, we get that for all $t \geq \left\|\Sigma^{1/2}A^\top\right\|_{HS}$ with probability at least $1 - \exp\left(-ct^2/\left\|\Sigma^{1/2}A^\top\right\|_{op}^2\right)$, $\|AX\|_2 \leq t$. The latter statement coincides with point *(c)* above for $\alpha = 2$, $K_3 \sim \left\|\Sigma^{1/2}A^\top\right\|_{op}$ and $K_3' \sim \left\|\Sigma^{1/2}A^\top\right\|_{HS}$ meaning that $\|AX\|_2$ is a subgaussian variable with subgaussian norm satisfying

$$\left\|\|AX\|_2\right\|_{\psi_2} \lesssim \left\|\Sigma^{1/2}A^\top\right\|_{HS}.$$

This follows from the equivalence between *(a)* and *(c)* above. As a consequence,

$$\left\|\|AX\|_2^2\right\|_{\psi_1} \lesssim \left\|\Sigma^{1/2}A^\top\right\|_{HS}^2 = \mathbb{E}\,\|AX\|_2^2,$$

and so it follows from (3.46) that for all $t > 0$, with probability at least $1-\exp(-ct)$,

$$\left|\frac{1}{N}\sum_{i=1}^N \|AX_i\|_2^2 - \mathbb{E}\,\|AX\|_2^2\right| \leq c\mathbb{E}\,\|AX\|_2^2\left(\sqrt{\frac{t}{N}} + \frac{t}{N}\right).$$

(Note that this results holds even if $X$ is not centered and does not have independent coordinates unlike the Hansen-Wright inequality from Theorem 6.2.1 in [Ver18]). For $t \sim N$ we just proved the following result.

**Lemma 15.** *There exists an absolute constant $c > 0$ such that the following holds. Let $X$ be a sub-gaussian vector in $\mathbb{R}^p$ and denote $\Sigma = \mathbb{E}XX^\top$ ($X$ is not necessarily centered). Let $A$ be a matrix in $\mathbb{R}^{p\times d}$. With probability at least $1-\exp(-cN)$,*

$$\frac{1}{2}\left\|\Sigma^{1/2}A^\top\right\|_{HS}^2 \leq \frac{1}{N}\sum_{i=1}^N \|AX_i\|_2^2 \leq \frac{3}{2}\left\|\Sigma^{1/2}A^\top\right\|_{HS}^2.$$

### 3.5.1 Proof of Corollary 4

By the proof of Proposition 7 of [P2], if $t^{-1} \sim N^{-\frac{\alpha}{1+s\alpha}}$, regardless of the relationship between $s$ and 2, we always have

$$\sigma_\xi^2 \frac{|J_*|}{N} + \sigma_\xi^2 \frac{N \operatorname{Tr}(\Sigma_{J_*^c}^2)}{(Nt^{-1})^2} \sim \sigma_\xi^2 N^{-\frac{\alpha s}{1+\alpha s}}, \quad \text{and} \quad \|\Sigma_{J_*^c}^{\frac{1}{2}} \boldsymbol{\beta}_{J_*^c}^*\|_2^2 \sim N^{-\frac{\alpha s}{1+\alpha s}}.$$

The difference is that for ridge, $\psi_t^{(B)}(x) = \frac{1}{xt+1}$, hence by the proof of Proposition 7 of [P2],

$$\left\| \Sigma_{J_*}^{\frac{1}{2}} \psi_t^{(B)}(\Sigma) \boldsymbol{\beta}_{J_*}^* \right\|_2^2 \sim N^{-\frac{\alpha \tilde{s}}{1+\alpha \tilde{s}}}.$$

On the other hand, by Definition 6, item *2.*, we have

$$\begin{aligned} \left\| \Sigma_{J_*}^{\frac{1}{2}} \psi_t^{(A)}(\Sigma) \boldsymbol{\beta}_{J_*}^* \right\|_2^2 &= \sum_{j \le k^*_{t^{-1},b}} \sigma_j^s (\psi_t^{(A)}(\sigma_j))^2 \sigma_j^{1-s} \langle \boldsymbol{\beta}^*, \boldsymbol{e}_j \rangle^2 \\ &\le C_{28}^2 t^{-s} \left\| \Sigma_{J_*}^{\frac{1-s}{2}} \boldsymbol{\beta}_{J_*}^* \right\|_2^2 \lesssim N^{-\frac{\alpha s}{1+\alpha s}}. \end{aligned}$$

As the choice of $t$ is optimal over the class $\mathfrak{R}_{\mathrm{Sob}}(s,\alpha)$ (see, for instance, [LGSL24]), we conclude that $\{\varphi^{(A)}\}_{t\ge 1} \preceq_{\mathcal{R}} \{\varphi^{(B)}\}_{t\ge 1}$ for any $\mathcal{R} \in \mathfrak{R}_{\mathrm{Sob}}(s,\alpha)$.

### 3.5.2 Proof of Corollary 5

For any $t$ in the interval $I = \{t : b^{-1}\varepsilon \le t^{-1} < \sigma\}$, it is easy to verify that $k^*_{t^{-1},b} = k$. Moreover, since we have assumed that for any $1 \le j \le k$, there holds $|\langle \boldsymbol{\beta}^*, \boldsymbol{e}_j \rangle| = \alpha_*$, and for any $j > k$, $\langle \boldsymbol{\beta}^*, \boldsymbol{e}_j \rangle = 0$, we have $\|\Sigma_{J_*^c}^{1/2} \boldsymbol{\beta}_{J_*^c}^*\|_2 = 0$. Moreover, $\|\Sigma_{J_*}^{1/2} \psi_t(\Sigma) \boldsymbol{\beta}_{J_*}^*\|_2 = (\sum_{j\le k} \sigma \psi_t^2(\sigma) \alpha_*^2)^{1/2} = \alpha_* \psi_t(\sigma) \sqrt{k\sigma}$, and $\sigma_\xi t \sqrt{\operatorname{Tr}(\Sigma_{J_*^c}^2)/N} = \sigma_\xi \varepsilon t \sqrt{(p-k)/N}$. We compute that

$$\mathrm{SNR} = \frac{\alpha_*}{\sigma_\xi} \frac{\sigma^{3/2}}{\varepsilon} \sqrt{\frac{kN}{p-k}}.$$

1. When $\psi_t(x) = \psi_t^{(\mathrm{Ridge})}(x) = \frac{1}{xt+1}$. Then

$$\min_{t\in I} r^{(\mathrm{Ridge})}(V_{J_*}, V_{J_*^c}) = \sigma_\xi \sqrt{\frac{k}{N}} + \min_{t\in I} \left( \sigma_\xi \varepsilon t \sqrt{\frac{p-k}{N}} + \alpha_* \frac{\sqrt{k\sigma}}{\sigma t + 1} \right).$$

   Under the assumption that

$$4 < \frac{\alpha_*}{\sigma_\xi} \frac{\sigma^{3/2}}{\varepsilon} \sqrt{\frac{kN}{p-k}} < \frac{\sigma}{\varepsilon} b \le \left(1 + \frac{\sigma}{\varepsilon} b\right)^2,$$

   the minimum is given by

$$\min_{t\in I} r^{(\mathrm{Ridge})}(V_{J_*}, V_{J_*^c}) = \sigma_\xi \sqrt{\frac{k}{N}} + \frac{\sigma_\xi}{\sigma} \varepsilon \sqrt{\frac{p-k}{N}} \left(2\sqrt{\mathrm{SNR}} - 1\right). \tag{3.48}$$

2. When $\psi_t(x) = \psi_t^{(\mathrm{GF})}(x) = \exp(-tx)$. Then

$$\min_{t\in I} r^{(\mathrm{GF})}(V_{J_*}, V_{J_*^c}) = \sigma_\xi \sqrt{\frac{k}{N}} + \min_{t\in I} \left( \sigma_\xi \varepsilon t \sqrt{\frac{p-k}{N}} + \alpha_* \sqrt{k\sigma} \exp(-t\sigma) \right).$$

   Under the assumption that

$$e < \frac{\alpha_*}{\sigma_\xi} \frac{\sigma^{3/2}}{\varepsilon} \sqrt{\frac{kN}{p-k}} < \frac{\sigma}{\varepsilon} b \le \exp\left(\frac{\sigma}{\varepsilon} b\right),$$

the minimum is given by

$$\min_{t\in I} r^{(\mathrm{GF})}(V_{J_*}, V_{J_*^c}) = \sigma_\xi\sqrt{\frac{k}{N}} + \frac{\sigma_\xi}{\sigma}\varepsilon\sqrt{\frac{p-k}{N}}\,(1+\log(\mathrm{SNR}))\,. \tag{3.49}$$

Combining (3.48) and (3.49) and using the fact that $1+\log(\mathrm{SNR}) \le 2\sqrt{\mathrm{SNR}} - 1$ for any $\mathrm{SNR} \ge 1$, we know that

$$\min_{t\in I} r^{(\mathrm{GF})}(V_{J_*}, V_{J_*^c}) \le \min_{t\in I} r^{(\mathrm{Ridge})}(V_{J_*}, V_{J_*^c}).$$

Moreover, when $R \to \infty$, $\{\varphi_t^{(\mathrm{Ridge})}\}_{t\in I} \prec_{\mathcal{R}} \{\varphi_t^{(\mathrm{GF})}\}_{t\in I}$.

### 3.5.3 Definition of the contour $\mathcal{C}_t$ and proof of Lemma 10

In this section, we construct the family of contours $(\mathcal{C}_t)_{t\ge 1}$ used in the formulae (3.18). This formulae follows from the residue theorem, but, in order to apply this theorem, we need the contour $\mathcal{C}_t$ to surround both spectra of $\Sigma$ and $\hat{\Sigma}$. By definition, the spectrum of $\Sigma$ lies in $[0, \sigma_1]$ and the one of $\hat{\Sigma}$ lies in $[0, \hat{\sigma}_1]$. Moreover, thanks to Lemma 9, we know that on $\Omega_t$, we have $\hat{\sigma}_1 \le 4(\sigma_1 + t^{-1})$. As a consequence, formulae (3.18) is valid on $\Omega_t$ if we construct a contour $\mathcal{C}_t$ in such a way that it contains $[0, 4(\sigma_1 + t^{-1})]$. Moreover, we also need to choose $\mathcal{C}_t$ so that Lemma 10 and 11 hold on $\Omega_t$.

We follow [LGSL24] for the construction of such a contour: for all $t \ge 1$, define $\mathcal{C}_t = \mathcal{C}_{t,1} \cup \mathcal{C}_{t,2} \cup \mathcal{C}_{t,3}$ where $\mathcal{C}_{t,k}, k = 1,2,3$ are defined now. We let $L : x \in \mathbb{R} \to \alpha x + \beta$, where

$$\alpha = \frac{5(\sigma_1 + t^{-1})}{\sigma_1 + t^{-1}/2}, \text{ and } \beta = \frac{\alpha}{2t}.$$

Note that $L(-1/(2t)) = 0$ and $L(\sigma_1) = 5(\sigma_1 + t^{-1})$ so that by setting

$$\begin{aligned}
\mathcal{C}_{t,1} &= \{x + L(x)i : x \in [-1/(2t), \sigma_1]\}\,,\\
\mathcal{C}_{t,2} &= \{x - L(x)i : x \in [-1/(2t), \sigma_1]\}\,,\\
\mathcal{C}_{t,3} &= \left\{z \in \mathbb{C} : |z - \sigma_1| = 5(\sigma_1 + t^{-1}), \mathrm{Re}(z) \ge \sigma_1\right\},
\end{aligned} \tag{3.50}$$

the union $\cup_{k=1,2,3}\mathcal{C}_{t,k}$ is well defining a contour in $\mathbb{C}$; this is the one we call $\mathcal{C}_t$ depicted in Figure 3.1.

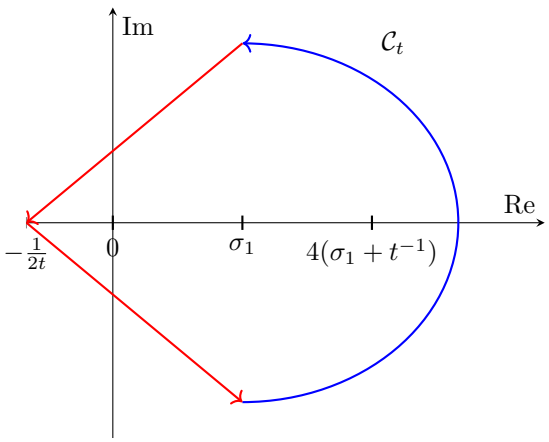


Figure 3.1: The contour $\mathcal{C}_t$ defined in (3.50) surrounds both spectra of $\Sigma$ and of $\hat{\Sigma}$ on $\Omega_t$ since, on that event, $\hat{\sigma}_1 \le 4(\sigma_1 + t^{-1})$ thanks to Lemma 9.

#### Proof of Lemma 10

*Proof.* Let $z \in \mathcal{C}_t$. We first show that $\left\|\Sigma_t^{\frac{1}{2}}\left(\hat{\Sigma} - zI_p\right)^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{\mathrm{op}} \le 3C$. To that end, we first bound $\|\hat{\Sigma}_t^{\frac{1}{2}}(\hat{\Sigma} - zI_p)^{-1}\hat{\Sigma}_t^{\frac{1}{2}}\|_{\mathrm{op}}$ from above and then we will conclude using Lemma 11. Using SVD, we have

$$\left\|\hat{\Sigma}_t^{\frac{1}{2}}(\hat{\Sigma} - zI_p)^{-1}\hat{\Sigma}_t^{\frac{1}{2}}\right\|_{\mathrm{op}} = \sup_{\sigma\in\sigma(\hat{\Sigma})}\left|\frac{\sigma + t^{-1}}{\sigma - z}\right|$$

where $\sigma(\hat{\Sigma})$ denotes the spectrum of $\hat{\Sigma}$. We recall that $\hat{\sigma}_1$ denotes the largest singular values of $\hat{\Sigma}_1$ so that $\sigma(\hat{\Sigma}) \subset [0, \hat{\sigma}_1]$. Moreover, by Lemma 9, $\hat{\sigma}_1 < 4(\sigma_1 + t^{-1})$ on $\Omega_t$. As a consequence, on $\Omega_t$,

$$\left\| \hat{\Sigma}_t^{\frac{1}{2}} (\hat{\Sigma} - zI_p)^{-1} \hat{\Sigma}_t^{\frac{1}{2}} \right\|_{\text{op}} \leq \sup_{0 \leq \sigma \leq 4(\sigma_1 + t^{-1})} \left| \frac{\sigma + t^{-1}}{\sigma - z} \right|.$$

We are now considering two cases: either $z$ belongs to the 'linear' section $\mathcal{C}_{t,1} \cup \mathcal{C}_{t,2}$ of the contour $\mathcal{C}_t$ or to the semi-circle section $\mathcal{C}_{t,3}$, see the definitions in (3.50). We start with the linear section.

**First case, when** $z = x \pm L(x)i \in \mathcal{C}_{t,1} \cup \mathcal{C}_{t,2}$, where $x \in [-1/(2t), \sigma_1]$, we get

$$\sup_{\sigma \in \sigma(\hat{\Sigma})} \left| \frac{\sigma + t^{-1}}{\sigma - z} \right|^2 \leq \sup_{\sigma \geq 0} \left| \frac{\sigma + t^{-1}}{\sigma - z} \right|^2.$$

Let $y = \sigma + t^{-1}$, $B = x + t^{-1}$, and $C = B^2 + L(x)^2$. Then $|\sigma - z|^2 = (\sigma - x)^2 + L(x)^2 = (y - B)^2 + C - B^2$, thus

$$\left| \frac{\sigma + t^{-1}}{\sigma - z} \right|^2 = \frac{y^2}{y^2 - 2By + C}.$$

The function $y \mapsto \frac{y^2}{y^2 - 2By + C}$ is maximized at $y = \max\{\frac{C}{B}, t^{-1}\}$. Therefore when $\frac{C}{B} > t^{-1}$, we have the maximum $\frac{C}{C - B^2}$, otherwise we have the maximum when $y = t^{-1}$, when $\sigma = 0$. Solving $t^{-1} = \frac{C}{B}$ gives $x_0 = -\frac{1}{2t} + \frac{1}{2t\sqrt{1+\alpha^2}}$.

- If $\frac{C}{B} > t^{-1}$, combined with $x > -\frac{1}{2t}$, we have $x > -\frac{2-\sqrt{2}}{4} t^{-1}$, and the maximum is given by $\frac{C}{C - B^2} = 1 + \frac{(x + t^{-1})^2}{\alpha^2 (x + \frac{1}{2t})^2}$. Let $\delta = tx$, then

$$\sup_{\sigma \geq 0} \left| \frac{\sigma + t^{-1}}{\sigma - z} \right|^2 = \sup \left( 1 + \frac{1}{\alpha^2} \frac{(\delta + 1)^2}{(\delta + \frac{1}{2})^2} : -\frac{1}{2} \leq \delta \leq t\sigma_1 \right).$$

  One may show that the maximum is achieved when $\delta = tx_0$, and

$$\sup_{\sigma \geq 0} \left| \frac{\sigma + t^{-1}}{\sigma - z} \right|^2 = 2 + \frac{2}{\alpha^2} \left( 1 + \sqrt{1 + \alpha^2} \right), \quad \text{where } \alpha = \frac{5(\sigma_1 + t^{-1})}{\sigma_1 + t^{-1}/2}.$$

- Else, the maximum is given by

$$\sup_{\sigma \geq 0} \left| \frac{\sigma + t^{-1}}{\sigma - z} \right|^2 = \frac{t^{-2}}{x^2 + \alpha^2 \left( x + \frac{1}{2t} \right)^2} \leq \frac{5(1 + \alpha^2)}{\alpha^2}.$$

As a consequence, when $z \in \mathcal{C}_{t,1} \cup \mathcal{C}_{t,2}$, we have

$$\sup_{\sigma \in \sigma(\hat{\Sigma})} \left| \frac{\sigma + t^{-1}}{\sigma - z} \right|^2 \leq 8.$$

**Second case, when** $z \in \mathcal{C}_{t,3}$**.** We have $|\sigma - z| \geq 2\sigma_1 + t^{-1}$ for $\sigma \in \sigma(\hat{\Sigma}) \subseteq [0, \hat{\sigma}_1]$, so, on $\Omega_t$, it follows from Lemma 9 that $\hat{\sigma}_1 < 4(\sigma_1 + t^{-1})$ and so

$$\sup_{\sigma \in \sigma(\hat{\Sigma})} \left| \frac{\sigma + t^{-1}}{\sigma - z} \right| \leq \frac{4\sigma_1 + 5t^{-1}}{2\sigma_1 + 5t^1} \leq 5.$$

Recall that from Lemma 11,

$$\|\Sigma_t^{-\frac{1}{2}} \hat{\Sigma}_t^{\frac{1}{2}}\|_{\text{op}}^2 \leq 2, \quad \text{and} \quad \|\Sigma_J^{\frac{1}{2}} \hat{\Sigma}_t^{-\frac{1}{2}}\|_{\text{op}}^2 \leq 2.$$

The upper bound of $\|\Sigma_J^{\frac{1}{2}} \left( \hat{\Sigma} - zI_p \right)^{-1} \Sigma_J^{\frac{1}{2}}\|_{\text{op}}$ is given by:

$$\left\| \Sigma_J^{\frac{1}{2}} \left( \hat{\Sigma} - zI_p \right)^{-1} \Sigma_J^{\frac{1}{2}} \right\|_{\text{op}} < \left\| \Sigma_J^{\frac{1}{2}} \hat{\Sigma}_t^{-\frac{1}{2}} \right\|_{\text{op}} \left\| \hat{\Sigma}_t^{\frac{1}{2}} (\hat{\Sigma} - zI_p)^{-1} \hat{\Sigma}_t^{\frac{1}{2}} \right\|_{\text{op}} \left\| \Sigma_J^{\frac{1}{2}} \hat{\Sigma}_t^{-\frac{1}{2}} \right\|_{\text{op}} < 3C,$$

for some absolute constant $C > 1$.

The upper bound for $\|\Sigma_t^{\frac{1}{2}}(\Sigma - zI_p)^{-1}\Sigma_t^{\frac{1}{2}}\|_{\text{op}}$ is similar but simpler since

$$\left\|\Sigma_t^{\frac{1}{2}}(\Sigma - zI_p)^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{\text{op}} = \sup_{\sigma\in\sigma(\Sigma)}\left|\frac{\sigma + t^{-1}}{\sigma - z}\right|$$

and $\sigma(\Sigma) \subset [0, \sigma_1]$, so we omit it.

Finally, we move to the integral of the holomorphic extensions of the filter and residual functions. We have

$$\oint_{\mathcal{C}_t} |\varphi_t(z)\mathrm{d}z| \leq C \oint_{\mathcal{C}_t} \frac{1}{|z + t^{-1}|}|\mathrm{d}z|.$$

Now we focus on the latter integral. For $z \in \mathcal{C}_{t,1}$, we have $|z + t^{-1}| \geq \sqrt{17}t^{-1}$ and thus

$$\int_{\mathcal{C}_{t,1}} \frac{1}{|z + t^{-1}|}|\mathrm{d}z| \leq \frac{1}{\sqrt{17}} t^{-1}\,|\mathcal{C}_{t,1}| \leq C$$

for some absolute constant $C > 1$, where we notice that $|\mathcal{C}_{t,1}| \leq Ct^{-1}$. For $\mathcal{C}_{t,2}$, we have

$$\begin{aligned}\int_{\mathcal{C}_{t,2}} \frac{1}{|z + t^{-1}|}|\mathrm{d}z| &= 2\int_0^{\sigma_1} \frac{1}{|x + (x + t^{-1}/2)i + t^{-1}|}\sqrt{2}\,\mathrm{d}x \\ &\leq C\int_0^{\sigma_1} \frac{1}{x + t^{-1}}\mathrm{d}x \leq C\log(t),\end{aligned}$$

where we have used that assumption that $\sigma_1$ is at most a constant. For $z \in \mathcal{C}_{t,3}$, we have $|z + t^{-1}| \geq \sqrt{17}(\sigma_1 + t^{-1})$ and thus

$$\int_{\mathcal{C}_{t,3}} \frac{1}{|z + t^{-1}|}|\mathrm{d}z| \leq \frac{1}{\sqrt{17}(\sigma_1 + t^{-1})}\,|\mathcal{C}_{t,3}| \leq C,$$

for some absolute constant.

■

### 3.5.4 Proof of Lemma 11

*Proof.* On the event $\Omega_t$, we have

$$\begin{aligned}\left\|\Sigma_t^{-\frac{1}{2}}\hat{\Sigma}_t^{\frac{1}{2}}\right\|_{\text{op}}^2 &= \left\|\Sigma_t^{-\frac{1}{2}}\hat{\Sigma}_t\Sigma_t^{-\frac{1}{2}}\right\|_{\text{op}} = \left\|\Sigma_t^{-\frac{1}{2}}\left(\hat{\Sigma} + t^{-1}I\right)\Sigma_t^{-\frac{1}{2}}\right\|_{\text{op}} \\ &= \left\|\Sigma_t^{-\frac{1}{2}}\left(\hat{\Sigma} - \Sigma + \Sigma + t^{-1}I\right)\Sigma_t^{-\frac{1}{2}}\right\|_{\text{op}} \\ &\leq \left\|\Sigma_t^{-\frac{1}{2}}\left(\hat{\Sigma} - \Sigma\right)\Sigma_t^{-\frac{1}{2}}\right\|_{\text{op}} + 1 \leq \square + 1 \leq 2.\end{aligned}$$

Let us now move to the first statemant of Lemma 11. On the event $\Omega_t$ we have

$$\left\|\Sigma_t^{-1/2}(\hat{\Sigma} - \Sigma)\Sigma_t^{-1/2}\right\|_{\text{op}} \leq \square < 1,$$

as a consequence, we have on that event

$$\left\|\left(I - \Sigma_t^{-\frac{1}{2}}\left(\hat{\Sigma} - \Sigma\right)\Sigma_t^{-\frac{1}{2}}\right)^{-1}\right\|_{\text{op}} \leq \sum_{k=0}^{\infty}\left\|\Sigma_t^{-\frac{1}{2}}\left(\hat{\Sigma} - \Sigma\right)\Sigma_t^{-\frac{1}{2}}\right\|_{\text{op}}^k \leq \sum_{k=0}^{\infty}\square^k \leq 2$$

because $\square \leq 1/2$. Next, using (3.11), we observe that

$$\begin{aligned}\left\|\Sigma_J^{\frac{1}{2}}\hat{\Sigma}_t^{-\frac{1}{2}}\right\|_{\text{op}}^2 &\leq \left\|\Sigma_J^{\frac{1}{2}}\Sigma_t^{-\frac{1}{2}}\right\|_{\text{op}}^2\left\|\Sigma_t^{\frac{1}{2}}\hat{\Sigma}_t^{-\frac{1}{2}}\right\|_{\text{op}}^2 \leq \left\|\Sigma_t^{\frac{1}{2}}\hat{\Sigma}_t^{-\frac{1}{2}}\right\|_{\text{op}}^2 = \left\|\Sigma_t^{\frac{1}{2}}\hat{\Sigma}_t^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{\text{op}} \\ &= \left\|\left(\Sigma_t^{-\frac{1}{2}}\hat{\Sigma}_t\Sigma_t^{-\frac{1}{2}}\right)^{-1}\right\|_{\text{op}} = \left\|\left(I - \Sigma_t^{-\frac{1}{2}}\left(\hat{\Sigma} - \Sigma\right)\Sigma_t^{-\frac{1}{2}}\right)^{-1}\right\|_{\text{op}}.\end{aligned}$$

■

## 3.6 Statistical analysis of PCR: proof of Theorem 9

In this section we prove Theorem 9. We recall that the Principle Component Regression (PCR) estimator $\hat{\boldsymbol{\beta}} = \frac{1}{N}\varphi_t(\hat{\Sigma})\mathbb{X}^\top \boldsymbol{y}$ is obtained for the filter function and its associated residual function given for $t \geq 1$ by

$$\varphi_t : x > 0 \mapsto x^{-1}\mathbb{1}(x \geq bt^{-1}), \text{ and } \psi_t : x \in \mathbb{R} \mapsto 1 - x\varphi_t(x) = \mathbb{1}(x < bt^{-1})$$

where $0 < b < 1$ is the same parameter used in the definition of

$$k^* := \min\left(k \in [p] : \sigma_{k+1} \leq bt^{-1}\right),$$

the estimation dimension. In this section, we also denote $J = J_* = [k^*]$.

First note that for all $t \geq 1$ and $x > 0$, we have

$$\varphi_t(x) = \frac{1}{x}\mathbb{1}(x \geq bt^{-1}) \leq \frac{C_1}{x + t^{-1}} \text{ for } C_1 = \frac{b+1}{b} \tag{3.51}$$

so that Assumption 5 is satisfied by PCR's filter function for $c_1 = 0$ and $C_1 = (b+1)/b$.

A key observation in the analysis of PCR estimator is that, for a given SDP matrix $M$, $\psi_t(M)$ is the orthogonal projection on the eigenspace of $M$ spanned by all eigenvectors associated with eigenvalues less than $bt^{-1}$. In particular, $\psi_t(\Sigma) = P_{J^c} = \sum_{j \in J^c} \boldsymbol{e}_j \otimes \boldsymbol{e}_j$ and so for all $\boldsymbol{\beta} \in V_J, \psi_t(\Sigma)\boldsymbol{\beta} = 0$. We also observe that $x\varphi_t(x) = \mathbb{1}(x \geq bt^{-1})$ so that $\Sigma\varphi_t(\Sigma) = P_J$; in particular, for $\tilde{\boldsymbol{\beta}}_J$ defined in Section 3.3.2 we have $\tilde{\boldsymbol{\beta}}_J = \varphi_t(\Sigma)\Sigma\boldsymbol{\beta}^* = P_J\boldsymbol{\beta}^* = \boldsymbol{\beta}^*_J$. As a consequence, the risk decomposition of the estimation part from Section 3.3.2 can be made simpler in the PCR case.

Let us now start the risk analysis of the PCR estimator. As in (3.10), we recall the risk decomposition that follows from the FSD method:

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

Next, as mentioned above, the risk decomposition of the estimation part is simpler for the PCR estimator than in (3.19) since we have

$$\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J)\right\|_2 \leq \left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}^*_J) - \boldsymbol{\beta}^*_J)\right\|_2 + \left\|\Sigma_J^{1/2}\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}^*_{J^c} + \boldsymbol{\xi})\right\|_2.$$

Now, we upper bound the two terms from this sum. For the first term, we have on $\Omega_t$

$$\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}^*_J) - \boldsymbol{\beta}^*_J)\right\|_2 = \left\|\Sigma_J^{1/2}\left(\psi_t(\hat{\Sigma}) - \psi_t(\Sigma)\right)\boldsymbol{\beta}^*_J\right\|_2 \lesssim \frac{\square}{\theta^2}\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}^*_J\right\|_2$$

where the last inequality follows from an adaptation of the argument used in (3.26) to the PCR case for the contour $\mathcal{C}_t$ defined in (3.54): thanks to (3.53), we indeed have

$$\begin{aligned}
&\left\|\Sigma_J^{\frac{1}{2}}\left(\psi_t(\Sigma) - \psi_t(\hat{\Sigma})\right)\boldsymbol{\beta}^*_J\right\|_2 = \frac{1}{2\pi}\left\|\oint_{\mathcal{C}_t}\Sigma_J^{\frac{1}{2}}(\hat{\Sigma} - zI_p)^{-1}(\hat{\Sigma} - \Sigma)(\Sigma - zI_p)^{-1}\boldsymbol{\beta}^*_J dz\right\|_2 \\
&\leq \frac{1}{2\pi}\oint_{\mathcal{C}_t}\left\|\Sigma_t^{\frac{1}{2}}\left(\hat{\Sigma} - zI_p\right)^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{op}\left\|\Sigma_t^{-\frac{1}{2}}\left(\hat{\Sigma} - \Sigma\right)\Sigma_t^{-\frac{1}{2}}\right\|_{op}\left\|\Sigma_t^{\frac{1}{2}}(\Sigma - zI_p)^{-1}\Sigma_J^{\frac{1}{2}}\right\|_{op} \\
&\qquad \cdot \left\|\Sigma_J^{-1/2}\boldsymbol{\beta}^*_J\right\|_2 |dz| \\
&\lesssim \frac{\square}{\theta^2}\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}^*_J\right\|_2 \oint_{\mathcal{C}_t}|dz| \lesssim \frac{\square}{\theta^2}\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}^*_J\right\|_2.
\end{aligned} \tag{3.52}$$

For the second term, we use exactly the same arguments as in Section 3.3.2: with probability at least $1 - \exp(-c|J|) - \mathbb{P}[\Omega_t^c]$,

$$\|\Sigma_J^{1/2}\hat{\boldsymbol{\beta}}_J(\mathbb{X}\boldsymbol{\beta}^*_{J^c} + \boldsymbol{\xi})\|_2 \lesssim \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2 + \sigma_\xi\sqrt{\frac{|J|}{N}}.$$

As a consequence, we conclude that for the estimation part, we have with probability at least $1 - \exp(-c|J|) - \mathbb{P}[\Omega_t^c]$,

$$\left\|\Sigma_J^{1/2}\left(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J\right)\right\|_2 \lesssim \frac{\square}{\theta^2}\left\|\Sigma_J^{-1/2}\boldsymbol{\beta}^*_J\right\|_2 + \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2 + \sigma_\xi\sqrt{\frac{|J|}{N}}.$$

Now, we prove a high probability upper bound on the 'noise absorption' part of the PCR estimator, i.e. on the quantity $\left\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\right\|_2$. We follow the same analysis as in Section 3.3.3 but for the contour $\mathcal{C}_t$ specially designed for the PCR estimator, i.e. the one from (3.54) and where we use Lemma 16 instead of Lemma 10: with probability at least $1-2\exp(-c|J|)-\mathbb{P}[\Omega_t^c]$,

$$\left\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\right\|_2 \lesssim \frac{\square}{\theta^2}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}^*_J\right\|_2 + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\right\|_2 + \sigma_\xi\sqrt{\frac{|J|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}}.$$

Gathering both controls on the estimation part and the noise absorption part in the risk decomposition of the PCR estimator that follows from the FSD method, we obtain that with probability at least $1-c\exp(-|J|/c)-c\exp(-\square^2 N/c)$,

$$\begin{aligned}\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right\|_2 &\lesssim \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\right\|_2 + \sigma_\xi\sqrt{\frac{|J|}{N}} + \sigma_\xi t\sqrt{\frac{\mathrm{Tr}(\Sigma_{J^c}^2)}{N}} + \frac{\square}{\theta^2}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}^*_J\right\|_2 \\ &\lesssim r(V_J,V_{J^c}) + \frac{\square}{\theta^2}\left\|\Sigma_J^{-\frac{1}{2}}\boldsymbol{\beta}^*_J\right\|_2.\end{aligned}$$

### 3.6.1 Construction and properties of the contour for the analysis of PCR

Let $\mathcal{C}_t \subset \mathbb{C}$ be a contour such that:

(i) $\mathcal{C}_t$ surrounds the set of all singular values of $\Sigma$ and $\hat{\Sigma}$ below $bt^{-1}$, i.e. the set $\left[\sigma(\Sigma)\cup\sigma(\hat{\Sigma})\right]\cap[0,bt^{-1}]$,

(ii) all singular values of $\Sigma$ and $\hat{\Sigma}$ above $bt^{-1}$, i.e. the set $\left[\sigma(\Sigma)\cup\sigma(\hat{\Sigma})\right]\cap[bt^{-1},+\infty]$ are 'outside' $\mathcal{C}_t$.

For a contour $\mathcal{C}_t$ satisfying the two points above, it follows from [Kat95, pp. 39], see also [KL16, pp. 1984] that

$$\begin{aligned}\psi_t(\Sigma)-\psi_t(\hat{\Sigma}) = P_{J^c}-\hat{P} &= \frac{1}{2\pi i}\oint_{\mathcal{C}_t}\left[(\hat{\Sigma}-zI)^{-1}-(\Sigma-zI)^{-1}\right]dz \\ &= -\frac{1}{2\pi i}\oint_{\mathcal{C}_t}(\hat{\Sigma}-zI)^{-1}(\hat{\Sigma}-\Sigma)(\Sigma-zI)^{-1}dz \qquad (3.53)\end{aligned}$$

where $\hat{P}$ is the orthogonal projection onto the space spanned by all singular vectors of $\hat{\Sigma}$ associated with a singular value less than $bt^{-1}$. In particular, we recover a formulae similar to (3.18) but for $\psi_t$.

Now we define a contour that to satisfies the two requirements above. This contour is a counterclockwise rectangle $\mathcal{C}_t = \mathcal{C}_{t,1}\sqcup\mathcal{C}_{t,2}\sqcup\mathcal{C}_{t,3}\sqcup\mathcal{C}_{t,4}$ made of the four segments:

$$\begin{aligned}&\mathcal{C}_{t,1}=\{-1+iy : -1\le y\le 1\},\ \mathcal{C}_{t,2}=\left\{bt^{-1}+iy : -1\le y\le 1\right\}, \\ &\mathcal{C}_{t,3}=\left\{x+i : -1\le x\le bt^{-1}\right\} \text{ and } \mathcal{C}_{t,4}=\left\{x-i : -1\le x\le bt^{-1}\right\}.\end{aligned} \qquad (3.54)$$

It is clear from the definition of $\mathcal{C}_t$ that the two conditions *(i)* and *(ii)* are satisfied by this contour. Let us now turn to properties of $\mathcal{C}_t$ that will be useful for the statistical analysis of PCR, i.e. to results similar to the one from Lemma 10. We first recall that the $k^*$-th spectral gap of $\Sigma$ is the quantity $\gamma_{k^*}=\sigma_{k^*}-\sigma_{k^*+1}$. The following result requires $\gamma_{k^*}$ to be large enough so that $\theta>0$ where we recall that

$$\theta := \min\left(bt^{-1}-\left(\sigma_{k^*+1}+\square(\sigma_{k^*+1}+t^{-1})\right),\left(\sigma_{k^*}-\square(\sigma_{k^*}+t^{-1})\right)-bt^{-1}\right)$$

**Lemma 16.** *Let $t\ge 1$, $0<\square<1/9$ and $0<b<1$ be such that $\theta>0$. Let $\mathcal{C}_t$ be the contour defined in (3.54). For all $z\in\mathcal{C}_t$, we have*

$$\left\|\Sigma_t^{\frac{1}{2}}(\Sigma-zI_p)^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{op}\le\frac{2}{\theta} \text{ and } \oint_{\mathcal{C}_t}|dz|\le 6.$$

*Moreover, on $\Omega_t$ we have for all $z\in\mathcal{C}_t$, $\left\|\Sigma_t^{1/2}\left(\hat{\Sigma}-zI_p\right)^{-1}\Sigma_t^{1/2}\right\|_{op}\le 2/\theta$.*

*Proof.* Let $z \in \mathcal{C}_t$. We have

$$\left\|\Sigma_t^{\frac{1}{2}}(\Sigma - zI_p)^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{\mathrm{op}} = \max\left(\left|\frac{\sigma_j + t^{-1}}{\sigma_j - z}\right| : j \in J\right) \leq \max\left(\left|\frac{\sigma_j + t^{-1}}{\sigma_j - bt^{-1}}\right| : j \in J\right) \leq \frac{2}{\theta}.$$

Given that $bt^{-1} \leq 1$, the length of $\mathcal{C}_t$ is at most 6 and so $\oint_{\mathcal{C}_t} |dz| \leq 6$. Next, we have

$$\left\|\Sigma_t^{\frac{1}{2}}\left(\hat{\Sigma} - zI_p\right)^{-1}\Sigma_t^{\frac{1}{2}}\right\|_{\mathrm{op}} \leq \frac{\sigma_1 + t^{-1}}{\min_j |\hat{\sigma}_j - z|} \leq \frac{2}{\min_j |\hat{\sigma}_j - bt^{-1}|}.$$

On the event $\Omega_t$, it follows from (3.16) that for all $\boldsymbol{u} \in \mathbb{R}^p$,

$$(1 - \square)\left\|\Sigma^{1/2}\boldsymbol{u}\right\|_2^2 - \square t^{-1}\|\boldsymbol{u}\|_2^2 \leq \left\|\hat{\Sigma}^{1/2}\boldsymbol{u}\right\|_2^2 \leq (1 + \square)\left\|\Sigma^{1/2}\boldsymbol{u}\right\|_2^2 + \square t^{-1}\|\boldsymbol{u}\|_2^2. \tag{3.55}$$

As a consequence, for all $\boldsymbol{u} \in V_{J_*}$, we have

$$\left\|\hat{\Sigma}\boldsymbol{u}\right\|_2 \geq \left[(1 - \square)\sigma_{k^*} - \square t^{-1}\right]\|\boldsymbol{u}\|_2 \tag{3.56}$$

and for all $\boldsymbol{u} \in V_{J_*^c}$,

$$\left\|\hat{\Sigma}\boldsymbol{u}\right\|_2 \leq \left[(1 + \square)\sigma_{k^*+1} + \square t^{-1}\right]\|\boldsymbol{u}\|_2. \tag{3.57}$$

Given that $V_{J_*}$ is of dimension $k^*$ (and so $V_{J_*^c}$ is of dimension $p - k^*$), it follows from (3.56), (3.57) and the Courant-Fischer minimax variational formulas (see for instance Theorem 4.2.1 in [CGLP12]) that

$$\hat{\sigma}_{k^*} = \max_{V:\dim(V)=k^*}\min_{\boldsymbol{u}\in V:\|\boldsymbol{u}\|_2=1}\left\|\hat{\Sigma}\boldsymbol{u}\right\|_2 \geq \min_{\boldsymbol{u}\in V_{J_*}:\|\boldsymbol{u}\|_2=1}\left\|\hat{\Sigma}\boldsymbol{u}\right\|_2 \geq \sigma_{k^*} - \square\left[\sigma_{k^*} + t^{-1}\right].$$

and

$$\hat{\sigma}_{k^*+1} = \min_{V:\dim(V)=p-k^*}\max_{\boldsymbol{u}\in V:\|\boldsymbol{u}\|_2=1}\left\|\hat{\Sigma}\boldsymbol{u}\right\|_2 \leq \max_{\boldsymbol{u}\in V_{J_*^c}:\|\boldsymbol{u}\|_2=1}\left\|\hat{\Sigma}\boldsymbol{u}\right\|_2 \leq \sigma_{k^*+1} + \square\left[\sigma_{k^*+1} + t^{-1}\right].$$

As a consequence, on $\Omega_t$, we obtain that

$$\min_j\left|\hat{\sigma}_j - bt^{-1}\right| \geq \theta$$

and so the result follows. ∎

# Chapter 4

# Benign overfitting property of the minimum $\ell_q$ norm interpolant estimator via Feature Space Decomposition

*"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.*

In this chapter, we apply the FSD method introduced in Section 1.5 to study the benign overfitting phenomenon in the minimum $\ell_q$-norm interpolant estimator in linear regression and the minimum $\ell_2$-norm interpolant classifier in linear classification.

## 4.1 Introduction

We consider the linear model in this chapter, where the function class is given by $\mathcal{F} = \{f_{\boldsymbol{\beta}}(\cdot) = \langle \boldsymbol{\beta}, \cdot \rangle : \boldsymbol{\beta} \in \mathbb{R}^p\}$. In the following, we will identify $f_{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}$. In this chapter, all inner products $\langle \cdot, \cdot \rangle$ are Euclidean inner products. In this chapter, we always assume that $N < p$ and that $\sigma_\xi$ is a constant independent of $N$ and $p$. Under the linear model, the regression and classification problems can be formulated as follows.

- Regression problem. Let $\sigma_\xi > 0$ denote a positive real number, referred to as the noise level, and let $\xi \in \mathbb{R}$ be a centered random variable with variance $\sigma_\xi^2$, independent of $X$. Assume there exists $\boldsymbol{\beta}^* \in \mathbb{R}^p$, referred to as the signal, such that $Y = \langle \boldsymbol{\beta}^*, X \rangle + \xi$. It is straightforward to verify that $f_{\boldsymbol{\beta}^*}$ is the minimizer of $P\ell_f^{(2)}$ and that $P\mathcal{L}_{\hat{\boldsymbol{\beta}}}^{(2)} = \|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle\|_{L^2(\mu_X)}^2$, where we use $\mathcal{L}^{(2)}$ to emphasize the excess risk for the squared loss.
- Classification problem. Assume that $Y \in \{-1, 1\}$. Let $\eta : \boldsymbol{x} \in \mathbb{R}^p \mapsto \mathbb{P}(Y = 1 \mid X = \boldsymbol{x})$ denote the posterior distribution function. It is straightforward to verify that, in the classification problem, the Bayes rule is given by $\boldsymbol{x} \in \mathbb{R}^p \mapsto \operatorname{sign}(\eta(\boldsymbol{x}) - \frac{1}{2})$, where $\operatorname{sign}(t) = 1$ if $t > 0$; $\operatorname{sign}(t) = -1$ if $t < 0$; and $\operatorname{sign}(t) \in [-1, 1]$ if $t = 0$. This Bayes rule typically does not belong to the linear model. In classification problems we usually compare with the population excess risk of the Bayes rule. In this case, the population excess risk is

$$P\mathcal{L}_{\hat{\boldsymbol{\beta}}}^{\{0,1\}} = \mathbb{P}\left(\operatorname{sign}\left(\langle \hat{\boldsymbol{\beta}}, X \rangle\right) \neq Y \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}(\operatorname{sign}(\eta(X) - \frac{1}{2}) \neq Y). \tag{4.1}$$

  Here we use $\mathcal{L}_{\boldsymbol{\beta}}^{\{0,1\}}$ to emphasize that this is the excess risk for the 0-1 loss.

**Minimum Norm Interpolant Estimator.** In this chapter, the estimators $\hat{\boldsymbol{\beta}}$ considered are minimum norm interpolant estimators for various choices of norms. We now introduce the necessary notation to define $\hat{\boldsymbol{\beta}}$. For any $1 \leq q < \infty$, let $\|\cdot\|_q : \boldsymbol{v} \in \mathbb{R}^p \mapsto (\sum_{j=1}^p |\langle \boldsymbol{v}, \boldsymbol{e}_j \rangle|^q)^{1/q}$ be the $\ell_q^p$ norm with respect to the canonical basis $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_p\}$

and $\|\cdot\|_\infty : \boldsymbol{v} \in \mathbb{R}^p \mapsto \max(|\langle \boldsymbol{v}, \boldsymbol{e}_j\rangle| : j \in [p])$ be the $\ell_\infty^p$ norm, where $[p] = \{1, \cdots, p\}$. Suppose $q$ is independent with $N$ and $p$. Sometimes, we also use notation $\|\cdot\|_2$ to denote the $\ell_2$ norm in $\mathbb{R}^N$, but this is usually clear from the context. In regression and classification problems, we respectively define the minimum $\ell_q$ norm interpolant estimator as follows:

- In the linear regression model, the minimum $\ell_q$-norm interpolant estimator is

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left( \|\boldsymbol{\beta}\|_q : \forall i \in [N],\, \langle X_i, \boldsymbol{\beta}\rangle = Y_i \right). \tag{4.2}$$

- In the classification mode, the minimum $\ell_2$-norm interpolant estimator is

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left( \|\boldsymbol{\beta}\|_2 : \forall i \in [N],\, Y_i \langle X_i, \boldsymbol{\beta}\rangle \geq 1 \right), \tag{4.3}$$

  which is called the minimum $\ell_2$-norm/max-margin interpolant classifier, also known as hard margin support vectors machine.

That is, $\hat{\boldsymbol{\beta}}$ achieves zero training error on the training data with $P_N \ell^{(2)}_{\hat{\boldsymbol{\beta}}} = 0$ or $P_N \ell^{\{0,1\}}_{\hat{\boldsymbol{\beta}}} = 0$ and among all estimator having this property $\hat{\boldsymbol{\beta}}$ is the one with the smallest norm (for some given norm).

**Benign overfitting.** When an estimator $\hat{f}$ satisfies $P_N \ell_{\hat{f}} = 0$, we refer to it as an *interpolant estimator* or *overfitting estimator*. The research objective of benign overfitting is to investigate the necessary and sufficient conditions under which overfitting is harmless to generalization, i.e., to determine under what conditions $P\ell_{\hat{f}}$ remains small even though it interpolates the data - a property previously considered as an obstacle to generalization. Therefore, we define the following two types of benign overfitting.

- We say that an overfitting estimator $\hat{f}$ exhibits **exact benign overfitting** if, as $N, p \to \infty$, the population excess risk $P\mathcal{L}_{\hat{f}} = P\ell_{\hat{f}} - P\ell_{f^*}$ converges to zero.

- We say that an overfitting estimator $\hat{f}$ exhibits **non-exact benign overfitting** if there exists $0 < \varepsilon < 1$ such that, as $N, p \to \infty$, the non-exact population excess risk satisfies $\limsup P\ell_{\hat{f}} - (1+\varepsilon)P\ell_{f^*} \leq 0$.

In the existing literature, there exists yet another definition. To name a few, [CGB22, FVBS23, FCB22, XWF+23, WMT21, KCCG23, MZC23, KYS23, JHZ+24]. Although these works also refer to it as "benign overfitting", we shall distinguish it in this chapter by calling it "test error benign overfitting", defined as follows. We say that an overfitting estimator $\hat{f}$ exhibits **test error benign overfitting** if, as $N, p \to \infty$, the population risk $P\ell_{\hat{f}} \to 0$.

In what follows, unless otherwise specified, benign overfitting refers to exact benign overfitting. To the best of our knowledge, non-exact benign overfitting is a novel concept, motivated by the notion of non-exact oracle inequalities, [Kol11]. Strictly speaking, non-exact benign overfitting belongs to the class of **tempered overfitting** [MSA+22] — specifically, where $P\mathcal{L}_{\hat{f}}$ is finite but non-zero. Yet it delivers more information than tempered overfitting since $\varepsilon$ is small.

If one aims to demonstrate the smallness of $P\ell_{\hat{f}}$, in many cases it suffices to compare it to $(1+\varepsilon)P\ell_{f^*}$ – this is precisely the motivation of non-exact oracle inequality, [Kol11]. While this is trivial in regression settings when $\varepsilon$ is a constant, it is often effective in classification problems. Test error benign overfitting constitutes precisely such an example. As test error benign overfitting requires $P\ell_{\hat{f}} \to 0$, the comparison with $(1+\varepsilon)P\ell_{f^*}$ in non-exact benign overfitting implies test error benign overfitting, since the test error benign overfitting implicitly implies that $P\ell_{f^*} \to 0$. In contrast, if we only know that $P\mathcal{L}_{\hat{f}}$ is finite (i.e., tempered overfitting), then we cannot conclude $P\ell_{\hat{f}} \to 0$ even though $P\ell_{f^*}$ may go to 0. On the other hand, the key distinction between non-exact benign overfitting and test error benign overfitting lies in whether $P\ell_{f^*}$ tends to zero or not. In classical mathematical statistics, $P\ell_{f^*}$ typically measures the difficulty level of a statistical problem as it measures the 'size' of the noise. If $P\ell_{f^*} \to 0$, it indicates that the statistical problem becomes relatively simple as $p \to +\infty$. Therefore, we argue that non-exact benign overfitting represents an intermediate phenomenon between exact benign overfitting and test error benign overfitting. Compared to exact benign overfitting, it is more likely to hold, while in more challenging statistical problems, it can provide richer information about $P\ell_{\hat{f}}$ when test error benign overfitting does not hold because $P\ell_{f^*}$ may not tend to 0.

As a brief remark, when $\varepsilon$ is not a fixed constant but instead satisfies $\varepsilon P\ell_{f^*} = o(1)$ as $N, p \to \infty$, non-exact benign overfitting in fact implies benign overfitting.

### 4.1.1 Our Contributions

1. We develop a feature space decomposition analytical framework for supervised regression and classification problems. As an improvement over uniform convergence arguments, this method can be applied to any supervised learning problem to obtain sharper high-probability upper bounds for an estimator's population excess risk. This approach potentially refines one of the most fundamental methodologies in mathematical statistics—the uniform convergence argument.

2. As concrete applications, we investigate the self-regularization properties of minimum $\ell_q$-norm interpolant estimators for all $q \geq 1$, establishing non-asymptotic high-probability upper bounds for: (i) the estimation error of minimum $\ell_q$-norm interpolant estimators in linear regression, and (ii) the $0-1$ excress risk of the minimum $\ell_2$-norm classifiers in linear classification. This yields sufficient conditions for benign overfitting for these interpolant estimators. Our self-regularization argument may serve as a foundational method for analyzing statistical properties of minimum $\ell_q$ norm interpolant estimators. Remarkably, through self-regularization, we discover that projections of minimum norm interpolants correspond to classical estimators like the square-root LASSO, squared hinge loss support vector machines, and regularized M-estimators—a novel phenomenon previously unrecognized in the literature.

3. We propose a new characterization of benign overfitting called *non-exact benign overfitting*. This overlooked phenomenon provides more information than test-error-based benign overfitting while being more broadly applicable than exact benign overfitting. Consequently, non-exact benign overfitting may emerge as a new evaluative metric for overfitting estimators. We establish sufficient conditions for the occurrence of non-exact benign overfitting for the aformentioned minimum $\ell_q/\ell_2$-norm interpolant estimators.

4. We extend the Dvoretzky–Milman theorem to the $\|\cdot\|_{q'}$-norm under general probability measures. This is a result of independent interest.

### 4.1.2 Notation

For any $1 \leq q < \infty$, we denote $B_q^p$ is the unit ball of $\|\cdot\|_q$, $S_q^p$ is the unit sphere of it, except $S_2^{p-1}$ is preserved for the unit sphere of $\|\cdot\|_2$. For any orthogonal projection $P_J$ onto some subspace $V_J \subset \mathbb{R}^p$, we denote $B_q^J = P_J B_q^p$, $S_q^J = P_J S_q^p$, and $S_2^J = P_J S_2^{p-1}$. We denote $I_{V_J} : \boldsymbol{v} \in V_J \mapsto \boldsymbol{v}$ as the identical operator. We use $\lesssim$ (or $\gtrsim$) to denote inequality up to multiplicative constant. We say a random vector is sub-Gaussian, if it satisfies [Ver18, Definition 3.4.1]. For a deterministic vector $\boldsymbol{v} \in \mathbb{R}^p$ and a P.S.D. matrix $A \in \mathbb{R}^{p\times p}$, we denote by $\mathcal{N}(\boldsymbol{v}, A)$ the standard Gaussian random vector whose mean is $\boldsymbol{v}$ and covariance matrix is $A$. We write $\dim(V_J)$ as the linear dimension of $V_J$. For a vector $\boldsymbol{\beta} \in \mathbb{R}^p$, we write $\beta_j = \langle \boldsymbol{\beta}, \boldsymbol{e}_j \rangle$ for each $j \in [p]$. For any convex body $K \subset \mathbb{R}^p$, we define $\ell_*(K) = \mathbb{E}(\sup\langle \boldsymbol{v}, G\rangle : \boldsymbol{v} \in K)$ as the Gaussian mean width of $K$, where $G \in \mathbb{R}^p$ is a standard Gaussian random vector. We denote $K^\circ = \{\boldsymbol{v} \in \mathbb{R}^p : \langle \boldsymbol{v}, \boldsymbol{u}\rangle \leq 1, \forall \boldsymbol{u} \in K\}$ as the polar body of $K$. We let $\operatorname{diam}(K) = \max(\|\boldsymbol{v}\|_2 : \boldsymbol{v} \in K) = \max(\|\boldsymbol{u}\|_{K^\circ} : \boldsymbol{u} \in B_2^p)$ be the $\ell_2$ diameter of $K$, where $\|\cdot\|_{K^\circ}$ is the norm whose unit ball is $K^\circ$. Denote $\ell_*(K) = \mathbb{E}\sup(\langle \boldsymbol{v}, G\rangle : \boldsymbol{v} \in K)$ by the Gaussian mean width of $K$, where $G$ is a standard Gaussian random vector. Denote $d_*(K) = (\ell_*(K^\circ)/\operatorname{diam}(K^\circ))^2$ to be the Dvoretzky dimension of $K$. If $\zeta_1, \cdots$ are random variables, we denote $\mathbb{E}_{\zeta_1}$ as the conditional expectation given all other random variables. We say a centered random vector $\boldsymbol{\zeta} \in \mathbb{R}^p$ is isotropic, if $\mathbb{E}[\boldsymbol{\zeta} \otimes \boldsymbol{\zeta}] = I_{\mathbb{R}^p}$. We say $\boldsymbol{\zeta}$ is centered, if $\mathbb{E}[\boldsymbol{\zeta}] : \boldsymbol{v} \in \mathbb{R}^p \mapsto \mathbb{E}[\langle \boldsymbol{\zeta}, \boldsymbol{v}\rangle] = 0$. We denote by $L^q$ the corresponding $L^q$ space, where the underlying probability measure is usually clear from the context. Let $\Sigma = \mathbb{E}[X \otimes X] : \boldsymbol{v} \in \mathbb{R}^p \mapsto \mathbb{E}[X\langle X, \boldsymbol{v}\rangle] \in \mathbb{R}^p$ be the population covariance matrix of $X$. Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$ be eigenvalues of $\Sigma$. For any matrix $A \in \mathbb{R}^{p\times p}$, let $\operatorname{Tr}(A)$ be its trace, and $\|A\|_{\ell_q \to \ell_2} = \sup(\|A\boldsymbol{v}\|_2 : \|\boldsymbol{v}\|_q = 1)$ be the $\ell_q \to \ell_2$ operator norm. In particular, the $\ell_2 \to \ell_2$ operator norm is denoted by $\|\cdot\|_{\mathrm{op}}$.

### 4.1.3 Structure of this chapter

We present the principal findings of this work in Section 4.2. Before delving into the technical proofs, Section 4.3 examines the phenomenon of self-regularization — the key mechanism enabling benign overfitting — that is explored in the 'estimation' subspace $V_J$ when applying the features space decomposition method. Section 4.4 then investigates the noise absorption phenomenon in high-dimensional subspaces. Together, these two components form the analytical foundation of our main results. The corresponding proofs can be found in Section 4.6, Section 4.7, and Section 4.8, respectively. Section 4.5 contains the conclusions of this chapter and some future directions.

Since this chapter focuses on the minimum $\ell_2$-norm interpolating classifier, we select $\boldsymbol{\beta}_J^*$ in the following manner. We first define several quantities that will be used throughout the discussion. The significance of these quantities

will be explained later. Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the squared hinge loss (also called the truncated quadratic loss), that is,

$$\ell : (y, y') \in (\mathbb{R} \times \mathbb{R}) \mapsto (1 - yy')_+^2, \text{ where } (\cdot)_+ : x \in \mathbb{R} \mapsto (x)_+ = \max\{x, 0\}. \tag{4.4}$$

Given an **arbitrary** decomposition of $\mathbb{R}^p$ as $V_J \oplus^\perp V_{J^c}$ (in particular, this does not imply that $V_J$ or $V_{J^c}$ must be spanned by the eigenvectors of $\Sigma$), we define the oracle in $V_J$ for the squared hinge loss as

$$\boldsymbol{\beta}_J^* \in \operatorname{argmin} \left( P\ell_{\boldsymbol{\beta}_J}(X, Y) : \boldsymbol{\beta}_J \in V_J \right), \text{ where } \ell_{\boldsymbol{\beta}}(X, Y) = \ell(\langle X, \boldsymbol{\beta} \rangle, Y). \tag{4.5}$$

In Section 4.3.3 later, we show that $\hat{\boldsymbol{\beta}}_J$ is the RERM corresponding to the squared hinge loss, thereby explaining why this oracle is chosen. We continue with the main idea of FSD: $\hat{\boldsymbol{\beta}}_J$ is used for estimation, while $\hat{\boldsymbol{\beta}}_{J^c}$ serves as noise interpolation. Thus, we should treat the 0-1 risk of $\hat{\boldsymbol{\beta}}_J$ relative to the oracle $\boldsymbol{\beta}_J^*$ as the estimation error, and consider $\hat{\boldsymbol{\beta}}_{J^c}$ as the model noise. Additionally, we approximate $\operatorname{sign}(\eta(X) - 1/2)$ by $\operatorname{sign}(\langle \boldsymbol{\beta}_J^*, X \rangle)$ as the approximation error of the model. Our intuition suggests that $V_J$ is the subspace spanned by the optimal linear classifier, and we anticipate that the oracle $\boldsymbol{\beta}_J^*$ in this subspace aligns with the optimal linear classifier. Consequently, this would render the estimation error (1.18) equal to zero.

## 4.2 Main Results

In this section, we present our main results on exact and non-exact benign overfitting for minimum norm interpolant estimators in both linear regression and classification problems.

### 4.2.1 Regression problem

We now present the first main result of this chapter: sufficient conditions for the minimum $\ell_q$-norm interpolant estimator to exhibit non-exact benign overfitting properties. We begin by introducing some notation and terminology.

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

$$\alpha_q(x, y) = \begin{cases} \frac{q}{2} x^{q-2} y^2, & \text{if } |y| \leq x \\ |y|^q + \left(\frac{q}{2} - 1\right) x^q, & \text{otherwise.} \end{cases} \tag{4.6}$$

This function was introduced in [BCLL18]. For any $\rho > 0$, we define the set $\rho K_{\text{model}} = \{\boldsymbol{v} \in V_J : \sum_{j \in J} \alpha_q(|\beta_j^*|, v_j) \leq \rho^q\}$ when $q < 2$; while $\rho K_{\text{model}} = \rho B_q^J$ otherwise. Define

$$|||\boldsymbol{\beta}||| = \sup \left( \langle \boldsymbol{\beta}, \boldsymbol{u} \rangle : \boldsymbol{u} \in \frac{C_{29}\sqrt{N}}{\ell_*(\Sigma_{J^c}^{1/2} B_q^p)} K_{\text{model}} \cap \Sigma_J^{-1/2} B_2^J \right), \tag{4.7}$$

where $C_{29} = C_{29}(q)$ is some absolute constant.

**Remark 8.** *$K_{\text{model}}$ is a non-empty convex set containing $\boldsymbol{\beta}_J^*$.*

Define

$$r(V_J, V_{J^c}) = \begin{cases} \sigma_\xi \left(\frac{|J|}{N}\right)^{\frac{1}{2(q-1)}} + \|\Sigma_{J^c}^{1/2} \boldsymbol{\beta}_{J^c}^*\|_2 + |||\boldsymbol{\beta}_J^* \odot |\boldsymbol{\beta}_J^*|^{\odot(q-2)}|||^{\frac{1}{q-1}} \frac{\ell_*^{\frac{q}{q-1}}(\Sigma_{J^c}^{1/2} B_q^p)}{N^{\frac{q}{2(q-1)}}} & , q \geq 2 \\ \sigma_\xi^{\frac{1}{3}} \left(\frac{|J|}{N}\right)^{\frac{1}{6}} + \|\Sigma_{J^c}^{1/2} \boldsymbol{\beta}_{J^c}^*\|_2 + \sigma_\xi^{q-2} |||\boldsymbol{\beta}_J^* \odot |\boldsymbol{\beta}_J^*|^{\odot(q-2)}||| \frac{\ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)}{N^{\frac{q}{2}}} & , 1 < q < 2 \\ \sqrt{\frac{|J|}{N}} + \|\Sigma_{J^c}^{1/2} \boldsymbol{\beta}_{J^c}^*\|_2 + \varepsilon_1 \sigma_\xi & , q = 1. \end{cases} \tag{4.8}$$

Let $\operatorname{Log}(x) = \max\{1, \ln(x)\}$. We say that an FSD $\mathbb{R}^p = V_J \oplus V_{J^c}$ is admissible if the following conditions are satisfied:

1. $V_J = \operatorname{span}(\boldsymbol{e}_j : j \in J)$.
2. $X_J$ and $X_{J^c}$ are independent.
3. There exist $0 < \varepsilon_1, \kappa_{RIP}, \kappa_{DM} < 1$ such that $\kappa_{RIP}^{-1}|J| \leq N \leq \kappa_{DM} \varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p) \operatorname{Log}^{-2}(|J^c|^{1/q'} / d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p))$ when $q \geq 2$ and $\kappa_{RIP}^{-1}|J| \leq N \leq \kappa_{DM} \varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p)$ when $1 \leq q < 2$. When $X_{J^c}$ follows a Gaussian measure, this logarithmic factor can be removed.

The main theorem of this section is stated below. The proof of Theorem 10 can be found in Section 4.7.

**Theorem 10.** *Suppose $\xi$ is independent with $X$, $\mathbb{E}[\xi]=0$, $\mathbb{E}[\xi^2]=\sigma_\xi^2$ and $\mathbb{E}[X]=\mathbf{0}$. There exists an absolute constant $C_{30}=C_{30}(q)>1$ such that the following holds. For any admissible FSD $(V_J,V_{J^c})$, the following holds with probability at least as specified in (4.43) later: $\|\langle\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*,X\rangle\|_{L^2(\mu_X)}^2\le 2\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}^*_J)\|_2^2+2\|\Sigma_{J^c}^{1/2}(\hat{\boldsymbol{\beta}}_{J^c}-\boldsymbol{\beta}^*_{J^c})\|_2^2$ where*

$$\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}^*_J)\|_2\le C_{30}\,r(V_J,V_{J^c})\ \text{and}\ \|\Sigma_{J^c}^{1/2}(\hat{\boldsymbol{\beta}}_{J^c}-\boldsymbol{\beta}^*_{J^c})\|_2\le C_{30}\,r(V_J,V_{J^c})+C_{30}\sqrt{\frac{N}{d_*(\Sigma_{J^c}^{-1/2}B_{q'}^p)}}\,\sigma_\xi,\tag{4.9}$$

*provided that, in each of the regimes $q=1$, $1<q<2$, and $q\ge 2$, the corresponding assumptions ensuring these bounds are satisfied.*

1. *When $q\ge 2$,*
   (a) *either there exists a sub-Gaussian random vector $Z\in V_{J^c}$ with i.i.d. coordinates and a diagonal matrix $\Sigma_{J^c}$ such that $X_{J^c}=\Sigma_{J^c}^{1/2}Z$;*
   (b) *or $X$ is a Gaussian random vector and $\Sigma_{J^c}$ is not necessarily to be diagonal.*
2. *When $1<q<2$, assume that $X_J\sim\mathcal{N}(\mathbf{0},\Sigma_J)$, $\xi\sim\mathcal{N}(0,\sigma_\xi^2)$. Suppose*
   (a) *either $\boldsymbol{\beta}^*_{J^c}=\mathbf{0}$, then the assumption on $X_{J^c}$ is the same as in the case $q\ge 2$;*
   (b) *or $X_{J^c}\sim\mathcal{N}(\mathbf{0},\Sigma_{J^c})$.*
3. *When $q=1$, assume that $\Sigma$ is diagonal, $X_J$ is a sub-Gaussian random vector, and $X_{J^c}\sim\mathcal{N}(\mathbf{0},\Sigma_{J^c})$. Suppose there exists an absolute constant $c_{15}<1$ such that $(\ell^*(\Sigma_{J^c}^{1/2}B_1^p))^2\le c_{15}N\left(\sum_{j\in J\cap\mathrm{supp}(\boldsymbol{\beta}^*)}\sigma_j^{-1}\right)^{-1}$. In particular, when $\Sigma_{J^c}=I_{V_{J^c}}$, the condition $N\le\kappa_{DM}\varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2}B_\infty^p)$ can be improved to $N\le\kappa_{DM}\frac{\varepsilon_1}{\log(1/\varepsilon_1)}\log(|J^c|)$.*

*Furthermore, if there exists a random vector $Z\in V_{J^c}$ satisfying Assumption 8 later, and a diagonal matrix $\Sigma_{J^c}$ such that $X_{J^c}=\Sigma_{J^c}^{1/2}Z$, then*

1. *for $q\ge 2$, with the same probability, we have $\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}^*_J)\|_2\le C_{30}r(V_J,V_{J^c})\mathrm{Log}^{\frac{q}{q-1}}(|J^c|)$, and $\|\Sigma_{J^c}^{1/2}(\hat{\boldsymbol{\beta}}_{J^c}-\boldsymbol{\beta}^*_{J^c})\|_2\le C_{30}\,r(V_J,V_{J^c})\mathrm{Log}^{\frac{q}{q-1}}(|J^c|)+C_{30}\sqrt{\frac{N}{d_*(\Sigma_{J^c}^{-1/2}B_{q'}^p)}}\,\sigma_\xi$;*
2. *for $1<q<2$, (4.9) still holds under the same conditions and with the same probability.*

Therefore, for any $q\ge 1$, sufficient conditions for benign overfitting to occur is the existence of an admissible FSD $\mathbb{R}^p=V_J\oplus^\perp V_{J^c}$ such that $r(V_J,V_{J^c})=o(1)$ and $N=o(d_*(\Sigma_{J^c}^{-1/2}B_{q'}^p))$. Sufficient conditions for non-exact benign overfitting can be obtained in a similar manner. By [vH18], $\ell_*(\Sigma_{J^c}^{1/2}B_1^p)^2\sim\max_{j\in J^c}\sigma_j\log(j+1)$. Therefore, a sufficient condition for $(\ell^*(\Sigma_{J^c}^{1/2}B_1^p))^2\le c_{15}N\left(\sum_{j\in J\cap\mathrm{supp}(\boldsymbol{\beta}^*)}\sigma_j^{-1}\right)^{-1}$ is $\max_{j\in J^c}\sigma_j\log(j+1)\lesssim N\left(\sum_{j\in J\cap\mathrm{supp}(\boldsymbol{\beta}^*)}\sigma_j^{-1}\right)^{-1}$. In particular, if $J=\mathrm{supp}(\boldsymbol{\beta}^*)$ and if there exist $\eta_1,\eta_2>0$, $\frac{\eta_2}{\eta_1}\lesssim\frac{N}{|\mathrm{supp}(\boldsymbol{\beta}^*)|\log(|J^c|)}$, such that $\Sigma=\eta_1 I_{V_J}\oplus\eta_2 I_{V_{J^c}}$, then this condition is satisfied. In this case, $r(V_J,V_{J^c})\lesssim\sigma_\xi\sqrt{\frac{|\mathrm{supp}(\boldsymbol{\beta}^*)|}{N}}+\sigma_\xi\frac{N}{\log(|J^c|)}$.

In particular, when $q=2$, our upper bound on $\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}^*_J)\|_2$ recovers the optimal result of [P4], while the part concerning $\|\Sigma_{J^c}^{1/2}(\hat{\boldsymbol{\beta}}_{J^c}-\boldsymbol{\beta}^*_{J^c})\|_2$ does not. This discrepancy arises because, for $q\ne 2$, we use $\|\Sigma_{J^c}^{1/2}\|_{\ell_q\to\ell_2}$ together with the upper bound for $\|\hat{\boldsymbol{\beta}}_{J^c}\|_q$. Note that when $q\ne 2$, the operator $\hat{\boldsymbol{\beta}}_{J^c}[\cdot]$ becomes nonlinear, and hence the method based on the upper side of Dvoretzky–Milman theorem used applied for $\|\Sigma_{J^c}\cdot\|_2$ no longer applies. Therefore, we regard the convergence rate of $\|\Sigma_{J^c}^{1/2}(\hat{\boldsymbol{\beta}}_{J^c}-\boldsymbol{\beta}^*_{J^c})\|_2$ characterized in Theorem 10 as quantitatively suboptimal, the reason being that we currently lack suitable tools to bound $\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$; see the discussion in Section 4.5.

For completeness, we include an estimate of $\ell_*(\Sigma_{J^c}^{1/2}B_q^p)$. The proof of the following Lemma 17 may be found in Section 4.9.6.

**Lemma 17.** *Suppose $\Sigma_{J^c}$ is diagonal. Then there exist absolute constants $c_{16}<1$ and $C_{31}=C_{31}(q)>1$ such that $c_{16}(\sum_{j\in J^c}\sigma_j^{q'/2})^{1/q'}\le\ell_*(\Sigma_{J^c}^{1/2}B_q^p)\le C_{31}(\sum_{j\in J^c}\sigma_j^{q'/2})^{1/q'}$. Moreover, when $q\le 2$, $\mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)=\max\{\sigma_j:j\in J^c\}$; when $q>2$, $\mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)=\|\boldsymbol{\sigma}_{J^c}\|_{\frac{2q'}{2-q'}}$ where $\boldsymbol{\sigma}_{J^c}=(\sigma_j)_{j\in J^c}$.*

### 4.2.2 Classification problem

Before presenting the main results for the classification problem, we first introduce two commonly studied models in supervised classification.

**Definition 19** (Gaussian Mixture classification model and logistic model)**.** *Let $\boldsymbol{\mu} \in \mathbb{R}^p$ be called the signal, and $\Lambda \in \mathbb{R}^{p\times p}$ be a P.S.D. matrix.*

1. ***Gaussian Mixture classification model****. Let $Y$ be a Rademacher random variable, i.e., $\mathbb{P}(Y=1)=\mathbb{P}(Y=-1)=1/2$. Then define the conditional distribution $X \mid Y \sim \mathcal{N}(Y\boldsymbol{\mu}, \Lambda)$. This model is called the Gaussian Mixture Model (GMM), [Ver18, Section 4.7.1].*
2. ***Logistic model****. Let $X \sim \mathcal{N}(\mathbf{0}, \Lambda)$ be a Gaussian random vector with mean $\mathbf{0}$ and covariance matrix $\Lambda$. By defining $\eta(\boldsymbol{x}) = \mathbb{P}(Y=1 \mid X=\boldsymbol{x}) = 1/(1+\exp(-2\langle \Lambda^{-1}\boldsymbol{\mu}, \boldsymbol{x}\rangle))$ and $\mathbb{P}(Y=-1 \mid X=\boldsymbol{x}) = 1-\eta(\boldsymbol{x})$, we specify the distribution of $Y$. This problem is called the logistic model, [Gir14, Section 11.1.3]. Without loss of generality, we assume $\|\Lambda^{-1}\boldsymbol{\mu}\|_2 > 1$.*

It is straightforward to verify that in both models, the Bayes classifier $f^*(\cdot) = \operatorname{sign}(\langle \Lambda^{-1}\boldsymbol{\mu}, \cdot\rangle)$ is "collinear" to $\langle \Lambda^{-1}\boldsymbol{\mu}, \cdot\rangle$. Consequently, the optimal linear classifier $\boldsymbol{\beta}^*$ mentioned in Section 1.5.1 can be taken as $\Lambda^{-1}\boldsymbol{\mu}$ (up to a positive multiplicative constant). When selecting $V_J = \operatorname{span}(\boldsymbol{\beta}^*)$, $\boldsymbol{\beta}^*_J$ becomes collinear with $\boldsymbol{\beta}^*$ in the same direction, in which case (1.18) equals 0, see also Section 4.9.3.

For the classification problem, our main results rely on a local Bernstein condition. Here, the loss function is taken to be the squared hinge loss, where $\boldsymbol{\beta}^*_J$ serves as the oracle for this loss function, as defined in (4.5).

**Assumption 7** (Local Bernstein's Condition)**.** *There exist absolute constants $\kappa_2 \geq 1$ and $L_1 > 0$ and parameters $\rho, r(\rho) > 0$ defined in (4.67), such that for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}^*_J + r(\rho)\Sigma_J^{-1/2} S_2^J \cap \rho B_2^J$, we have*

$$P\mathcal{L}_{\boldsymbol{\beta}_J} \geq L_1 \left\| \Sigma_J^{1/2}(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J) \right\|_2^{2\kappa_2} \tag{4.10}$$

*where $\ell_{\boldsymbol{\beta}_J}(X,Y) = (1 - Y\langle X, \boldsymbol{\beta}_J\rangle)_+^2$ and $\mathcal{L}_{\boldsymbol{\beta}_J} = \ell_{\boldsymbol{\beta}_J} - \ell_{\boldsymbol{\beta}^*_J}$.*

We verify this condition in Lemma 28. In classification problems, we similarly define an interpolation norm $|||\cdot||| = \sup(\langle \boldsymbol{\beta}, \boldsymbol{u}\rangle : \boldsymbol{u} \in \frac{\rho}{r(\rho)} B_2^J \cap \Sigma_J^{-1/2} B_2^J)$, where $\rho$ and $r(\rho)$ will be defined in (4.67). The proof of the following Theorem 11 may be found in Section 4.8.

**Theorem 11.** *Grant Assumption 7 with constants $L_1 > 0$ and $\kappa_2 = 1$. There exist absolute constants $\kappa_{RIP}, \kappa_{DM} < 1$ such that the following holds. Suppose $XY$ is a centered sub-Gaussian random vector. Let $0 < \delta_4 < 1$ and define*

$$r(V_J, V_{J^c}) = L_1^{-1} P\ell_{\boldsymbol{\beta}^*_J} \sqrt{\frac{\dim(V_J)}{N}} + L_1^{-1} \frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J||| + L_1^{-1}\delta_4 P\ell^{\frac{1}{2}}_{\boldsymbol{\beta}^*_J}. \tag{4.11}$$

*For any $0 < \bar{\delta} < 1$ depending on $\delta_4$ (see (1.29) for explicit dependence), suppose the decomposition $\mathbb{R}^p = V_J \oplus V_{J^c}$ satisfies that*

$$\kappa_{RIP}^{-1} \dim(V_J) \leq N \leq \kappa_{DM}\bar{\delta}^2 \frac{\operatorname{Tr}(\Sigma_{J^c})}{\|\Sigma_{J^c}\|_{op}}. \tag{4.12}$$

*Then there exist absolute constants $c_{17}$, $c_{18} < 1$, $C_{32}$, $C_{33}$, $C_{34} > 1$ such that with probability at least*

$$1 - \bar{p}_{DM}(\delta_4) - c_{17}\frac{\log^4(N)}{N} - \exp(-C_{32}\dim(V_J)) - \exp\left(-c_{18}\frac{N}{\max\{\|1-\langle Y_iX_i, \boldsymbol{\beta}^*_J\rangle\|_{\psi_2}^2, \|1-\langle Y_iX_i, \boldsymbol{\beta}^*_J\rangle\|_{\psi_2}^4\}}\right), \tag{4.13}$$

*we have $\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J)\right\|_2 \leq C_{33} r(V_J, V_{J^c})$, and $P\mathcal{L}_{\hat{\boldsymbol{\beta}}_J} \leq C_{34}(r(V_J, V_{J^c}))^2$ where $P\mathcal{L}_{\hat{\boldsymbol{\beta}}_J}$ is the excess risk with respect to the squared hinge loss defined in (4.4) and $\bar{p}_{DM}(\delta_4)$ may be found in Theorem 4.*

*Moreover, Assumption 7 is verified with some $L_1 > 0$, $\kappa_2 = 1$ for both the Gaussian mixture classification model and the logistic model, if $V_J = \operatorname{span}(\Lambda^{-1}\boldsymbol{\mu})$. Suppose (4.12) holds, $\frac{1}{2} < \|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2$ and $r(V_J, V_{J^c}) < \frac{1}{4C_{33}}(4\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 + 1 - \sqrt{16\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 + 1})$. Then there exists an absolute constant $C_{35} > 1$ such that for any*

$$\frac{N\|\Sigma_{J^c}\|_{op}}{\operatorname{Tr}(\Sigma_{J^c})} \leq \frac{(1-\delta_4)^2}{C_{35}} \frac{(\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 - C_{33}r(V_J, V_{J^c}))^2 - \frac{1}{2}(\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 + C_{33}r(V_J, V_{J^c}))}{P\ell_{\boldsymbol{\beta}^*_J}}, \tag{4.14}$$

*under the same probability, there exist absolute constants* $C_{36}$, $C_{37}$ *and* $C_{38} > 1$, *such that we have the* $P\mathcal{L}_{\hat{\boldsymbol{\beta}}}^{\{0,1\}} \leq$ (1.16) + (1.17) + (1.18), *where for*

1. *gaussian mixture classification model,*

$$(1.16) \leq C_{38}\sqrt{\frac{N\|\Sigma_{J^c}\|_{op}}{\mathrm{Tr}(\Sigma_{J^c})}P\ell_{\boldsymbol{\beta}_J^*}},\ (1.17) \leq C_{36}r^{\frac{2}{3}}(V_J, V_{J^c}),\ \ \textit{and}\ (1.18) = 0,$$

2. *and for logistic model,*

$$(1.16) \leq C_{38}\sqrt{\frac{N\|\Sigma_{J^c}\|_{op}}{\mathrm{Tr}(\Sigma_{J^c})}P\ell_{\boldsymbol{\beta}_J^*}},\ (1.17) \leq C_{37}\|\Lambda^{-1/2}\boldsymbol{\mu}\|_2^{-3}r^2(V_J, V_{J^c}),\ \ \textit{and}\ (1.18) = 0.$$

Condition (4.14) is of a technical nature. In general classification problems, the quantity $\|\Sigma^{1/2}\boldsymbol{\beta}_J^*\|_2$ is at least of constant order. Hence, when $N$ is sufficiently large, there always exists a value of $\bar{\delta}$ for which (4.14) holds.

We note that in some literature, the study of benign overfitting for the minimum norm interpolant classifier concerns the case where the interpolant classifier coincides with the minimum norm interpolant regression solution, i.e., when $\hat{\boldsymbol{\beta}} \in \mathrm{argmin}(\|\boldsymbol{\beta}\|_q : \mathbb{X}\boldsymbol{\beta} = \boldsymbol{y})$. This corresponds to a regression problem with response vector in $\{-1, 1\}^N$, as in [CGB22, TCFB25]. For convenience, we refer to this as the *proliferated classifier*. The hard margin support vectors machine studied in this chapter and the proliferated classifier exhibit different forms of self-regularization. Indeed, as observed in the regression setting, the proliferated classifier is self-regularized by the ridge penalty, [LR21], rather than the squared hinge loss, and hence these are fundamentally different types of classifiers. While the proliferated classifier coincides with the hard margin support vectors machine under certain special conditions, [ASH21, HMX21], the conclusions of this chapter and those of [CGB22, TCFB25]—though sharing similar terms—are essentially distinct in nature and require different analysis.

## 4.3 Self-regularization: $\hat{\boldsymbol{\beta}}_J$ is a regularized estimator

The FSD assigns distinct roles to the two projections of $\hat{\boldsymbol{\beta}}$. The aforementioned decomposition holds for arbitrary estimators. We now apply it specifically to minimum norm interpolant estimators. In this section, we investigate the estimation properties of $\hat{\boldsymbol{\beta}}_J$. We present the key technique for proving the main results – the self-regularization argument. We emphasize that the development of the features space decomposition and the self-regularization argument constitutes an important contribution of this chapter, in addition to the main results.

### 4.3.1 Self-regularization: identify $\hat{\boldsymbol{\beta}}_J$ as a regularized, generalized M-estimator.

In the current paragraph, we will establish the relationship between these two projections through self-regularization. First, we establish the notational conventions to be used throughout. Let $J \subset [p]$ be a set of indices, and $V_J = \mathrm{span}(\boldsymbol{e}_j : j \in J)$, where $\boldsymbol{e}_1, \cdots, \boldsymbol{e}_p$ is the canonical basis. Define $\mathbb{X}_J = [P_J X_1|\cdots|P_J X_N]^\top : \boldsymbol{v} \in \mathbb{R}^p \mapsto (\langle P_J X_i, \boldsymbol{v}\rangle)_{i=1}^N \in \mathbb{R}^N$ and $\mathbb{X}_{J^c} = [P_{J^c} X_1|\cdots|P_{J^c} X_N]^\top : \boldsymbol{v} \in \mathbb{R}^p \mapsto (\langle P_{J^c} X_i, \boldsymbol{v}\rangle)_{i=1}^N \in \mathbb{R}^N$. Define $\mathbb{X}_{\boldsymbol{y}} = [Y_1 X_1|\cdots|Y_N X_N]^\top : \boldsymbol{\beta} \in \mathbb{R}^p \mapsto (Y_i\langle X_i, \boldsymbol{\beta}\rangle)_{i=1}^N \in \mathbb{R}^N$. We equip a partial order on $\mathbb{R}^N$ by $\boldsymbol{a} \succeq \boldsymbol{b}$ if and only if $a_i \geq b_i$ for any $i \in [N]$ where $(a_i)_i, (b_i)_i$ are coordinates of $\boldsymbol{a}, \boldsymbol{b}$. Let $q'$ be the conjugate index of $q$, that is, $q'$ such that $1/q' + 1/q = 1$.

**In the linear regression model**. The following non-linear operator plays a key role in our analysis. Define

$$\mathcal{A} : \boldsymbol{\mu} \in \mathbb{R}^N \mapsto \mathcal{A}[\boldsymbol{\mu}] \in \operatorname*{argmin}_{\boldsymbol{\nu} \in \mathbb{R}^p} \left(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{\mu}\right). \tag{4.15}$$

By standard duality argument, $\|\mathcal{A}[\boldsymbol{\mu}]\|_q = \min(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{\mu}) = \max(\langle\boldsymbol{\lambda}, \boldsymbol{\mu}\rangle : \|\mathbb{X}_{J^c}^\top\boldsymbol{\lambda}\|_{q'} \leq 1)$. Then by (4.2) we know that $\hat{\boldsymbol{\beta}}_{J^c} = \mathrm{argmin}_{\boldsymbol{\beta} \in V_{J^c}}(\|\boldsymbol{\beta}\|_q : \mathbb{X}_{J^c}\boldsymbol{\beta} = \boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J) = \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]$. Therefore,

$$\hat{\boldsymbol{\beta}}_J = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left(\|\boldsymbol{\beta}_J\|_q^q + \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q\right). \tag{4.16}$$

We emphasize that the decomposition in (4.16) must be aligned with the fixed canonical basis $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_p\}$ of $\mathbb{R}^p$. We denote $P_N\ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$ as the empirical risk of this random loss function—termed "random" because the loss function $\|\mathcal{A}[\cdot]\|_q^q$ depends on the random matrix $\mathbb{X}_{J^c}$. Let $P\ell_{\boldsymbol{\beta}_J} = \mathbb{E}_{\mathbb{X}_J, \boldsymbol{\xi}}\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$ denote the conditional

expectation given $\mathbb{X}_{J^c}$. This definition is justified since $\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J$ represents the residual of $\boldsymbol{\beta}_J$ with respect to the response vector $\boldsymbol{y}$. As shown later in Lemma 21, $\boldsymbol{\beta}^*_J$ precisely minimizes $P\ell_{\boldsymbol{\beta}_J}$—making it the oracle for this random loss function. Thus, statistically speaking, (4.16) demonstrates that $\hat{\boldsymbol{\beta}}_J$ is a regularized estimator for the regression problem where:

- $\boldsymbol{\beta}^*_J$ serves as the signal,
- $\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}^*_J = \mathbb{X}_{J^c}\boldsymbol{\beta}^*_{J^c} + \boldsymbol{\xi}$ constitutes the noise, and
- $P_N\ell_{\boldsymbol{\beta}_J} : (\boldsymbol{y}, \mathbb{X}_J) \to \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$ is the empirical loss function.

**In the classification model**. Let $\mathbb{1} = (1, 1, \cdots, 1) \in \mathbb{R}^N$, and $\mathbb{X}_{\boldsymbol{y},J^c} = [Y_1P_{J^c}X_1|\cdots|Y_NP_{J^c}X_N]^\top$. We have that (4.3) is equivalent to

$$\hat{\boldsymbol{\beta}} \in \operatorname{argmin}\left(\|\boldsymbol{\beta}\|_2 : \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta} \succeq \mathbb{1}\right) = \operatorname{argmin}\left(\|\boldsymbol{\beta}\|_2^2 : \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta} \succeq \mathbb{1}\right). \tag{4.17}$$

Define

$$\mathcal{B} : \boldsymbol{\mu} \in \mathbb{R}^N \mapsto \mathcal{B}[\boldsymbol{\mu}] \in \operatorname*{argmin}_{\boldsymbol{\nu}\in\mathbb{R}^p}\left(\|\boldsymbol{\nu}\|_2 : \mathbb{X}_{\boldsymbol{y},J^c}\boldsymbol{\nu} \succeq \boldsymbol{\mu}\right). \tag{4.18}$$

It is straightforward to see that

$$\hat{\boldsymbol{\beta}}_{J^c} \in \operatorname{argmin}\left(\|\boldsymbol{\beta}_{J^c}\|_2 : \mathbb{X}_{\boldsymbol{y},J^c}\boldsymbol{\beta}_{J^c} \succeq \mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\hat{\boldsymbol{\beta}}_J\right) = \mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\hat{\boldsymbol{\beta}}_J]. \tag{4.19}$$

Moreover, since $\|\hat{\boldsymbol{\beta}}\|_2^2 = \|\hat{\boldsymbol{\beta}}_J\|_2^2 + \|\hat{\boldsymbol{\beta}}_{J^c}\|_2^2$ and $\hat{\boldsymbol{\beta}}$ has the minimum $\|.\|_2$ norm, we have

$$\hat{\boldsymbol{\beta}}_J \in \operatorname{argmin} L(\boldsymbol{\beta}_J), \text{ where } L(\boldsymbol{\beta}_J) = \|\boldsymbol{\beta}_J\|_2^2 + \|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y},J}\boldsymbol{\beta}_J]\|_2^2. \tag{4.20}$$

We emphasize that, due to the rotational invariance of the $\ell_2$ norm, (4.20) does not need to be aligned with any canonical basis. If we wish to handle more general $\ell_q$ norm, we need to choose $V_J$ to be well aligned with the basis associated with these norms, since such norms are not basis-independent. In this chapter, we restrict our attention to the case $q = 2$.

In other words, $\hat{\boldsymbol{\beta}}_J$ can be defined as an estimator that minimizes the random loss function $\ell_{\boldsymbol{\beta}_J} : (\boldsymbol{y}, \mathbb{X}_{\boldsymbol{y}}) \to \|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]\|_2^2$ with the regularization term $\|\boldsymbol{\beta}_J\|_2^2$—this is what we refer to as self-regularization. Compared to the regression problem, identifying the oracle proves challenging in this setting. However, as we shall demonstrate in Section 4.3.3, thanks to the Dvoretzky-Milman theorem that will be introduced in the next section, the isometric profile of this loss function can be characterized by the squared hinge loss.

Before concluding this section, let us emphasize that self-regularization and the FSD framework can also handle the minimum $\|\cdot\|$-norm interpolant estimator defined with respect to general norms. Specifically, when there exist normed spaces $(V_J, |||\cdot|||)$ and $(V_{J^c}, ||||\cdot||||)$ such that $(\mathbb{R}^p, \|\cdot\|) = (V_J, |||\cdot|||) \oplus (V_{J^c}, ||||\cdot||||)$, that is, when the feature space admits a direct-sum decomposition in the sense of normed spaces, we can treat this setting as a block decomposition.

### 4.3.2 Dvoretzky-Milman theorem.

Up to this point, (4.16) and (4.20) are merely reformulation, since this loss function is defined via the highly complex random nonlinear maps $\mathcal{A}$ and $\mathcal{B}$.

In this subsection, we introduce the Dvoretzky-Milman theorem, which allows us to understand this stochastic nonlinear loss function. Below we present the standard Dvoretzky-Milman theorem, along with its isometric extension to general probability measures, which are respectively applied in regression and classification problems. More specifically, we need to apply the Dvoretzky-Milman theorem to approximately solve the two convex optimization problems defined by $\|\mathcal{A}[\cdot]\|_q$ and $\|\mathcal{B}[\cdot]\|_2$ – by characterizing an isomorphy of the norm $\|\mathbb{X}_{J^c}^\top \cdot\|_{q'}$ ($\|\mathbb{X}_{\boldsymbol{y}}^\top P_{J^c} \cdot\|_2$ respectively) to simplify the feasible set of this optimization problem.

**Dvoretzky-Milman theorem.** Below is Milman's version of Dvoretzky's theorem; see [Pis89].

**Theorem 2 (recall).** *There are absolute constants $\kappa_{DM} \le 1$ and $c_1$ such that the following holds. Let $||||\cdot||||$ be some norm on $\mathbb{R}^p$ and denote by $B$ its unit ball. Denote by $\mathbb{G} := \mathbb{G}^{(N\times p)}$, the $N \times p$ standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ Gaussian entries. Given any $0 < \varepsilon_1 \le 1$. Assume that $N \le \kappa_{DM}\varepsilon_1^2 d_*(B)$. Then with probability at least $1 - \exp(-c_1\varepsilon_1^2 d_*(B))$, for every $\boldsymbol{\lambda} \in \mathbb{R}^N$,*

$$(1 - \varepsilon_1)\|\boldsymbol{\lambda}\|_2 \ell_*(B^*) \le ||||\mathbb{G}^\top\boldsymbol{\lambda}|||| \le (1 + \varepsilon_1)\|\boldsymbol{\lambda}\|_2 \ell_*(B^*). \tag{1.21}$$

For all $0 < \varepsilon_1 < 1$, we define the event

$$\Omega_{\mathrm{DM,reg}}(\varepsilon_1) := \left\{\forall \boldsymbol{\lambda} \in \mathbb{R}^N : \|\boldsymbol{\lambda}\|_2 (1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p) \leq \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \leq \|\boldsymbol{\lambda}\|_2 (1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)\right\} \tag{1.22}$$

$$\subset \left\{\forall \boldsymbol{\mu} \in \mathbb{R}^N : \frac{\|\boldsymbol{\mu}\|_2}{(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)} \leq \|\mathcal{A}[\boldsymbol{\mu}]\|_q \leq \frac{\|\boldsymbol{\mu}\|_2}{(1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2} B_q^p)}\right\}. \tag{1.23}$$

It follows from Theorem 2 applied to the norm $\|\!|\!|\cdot|\!|\!| = \|\Sigma_{J^c}^{1/2} \cdot\|_{q'}$ that, if $X_{J^c}$ is a Gaussian random vector and $\kappa_{DM}\varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p) \geq N$, then $\mathbb{P}(\Omega_{\mathrm{DM,reg}}(\varepsilon_1)) \geq 1 - \exp(-c_1\varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2} B_{q'}^p))$. The inclusion from (1.23) follows from strong duality: for all $\boldsymbol{\mu} \in \mathbb{R}^N$,

$$\|\mathcal{A}[\boldsymbol{\mu}]\|_q = \min\left(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}^\top \boldsymbol{\nu} = \boldsymbol{\mu}\right) = \max\left(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle : \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \leq 1\right). \tag{1.24}$$

Even though $\mathcal{A} : (\mathbb{R}^N, \ell_2) \to (V_{J^c}, \ell_q)$ is a non-linear metric embedding (except when $q = 2$), it satisfies a DM theorem inherited from $\mathbb{X}_{J^c}^\top$. Since our loss functions in the estimation part of the features space depend on $\mathcal{A}$ in the regression problem, working on the event $\Omega_{\mathrm{DM,reg}}(\varepsilon_1)$ will allow us to greatly simplify its expression because now it is isomorphic to the $\ell_2^N$-norm and so we will work with the classical squared loss function. That is the reason why DM theorem plays a crucial role in our analysis: we use this isomorphic property from DM to greatly simplify the loss function appearing in $V_J$ and then go back to the classical analysis of regularized ERM with respect to the squared loss on $V_J$. Of course, if one wants to go beyond the Gaussian design case, one needs to extend DM theorem beyond that case.

**Dvoretzky-Milman theorem for $\ell_{q'}$ norm under general probability measure assumptions.** Theorem 2 provides the Dvoretzky–Milman theorem for Gaussian measures. Because we need to study the case where $X_{J^c}$ is distributed according to a general probability measure, we require an extension of the Dvoretzky–Milman theorem for $\|\cdot\|_{q'}$-norms. Extensions of the Dvoretzky–Milman theorem to general probability measures already exist in a substantial body of literature, e.g., [GLPTJ07, MTJ08, BM22a, BM22b, Men22, BM24]. In these works the random embedding operator is usually induced by row-independent random matrices or by more complex random-matrix models; however, in $\Omega_{\mathrm{DM,reg}}(\varepsilon)$ we need a column-independent random-matrix model. Hence, an entirely new Dvoretzky–Milman theorem for such random matrices is required. The following theorem is a contribution to GAFA that was motivated precisely by the FSD method. Its proof may be found in Section 5.1.

**Assumption 8.** *$\boldsymbol{\zeta} = (\zeta_j)_{j=1}^p$ is a centered, isotropic random vector in $\mathbb{R}^p$ with i.i.d. coordinates, satisfying $\mathbb{E}[\zeta_1^2] = 1$, and there exist absolute constants $0 < \kappa \leq 1$ and $\varepsilon > 0$ such that $\mathbb{E}|\zeta_1|^{\max\{4,2q+\varepsilon\}} \leq \kappa^{\max\{4,2q+\varepsilon\}}$.*

**Theorem 3 (recall).** *Let $\boldsymbol{\zeta}$ be a random vector satisfying Assumption 8, and let $\Sigma$ be a positive definite diagonal matrix on $\mathbb{R}^p$, $\Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_p)$. Let $X = \Sigma^{1/2}\boldsymbol{\zeta}$, and let $X_1, \cdots, X_N$ be independent copies of $X$, forming the random matrix $\mathbb{X} = [X_1|\cdots|X_N]^\top = [Z_1|\cdots|Z_p]$, where $(Z_j)_{j=1}^p$ are the column vectors of $\mathbb{X}$. Denote $\ell_* = \ell_*(\Sigma^{1/2} B_q^p)$ and $d_* = d_*(\Sigma^{-1/2} B_{q'}^p)$. Without loss of generality, assume that $d_* \geq 1$. There then exists an absolute constant $0 < \theta < 1$ such that for any $\boldsymbol{\lambda} \in S_2^{N-1}$, $\mathbb{P}(|\langle Z_j, \boldsymbol{\lambda}\rangle| \geq \theta) \geq \kappa$. Moreover, there exist absolute constants $c, c', C, C', C'', \kappa_{DM}, \varepsilon_0 > 0$ such that the following facts hold.*

1. *When $q \geq 2$. If $N \leq \kappa_{DM} d_* \mathrm{Log}^{-2}(p^{\frac{1}{q'}}/d_*)$, then with probability at least*

$$1 - C'\mathrm{Log}\left(\frac{p^{\frac{1}{q'}}}{d_*}\right)\exp\left(-C''\kappa_{DM}\frac{d_*^\theta}{\mathrm{Log}^{2\theta}\left(\frac{p^{\frac{1}{q'}}}{d_*}\right)}\right) - 2\exp\left(-C' d_*\right) - C' d_*^{-c\min\{\varepsilon,\varepsilon_0\}} =: 1 - \bar{p}_{DM},$$

*there holds for any $\boldsymbol{\lambda} \in S_2^{N-1}$,*

$$c\ell_* \leq \left\|\mathbb{X}^\top \boldsymbol{\lambda}\right\|_{q'} \leq C\mathrm{Log}(p)\ell_*.$$

2. *When $q < 2$. If $N \leq \kappa_{DM} d_*(\Sigma^{-1/2} B_{q'}^p)$, then*

$$\mathbb{P}\left(\forall \boldsymbol{\lambda} \in S_2^{N-1}, c\ell_* \leq \|\mathbb{X}^\top \boldsymbol{\lambda}\|_{q'} \leq C\ell_*\right) \geq 1 - 3\exp(-c' d_*) - C' d_*^{-\frac{q'-2}{4}} =: 1 - \bar{p}_{DM}.$$

Theorem 3 establishes that, under Assumption 8, the linear span of $N$ independent copies of $\Sigma^{1/2}\boldsymbol{\zeta}$ provides a generalization (up to a logarithmic factor when $q \geq 2$) of the Dvoretzky-Milman theorem for the convex body $B_{q'}^p$ under a general probability measure. We emphasize here that if one focuses solely on the sub-Gaussian case, then for $q \geq 2$, the $\mathrm{Log}(p)$ factor in the uniform upper bound for $\|\mathbb{X}^\top \boldsymbol{\lambda}\|_{q'}$ can be removed. To the best of our knowledge, this theorem is the first generalization of the Dvoretzky-Milman theorem for the $\|\Sigma^{1/2} \cdot\|_{q'}$ norm under such broad (almost the most general) conditions.

**Isometric Dvoretzky-Milman theorem for ellipsoid norm.** For the case $|||\cdot|||$ being a Hilbert norm, [P2] established that the isometric Dvoretzky-Milman theorem still holds for random vectors satisfying the concentration property of $\|X_{J^c}\|_2$ and an $L^{4+\varepsilon} - L^2$ equivalence condition on the marginal distributions. Notably, this allows for significant dependencies among the coordinates of $X_{J^c}$, and this is precisely what is required for the classification problem. Here we employ only the weak sub-Gaussian condition form of the conclusion from [P2]. We have the following proposition, whose proof can be found in Section 4.8.1.

**Proposition 26.** *Let $YX_{J^c}$ be a centered sub-Gaussian random vector, and suppose there exists $0 < \bar{\delta} < 1$ such that $N \leq \kappa_{DM}\bar{\delta}^2 \frac{\mathrm{Tr}(\Sigma_{J^c})}{\|\Sigma_{J^c}\|_{op}}$, then with probability at least $1 - \bar{p}_{DM}(\delta_4)$ (where $\delta_4$ is defined in (1.29) and $\bar{p}_{DM}$ may be found in Theorem 4 later), the following event happens:*

$$\Omega_{DM,class}(\delta_4) := \left\{\forall \boldsymbol{\lambda} \in \mathbb{R}^N : \|\boldsymbol{\lambda}\|_2 (1-\delta_4)\sqrt{\mathrm{Tr}(\Sigma_{J^c})} \leq \left\|\mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}\right\|_2 \leq \|\boldsymbol{\lambda}\|_2 (1+\delta_4)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}\right\} \tag{1.31}$$

$$\subseteq \left\{\forall \boldsymbol{\mu} \in \mathbb{R}^N : \frac{\|[\boldsymbol{\mu}]_+\|_2}{(1+\delta_4)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}} \leq \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \leq \frac{\|[\boldsymbol{\mu}]_+\|_2}{(1-\delta_4)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}}\right\}, \tag{1.32}$$

*where $[\boldsymbol{\mu}]_+ = (\max(\mu_i, 0))_{i=1}^N$, and $\mathbb{X}_{\boldsymbol{y},J^c}^\top = [P_{J^c}X_1Y_1|\cdots|P_{J^c}X_NY_N]$.*

It is worth noting that the passage from (1.31) to (1.32) can be made. We include it below. We include it below. By standard duality argument, see, for instance, [BV14, Equation 5.11], we obtain that

$$\|\mathcal{B}[\boldsymbol{\mu}]\|_2 = \max\left(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle : \boldsymbol{\lambda} \succeq \mathbf{0},\ \left\|\mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}\right\|_2 \leq 1\right).$$

Condition on $\Omega_{\mathrm{DM,class}}(\delta_4)$, see (1.31), we have

$$\max_{\boldsymbol{\lambda}\succeq\mathbf{0}}\left(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle : \|\boldsymbol{\lambda}\|_2 \leq \frac{1}{(1+\delta_4)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}}\right) \leq \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \leq \max_{\boldsymbol{\lambda}\succeq\mathbf{0}}\left(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle : \|\boldsymbol{\lambda}\|_2 \leq \frac{1}{(1-\delta_4)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}}\right).$$

Let $H(\boldsymbol{\mu}) := \{i \in [N] : \mu_i < 0\}$ and let $\boldsymbol{\lambda}^-$ be the maximizer of the left-hand-side maximization problem and $\boldsymbol{\lambda}^+$ be the maximizer of the right-hand-side maximization problem. We prove that if $i \in H(\boldsymbol{\mu})$, then $\lambda_i^- = 0$. We prove this by contradiction. Suppose $i \in H(\boldsymbol{\mu})$ but $\lambda_i^- > 0$, then by letting $\tilde{\boldsymbol{\lambda}}^- = (\lambda_1^-, \cdots, \lambda_{i-1}^-, 0, \lambda_{i+1}^-, \cdots, \lambda_N^-)$, we know that $\|\tilde{\boldsymbol{\lambda}}^-\|_2 < \|\boldsymbol{\lambda}^-\|_2 \leq \frac{1}{(1+\delta_4)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}}$. Moreover, $\langle \boldsymbol{\mu}, \tilde{\boldsymbol{\lambda}}^-\rangle = \sum_{i'\neq i}\mu_{i'}\lambda_{i'}^- > \sum_{i'=1}^N \mu_{i'}\lambda_{i'}^- = \langle \boldsymbol{\mu}, \boldsymbol{\lambda}^-\rangle$ since $\mu_i\lambda_i^- < 0$. This implies that $\tilde{\boldsymbol{\lambda}}^-$ is a feasible solution but with larger objective function value, hence contradicting the assumption that $\boldsymbol{\lambda}^-$ is the maximizer. Recalling the constraint that $\boldsymbol{\lambda} \succeq \mathbf{0}$, we have: for any $i \in H(\boldsymbol{\mu})$, we necessarily have $\lambda_i^- = 0$. The same also holds for $\boldsymbol{\lambda}^+$. Now, by Cauchy-Schwartz, we have $\boldsymbol{\lambda}^- = (\boldsymbol{\mu}/(\|\boldsymbol{\mu}\|_2(1+\delta_4)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}))_+$, and $\boldsymbol{\lambda}^+ = (\boldsymbol{\mu}/(\|\boldsymbol{\mu}\|_2(1-\delta_4)\sqrt{\mathrm{Tr}(\Sigma_{J^c})}))_+$. Therefore, condition on $\Omega_{\mathrm{DM,class}}(\delta_4)$, (1.32) follows.

On the event $\Omega_{\mathrm{DM,class}}(\delta_4)$, the a priori complicated loss function $(\boldsymbol{y}, \mathbb{X}_J) \to \|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]\|_2^2$ appearing in the estimation space $V_J$ will be greatly simplified since it will behave as the square hinge loss. As a consequence, the analysis of the estimator on the estimation part of the features space will boil down to the study of a regularized ERM with respect to the square hinge loss function (see more detail in the section below).

### 4.3.3 Identifying $\hat{\boldsymbol{\beta}}_J$ as a regularized ERM

We now employ the Dvoretzky-Milman theorem to analyze the stochastic nonlinear loss functions $\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$ and $\|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]\|_2^2$. In this section, we demonstrate the fruitful consequences yielded by combining the Dvoretzky-Milman theorem with features space decomposition – we show that $\hat{\boldsymbol{\beta}}_J$ can be identified as classical estimators in both regression and classification problems, including: squared hinge loss support vector machine; square-root LASSO; and $\|\cdot\|_q^q$-regularized M-estimators.

**A $\ell_2$-self-regularized ERM w.r.t. the squared hinge loss in classification**

In this section, we establish the self-regularization property of $\hat{\boldsymbol{\beta}}_J$ defined by (4.20) by using Dvoretzky-Milman theorem. Applying (1.32) to $\boldsymbol{\mu} = \mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J$, we obtain that $L(\boldsymbol{\beta}_J)$ defined by (4.20) satisfies that: for any $\boldsymbol{\beta}_J \in V_J$,

$$\|\boldsymbol{\beta}_J\|_2^2 + \frac{\|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]_+\|_2^2}{(1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c})} \le L(\boldsymbol{\beta}_J) \le \|\boldsymbol{\beta}_J\|_2^2 + \frac{\|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]_+\|_2^2}{(1-\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c})}. \tag{4.21}$$

Recalling that $\|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]_+\|_2^2 = \sum_{i=1}^N (1 - Y_i\langle X_i, \boldsymbol{\beta}_J\rangle)_+^2$, we know that the estimator $\hat{\boldsymbol{\beta}}_J$ defined in (4.20) is almost a support vectors machine with squared hinge loss, and tuning parameter of the order of $\operatorname{Tr}(\Sigma_{J^c})/N$. To the best of our knowledge, the connection between the minimum $\ell_2$-norm interpolant classifier and the squared hinge loss was first noted by [Sha22]. However, [Sha22] only obtained asymptotic results. [ZKS+22] also investigated benign overfitting for the squared hinge loss, but did not establish its connection with the minimum $\ell_2$-norm interpolant classifier. Leveraging the Dvoretzky-Milman theorem, we reveal the statistical significance of this relationship through the isometric profile (4.21) of empirical risk $\|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]\|_2^2$, see [Lec11, pp. 41] for the definition of isometric profile.

Equation (4.21) establishes the isometric profile of the regularized empirical risk (rather than excess risk) for the loss function $\|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\cdot]\|_2^2$. This explains why we obtain a non-exact oracle inequality. However, we emphasize that this non-exact oracle inequality stems from the distortion $\delta_4$ in the Dvoretzky-Milman theorem. For the minimum $\ell_2$-norm interpolant classifier, we have access to the isometric Dvoretzky-Milman theorem - specifically, we can make $\delta_4$ in $\Omega_{\mathrm{DM,class}}(\delta_4)$ arbitrarily close to 0 thanks to the result from [P2]. Consequently, we can derive an exact oracle inequality from the non-exact version, thereby obtaining benign overfitting from non-exact benign overfitting. Nevertheless, we conjecture this approach may not be optimal. We hypothesize that a better method would be to directly investigate the isomorphic profile of the excess risk for the loss function $\|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\cdot]\|_2^2$ itself without going through the squared hinge loss as in (4.21).

**A $\ell_2$-self-regularized ERM w.r.t. the squared loss in linear regression**

Similarly, by applying (1.22), we can show that $\hat{\boldsymbol{\beta}}_J$ defined in (4.16) can be identified as a (generalized) M-estimator $\operatorname{argmin}(\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J\|_2^q + \ell_*(\Sigma_{J^c}^{1/2} B_q^p)^q \|\boldsymbol{\beta}_J\|_q^q)$. In particular, when $q = 1$, $\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_1 \sim \|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}\|_2 \ell_*$. Consequently, in this case $\hat{\boldsymbol{\beta}}_J$ can be recognized as a square-root LASSO estimator, meaning it is "almost" the solution to the following problem: $\operatorname{argmin}(\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J\|_2 + \ell_*\|\boldsymbol{\beta}_J\|_1)$.

### 4.3.4 Uniform convergence on the low-dimensional subspace $V_J$ in regression problem

In the previous subsection, owing to the Dvoretzky-Milman theorem, we have established an understanding of the statistical problem defined by $\hat{\boldsymbol{\beta}}_J$ in both regression (4.16) and classification (4.20) settings. In the current section, we derive the estimation error of $\hat{\boldsymbol{\beta}}_J$ in regression problem, that is, $\hat{\boldsymbol{\beta}}_J$ defined in (4.16). Our fundamental approach is the uniform convergence argument on $V_J$. Let $\rho, r(\rho) > 0$ to be determined later. Recall that in regression problem, $P\ell_{\boldsymbol{\beta}_J} = \mathbb{E}_{\mathbb{X}_J,\boldsymbol{\xi}}\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$ and $P_N\ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$. Let $P\mathcal{L}_{\boldsymbol{\beta}_J} = P\ell_{\boldsymbol{\beta}_J} - P\ell_{\boldsymbol{\beta}_J^*}$ and $P_N\mathcal{L}_{\boldsymbol{\beta}_J} = P_N\ell_{\boldsymbol{\beta}_J} - P_N\ell_{\boldsymbol{\beta}_J^*}$. The crux of our analysis for the minimum $\ell_q$-norm interpolant estimator lies in applying the uniform convergence argument specifically to $\hat{\boldsymbol{\beta}}_J$ within the subspace $V_J$ (rather than to the full estimator $\hat{\boldsymbol{\beta}}$ across the entire feature space $\mathbb{R}^p$). We need to select two real numbers $\rho_*, r_*(\rho_*) > 0$ and prove that (with high probability) $\hat{\boldsymbol{\beta}}_J \in \boldsymbol{\beta}_J^* + (\rho_* K_{\mathrm{model}} \cap r_*(\rho_*)\Sigma_J^{-1/2} B_2^J)$. The uniform convergence argument employs the following strategy. From (4.16), if $P_N\mathcal{L}_{\boldsymbol{\beta}_J} + (\|\boldsymbol{\beta}_J\|_q^q - \|\boldsymbol{\beta}_J^*\|_q^q) > 0$ for all $\boldsymbol{\beta}_J \notin \boldsymbol{\beta}_J^* + (\rho_* K_{\mathrm{model}} \cap r_*(\rho_*)\Sigma_J^{-1/2} B_2^J)$ then it must hold that $\hat{\boldsymbol{\beta}}_J \in \boldsymbol{\beta}_J^* + (\rho_* K_{\mathrm{model}} \cap r_*(\rho_*)\Sigma_J^{-1/2} B_2^J)$. The aforementioned lower bound for $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ must hold uniformly over all $\boldsymbol{\beta}_J \notin \boldsymbol{\beta}_J^* + (\rho_* K_{\mathrm{model}} \cap r_*(\rho_*)\Sigma_J^{-1/2} B_2^J)$, hence termed the uniform convergence argument. In this problem, since $\ell_{\boldsymbol{\beta}_J}$ differs from $\ell_{\boldsymbol{\beta}_J}^{(2)}$ (the squared loss function), we additionally require a Bernstein-type property—namely, a lower bound for $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ involving $P_N\mathcal{L}_{\boldsymbol{\beta}_J}^{(2)}$. Finally, we will utilize the isomorphic property between $P_N\mathcal{L}_{\boldsymbol{\beta}_J}^{(2)} = \frac{1}{N}\|\mathbb{X}(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)\|_2^2$ and $P\mathcal{L}_{\boldsymbol{\beta}_J}^{(2)} = \|\Sigma_J^{1/2}(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)\|_2^2$ to obtain the estimation error. Up to this point, we have only applied uniform convergence to $\hat{\boldsymbol{\beta}}_J$ within $V_J$. Below we present the advantages of employing uniform convergence only on $V_J$ rather than on the entire space. We will deal later with $V_{J^c}$ onto which another argument is used, namely the DM theorem.

**FSD reduces the RIP complexity fixed point**

In this section, we investigate the isomorphy between $P_N\mathcal{L}^{(2)}_{\boldsymbol{\beta}_J}$ and $P\mathcal{L}^{(2)}_{\boldsymbol{\beta}_J}$. We define the following random event:

$$\Omega_{\text{RIP}} := \left\{ \forall \boldsymbol{\beta}_J \in \boldsymbol{\beta}^*_J + \left[ \rho K_{\text{model}} \cap r(\rho)\Sigma_J^{-1/2} S_2^{p-1} \right] : c_{19} r(\rho) \leq \frac{1}{\sqrt{N}} \|\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J)\|_2 \leq C_{39} r(\rho) \right\}, \tag{4.22}$$

where $c_{19} < 1$ and $C_{39} > 1$ are absolute constants, and $r(\rho) = C_{30} r(V_J, V_{J^c})$ where $r(V_J, V_{J^c})$ defined in (4.8) and $C_{30}$ is some absolute constant. The property characterized by the random event $\Omega_{\text{RIP}}$ is called the Restricted Isomorphic Property (RIP) in [P4] because it extends the classical RIP introduced in [CT05] to other subset of the $\ell_2$-sphere. This property establishes the concentration between the population excess risk $P\mathcal{L}^{(2)}_{\boldsymbol{\beta}} = \|\Sigma_J^{1/2}(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J)\|_2^2$ (for the squared loss) and its empirical counterpart $P_N\mathcal{L}^{(2)}_{\boldsymbol{\beta}} = \frac{1}{N}\|\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J)\|_2^2$ uniformly over a subset of a $\ell_2$-sphere centered at $\boldsymbol{\beta}^*_J$. For a given radius $\rho$, if $|J| \gtrsim N$, there exists a smallest real number $r_{\text{RIP}}(\rho) > 0$, called the RIP fixed point, such that for all $r(\rho) > r_{\text{RIP}}(\rho)$, $\Omega_{\text{RIP}}$ holds with high probability. If $|J| \leq \kappa_{RIP} N$ for some absolute constant $0 < \kappa_{RIP} < 1$, then $r_{\text{RIP}}(\rho) = 0$ (see Section 2.2 of [P4]). In the latter case, $\mathbb{X}_J$ behaves as an isometry on the entire sphere without restriction.

It is well-known that the RIP fixed point constitutes part of the upper bound for the excess risk of ERM, RERM, and their generalizations, [Men18, LM18, CLL20]. However, since we have the freedom to select $J$, we may choose $|J| \leq \kappa_{RIP} N$, thereby reducing the RIP fixed point on $V_J$ to 0 – meaning the RIP holds on the entire space $V_J$, which we sometimes refer to as the isomorphic property. In other words, the features space decomposition method enables us to obtain a smaller RIP complexity fixed point $r_{\text{RIP}}(\rho)$ and so a better convergence rate for $\hat{\boldsymbol{\beta}}_J$. However, the choice of $V_J$ such that $|J| \lesssim N$ may not be the optimal one. In that case, one may have $|J| \gtrsim N$ and so $\mathbb{X}_J$ is an isomorphy only on a restricted cone in $V_J$. We will not explore that case in this work but it is possible to do it and we refer the reader to [P4] where this analysis was done for the minimum $\ell_2$-norm interpolant estimator in linear regression.

This features space decomposition method holds not only for interpolant estimators. We emphasize that for ridge regression [P2] and more generally spectral algorithms (such as gradient descent/flow) [P3], the same method achieves an exact description of the excess risk (up to multiplicative constant) by eliminating the RIP fixed point when possible (that is when it is possible to choose $|J| \lesssim N$) or at least by reducing it (since it is a property of $\mathbb{X}$ on $V_J$ a smaller set than $\mathbb{R}^p$). We expect the features space decomposition method to be useful for the analysis of other type of estimators in particular when the cost of uniform convergence over the entire features space is too high, see other papers in this series [P2],[P3] for implementations of this idea.

**FSD reduces multiplier fixed point**

As mentioned in the beginning of Section 4.3.4, we need to establish a lower bound for $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ that incorporates $P_N\mathcal{L}^{(2)}_{\boldsymbol{\beta}_J}$. Here, the lower bound for $P_N\mathcal{L}_{\boldsymbol{\beta}_J} = P_N\ell_{\boldsymbol{\beta}_J} - P_N\ell_{\boldsymbol{\beta}^*_J}$ consists of two components, originating from the linear and quadratic terms in the Taylor expansion of $P_N\ell_\bullet$ at $\boldsymbol{\beta}^*_J$, where the linear term is referred to as the multiplier process, that is, $\langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J \rangle$ where $\boldsymbol{g} \in (\partial^- P_N\ell_\bullet)(\boldsymbol{\beta}^*_J)$ is a sub-gradient of $P_N\ell_\bullet$ evaluated at $\boldsymbol{\beta}^*_J$. We will compute this gradient in (4.29) of Lemma 21. First, we emphasize that this random loss function is a well-defined loss function, in the sense that $\boldsymbol{\beta}^*_J$ is a minimizer of its population risk, that is, $(\nabla P\ell_\bullet)(\boldsymbol{\beta}^*_J) = \mathbf{0}$. We will prove this claim in Lemma 21.

The multiplier fixed point is the fixed point used in supervised learning theory to control the first-order expansion, [LM17]. Roughly speaking, for any $0 < \delta < 1$ (typically $\delta > 1/2$), the multiplier fixed point $r_M(\rho, \delta)$ is defined as follows. Let $\theta_1 = \frac{1}{4}c_{23}$ for some absolute constant $c_{23}$ when $1 < q < 2$ and $\theta_1 = \frac{(q-1)c_{19}^q}{2q2^q(1+\varepsilon_1)^q}$ when $q \geq 2$ where $c_{19}$ is the absolute constant in $\Omega_{\text{RIP}}$, see (4.22). Let $\square = N^{\frac{q}{2}}\ell_*^{-q}(\Sigma^{1/2}_{J^c}B_q^p)$ when $q \geq 2$; and $\square = N^{\frac{q}{2}}\sigma_\xi^{q-2}\ell_*^{-q}(\Sigma^{1/2}_{J^c}B_q^p)$ when $1 < q < 2$. Define

$$r_M(\rho, \delta) = \min_{r>0} \left( \mathbb{P}\left( \sup_{\boldsymbol{u} \in \rho K_{\text{model}} \cap r\Sigma_J^{-1/2} B_2^J} \inf_{\boldsymbol{g} \in P_N\ell_\bullet(\boldsymbol{\beta}^*_J)} |\langle \boldsymbol{g}, \boldsymbol{u} \rangle| \leq \theta_1 \square r^{\min\{q,2\}} \right) \geq 1 - \delta \right).$$

Here, the multiplier fixed point we define compares an upper bound of a multiplier process with $\square r^{\min\{q,2\}}$. The classical multiplier fixed point is used for the squared loss and therefore is usually compared with $r^2$, [LM16]. The choice of $\square$ and the exponent $\min\{q, 2\}$ is made so that the upper bound of the multiplier process does not exceed the lower bound of the empirical excess risk $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ when $\boldsymbol{\beta}_J$ belongs to the set defined in (4.22). In many

supervised learning theories, the multiplier fixed point is typically large. However, thanks to FSD, we now restrict it to a low-dimensional subspace $V_J$, and thereby reduce its magnitude. Note that if we take $V_J$ to be the entire feature space $\mathbb{R}^p$, then $r_M(\rho,\delta)$ reduces to the classical multiplier fixed point in supervised learning theory. In other words, choosing this trivial FSD returns us to the standard setting, where the entire space is used to approximate $\boldsymbol{\beta}^*$. Thanks to FSD, however, we may select a suitable subspace $V_J$ to approximate $\boldsymbol{\beta}^*$, thereby reducing the multiplier fixed point. In fact, in Lemma 23 we will show that there exists a sequence of $(\delta_N)_N$, tending to 0 (almost exponentially) as $N$ grows, for which, for any $N$, $r_M(\rho,\delta_N) \lesssim_q (\sigma_\xi + \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2)(\frac{|J|}{N})^{\frac{1}{2(q-1)}}$ when $q \geq 2$ and $r_M(\rho,\delta_N) \lesssim \sqrt{\frac{|J|}{N}}(\sigma_\xi + \sigma_\xi^{2-q}\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2^{q-1})$ when $1 < q < 2$. In fact, we will prove that there exists an absolute constant $C_{40} = C_{40}(q) \in (1,\infty)$ for any $1 \leq q < \infty$, such that for any $r, \rho > 0$, the following random event holds with high probability (see Proposition 30 for the exact probability deviation).

$$\Omega_{\text{multi}}^{q>1} := \left\{ \sup\left( \inf_{\boldsymbol{g}\in\partial^- P_N\ell_\bullet(\boldsymbol{\beta}^*_J)} |\langle \boldsymbol{g}, \boldsymbol{u}\rangle| : \boldsymbol{u} \in r\Sigma_J^{-1/2}B_2^J \cap \rho K_{\text{model}} \right) \leq C_{40}\frac{(\sigma_\xi^{q-1} + \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2^{q-1})N^{\frac{q-1}{2}}\sqrt{|J|}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)} r \right\}, \tag{4.23}$$

where $\partial^- P_N\ell_\bullet(\boldsymbol{\beta}^*_J)$ is computed later in Lemma 21. This explains the origin of the term $\sigma_\xi(\frac{|J|}{N})^{\frac{1}{2(q-1)}}$ in $r(V_J, V_{J^c})$ when $q \geq 2$. When $1 < q < 2$, the contribution of the multiplier fixed point becomes a higher-order term relative to the corresponding quadratic fixed point, and therefore does not appear explicitly in $r(V_J, V_{J^c})$.

**FSD reduces quadratic fixed point**

The first-order Taylor expansion of $P_N\ell_{\boldsymbol{\beta}_J}$ is insufficient to obtain $P\mathcal{L}^{(2)}_{\boldsymbol{\beta}_J}$, hence this subsection considers its second-order Taylor expansion. This is commonly referred to as the quadratic fixed point/ Bernstein-type property for historical reasons—specifically, a lower bound for $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ involving $P_N\mathcal{L}^{(2)}_{\boldsymbol{\beta}_J}$, or a lower bound for $P\mathcal{L}_{\boldsymbol{\beta}_J}$ involving $P\mathcal{L}^{(2)}_{\boldsymbol{\beta}_J}$.

Here, we only consider the local Bernstein's condition developed in [CLL21, CLL20], which holds on the set $\boldsymbol{\beta}^*_J + (r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho K_{\text{model}})$ – this suffices for us to obtain an upper bound for $P\mathcal{L}^{(2)}_{\hat{\boldsymbol{\beta}}_J}$. Generally speaking, the local Bernstein's condition describes a property on the local curvature of the excess risk around $\boldsymbol{\beta}^*_J$, [LN24] : there exist some scaling factor $\alpha > 0$ and some constant $\kappa \geq 1$ such that for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}^*_J + (r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho K_{\text{model}})$, we have $P\mathcal{L}_{\boldsymbol{\beta}_J} \geq \alpha\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J)\|_2^{2\kappa}$. Such a property may follow from a lower bound on the smallest eigenvalue of the Hessian of the population excess risk $P\mathcal{L}_\bullet$ at $\boldsymbol{\beta}^*_J$.

Analogously to the multiplier fixed point associated with the first-order condition, supervised learning theory features another fixed point, called the quadratic fixed point, which characterizes the minimal $r(\rho)$ for which $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ can be compared with its quadratic approximation $P_N\mathcal{L}^{(2)}_{\boldsymbol{\beta}_J}$, defined as

$$r_Q(\rho,\delta) = \min_{r>0}\left( \mathbb{P}\left( \forall \boldsymbol{u} \in \rho K_{\text{model}} \cap r\Sigma_J^{-1/2}S_2^J,\, P_N\mathcal{L}_{\boldsymbol{\beta}_J} \geq \triangle r^\kappa + \sup_{\boldsymbol{g}\in\partial^- P_N\ell_\bullet(\boldsymbol{\beta}^*_J)} \langle \boldsymbol{g}, \boldsymbol{u}\rangle \right) \geq 1-\delta \right),$$

where, $\triangle$ is some parameter and $\kappa > 0$ is some real number. We will see below that, once again thanks to FSD, when $q \geq 2$, for any $\rho > 0$ there exists a sequence $(\delta_N)_N$ tending to 0 at a nearly exponential rate as $N$ increases, such that for every $\delta_N$ in this sequence, one may take $r_Q(\rho,\delta_N) = 0$. When $1 < q < 2$, we will require the following local Bernstein condition together with a quadratic fixed point. For $q = 1$, we will directly employ the D-M to study the squared loss, for which the loss is itself quadratic. Here we focus exclusively on the case $q > 1$ which is more problematic. In this section, we do not consider the Bernstein condition for the squared hinge loss in classification problems, namely Assumption 7, which will be proved later in Lemma 28.

**When $q \geq 2$.** The following lemma shows that there exist $\delta$ and $\kappa = q$ such that, for any $\rho > 0$, one may take $r_Q(\rho,\delta) = 0$ — once again illustrating the strength of FSD. The proof of the following Lemma 18 may be found in Section 4.7.2.

**Lemma 18.** *Suppose $q \geq 2$. For any $r > 0$, condition on the event $\Omega_{DM,reg}(\varepsilon_1) \cap \Omega_{RIP}$, we have for all $\boldsymbol{\beta}_J \in \boldsymbol{\beta}^*_J + \left[\rho K_{\text{model}} \cap r\Sigma_J^{-1/2}S_2^{p-1}\right]$,*

$$P_N\mathcal{L}_{\boldsymbol{\beta}_J} \geq \frac{q-1}{q2^q}\frac{c_{19}^q N^{\frac{q}{2}} r^q}{(1+\varepsilon_1)^q \ell_*(\Sigma_{J^c}^{1/2}B_q^p)^q} + \sup_{\boldsymbol{g}\in\partial^- P_N\ell_\bullet(\boldsymbol{\beta}^*_J)} \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J\rangle. \tag{4.24}$$

The proof relies on structural properties of the loss function $P_N\ell_\bullet$, in particular on the fact that the norm $\|\cdot\|_q$ is $q$-uniformly convex in this regime, [Pis16, Section 10.1]. Here, $\triangle = \frac{q-1}{q2^q}\frac{c_{19}^q N^{\frac{q}{2}}}{(1+\varepsilon_1)^q \ell_*(\Sigma_{J^c}^{1/2}B_q^p)^q}$, which also explains the choices of $\theta_1$ and $\square$ in the definition of $r_{\mathrm{M}}(\rho,\delta)$.

**When $1<q<2$.** When $1<q<2$, the norm $\|\cdot\|_q$ is no longer $q$-uniformly convex. In this case $r_Q$ is no longer zero, but with the aid of the local Bernstein assumption below, we can still identify $r_Q(\rho,\delta)$.

**Assumption 9** (local Bernstein's condition for random loss generated by $\ell_q$ norm)**.** *There exist absolute constants $0<c<1$, $\kappa\geq 1$ and parameters $\rho, r(\rho)>0$ defined by (4.8) such that for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho B_1^J)$, there holds*

$$P\mathcal{L}_{\boldsymbol{\beta}_J} \geq c\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)} r^{2\kappa}(\rho) = c\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}\|\Sigma_J^{1/2}(\boldsymbol{\beta}_J-\boldsymbol{\beta}_J^*)\|_2^{2\kappa}, \tag{4.25}$$

*where $P\ell_{\boldsymbol{\beta}_J} = \mathbb{E}_{\mathbb{X}_J,\boldsymbol{\xi}}\|\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$ and $P\mathcal{L}_{\boldsymbol{\beta}_J} = P\ell_{\boldsymbol{\beta}_J} - P\ell_{\boldsymbol{\beta}_J^*}$.*

In the following Lemma 19, we prove that if $X_J$ is a Gaussian random vector, then conditioned on $\Omega_{\mathrm{DM,reg}}(\varepsilon_1)$, Assumption 9 holds. The proof of Lemma 19 may be found in Section 4.9.5.

**Lemma 19.** *Suppose $X_J\sim\mathcal{N}(\mathbf{0},\Sigma_J)$, $\boldsymbol{\xi}\sim\mathcal{N}(\mathbf{0},\sigma_\xi^2 I_N)$, and suppose $X_J$ is independent with $\boldsymbol{\xi}$. Suppose either $X_{J^c}\sim\mathcal{N}(\mathbf{0},\Sigma_{J^c})$, or $\boldsymbol{\beta}_{J^c}^*=\mathbf{0}$. Condition on the event $\Omega_{DM,reg}(\varepsilon_1)$, then Assumption 9 holds with $\kappa=1$ and $c=\frac{1}{2C_{41}}$ for some absolute constant $C_{41}$.*

We conjecture that Assumption 9 remains valid for more general probability measures than the Gaussian one. Its proof would require studying the lower bound of the smallest eigenvalue of the Hessian of $\|\mathcal{A}[\cdot]\|_q^q$ around $\boldsymbol{\xi}$. This constitutes an interesting problem in stochastic geometry in its own right, but lies beyond the scope of the present work.

Define

$$r_Q(\rho) := \min\left(r>0:\ \ell_*(r(\rho)S_2^J\cap\rho\Sigma_J^{1/2}B_1^J) < \kappa_3\sigma_\xi^{-2}r^4(\rho)\sqrt{N}\right).$$

We prove in Proposition 29 that under some assumptions, when $r>r_Q(\rho)$, then with high probability, for any $\boldsymbol{\beta}_J\in\boldsymbol{\beta}_J^*+(r(\rho)\Sigma_J^{-1/2}S_2^J\cap\rho B_1^J)$, there holds $P_N\mathcal{L}_{\boldsymbol{\beta}_J}\geq\frac{1}{2}c_{23}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho)$ for some absolute constant $c_{23}$. This explains the choice of $\theta_1$ and $\square$ in the definition of $r_M(\rho,\delta)$. When $1<q<2$, the estimate for $r_{\mathrm{Q}}$ yields $\sigma_\xi^{\frac{1}{3}}(\frac{|J|}{N})^{\frac{1}{6}}$, which explains the origin of the corresponding term in $r(V_J,V_{J^c})$ for $1<q<2$—namely, this is the quadratic fixed point.

At this point, all the ingredients required to handle $\hat{\boldsymbol{\beta}}_J$ have been introduced. Before proceeding to the next section, where we analyze $\hat{\boldsymbol{\beta}}_{J^c}$, let us briefly summarize these ingredients: the Dvoretzky–Milman theorem provides a simplification of the nonlinear loss function $\ell_{\boldsymbol{\beta}_J}$; the multiplier process with dependent multipliers captures the first-order information of the excess risk; and Bernstein's property provides its second-order information. In regression problems, the second-order information corresponds to the population excess risk (or estimation error). Therefore, on the space $V_J$, we can apply the uniform convergence argument together with the localization technique to obtain a high-probability upper bound on the population excess risk $\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}_J^*)\|_2^2$.

## 4.4 Price for overfitting of $\hat{\boldsymbol{\beta}}_{J^c}$

In this section, we consider the other subspace $V_{J^c}$ appearing in the features space decomposition. As pointed out in Section 1.5.1, in this subspace, $\hat{\boldsymbol{\beta}}_{J^c}$ absorbs noise rather than estimates $\boldsymbol{\beta}_{J^c}^*$. This space is not considered in the classical uniform convergence analysis of estimators. It requires the use of Dvoretzky-Milman theorem which has been introduced in Section 1.5.3.

For regression problems, we directly apply the triangle inequality: $\|\Sigma_{J^c}^{1/2}(\hat{\boldsymbol{\beta}}_{J^c}-\boldsymbol{\beta}_{J^c}^*)\|_2 \leq \|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2 + \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\|_2$, and thus in this section we investigate the high-probability upper bound of $\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$. This quantifies the impact of $\hat{\boldsymbol{\beta}}_{J^c}$ absorbing noise (rather than estimating the signal) on $P\mathcal{L}_{\hat{\boldsymbol{\beta}}}^{(2)}$. We establish its upper bound in Section 4.4.1. Using the triangle inequality as we did above seems inefficient if $\hat{\boldsymbol{\beta}}_{J^c}$ was expected to be an estimator of $\boldsymbol{\beta}_{J^c}^*$ but, as we said, this is not the case: $\hat{\boldsymbol{\beta}}$ uses the space $V_{J^c}$ (via $\hat{\boldsymbol{\beta}}_{J^c}$) to absorbs noise (in particular, to make $\hat{\boldsymbol{\beta}}$ interpolating

the response vector $\boldsymbol{y}$) and does not aim to estimate the signal on that part of the space. This decomposition appears to be optimal in the $q = 2$ case (see the lower bound in [P4]), in the ridge regression case [P2] and the more general spectral methods case [P3].

For the classification problem, we use (1.16) to characterize how $\hat{\boldsymbol{\beta}}_{J^c}$ absorbing noise may potentially flip the prediction sign($\langle X, \hat{\boldsymbol{\beta}}_J\rangle$), and consequently its effect on $P\mathcal{L}_{\hat{\boldsymbol{\beta}}}^{\{0,1\}}$. We establish its upper bound in Section 4.4.2.

### 4.4.1 Price for Overfitting in Regression problem

When the Dvoretzky-Milman condition $N \le \kappa_{DM}\varepsilon_1^2 d_*(\Sigma_{J^c}^{1/2}B_{q'}^p)$ holds, since $\|\Sigma_{J^c}^{1/2}\|_{\ell_q\to\ell_2} = \mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)$, on $\Omega_{\mathrm{DM,reg}}$, there holds $\|\Sigma_{J^c}^{1/2}\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]\|_2 \le \mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]\|_q \lesssim \frac{\mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)}{\ell_*(\Sigma_{J^c}^{1/2}B_{q'}^p)}\|\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J\|_2$. Then the upper bound for $\|\boldsymbol{y}-\mathbb{X}_J\hat{\boldsymbol{\beta}}_J\|_2 = \|\mathbb{X}_J(\boldsymbol{\beta}_J^*-\hat{\boldsymbol{\beta}}_J)+\mathbb{X}_{J^c}\boldsymbol{\beta}_{J^c}^*+\boldsymbol{\xi}\|_2$ follows from RIP, the upper bound for $\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}_J^*)\|_2$.

### 4.4.2 Price for Overfitting in Classification problem

In this section, we investigate a high-probability upper bound for (1.16). This will demonstrate the cost of the noise interpolation behavior of $\hat{\boldsymbol{\beta}}_{J^c}$ in terms of $P\mathcal{L}_{\hat{\boldsymbol{\beta}}}^{\{0,1\}}$. In this section, all probabilities are understood to be conditional on $(X_i, Y_i)_{i=1}^N$.

The analysis remains centered on the marginal behavior of $\hat{\boldsymbol{\beta}}_{J^c}$, specifically on the magnitude of $\langle X, \hat{\boldsymbol{\beta}}_{J^c}\rangle$. Notably, this quantity concentrates around $\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$ — just as in the case of regression problems. We use Proposition 15.

If we do not impose any conditions on $X$, it is clear that we cannot obtain an upper bound for $\mathbb{P}(|\langle X, \hat{\boldsymbol{\beta}}_{J^c}\rangle| > |\langle X, \hat{\boldsymbol{\beta}}_J\rangle|)$ that decays to zero as $\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2 \to 0$. Therefore, in what follows we restrict our analysis to the setup of the Gaussian mixture classification and logistic models. The proof of the following Proposition 27 can be found in Section 4.8.3.

**Proposition 27.** *In Gaussian mixture classification model and logistic model, we have*

$$\mathbb{P}(|\langle X, \hat{\boldsymbol{\beta}}_{J^c}\rangle| > |\langle X, \hat{\boldsymbol{\beta}}_J\rangle||\mathcal{D}) \le \frac{2}{\pi}\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2\frac{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2}{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2^2 - \|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2^2}$$

*where* $\mathcal{D} = (\boldsymbol{y}, \mathbb{X})$.

Up to this point, we observe the following fact: the impact of noise interpolation / absorption by $\hat{\boldsymbol{\beta}}_{J^c}$ on prediction primarily depends on whether $\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$ is sufficiently small as well as if $\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$ is significantly smaller (like half) than $\|\Sigma_J^{1/2}\hat{\boldsymbol{\beta}}_J\|_2$. Next, we analyze an upper bound for $\|\Sigma_{J^c}^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$. Here, $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\hat{\boldsymbol{\beta}}_J]$ constitutes a nonlinear operator, and consequently we employ only the simplest upper bound on its operator norm, that is, on $\Omega_{\mathrm{DM,class}}(\delta_4)$, $\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2 \le \|\Sigma_{J^c}\|_{\mathrm{op}}^{1/2}\|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\hat{\boldsymbol{\beta}}_J]\|_2$.

**Proposition 28.** *Suppose the choice of $V_J$ satisfies that $N\|\Sigma_{J^c}\|_{op} \le \kappa_{DM}\bar{\delta}^2\,\mathrm{Tr}(\Sigma_{J^c})$. Suppose $\frac{1}{2} < \|\Sigma^{1/2}\boldsymbol{\beta}_J^*\|_2$ and $r(V_J, V_{J^c}) < \frac{1}{4C_{33}}(4\|\Sigma^{1/2}\boldsymbol{\beta}_J^*\|_2 + 1 - \sqrt{16\|\Sigma^{1/2}\boldsymbol{\beta}_J^*\|_2 + 1})$, where $r(V_J, V_{J^c})$ and $C_{33}$ are defined in Theorem 11. Then for $C_{35}$ defined in Theorem 11, in the Gaussian mixture classification model and the logistic model, for any $\bar{\delta}$ satisfying (4.14), with the same probability as in (4.13) and for $C_{38}$ in Theorem 11, we have*

$$(1.16) = \mathbb{P}\left(Y\langle X, \hat{\boldsymbol{\beta}}\rangle < 0|\mathcal{D}\right) - \mathbb{P}\left(Y\langle X, \hat{\boldsymbol{\beta}}_J\rangle < 0|\mathcal{D}\right) \le C_{38}\bar{\delta}\sqrt{P\ell_{\boldsymbol{\beta}_J^*}}. \tag{4.26}$$

The proof of Proposition 28 may be found in Section 4.8.3.

## 4.5 Conclusions and Research Perspectives

**Conclusions.**

1. This chapter establishes an analytical framework for the features space decomposition method and, through the self-regularization property, applies it to derive non-asymptotic upper bounds for: (i) the population excess risk of minimum $\ell_q$-norm interpolant estimators in linear regression problems, and (ii) minimum $\ell_2$-norm interpolant classifiers in linear classification problems. This method has the potential to improve one of the most fundamental approaches in mathematical statistics — the uniform convergence argument.

2. We introduce a new class of benign overfitting phenomena, termed *non-exact benign overfitting*. Building upon our main results, we provide sufficient conditions for: (a) minimum $\ell_q$-norm interpolant estimators to exhibit non-exact and exact benign overfitting in regression settings, and (b) minimum $\ell_2$-norm interpolant classifiers to achieve both non-exact and exact benign overfitting in classification problems. This phenomenon lies between benign overfitting and test-error benign overfitting, and it still provides useful information about the population risk of overfitting estimators. We believe that this phenomenon deserves further attention.

3. Our technical approach deliberately avoids the convex min-max theorem that requires the Gaussian assumption. Instead, we incorporate geometric tools from GAFA combined with the features space decomposition method, thereby offering a geometric perspective for understanding benign overfitting phenomena in minimum-norm interpolant estimators. This suggests that in supervised learning problems, particularly in high-dimensional statistics, it is necessary to introduce more tools from GAFA.

We now highlight several potential future research directions related to the problem of benign overfitting, as well as new questions in probability theory and random geometry that this line of inquiry may inspire.

**Research Perspectives.**

1. Beyond the minimum $\ell_q$ norm interpolant estimator, other interesting interpolant estimators motivated by neural network theory include the minimum Schatten–1 norm interpolant estimator, which corresponds to the implicit regularization of shallow linear neural networks, and the minimum Schatten–$q$ quasi-norm interpolant estimator, which corresponds to the implicit regularization of depth-$L = \lfloor 2/q \rfloor$ deep linear neural networks, [GLSS18, SLS$^+$20, RBD25]. Investigating the exact and non-exact benign overfitting properties of these interpolant estimators would be of significant interest.

2. Moreover, this chapter establishes only sufficient conditions for the occurrence of exact and non-exact benign overfitting, and studying the necessary conditions represents another important research direction. For instance, in our analysis of the minimum $\ell_2$ norm interpolant estimator, we did not, unlike previous works [MRSS23, MRSY25], assume that $V_J$ is chosen as the eigenspace of $\Sigma$. However, since the result obtained for the minimum $\ell_2$ norm interpolant classifier corresponds to non-exact benign overfitting, a direct comparison with existing works is not possible. Establishing necessary conditions would therefore help determine which subspace the feature space decomposition of the minimum $\ell_2$ norm interpolant classifier actually selects.

3. We still lack mathematical tools to handle $\|\Sigma_{J^c}^{1/2}\mathcal{A}[\boldsymbol{\xi}+\boldsymbol{\zeta}]\|_2^2$ in Section 4.4, where $\boldsymbol{\zeta} = \mathbb{X}_{J^c}\boldsymbol{\beta}^*_{J^c} + \mathbb{X}_J(\boldsymbol{\beta}^*_J - \hat{\boldsymbol{\beta}}_J)$ satisfies $\|\boldsymbol{\zeta}\|_2 = o(\sqrt{N})$ with high probability. In the case $q = 2$, the analysis in [P4] relies on the fact that $\mathcal{A}$ is a linear operator, and in this case, the "correct" tool is the upper side of Dvoretzky-Milman theorem applied to norm $\|\Sigma_{J^c}\cdot\|_2$. When $q \neq 2$, we aim to seek an alternative to capture the non-linearity of $\mathbb{E}_{\boldsymbol{\xi}}\|\Sigma_{J^c}^{1/2}\mathcal{A}[\boldsymbol{\xi}+\boldsymbol{\zeta}]_2^2$. This may require introducing certain tools from random geometry especially when $q = 1$, [Sch13]. When $q = 1$, we conjecture that under the assumptions of Theorem 10, there exists an absolute constant $C > 1$ such that, with high probability, one has $\|\mathcal{A}[\boldsymbol{\xi}+\boldsymbol{\zeta}]\|_2 \leq \frac{C}{\sqrt{N}}\|\mathcal{A}[\boldsymbol{\xi}+\boldsymbol{\zeta}]\|_1$. In other words, viewed as the solution to the basis pursuit problem, the vector $\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]$ is supported on $N$ coordinates, and the magnitudes of these $N$ nonzero coordinates are nearly of the same order. Since, with probability one, the vector $\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J$ does not lie in any subspace of dimension strictly smaller than $N$, the classical compressed sensing theory does not, to the best of our knowledge, provide any guarantees for such vectors [FR13]. Hence, the above conjecture lies outside the scope of existing compressed sensing results.

## Acknowledgments

We would like to thank Simon Foucart, Jaouad Mourtada, Johannes Schmidt-Hieber, Martin Wainwright, and Guillaume Wang for useful discussions. Part of this work was carried out during ZS's visits to RIKEN-AIP, Japan, and the Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw, Poland.

## 4.6 Proof: Properties of the nonlinear map $\mathcal{A}$ and $\mathcal{B}$

In this section, we gather all the properties we need on the nonlinear operators $\mathcal{A}$ and $\mathcal{B}$ appearing in the decomposition of $\hat{\boldsymbol{\beta}}$ in regression and classification. We start with $\mathcal{A}$ and we recall its definition and the associated dual

problem. Let $1 \leq q, q'$ be such that $1/q + 1/q' = 1$. For all $\boldsymbol{\mu} \in \mathbb{R}^N$, $\mathcal{A}[\boldsymbol{\mu}]$ is solution to the optimization problem $\min\left(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{y}\right)$, whose dual problem is

$$\max\left(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle : \|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\|_{q'} \leq 1\right). \tag{4.27}$$

We denote by $\boldsymbol{\lambda}^*[\boldsymbol{\mu}]$ a solution to the dual problem. We recall that by strong duality (see the von Neuman-Sion minmax theorem), we have

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

**Lemma 20.** *Let $1 \leq q < \infty$ and denote by $q'$ its conjugate number. The non-linear maps $\mathcal{A}[\cdot] : \mathbb{R}^N \to \mathbb{R}^p$ and $\boldsymbol{\lambda}^*[\cdot] : \mathbb{R}^N \to \mathbb{R}^N$ satisfy the following properties:*

1. *For any $\alpha > 0$ and $\boldsymbol{\mu} \in \mathbb{R}^N$, $\mathcal{A}[\alpha\boldsymbol{\mu}] = \alpha\mathcal{A}[\boldsymbol{\mu}]$, and $\mathcal{A}[\boldsymbol{\mu}] = \mathbf{0}$ if and only if $\boldsymbol{\mu} = \mathbf{0}$; $\boldsymbol{\lambda}^*[\alpha\boldsymbol{\mu}] = \boldsymbol{\lambda}^*[\boldsymbol{\mu}]$;*
2. *$\|\mathcal{A}[\cdot]\|_q$ is sub-additive and is a norm;*
3. *for all $\boldsymbol{\mu} \in \mathbb{R}^N$, $\mathcal{A}[-\boldsymbol{\mu}] = -\mathcal{A}[\boldsymbol{\mu}]$;*
4. *On the event $\Omega_{DM,reg}(\varepsilon_1)$ defined in (1.22), $\|\mathcal{A}[\cdot]\|_q$ is a Lipschitz function with Lipschitz constant $1/[(1-\varepsilon_1)\ell_*]$ where $\ell_* = \ell_*(\Sigma_{J^c}^{1/2}B_q^p)$. As a consequence, for any $\boldsymbol{\mu} \in \mathbb{R}^N$, $\|\mathcal{A}[\boldsymbol{\mu} + \cdot]\|_q$ is Lipschitz with the same constant.*
5. *Let $\boldsymbol{\mu} \neq \mathbf{0}$. For any solution $\boldsymbol{\lambda}^*[\boldsymbol{\mu}]$ of the dual problem, we have*

$$|\mathcal{A}[\boldsymbol{\mu}]|^{\odot(q-2)} \odot \mathcal{A}[\boldsymbol{\mu}] = \|\mathcal{A}[\boldsymbol{\mu}]\|_q^{q-1}\, \mathbb{X}_{J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}]. \tag{4.28}$$

   *In particular, when $q = 1$, $\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}] = \operatorname{sign}(\mathcal{A}[\boldsymbol{\mu}])$.*
6. *Let $\boldsymbol{\zeta} \in \mathbb{R}^N$ be a symmetric random vector independent of $\mathbb{X}_{J^c}$, then $\boldsymbol{\lambda}^*[\boldsymbol{\zeta}]$ and $\mathcal{A}[\boldsymbol{\zeta}]$ are symmetric random variables and, in particular, $\mathbb{E}[\boldsymbol{\lambda}^*[\boldsymbol{\zeta}]|\mathbb{X}_{J^c}] = \mathbf{0}$ and $\mathbb{E}[\mathcal{A}[\boldsymbol{\zeta}]|\mathbb{X}_{J^c}] = \mathbf{0}$.*
7. *For any $q \geq 1$ and any $R > 0$, on the random event $\Omega_{DM,reg}(\varepsilon_1)$, $\|\mathcal{A}[\cdot]\|_q^q$ is $\frac{2^{q-1}}{(1-\varepsilon_1)^q \ell_*^q} R^{q-1}$-Lipschitz in $RB_2^N$. Moreover, $(\boldsymbol{x}, \boldsymbol{y}) \in RB_2^N \mapsto \|\mathcal{A}[\boldsymbol{x} + \boldsymbol{y}]\|_q^q - \|\mathcal{A}[\boldsymbol{y}]\|_q^q$ is $\frac{C_q}{(1-\varepsilon_1)^q \ell_*^q} R^{q-1}$ Lipschitz for some absolute constant $C_q > 1$.*

**Proof.**

1. By the definition of $A$, we know that $\|\mathcal{A}[\alpha\boldsymbol{\mu}]\|_q = \min(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}\boldsymbol{\nu} = \alpha\boldsymbol{\mu}) = \min(\|\boldsymbol{\nu}\|_q : \frac{1}{\alpha}\mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{\mu})$. Let $\tilde{\boldsymbol{\nu}} = \frac{1}{\alpha}\boldsymbol{\nu}$, then $\boldsymbol{\nu} = \alpha\tilde{\boldsymbol{\nu}}$ and hence $\|\mathcal{A}[\alpha\boldsymbol{\mu}]\|_q = \min(\|\alpha\tilde{\boldsymbol{\nu}}\|_q : \mathbb{X}_{J^c}\tilde{\boldsymbol{\nu}} = \boldsymbol{\mu}) = \alpha \min(\|\tilde{\boldsymbol{\nu}}\|_q : \mathbb{X}_{J^c}\tilde{\boldsymbol{\nu}} = \boldsymbol{\mu}) = \alpha\|\mathcal{A}[\boldsymbol{\mu}]\|_q$. This implies that $\|\mathcal{A}[\cdot]\|_q$ is positive 1-homogeneous. When $\boldsymbol{\mu} = \mathbf{0}$, then $\operatorname{argmin}(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}\boldsymbol{\nu} = \mathbf{0}) = \mathbf{0}$; on the other side, since $\boldsymbol{\mu} = \mathbb{X}_{J^c}\mathcal{A}[\boldsymbol{\mu}]$, when $\mathcal{A}[\boldsymbol{\mu}] = \mathbf{0}$, we know that $\boldsymbol{\mu} = \mathbf{0}$. It is clear that $\boldsymbol{\lambda}^*[\alpha\boldsymbol{\mu}] = \boldsymbol{\lambda}^*[\boldsymbol{\mu}]$.
2. For any $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^N$, $\|\mathcal{A}[\boldsymbol{v}_1 + \boldsymbol{v}_2]\|_q \leq \|\mathcal{A}[\boldsymbol{v}_1] + \mathcal{A}[\boldsymbol{v}_2]\|_q$. This is because $\mathbb{X}_{J^c}\mathcal{A}[\boldsymbol{v}_1 + \boldsymbol{v}_2] = \boldsymbol{v}_1 + \boldsymbol{v}_2 = \mathbb{X}_{J^c}\mathcal{A}[\boldsymbol{v}_1] + \mathbb{X}_{J^c}\mathcal{A}[\boldsymbol{v}_2]$. Hence $\mathcal{A}[\boldsymbol{v}_1] + \mathcal{A}[\boldsymbol{v}_2]$ belongs to $\{\boldsymbol{\nu} : \mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{v}_1 + \boldsymbol{v}_2\}$, the feasible set in the definition of $\mathcal{A}[v_1 + v_2]$. Since $\mathcal{A}[\boldsymbol{v}_1 + \boldsymbol{v}_2]$ has the smallest $\|\cdot\|_q$ norm on this feasible set we get that $\|\mathcal{A}[v_1 + v_2]\|_q \leq \|\mathcal{A}[v_1] + \mathcal{A}[v_2]\|_q$. The sub-aditivity of $\|\mathcal{A}[\cdot]\|_q$ follows from the triangle inequality. Together with convexity item *1.*, we conclude that $\|\mathcal{A}[\cdot]\|_q$ is a norm and so it is convex.
3. This point is clear from by definition of $\mathcal{A}[\cdot]$.
4. By standard functional analysis, we only need to prove that any sub-gradient of $\|\mathcal{A}[\cdot]\|_q$ has its $\ell_2$-norm bounded by $1/[(1-\varepsilon_1)\ell_*]$. We recall that by strong duality we obtained that $\|\mathcal{A}[\boldsymbol{\mu}]\|_q = \max\left(\langle \mu, \boldsymbol{\lambda}\rangle : \|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\|_{q'} \leq 1\right)$. Hence, $\boldsymbol{\mu} \to \|\mathcal{A}[\boldsymbol{\mu}]\|_q$ is the maximal function of a set of linear functions and so its sub-differential at a point $\boldsymbol{\mu}$ is given by all the $\boldsymbol{\lambda}$'s achieving this max, ie of the dual problem. As a consequence, we obtain $\partial^- \mathcal{A}[\boldsymbol{\mu}] = \{\boldsymbol{\lambda}^*[\boldsymbol{\mu}] : \boldsymbol{\lambda}^*[\boldsymbol{\mu}]$ is solution of the dual problem$\}$. Moreover, on the event $\Omega_{\mathrm{DM,reg}}(\varepsilon_1)$ we have

$$\|\boldsymbol{\lambda}^*[\boldsymbol{\mu}]\|_2\, (1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_q^p) \leq \|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}]\|_{q'} \leq 1$$

   where the last inequality follows from the fact that $\boldsymbol{\lambda}^*[\boldsymbol{\mu}]$ belongs to the feasible set of the dual problem. We conclude that $\|\boldsymbol{\lambda}^*[\boldsymbol{\mu}]\|_2 \leq 1/[(1-\varepsilon_1)\ell_*]$ and since it holds uniformly for all $\boldsymbol{\mu}$, this implies that $\|\mathcal{A}[\cdot]\|_q$ is Lipschitz with constant $1/[(1-\varepsilon_1)\ell_*]$ on the event $\Omega_{\mathrm{DM,reg}}(\varepsilon_1)$

5. Let us first start with two observations: let $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in V_{J^c}$ be such that $\|\boldsymbol{w}\|_q = 1$. *a)* if $\boldsymbol{u} \in \partial^- \|\cdot\|_q(\boldsymbol{w})$ then $\boldsymbol{w} \in \partial^- \|\cdot\|_{q'}(\boldsymbol{u})$; and *b)* if $\boldsymbol{u} \in \partial^- \|.\|_q(\boldsymbol{v})$ then $\boldsymbol{u} \in \partial^- \|.\|(\alpha \boldsymbol{v})$ for any $\alpha > 0$. These two observations easily follow from the characterization: $\boldsymbol{u} \in \partial^- \|.\|_q(\boldsymbol{v})$ iif $\|\boldsymbol{u}\|_{q'} = 1$ and $\langle \boldsymbol{v}, \boldsymbol{u} \rangle = \|\boldsymbol{v}\|_q$.

   Next, by first order condition of convex optimization problem and Lagrangian duality the following holds: let $\boldsymbol{\lambda}^*[\boldsymbol{\mu}]$ be a solution of the dual problem then $\mathcal{A}[\boldsymbol{\mu}] \in \operatorname{argmin}(\|\boldsymbol{\beta}_{J^c}\|_q : \mathbb{X}_{J^c}\boldsymbol{\beta}_{J^c} = \boldsymbol{\mu})$ iff the KKT conditions are satisfied: $\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}] \in \partial^- \|.\|_q(\mathcal{A}[\boldsymbol{\mu}])$ and $\mathbb{X}_{J^c}\mathcal{A}[\boldsymbol{\mu}] = \boldsymbol{\mu}$. Since $\boldsymbol{\mu} \neq \mathbf{0}$, by item *1.*, $\mathcal{A}[\boldsymbol{\mu}] \neq \mathbf{0}$ and hence $\partial^- \|.\|_q(\mathcal{A}[\boldsymbol{\mu}]) \subset S_{q'}^{J^c}$ and, as a result, $\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}] \neq \mathbf{0}$. Applying previous observation to $\boldsymbol{u} = \mathbb{X}_{J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}]$ and $\boldsymbol{v} = \mathcal{A}[\boldsymbol{\mu}]$, then $\boldsymbol{u} \in \partial^- \|.\|_q(\boldsymbol{v})$, and hence $\boldsymbol{u} \in \partial^- \|.\|_q(\boldsymbol{v}/\|\boldsymbol{v}\|_q)$. This further implies $\boldsymbol{v}/\|\boldsymbol{v}\|_q \in \partial^- \|.\|_{q'}(\boldsymbol{u})$, that is, $\mathcal{A}[\boldsymbol{\mu}] \in \|\mathcal{A}[\boldsymbol{\mu}]\|_q \partial^- \|.\|_{q'}(\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}])$.

   The sub-differential of $\|\cdot\|_{q'}$ at $\boldsymbol{v}$ follows from: $\langle g, \boldsymbol{v} \rangle = \sum_{j \in J^c} v_j^2 |v_j|^{q'-2} \|\boldsymbol{v}\|_{q'}^{1-q'} = \|\boldsymbol{v}\|_{q'}$ and $\|g\|_q = 1$.

   Finally, since $\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}] \in \partial^- \|.\|_q(\mathcal{A}[\boldsymbol{\mu}])$, we have $\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}]\|_{q'} = 1$ hence $\boldsymbol{\lambda}^*[\boldsymbol{\mu}]$ is feasible. By item *4.*, $\|\boldsymbol{\lambda}^*[\boldsymbol{\mu}]\|_2 \leq \frac{1}{(1-\varepsilon_1)\ell_*}$.

   Since $\partial^- \|\cdot\|_q(\mathcal{A}[\boldsymbol{\mu}]) = |\mathcal{A}[\boldsymbol{\mu}]|^{\odot(q-2)} \odot \mathcal{A}[\boldsymbol{\mu}] \|\mathcal{A}[\boldsymbol{\mu}]\|_q^{1-q}$ (because $\boldsymbol{\mu} \neq 0$ and so $\mathcal{A}[\boldsymbol{\mu}] \neq \mathbf{0}$). The result follows from these two observations.

6. By item *3.* and the assumption that $\boldsymbol{\zeta}$ is symmetric, $-\mathcal{A}[\boldsymbol{\zeta}] = \mathcal{A}[-\boldsymbol{\zeta}]$ has the same distribution as $\mathcal{A}[\boldsymbol{\zeta}]$ conditionally on $\mathbb{X}_{J^c}$. Hence, $\mathcal{A}[\boldsymbol{\zeta}]$ is a symmetric variable and so it is centered conditionally on $\mathbb{X}_{J^c}$. Similarly, one may check that $\boldsymbol{\lambda}^*[-\boldsymbol{\zeta}] = -\boldsymbol{\lambda}^*[\boldsymbol{\zeta}]$ and so the same result hold for $\boldsymbol{\lambda}^*[\boldsymbol{\zeta}]$.

7. For any $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in RB_2^N$, applying the Lagrange mean value theorem to the map $a \in \mathbb{R}_+ \mapsto a^q$ yields $\big|\|\mathcal{A}[\boldsymbol{\mu}_1]\|_q^q - \|\mathcal{A}[\boldsymbol{\mu}_2]\|_q^q\big| \leq q(\|\mathcal{A}[\boldsymbol{\mu}_1]\|_q^{q-1} + \|\mathcal{A}[\boldsymbol{\mu}_2]\|_q^{q-1}) \big|\|\mathcal{A}[\boldsymbol{\mu}_1]\|_q - \|\mathcal{A}[\boldsymbol{\mu}_2]\|_q\big|$. By the triangle inequality and item *4* of Lemma 20, we obtain $\big|\|\mathcal{A}[\boldsymbol{\mu}_1]\|_q^q - \|\mathcal{A}[\boldsymbol{\mu}_2]\|_q^q\big| \leq \frac{2^{q-1}R^{q-1}}{(1-\varepsilon_1)^q \ell_*^q} \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2$. Moreover, for any $(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \in RB_2^N \times RB_2^N$, $\big|\|\mathcal{A}[\boldsymbol{x}_1 + \boldsymbol{y}_1]\|_q^q - \|\mathcal{A}[\boldsymbol{y}_1]\|_q^q - (\|\mathcal{A}[\boldsymbol{x}_2 + \boldsymbol{y}_2]\|_q^q - \|\mathcal{A}[\boldsymbol{y}_2]\|_q^q)\big| \leq \big|\|\mathcal{A}[\boldsymbol{x}_1 + \boldsymbol{y}_1]\|_q^q - \|\mathcal{A}[\boldsymbol{x}_2 + \boldsymbol{y}_2]\|_q^q\big| + \big|\|\mathcal{A}[\boldsymbol{y}_1]\|_q^q - \|\mathcal{A}[\boldsymbol{y}_2]\|_q^q\big| \lesssim_q \frac{1}{(1-\varepsilon_1)^q \ell_*^q} R^{q-1} \|(\boldsymbol{x}_1, \boldsymbol{y}_1) - (\boldsymbol{x}_2, \boldsymbol{y}_2)\|_2$.

■

The following are corollaries of Lemma 20.

**Lemma 21.** *1. Let $q \geq 1$. We consider the empirical loss function $P_N \ell_\bullet : \boldsymbol{\beta}_J \in V_J \mapsto P_N \ell_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J]\|_q^q$ and denote by $\partial^- P_N \ell_\bullet$ its sub-differential. We have for all $\boldsymbol{\beta}_J \in V_J$*

$$\partial^- P_N \ell_\bullet(\boldsymbol{\beta}_J) = \left\{ -q \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J]\|_q^{q-1} \mathbb{X}_J^\top \boldsymbol{\lambda}^*[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J] : \boldsymbol{\lambda}^*[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J] \text{ is solution to the dual problem (4.27)} \right\}. \tag{4.29}$$

*2. For all $\boldsymbol{\beta}_J \in V_J$, we define the risk $P\ell_{\boldsymbol{\beta}_J} = \mathbb{E}[P_N \ell_{\boldsymbol{\beta}_J} | \mathbb{X}_{J^c}] = \mathbb{E}_{\mathbb{X}_J, \boldsymbol{\xi}} \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J]\|_q^q$. Suppose $\mathbb{X}_{J^c}\boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}$ is independent of $\mathbb{X}_J$ and $\mathbb{E}[\mathbb{X}_J] = \mathbf{0}$, then $\boldsymbol{\beta}_J^*$ is a minimizer of the risk function $\boldsymbol{\beta}_J \to P\ell_{\boldsymbol{\beta}_J}$ over $V_J$.*

**Proof.**

1. Equation (4.29) follows from the chain rule and the fact that

$$\partial^- \|\mathcal{A}[\cdot]\|_q(\boldsymbol{\mu}) = \{\boldsymbol{\lambda}^*[\boldsymbol{\mu}] \text{ solution to the dual problem (4.27)}\}.$$

2. By convexity of the risk function, to show that $\boldsymbol{\beta}_J^*$ is a minimum of the risk function over $V_J$, it is enough to show that $\mathbf{0}$ is a sub-gradient of the risk function at $\boldsymbol{\beta}_J^*$. First note that, by convexity, for all $\boldsymbol{\beta}_J \in V_J$, we have $\partial^- P\ell_{\boldsymbol{\beta}_J} = \mathbb{E}[\partial^- P_N \ell_{\boldsymbol{\beta}_J} | \mathbb{X}_{J^c}]$. Let $\boldsymbol{\lambda}^*[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J^*]$ be a solution to (4.27) for $\boldsymbol{\mu} = \boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J^*$. From the first item, We only need to show that

$$\mathbb{E}\left[ \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J^*]\|_q^{q-1} \mathbb{X}_J^\top \boldsymbol{\lambda}^*[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J^*] | \mathbb{X}_{J^c} \right] = \mathbf{0}.$$

   We note that $\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J^* = \mathbb{X}_{J^c} \boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}$ is independent of $\mathbb{X}_J$ hence,

$$\begin{aligned}
&\mathbb{E}\left[ \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J^*]\|_q^{q-1} \mathbb{X}_J^\top \boldsymbol{\lambda}^*[\boldsymbol{y} - \mathbb{X}_J \boldsymbol{\beta}_J^*] | \mathbb{X}_{J^c} \right] = \mathbb{E}_{\boldsymbol{\xi}} \mathbb{E}_{\mathbb{X}_J} \left[ \|\mathcal{A}[\mathbb{X}_{J^c} \boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}]\|_q^{q-1} \mathbb{X}_J^\top \boldsymbol{\lambda}^*[\mathbb{X}_{J^c} \boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}] \right] \\
&= \mathbb{E}_{\boldsymbol{\xi}} \left[ \|\mathcal{A}[\mathbb{X}_{J^c} \boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}]\|_q^{q-1} \mathbb{E}_{\mathbb{X}_J} \left[ \mathbb{X}_J^\top \right] \boldsymbol{\lambda}^*[\mathbb{X}_{J^c} \boldsymbol{\beta}_{J^c}^* + \boldsymbol{\xi}] \right] = \mathbf{0}
\end{aligned}$$

   because $\mathbb{X}_J$ is centered.

■

Next, we turn to the study of the map $\mathcal{B}$ that appears in the classification problem. For this problem, we only considered the minimum $\ell_2$-norm interpolant estimator. However, unlike the regression problem, where $\mathcal{A}$ is a linear operator for $q = 2$, this is not the case in classification: $\mathcal{B}$ is in general a non-linear map. We first recall its definition, the strong duality property it satisfies and the notation $\mathbb{X}_{\boldsymbol{y},J^c} = [Y_1 P_{J^c} X_1 | \cdots | Y_N P_{J^c} X_N]^\top$. For all $\boldsymbol{\mu} \in \mathbb{R}^N$, $\mathcal{B}[\boldsymbol{\mu}]$ is solution to the optimization problem

$$\min \left( \|\boldsymbol{\nu}\|_2 : \mathbb{X}_{\boldsymbol{y},J^c} \boldsymbol{\nu} \succeq \boldsymbol{\mu} \right)$$

whose dual problem is

$$\max \left( \langle \boldsymbol{\mu}, \boldsymbol{\lambda} \rangle : \boldsymbol{\lambda} \succeq 0, \left\| \mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda} \right\|_2 \leq 1 \right).$$

We denote by $\boldsymbol{\lambda}^*[\mu]$ a solution to the dual problem. We recall that by strong duality (see the von Neuman-Sion minmax theorem), we have

$$\|\mathcal{B}[\boldsymbol{\mu}]\|_2 = \min \left( \|\boldsymbol{\nu}\|_2 : \mathbb{X}_{\boldsymbol{y},J^c} \boldsymbol{\nu} \succeq \boldsymbol{\mu} \right) = \max \left( \langle \boldsymbol{\nu}, \boldsymbol{\lambda} \rangle : \boldsymbol{\lambda} \succeq 0, \left\| \mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda} \right\|_2 \leq 1 \right) = \langle \boldsymbol{\mu}, \boldsymbol{\lambda}^*[\mu] \rangle.$$

**Lemma 22.** *The non-linear maps $\mathcal{B}[\cdot] : \mathbb{R}^N \to V_{J^c}$ and $\boldsymbol{\lambda}^*[\cdot] : \mathbb{R}^N \to \mathbb{R}^N$ satisfy the following properties:*

1. *For any $\boldsymbol{\mu} \in \mathbb{R}^N$ and $\alpha > 0$, $\mathcal{B}[\alpha \boldsymbol{\mu}] = \alpha \mathcal{B}[\boldsymbol{\mu}]$, and $\mathcal{B}[\boldsymbol{\mu}] = \mathbf{0}$ if and only if $\boldsymbol{\mu} \preceq \mathbf{0}$; $\boldsymbol{\lambda}^*[\alpha \boldsymbol{\mu}] = \boldsymbol{\lambda}^*[\boldsymbol{\mu}]$;*
2. *$\|\mathcal{B}[\cdot]\|_2$ is sub-additive and convex;*
3. *On the event $\Omega_{DM,class}(\delta_4)$ (defined in Proposition 26), $\|\mathcal{B}[\cdot]\|_2$ is a Lipschitz function with Lipschitz constant $1/[(1-\varepsilon_1)\ell_*]$ where $\ell_* = \ell_*(\Sigma_{J^c}^{1/2} B_2^p)$. As a consequence, for any $\boldsymbol{\mu} \in \mathbb{R}^N$, $\|\mathcal{B}[\boldsymbol{\mu} + \cdot]\|_2$ is Lipschitz with the same constant.*
4. *Let $\boldsymbol{\mu}$ be any vector in $\mathbb{R}^N$ that does not satisfy $\boldsymbol{\mu} \preceq \mathbf{0}$, then $\boldsymbol{\lambda}^*[\boldsymbol{\mu}]$ satisfies that $\mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}] \neq \mathbf{0}$, and*

$$\mathcal{B}[\boldsymbol{\mu}] \in \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \, \partial^- \|\cdot\|_2 \left( \mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}] \right).$$

   *Moreover, on the event $\Omega_{DM,class}(\delta_4)$, $\|\boldsymbol{\lambda}^*[\boldsymbol{\mu}]\|_2 \leq \frac{1}{(1-\varepsilon_1)\ell_*}$.*
5. *Let $\boldsymbol{\mu}$ be any vector in $\mathbb{R}^N$ that does not satisfy $\boldsymbol{\mu} \preceq \mathbf{0}$, then*

$$\mathcal{B}[\boldsymbol{\mu}] = \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \, \mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}]. \tag{4.30}$$

6. *Let $\boldsymbol{\lambda}$ be any vector in $\mathbb{R}^N$ that does not satisfy $\boldsymbol{\mu} \preceq \mathbf{0}$, then*

$$\left\| \Sigma_{J^c}^{1/2} \mathcal{B}[\boldsymbol{\mu}] \right\|_2 \leq \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \min \left( \left\| \Sigma_{J^c}^{1/2} \boldsymbol{v} \right\|_2 : \boldsymbol{v} \in \partial^- \|\cdot\|_2 \left( \mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}] \right) \right).$$

**Proof.** The proof for item *1.*, *2.* and *3.* are similar to that of the proof of Lemma 20, where the only modification required is replacing the "=" with "$\succeq$". We now proceed directly to the proof of item *4.*, *5.*, and item *6.*. For item *4.*, compared to that of the proof of Lemma 20, Similar to the proof in Lemma 20, for the convex optimization problem $\|\mathcal{B}[\cdot]\|_q$, its KKT conditions still include the stationarity condition $\mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}] \in \partial^- \|\cdot\|_q(\mathcal{B}[\boldsymbol{\mu}])$, as well as the dual feasibility condition $\|\mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}]\|_2 = 1$ (when $\mathcal{B}[\boldsymbol{\mu}] \neq 0$, which follows from item *1.*, i.e., when $\boldsymbol{\mu}$ does not satisfy $\boldsymbol{\mu} \preceq \mathbf{0}$). By the Fenchel duality theorem, these two conditions imply that $\mathcal{B}[\boldsymbol{\mu}] \in \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \partial^- \|\cdot\|_2 (\mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}])$. Moreover, the computation of $\partial^- \|\cdot\|_2 (\mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}^*[\boldsymbol{\mu}])$ follows the same procedure as in Lemma 20. Item *5.* and item *6.* follow from item *4.* immediately. ∎

## 4.7 Proof of Theorem 10: Benign overfitting of the minimum $\ell_q$-norm interpolant estimator

In what follows, we take $r(\rho) = C_{30} r(V_J, V_{J^c})$, defined in (4.8), and $\rho = \rho_*$ defined as follows

$$\rho_* = C_{29} \frac{\sqrt{N}}{\ell_*(\Sigma_{J^c}^{1/2} B_q^p)} r(V_J, V_{J^c}). \tag{4.31}$$

### 4.7.1 Stochastic Argument for regression problem

**Noise concentration.** For any $0 < \delta_5 < 1$, we let

$$\Omega_{\text{noise}}(\delta_5) := \Big\{ (1-\delta_5)N\sigma_\xi^2 \le \|\boldsymbol{\xi}\|_2^2 \le (1+\delta_5)N\sigma_\xi^2, \text{ and } \\ (1-\delta_5)N \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\right\|_2^2 \le \|\mathbb{X}_{J^c}\boldsymbol{\beta}^*_{J^c}\|_2^2 \le (1+\delta_5)N \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\right\|_2^2 \Big\}. \tag{4.32}$$

By standard concentration inequality, c.f. [Ver18, Theorem 3.1.1], we know that when $X_{J^c}$ and $\boldsymbol{\xi}$ are sub-Gaussian, then $\mathbb{P}(\Omega_{\text{noise}}(\delta_5)) \ge 1 - \exp(-c_{20}\delta_5^2 N)$ for some absolute constant $c_{20}$.

**Isomorphic Property of $\mathbb{X}_J$.** For the sake of simplicity, let us assume that $|J| = \lfloor \delta_6 N \rfloor$ for some $1/N \le \delta_6 < 1/2$. Recall from (4.22) that

$$\Omega_{\text{RIP}} := \left\{ \forall \boldsymbol{\beta}_J \in \boldsymbol{\beta}^*_J + \left[\rho K_{\text{model}} \cap r(\rho)\Sigma_J^{-1/2} S_2^{p-1}\right] : c_{19} r(\rho) \le \frac{1}{\sqrt{N}} \|\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J)\|_2 \le C_{39} r(\rho) \right\}.$$

By the same argument as in Section 6.5.2 of [P2], there exists an absolute constant $\kappa_{RIP} < 1$, such that when $\delta_6 < \kappa_{RIP}$, then $\Omega_{\text{RIP}}$ is implied by the event $\{c_{19}\sqrt{N} \le \sigma_{|J|}(\mathbb{X}_J\Sigma_J^{-1/2}) \le \sigma_1(\mathbb{X}_J\Sigma_J^{-1/2}) \le C_{39}\sqrt{N}\}$. By [Ver18, Corollary 7.3.3, Exercise 7.3.4], we know that $\mathbb{P}(\Omega_{\text{RIP}}) \ge 1 - 2\exp(-c_{21}|J|)$ for some absolute constant $c_{21} < 1$.

**Multiplier process on $V_J$.**

Here we develop a novel probabilistic tool for bounding the multiplier process in cases where the multipliers may be dependent. The proof of the following Lemma 23 may be found in Section 4.9.1.

**Lemma 23.** *Let $F \subset L^2(\mu_X)$ be a functions class with sub-Gaussian increments with respect to $\|\cdot\|_{L^2(\mu_X)}$, that is, there exists an absolute constant $\theta_2 > 1$ such that for any $f, g \in F$, $\|f-g\|_{\psi_2} \le \theta_2 \|f-g\|_{L^2(\mu_X)}$. Let $\boldsymbol{w} = (w_i)_{i=1}^N \in \mathbb{R}^N$ be a deterministic vector. Let $X_1, \cdots, X_N$ be i.i.d. random vectors distributed as $\mu_X$. Suppose $\mathbf{0} \in F$. Then there exists an absolute constant $C_{42}$ depending only on $\theta_2$ such that for any $t > 0$, with probability at least $1 - 2\exp(-t^2)$,*

$$\sup\left( \left| \sum_{i=1}^N w_i \left(f(X_i) - \mathbb{E}[f(X)]\right) \right| : f \in F \right) \le C_{42} \|\boldsymbol{w}\|_2 \left( \gamma_2(F, d_{L^2(\mu_X)}) + t \operatorname{diam}(F, \|\cdot\|_{L^2(\mu_X)}) \right),$$

*where $\gamma_2(F, d_{L^2(\mu_X)})$ is the Talagrand's $\gamma_2$-functional of $F$ with respect to the distance generated by $\|\cdot\|_{L^2(\mu_X)}$ while $\operatorname{diam}(F, \|\cdot\|_{L^2(\mu_X)}) = \sup(\|f\|_{L^2(\mu_X)} : f \in F)$.*

**When $q = 1$.** Notice that $\mathbb{E}[(\xi + \langle X, \boldsymbol{\beta}^*_{J^c}\rangle)\langle X, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J\rangle] = \mathbb{E}[\xi\langle X, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J\rangle] + \mathbb{E}[\langle X, \boldsymbol{\beta}^*_{J^c}\rangle\langle X, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J\rangle] = 0$ because $X_J$ is independent with $X_{J^c}$. Applying Talagrand's majorizing measure theorem [Tal21, Theorem 2.10.1], Lemma 23 with $f(X) = f_{\boldsymbol{v}}(X) = \langle \boldsymbol{v}, X\rangle$, $w_i = \xi_i + \langle X_i, \boldsymbol{\beta}^*_{J^c}\rangle$, $F = \{\langle \cdot, \boldsymbol{v}\rangle : \boldsymbol{v} \in \rho K_{\text{model}} \cap r(\rho)\Sigma_J^{-1/2} B_2^J\}$, $t = \ell_*(\rho\Sigma_J^{1/2} K_{\text{model}} \cap r(\rho) B_2^J / r(\rho))$, we obtain that there exists an absolute constant$C_{43} > 1$ such that with probability at least $1 - 2\exp(-\ell_*^2(\rho\Sigma_J^{1/2} K_{\text{model}} \cap r(\rho)B_2^J/r^2(\rho)))$, the following event holds

$$\Omega^{1<q<2}_{\text{multi}} := \Bigg\{ \sup\left( \sum_{i=1}^N (\xi_i + \langle X_i, \boldsymbol{\beta}^*_{J^c}\rangle) \langle X_i, \boldsymbol{\beta}^*_J - \boldsymbol{\beta}_J\rangle : \boldsymbol{\beta}^*_J - \boldsymbol{\beta}_J \in \rho K_{\text{model}} \cap r(\rho)\Sigma_J^{-1/2} B_2^J \right) \\ \le \frac{1}{2} C_{43}\sqrt{N} \left( \sigma_\xi + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\right\|_2 \right) \ell_* \left( \rho\Sigma_J^{1/2} K_{\text{model}} \cap r(\rho) B_2^J \right) \Bigg\}. \tag{4.33}$$

Since $(\rho\Sigma_J^{1/2} K_{\text{model}} \cap r(\rho)B_2^J) \subset r(\rho)B_2^J$, we have $\ell_*\left(\rho\Sigma_J^{1/2} K_{\text{model}} \cap r(\rho) B_2^J\right) \le \ell_*(r(\rho)B_2^J) \lesssim r(\rho)\sqrt{|J|}$. Therefore there exist some absolute constants $\bar{\kappa}_{RIP} < 1$ and $C > 1$ such that the right-hand-side of (4.33) is less than

$$C|J| \left( \sigma_\xi + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\right\|_2 \right)^2 + \bar{\kappa}_{RIP} N r^2(\rho)$$

for some absolute constant $C>1$ (so that $\bar{\kappa}_{RIP}<1$). Therefore

$$\begin{aligned}&\mathbb{P}\bigg(\sup\bigg(\sum_{i=1}^N(\xi_i+\langle X_i,\boldsymbol{\beta}^*_{J^c}\rangle)\langle X_i,\boldsymbol{\beta}^*_J-\boldsymbol{\beta}_J\rangle:\boldsymbol{\beta}^*_J-\boldsymbol{\beta}_J\in\rho K_{\rm model}\cap r(\rho)\Sigma_J^{-1/2}B_2^J\bigg)\\&\le\delta_5^2\Big(\sigma_\xi+\Big\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\Big\|_2\Big)^2N+\bar{\kappa}_{RIP}Nr^2(\rho)\Big)\ge1-2\exp\big(-\frac{\ell_*^2(\rho\Sigma_J^{1/2}K_{\rm model}\cap r(\rho)B_2^J)}{r^2(\rho)}\big),\end{aligned}\tag{4.34}$$

where $\delta_5\sim\sqrt{C|J|/N}$.

**When $q>1$.** By (4.29) of Lemma 21, we consider the following multiplier process.

$$q\|\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}^*_J]\|_q^{q-1}\sup\bigg(\sum_{i=1}^N\boldsymbol{\lambda}_i^*[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}^*_J]\langle X_i,\boldsymbol{v}\rangle:\boldsymbol{v}\in\rho K_{\rm model}\cap r(\rho)\Sigma_J^{-1/2}B_2^J\bigg).$$

Apply Lemma 23 to $\boldsymbol{w}=\boldsymbol{\lambda}^*[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}^*_J]$, $F=\{f_{\boldsymbol{v}}(\cdot)=\langle\cdot,\boldsymbol{v}\rangle:\boldsymbol{v}\in\rho K_{\rm model}\cap r(\rho)\Sigma_J^{-1/2}B_2^J\}$. Recall that on $\Omega_{\rm DM,reg}(\varepsilon_1)\cap\Omega_{\rm noise}(\delta_5)$, by Lemma 20, there hold

$$\begin{aligned}&\|\boldsymbol{w}\|_2\le\frac{1}{(1-\varepsilon_1)\ell_*},\ \text{and}\\&\|\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}^*_J]\|_q^{q-1}\le\frac{\|\boldsymbol{\xi}+\mathbb{X}_{J^c}\boldsymbol{\beta}^*_{J^c}\|_2^{q-1}}{(1-\varepsilon_1)^{q-1}\ell_*^{q-1}}\lesssim_q\frac{N^{\frac{q-1}{2}}\big(\sigma_\xi^{q-1}+\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2^{q-1}\big)}{\ell_*^{q-1}}.\end{aligned}\tag{4.35}$$

Moreover, by Lemma 20 together with the independence between $\mathbb{X}_J$ and $\boldsymbol{\xi},\mathbb{X}_{J^c}$, $\mathbb{E}[w_i\langle Z_i,\boldsymbol{v}\rangle]=0$ for any $i\le N$. Notice that in Lemma 23, the probability measure is with respect to $\mu_{\mathbb{X}_J}$. Therefore, with probability at least $1-\mathbb{P}((\Omega_{\rm DM,reg}(\varepsilon_1)\cap\Omega_{\rm noise}(\delta_5))^c)-2\exp(-t^2)$, $\Omega^{q>1}_{\rm multi}$ defined in (4.23) holds, where $t=\ell_*(\rho\Sigma_J^{1/2}K_{\rm model}\cap r(\rho)B_2^J)/r(\rho)$.

**Remark 9.** *Since in the subsequent (4.51) we will remove the localization, we can still obtain an upper bound for the multiplier process indexed by $r(\rho)\Sigma_J^{-1/2}B_2^J$ even when $X_J$ satisfies only $\mathbb{E}\|\Sigma_J^{-1/2}X_J\|_2\lesssim\sqrt{|J|}$. Indeed, by (4.29), $\mathbb{E}_{\mathbb{X}_J}\sup\big(|\langle\boldsymbol{g},\boldsymbol{\beta}_J-\boldsymbol{\beta}^*_J\rangle|:\boldsymbol{\beta}_J-\boldsymbol{\beta}^*_J\in\rho K_{\rm model}\cap r(\rho)\Sigma_J^{-1/2}B_2^J\big)\le q\|\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}^*_J]\|_q^{q-1}r(\rho)\,\mathbb{E}_{\mathbb{X}_J}\sup\|\mathbb{X}_J^\top\boldsymbol{\lambda}^*[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}^*_J]\|_2\lesssim q\|\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}^*_J]\|_q^{q-1}r(\rho)\sqrt{|J|}$. Moreover, this does not affect the final bound on $\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}^*_J)\|_2$, but only changes the probability level to $\mathbb{P}(\Omega^{q>1}_{multi})\ge0.99$, while the constant $C_{40}$ is replaced by another absolute constant.*

**Restricted Isomorphic Property of random loss function when $1<q<2$.** The objective of this paragraph is to prove that with high probability, for any $\boldsymbol{\beta}_J\in\boldsymbol{\beta}^*_J+(r(\rho)\Sigma_J^{-1/2}S_2^J\cap\rho K_{\rm model})$ we have

$$P_N\mathcal{L}_{\boldsymbol{\beta}_J}=\|\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q-\|\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\boldsymbol{\beta}^*_J]\|_q^q\ge c\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho),$$

where $c<1$. In this paragraph, all the expectation is condition on $\mathbb{X}_{J^c}$.

**Proposition 29.** *Grant Assumption 9. Let $\boldsymbol{\xi}'=\boldsymbol{\xi}+\mathbb{X}_{J^c}\boldsymbol{\beta}^*_{J^c}$. Suppose for any $\boldsymbol{\beta}_J\in\boldsymbol{\beta}^*_J+(r(\rho)\Sigma_J^{-1/2}S_2^J\cap\rho K_{\rm model})$, the random vector $\boldsymbol{\zeta}=((r(\rho))^{-1}\mathbb{X}_J(\boldsymbol{\beta}_J-\boldsymbol{\beta}^*_J),\sigma_\xi^{-1}\boldsymbol{\xi}')\in\mathbb{R}^{2N}$ has Lipschitz concentration property, that is, for any Lipschitz function $f$, $f(\boldsymbol{\zeta})-\mathbb{E}f(\boldsymbol{\zeta})$ is a sub-Gaussian random variable whose sub-Gaussian norm is (up to multiplicative constant) the Lipschitz constant of $f$. There exist absolute constants $\kappa_3<1$ and $0<c<1$ such that the following hold. Suppose $\max\{\sigma_\xi\exp(-cN),r_Q(\rho)\}<r(\rho)<\sigma_\xi$ where:*

$$r_Q(\rho):=\min\Big(r>0:\ell_*(r(\rho)S_2^J\cap\rho\Sigma_J^{1/2}K_{\rm model})<\kappa_3\sigma_\xi^{-2}r^4(\rho)\sqrt{N}\Big).\tag{4.36}$$

*Then with probability at least*

$$1-\exp\Big(-c_{24}\sigma_\xi^{-2}Nr^4(\rho)\Big)-\mathbb{P}(\Omega_{RIP}^c)-\mathbb{P}(\Omega_{DM,reg}(\varepsilon_1)^c)-\mathbb{P}(\Omega_{noise}(\delta_5)^c),$$

*for any $\boldsymbol{\beta}_J\in\boldsymbol{\beta}^*_J+(r(\rho)\Sigma_J^{-1/2}S_2^J\cap\rho K_{\rm model})$, the lower isomorphic property of $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ holds, that is,*

$$\Omega_{iso}:=\bigg\{\forall\boldsymbol{\beta}_J\in\boldsymbol{\beta}^*_J+(r(\rho)\Sigma_J^{-1/2}S_2^J\cap\rho K_{\rm model}):P_N\mathcal{L}_{\boldsymbol{\beta}_J}\ge\frac{1}{2}c_{23}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho)\bigg\}.\tag{4.37}$$

When both $\mathbb{X}_J$ and $\boldsymbol{\xi}$ are Gaussian random vectors, the Lipschitz concentration property holds when either $\mathbb{X}_{J^c}$ is Gaussian or $\boldsymbol{\beta}^*_{J^c} = \mathbf{0}$. Indeed, when $X_{J^c} \sim \mathcal{N}(\mathbf{0}, \Sigma_{J^c})$, for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}^*_J + r(\rho)\Sigma_J^{-1/2} S_2^J \cap \rho K_{\text{model}}$, we have $(\mathbb{X}_J(\boldsymbol{\beta}^*_J - \boldsymbol{\beta}_J), \boldsymbol{\xi}') \sim \mathcal{N}(\mathbf{0}, \operatorname{diag}(r^2(\rho) I_N, (\sigma_\xi^2 + \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2^2) I_N))$; when $\boldsymbol{\beta}^*_{J^c} = \mathbf{0}$, we have $(\mathbb{X}_J(\boldsymbol{\beta}^*_J - \boldsymbol{\beta}_J), \boldsymbol{\xi}') \sim \mathcal{N}(\mathbf{0}, \operatorname{diag}(r^2(\rho) I_N, \sigma_\xi^2 I_N))$. Consequently, the Borell-TIS inequality (Gaussian Poincaré inequality) implies Lipschitz concentration. The fixed-point equation (4.36) characterizes the conditions under which the lower isomorphic property, that is, (4.37), holds. This is fundamentally different from the case when $q \geq 2$. For $q \geq 2$, the fixed-point equation holds for any $r(\rho)$. However, when $q < 2$, the fixed point for the lower isomorphic property of the random loss function becomes non-trivial. This may represent a significant distinction between the case of $q < 2$ and that of $q \geq 2$. The fixed point characterized by (4.36) should be regarded as an analogue of the quadratic fixed point.

**Proof.** Below we only treat the case $\boldsymbol{\beta}^*_{J^c} = \mathbf{0}$. When $\mathbb{X}_{J^c}$ is Gaussian, the analysis is similar; only the final constants change, and we omit it here. Fix $\boldsymbol{\beta}_J \in \boldsymbol{\beta}^*_J + (\rho K_{\text{model}} \cap r(\rho)\Sigma_J^{-1/2} S_2^J)$. Let $R = 4\sigma_\xi\sqrt{N}$, let $\pi_R : \boldsymbol{z} \in \mathbb{R}^{2N} \mapsto \operatorname{argmin}(\|\boldsymbol{z} - \boldsymbol{a}\|_2 : \boldsymbol{a} \in D_R)$, where $D_R = \{(\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^{2N} : \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \leq R, \|\boldsymbol{x}_2\|_2 \leq R\}$. Then $\pi$ is 1-Lipschitz. Let $F : (\boldsymbol{x}_1, \boldsymbol{x}_2) \in \mathbb{R}^N \times \mathbb{R}^N \mapsto \|\mathcal{A}[\boldsymbol{x}_1 - \boldsymbol{x}_2]\|_q^q - \|\mathcal{A}[\boldsymbol{x}_2]\|_q^q$, then $P_N\mathcal{L}_{\boldsymbol{\beta}_J} = F(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi})$. By Lemma 20, item 7, $F \circ \pi$ is Lipschitz on $\mathbb{R}^N$ with Lipschitz norm not larger than (up to multiplicative constant depending on $q$) $\frac{1}{\ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)} R^{q-1}$. We have

$$
\begin{aligned}
&P_N\mathcal{L}_{\boldsymbol{\beta}_J} - P\mathcal{L}_{\boldsymbol{\beta}_J} = F(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) - \mathbb{E}F(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \\
&= (F \circ \pi)(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) - \mathbb{E}[(F \circ \pi)(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi})] \qquad (4.38) \\
&+ F(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) - (F \circ \pi)(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \qquad (4.39) \\
&+ \mathbb{E}(F \circ \pi)(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) - \mathbb{E}F(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}). \qquad (4.40)
\end{aligned}
$$

Since $\pi$ is a projection, for any $a > 0$, $\mathbb{P}(|(4.39)| > a) \leq \mathbb{P}((\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \notin D_R)$. Since $R = 4\sigma_\xi\sqrt{N}$, $r(\rho) < \sigma_\xi$, on $\Omega_{\text{noise}}(\delta_5) \cap \Omega_{\text{RIP}}$, there exists an absolute constant $0 < c < 1$ such that $\mathbb{P}(|(4.39)| > a) \leq \exp(-cN)$. Then we deal with (4.40). By Jensen's inequality and triangular inequality,

$$
\begin{aligned}
|(4.40)| &= \left| \mathbb{E}\left[ \left( (F \circ \pi)(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) - F(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \right) \mathbb{1}((\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \notin D_R) \right] \right| \\
&\leq \mathbb{E}\left[ \left( |(F \circ \pi)(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi})| + |F((\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}))| \right) \mathbb{1}((\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \notin D_R) \right].
\end{aligned}
$$

By triangular inequality and Lemma 20, item 4, $|F(\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi})| \leq \frac{1}{(1-\varepsilon_1)^q \ell_*^q}(\|\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J) - \boldsymbol{\xi}\|_2^q + \|\boldsymbol{\xi}\|_2^q)$. Therefore, by Hölder's inequality, there exists an absolute constant $C_{44} > 1$ such that

$$
\begin{aligned}
|(4.40)| &\lesssim \frac{1}{\ell_*^q} \mathbb{E}\left[ \mathbb{1}((\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \notin D_R) \left( \|\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J) - \boldsymbol{\xi}\|_2^q + \|\boldsymbol{\xi}\|_2^q \right) \right] + \frac{1}{\ell_*^q} R^q \mathbb{P}((\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \notin D_R) \\
&\lesssim \frac{1}{\ell_*^q} \left( (\mathbb{P}((\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \notin D_R))^{\frac{1}{2}} \left( \mathbb{E}\left[ \left( \|\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J) - \boldsymbol{\xi}\|_2^q + \|\boldsymbol{\xi}\|_2^q \right)^2 \right] \right)^{\frac{1}{2}} + R^q \mathbb{P}((\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J), \boldsymbol{\xi}) \notin D_R) \right) \\
&\leq C_{44} \frac{\sigma_\xi^q}{\ell_*^q} N^{\frac{q}{2}} \exp(-cN).
\end{aligned}
$$

Combining (4.38), (4.39) and (4.40), for any $t > 2C_{44}\frac{\sigma_\xi^q}{\ell_*^q} N^{\frac{q}{2}} \exp(-cN)$,

$$
\left\{ \left| P_N\mathcal{L}_{\boldsymbol{\beta}_J} - P\mathcal{L}_{\boldsymbol{\beta}_J} \right| \geq t \right\} \subset \left\{ |(4.38)| \geq t - 2C_{44}\frac{\sigma_\xi^q}{\ell_*^q} N^{\frac{q}{2}} \exp(-cN) \right\} \cup \left\{ |(4.39)| > C_{44}\frac{\sigma_\xi^q}{\ell_*^q} N^{\frac{q}{2}} \exp(-cN) \right\}. \qquad (4.41)
$$

Let $c_{22} < 1$ be some absolute constant, and let $t = c_{22}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q} r^2(\rho)$. Because $r(\rho) > \sigma_\xi \exp(-cN)$, there holds $2C_{44}\frac{\sigma_\xi^q}{\ell_*^q} N^{\frac{q}{2}} \exp(-cN) < \frac{1}{2}t$. Taking $\mathbb{P}$ on both sides of (4.41) yields

$$
\mathbb{P}\left( \left| P_N\mathcal{L}_{\boldsymbol{\beta}_J} - P\mathcal{L}_{\boldsymbol{\beta}_J} \right| \geq c_{22}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)} r^2(\rho) \right) \leq \exp(-cN) + \mathbb{P}\left( |(4.38)| \geq \frac{1}{2} c_{22}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)} r^2(\rho) \right).
$$

We apply the Lipschitz concentration inequality to deal with (4.38).

$$\mathbb{P}\left(|(4.38)| \geq \frac{1}{2}c_{22}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho)\right) \lesssim \exp\left(-\frac{\frac{N^q\sigma_\xi^{2(q-2)}}{\ell_*^{2q}(\Sigma_{J^c}^{1/2}B_q^p)}r^4(\rho)}{\frac{\sigma_\xi^{2(q-1)}N^{(q-1)}}{\ell_*^{2q}(\Sigma_{J^c}^{1/2}B_q^p)}}\right) \sim \exp\left(-r^4(\rho)\sigma_\xi^{-2}N\right).$$

Take $c_{22}$ small enough, together with (4.25), there exist absolute constants $c_{23}$ and $c_{24}$ such that for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho K_{\text{model}})$,

$$\mathbb{P}\left(P_N\mathcal{L}_{\boldsymbol{\beta}_J} \geq c_{23}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho)\right) \geq 1-\exp\left(-c_{24}\sigma_\xi^{-2}Nr^4(\rho)\right).$$

Let $0<\varepsilon_2<1$ to be determined, and $V_{\varepsilon_2}$ be an $\varepsilon_2$-net with respect to $\|\Sigma_J^{1/2}\cdot\|_2$ on $\boldsymbol{\beta}_J^* + (r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho K_{\text{model}})$, that is, $V_{\varepsilon_2} = \{\pi\boldsymbol{\beta}_J : \boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho K_{\text{model}})\}$ such that for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho K_{\text{model}})$, there exists $\pi\boldsymbol{\beta}_J \in V_{\varepsilon_2}$ such that $\|\Sigma_J^{1/2}(\boldsymbol{\beta}_J - \pi\boldsymbol{\beta}_J)\|_2 \leq \varepsilon_2$. We determine $\varepsilon_2$ by determining the cardinality of $|V_{\varepsilon_2}|$ to be

$$|V_{\varepsilon_2}| = \left\lceil \exp\left(\frac{1}{2}c_{24}\sigma_\xi^{-2}Nr^4(\rho)\right)\right\rceil.$$

By Sudakov's inequality, see, for example, [Tal21, Lemma 2.10.2], there exists an absolute constant $C_{45}>1$ such that

$$\varepsilon_2 \leq C_{45}\frac{\ell_*(r(\rho)S_2^J \cap \rho\Sigma_J^{1/2}B_1^J)}{\sqrt{\log|V_{\varepsilon_2}|}} \leq C_{45}\sqrt{2c_{24}^{-1}}\frac{\sigma_\xi}{r^2(\rho)}\frac{\ell_*(r(\rho)S_2^J \cap \rho\Sigma_J^{1/2}B_1^J)}{\sqrt{N}}. \tag{4.42}$$

By a union bound, we obtain

$$\mathbb{P}\left(\forall\pi\boldsymbol{\beta}_J \in V_{\varepsilon_2},\, P_N\mathcal{L}_{\pi\boldsymbol{\beta}_J} \geq c_{23}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho)\right) \geq 1-\exp\left(-\frac{1}{2}c_{24}\sigma_\xi^{-2}Nr^4(\rho)\right).$$

By item 7 of Lemma 20, $\Omega_{\text{RIP}}$, $\Omega_{\text{noise}}(\delta_5)$ and the assumption that $r(\rho)<\sigma_\xi$, there exists an absolute constant $C_{46}>1$ such that

$$\begin{aligned}&\left|P_N\mathcal{L}_{\boldsymbol{\beta}_J} - P_N\mathcal{L}_{\pi\boldsymbol{\beta}_J}\right| = \left|\|\mathcal{A}[\mathbb{X}_J(\boldsymbol{\beta}_J-\boldsymbol{\beta}_J^*)-\boldsymbol{\xi}]\|_q^q - \|\mathcal{A}[\mathbb{X}_J(\pi\boldsymbol{\beta}_J-\boldsymbol{\beta}_J^*)-\boldsymbol{\xi}]\|_q^q\right| \\ &\lesssim \frac{2^{q-1}}{(1-\varepsilon_1)^q\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}\sigma_\xi^{q-1}N^{\frac{q-1}{2}}\|\mathbb{X}_J(\boldsymbol{\beta}_J-\pi\boldsymbol{\beta}_J)\|_2 \leq C_{46}\sigma_\xi^{q-1}\frac{N^{\frac{q}{2}}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}\varepsilon_2.\end{aligned}$$

As a result, for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho K_{\text{model}})$,

$$P_N\mathcal{L}_{\boldsymbol{\beta}_J} \geq c_{23}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho) - C_{46}\sigma_\xi^{q-1}\frac{N^{\frac{q}{2}}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}\varepsilon_2.$$

By (4.42) and (4.36), where we take $\kappa_3 = \frac{c_{23}\sqrt{c_{24}}}{4C_{46}C_{45}}$, we have $C_{46}\sigma_\xi^{q-1}\frac{N^{\frac{q}{2}}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}\varepsilon_2 < \frac{1}{2}c_{23}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho)$. Therefore,

$$\mathbb{P}\left(\forall\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho K_{\text{model}}),\, P_N\mathcal{L}_{\boldsymbol{\beta}_J} \geq \frac{1}{2}c_{23}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho)\right) \geq 1-\exp\left(-\frac{1}{2}c_{24}\sigma_\xi^{-2}Nr^4(\rho)\right).$$

■

**Summary of this subsection.** Summarizing the above stochastic arguments, we obtain the following proposition. The events $\Omega_{\mathrm{RIP}} \cap \Omega_{\mathrm{noise}}(\delta_5) \cap \Omega^{q>1}_{\mathrm{multi}}$ do not rely on the Gaussian assumption. When $q > 1$ and $X$ satisfies Assumption 8, replacing the Dvoretzky–Milman theorem by Theorem 3. When $X$ is not a Gaussian random vector but satisfies Assumption 8 and $q \geq 2$, we replace the term $(1+\varepsilon_1)\ell_*(\Sigma^{1/2}_{J^c}B^p_q)$ in (1.22) and (1.23) within $\Omega_{\mathrm{DM,reg}}$ with $(1+\varepsilon_1)\ell_*(\Sigma^{1/2}_{J^c}B^p_q)\mathrm{Log}(|V^c|)$.

**Proposition 30.** *Let $0 < \kappa_{DM}, \kappa_{RIP}, c_1, c_{20}, c_{24} < 1$ and $C_{47} > 1$ be some absolute constants. For any $0 < \varepsilon_1, \delta_5 < 1$, define stochastic events $\Omega_{reg,q=1}(\varepsilon_1,\delta_5) = \Omega_{DM,reg}(\varepsilon_1) \cap \Omega_{RIP} \cap \Omega_{\mathrm{noise}}(\delta_5) \cap \Omega^{q=1}_{multi}$, $\Omega_{reg,1<q<2}(\varepsilon_1,\delta_5) = \Omega_{DM,reg}(\varepsilon_1) \cap \Omega_{RIP} \cap \Omega_{\mathrm{noise}}(\delta_5) \cap \Omega^{q>1}_{multi} \cap \Omega_{iso}$ and $\Omega_{reg,q\geq 2}(\varepsilon_1,\delta_5) = \Omega_{DM,reg}(\varepsilon_1) \cap \Omega_{RIP} \cap \Omega_{noise}(\delta_5) \cap \Omega^{q>1}_{multi}$. Grant the same assumptions as Theorem 10.*

*Then $\Omega_{reg,1<q<2}$ or $\Omega_{reg,q\geq 2}(\varepsilon_1,\delta_5)$ holds respectively with probability at least*

$$\begin{aligned} &1 - 2\exp\Big(-\frac{\ell_*^2(\rho\Sigma_J^{1/2}K_{\mathrm{model}} \cap r(\rho)B_2^J)}{r^2(\rho)}\Big) - \exp(-\frac{1}{2}c_{24}\sigma_\xi^{-2}Nr^4(\rho))\mathbb{1}(1<q<2) \\ &- \bar{p}_{DM} - 2\exp(-c_{21}|J|) - \exp(-c_{20}\delta_5^2 N), \end{aligned} \tag{4.43}$$

*where $\bar{p}_{DM} = \exp(-c_1\varepsilon_1^2 d_*(\Sigma^{-1/2}_{J^c}B^p_{q'}))$ if $X$ is Gaussian, and otherwise $\bar{p}_{DM}$ is stated in Theorem 3.*

### 4.7.2 Deterministic Argument for regression problem

In what follows, we work on $\Omega_{\mathrm{reg},q=1}(\varepsilon_1,\delta_5)$, $\Omega_{\mathrm{reg},1<q<2}(\varepsilon_1,\delta_5)$, and $\Omega_{\mathrm{reg},q\geq 2}(\varepsilon_1,\delta_5)$, respectively. For convenience, we only present the result when $\Omega_{\mathrm{DM,reg}}$ itself holds—the proof is similar when an additional Log factor is present.

Recall that

$$|||\boldsymbol{\beta}||| = \sup\left(\langle\boldsymbol{\beta},\boldsymbol{u}\rangle : \boldsymbol{u} \in \frac{\rho}{r(\rho)}K_{\mathrm{model}} \cap \Sigma_J^{-1/2}B_2^J\right). \tag{4.44}$$

For the choice of $\rho$ in (4.31), both triple norm defined in (4.7) and in (4.44) are the same.

**Homogeneous argument** We recall the definition of $\hat{\boldsymbol{\beta}}_J$ from (4.16):

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

By Lemma 20, item *2.*, we know the optimization objective is convex, hence the homogeneous argument [CLL21] holds. We may restrict $\boldsymbol{\beta}_J$ to the boundary of $\boldsymbol{\beta}^*_J + \rho K_{\mathrm{model}} \cap r(\rho)\Sigma_J^{-1/2}B_2^p$. In fact, let $\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J \notin \rho K_{\mathrm{model}} \cap r(\rho)\Sigma_J^{-1/2}B_2^p$. There exists $\boldsymbol{\beta}^\circ_J \in \boldsymbol{\beta}^*_J + \partial(\rho K_{\mathrm{model}} \cap r(\rho)\Sigma_J^{-1/2}B_2^p)$ such that $\boldsymbol{\beta}_J = \boldsymbol{\beta}^*_J + \alpha(\boldsymbol{\beta}^\circ_J - \boldsymbol{\beta}^*_J)$ for some $\alpha \geq 1$. By convexity of $\boldsymbol{\beta}_J \mapsto \mathcal{L}_{\boldsymbol{\beta}_J} = \ell_{\boldsymbol{\beta}_J} - \ell_{\boldsymbol{\beta}^*_J}$ where $\ell_{\boldsymbol{\beta}_J} = \|\boldsymbol{\beta}_J\|_q^q + \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q$, we can check that $\mathcal{L}_{\boldsymbol{\beta}_J} \geq \alpha\mathcal{L}_{\boldsymbol{\beta}^\circ_J}$. As a consequence it is enough to prove that $\boldsymbol{\beta}_J \mapsto \mathcal{L}_{\boldsymbol{\beta}_J}$ is positive on the boarder of $\boldsymbol{\beta}^*_J + (\rho K_{\mathrm{model}} \cap r(\rho)\Sigma_J^{-1/2}B_2^p)$ to show that it is positive everywhere outside of $\boldsymbol{\beta}^*_J + (\rho K_{\mathrm{model}} \cap r(\rho)\Sigma_J^{-1/2}B_2^p)$.

For any $\boldsymbol{\beta}_J \in V_J$, we define the empirical excess regularized risk by

$$P_N\mathcal{L}^{(\mathrm{Reg})}_{\boldsymbol{\beta}_J} = \|\boldsymbol{\beta}_J\|_q^q + \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q - \left(\|\boldsymbol{\beta}^*_J\|_q^q + \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}^*_J]\|_q^q\right). \tag{4.45}$$

We have $P_N\mathcal{L}^{(\mathrm{Reg})}_{\boldsymbol{\beta}_J} = P_N\mathcal{L}_{\boldsymbol{\beta}_J} + \mathcal{R}_{\boldsymbol{\beta}_J}$ where

$$P_N\mathcal{L}_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q - \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}^*_J]\|_q^q, \text{ and} \tag{4.46}$$

$$\mathcal{R}_{\boldsymbol{\beta}_J} = \|\boldsymbol{\beta}_J\|_q^q - \|\boldsymbol{\beta}^*_J\|_q^q. \tag{4.47}$$

Our aim is to show that $P_N\mathcal{L}^{(\mathrm{Reg})}_{\boldsymbol{\beta}_J} > 0$ when $\boldsymbol{\beta}_J$ is far (w.r.t. the interpolationn norm $|||\cdot|||$ defined in (4.44)) from $\boldsymbol{\beta}^*_J$. We therefore need to lower bound the two terms $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ and $\mathcal{R}_{\boldsymbol{\beta}_J}$. To that end we obtain 'second' order lower bound for both terms. We start with the regularization term $\mathcal{R}_{\boldsymbol{\beta}_J}$.

**Lower bound for $\mathcal{R}_{\boldsymbol{\beta}_J}$** Applying (4.71) to $s = \beta_j^*$ and $t = \beta_j$ for each $j \in J$ and take sum over all $j \in J$, we obtain that

$$\mathcal{R}_{\boldsymbol{\beta}_J} = \|\boldsymbol{\beta}_J\|_q^q - \|\boldsymbol{\beta}_J^*\|_q^q = \sum_{j\in J} |\beta_j|^q - \left|\beta_j^*\right|^q \geq q \sum_{j\in J} \beta_j^* \left|\beta_j^*\right|^{q-2} \left(\beta_j - \beta_j^*\right) + \frac{q-1}{q2^q} \sum_{j\in J} \alpha_q \left(\left|\beta_j^*\right|, \beta_j - \beta_j^*\right). \tag{4.48}$$

When $q \geq 2$, one can show that $\sum_{j\in J} \alpha_q\big(|\beta_j^*|, \beta_j - \beta_j^*\big) \geq \|\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\|_q^q$ (this also follows from $q$-uniform convexity, see [Pis16, Section 10.1]). When $1 < q < 2$, we keep the term $\sum_{j\in J} \alpha_q\big(|\beta_j^*|, \beta_j - \beta_j^*\big)$ as it stands. Combining (4.48) with (4.44), we obtain that for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho K_{\text{model}} \cap r(\rho)\Sigma_J^{-1/2} B_2^J)$,

$$\mathcal{R}_{\boldsymbol{\beta}_J} \geq -q \left|\!\left|\!\left| \boldsymbol{\beta}_J^* \odot |\boldsymbol{\beta}_J^*|^{\odot(q-2)} \right|\!\right|\!\right| r(\rho) + \frac{q-1}{q2^q} \|\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\|_q^q. \tag{4.49}$$

We defer the treatment of the case $q = 1$ to Section 4.7.2.

**Proof of Lemma 18 and the lower bound for $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ when $q \geq 2$.** Applying (4.71) to $s = (\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*])_j$ and $t = (\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J])_j$ for each $j \in J^c$ and take sum over all $j \in J^c$, we obtain that

$$\begin{aligned}
&\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q - \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q^q = \sum_{j\in J^c} |(\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J])_j|^q - |(\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*])_j|^q \\
&\geq q \left\langle \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*] \odot |\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]|^{\odot(q-2)}, \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*] \right\rangle \\
&+ \frac{q-1}{q2^q} \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q^q.
\end{aligned}$$

Next, the second order term in the decomposition above is controlled on the event $\Omega_{\text{DM,reg}}(\varepsilon_1) \cap \Omega_{\text{RIP}}$; this will end the proof of Lemma 18. After, the first order term will be controlled on the event $\Omega_{\text{multi}}^{q>1}$.

By the definition of $\mathcal{A}[\mathbb{X}_J(\boldsymbol{\beta}_J^* - \boldsymbol{\beta}_J)]$ and the fact that $\mathbb{X}_{J^c}(\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]) = (\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J) - (\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*) = \mathbb{X}_J(\boldsymbol{\beta}_J^* - \boldsymbol{\beta}_J)$, we obtain that $\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q^q \geq \|\mathcal{A}[\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)]\|_q^q$. Moreover, from Lemma 21 together with Lemma 20, *item 5*, (4.28), we know that

$$q \left\langle \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*] \odot |\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]|^{\odot(q-2)}, \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*] \right\rangle = \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}_J^* \rangle,$$

where $\boldsymbol{g}$ is any sub-gradient in $(\partial^- P_N\ell_\bullet)(\boldsymbol{\beta}_J^*)$, hence condition on $\Omega_{\text{DM,reg}}(\varepsilon_1) \cap \Omega_{\text{RIP}}$, we have the following isomorphic profile for the empirical excess risk

$$P_N\mathcal{L}_{\boldsymbol{\beta}_J} \geq \frac{q-1}{q2^q} \frac{c_{19}^q N^{\frac{q}{2}} r^q(\rho)}{(1+\varepsilon_1)^q \ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)} + \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}_J^* \rangle. \tag{4.50}$$

**When $q \geq 2$**

**When $\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^* \in \rho K_{\text{model}} \cap r(\rho)\Sigma_J^{-1/2} S_2^{p-1}$.** Then there exists an absolute constant $c_{19} < 1$ such that for any such $\boldsymbol{\beta}_J$, we have $\|\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)\|_2 \geq c_{19} r(\rho)\sqrt{N}$. It follows from Lemma 20, equations (4.24), (4.22) and (1.22) and the control of the multiplier term on the event $\Omega_{\text{multi}}^{q>1}$ that

$$\begin{aligned}
&P_N\mathcal{L}_{\boldsymbol{\beta}_J} = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q - \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q^q \\
&\geq \frac{q-1}{q2^q} \|\mathcal{A}[\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)]\|_q^q - q \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q^{q-1} |\langle \boldsymbol{\lambda}^*[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*], \mathbb{X}_J(\boldsymbol{\beta}_J^* - \boldsymbol{\beta}_J)\rangle| \\
&\geq \frac{(q-1)c_{19}^q}{q2^{2q}} \frac{r^q(\rho) N^{\frac{q}{2}}}{\ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)} - qC_{40} r(\rho) \frac{(\sigma_\xi^{q-1} + \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\|_2^{q-1}) N^{\frac{q-1}{2}} \sqrt{|J|}}{\ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)}.
\end{aligned} \tag{4.51}$$

where we have used $\varepsilon_1 < 1$ to absorb $(1+\varepsilon_1)^q$ to $2^q$. Regarding the regularization term, by (4.49) we have

$$\mathcal{R}_{\boldsymbol{\beta}_J} = \|\boldsymbol{\beta}_J\|_q^q - \|\boldsymbol{\beta}_J^*\|_q^q \geq -q \left|\!\left|\!\left| \boldsymbol{\beta}_J^* \odot |\boldsymbol{\beta}_J^*|^{\odot(q-2)} \right|\!\right|\!\right| r(\rho),$$

where we have used the fact that $\alpha_q$ is non-negative. Therefore, we have $P_N\mathcal{L}^{(\mathrm{Reg})}_{\boldsymbol{\beta}_J} = P_N\mathcal{L}_{\boldsymbol{\beta}_J} + \mathcal{R}_{\boldsymbol{\beta}_J} > 0$ when

$$\frac{(q-1)c_{19}^q}{q2^{2q}}\frac{r^q(\rho)N^{\frac{q}{2}}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)} > qC_{40}r(\rho)\frac{(\sigma_\xi^{q-1}+\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2^{q-1})N^{\frac{q-1}{2}}\sqrt{|J|}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)} + q\Big|\Big|\Big|\boldsymbol{\beta}^*_J\odot|\boldsymbol{\beta}^*_J|^{\odot(q-2)}\Big|\Big|\Big|r(\rho).$$

This is equivalent to

$$r^{q-1}(\rho) > \frac{q2^{2q}}{(q-1)c_{19}^q}qC_{40}\frac{(\sigma_\xi^{q-1}+\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2^{q-1})\sqrt{|J|}}{\sqrt{N}} + q\Big|\Big|\Big|\boldsymbol{\beta}^*_J\odot|\boldsymbol{\beta}^*_J|^{\odot(q-2)}\Big|\Big|\Big|\frac{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}{N^{\frac{q}{2}}}\frac{q2^{2q}}{(q-1)c_{19}^q}.$$

There exists an absolute constant $C_{30}$ such that, for $r(\rho) = C_{30}\, r(V_J, V_{J^c})$, where $r(V_J, V_{J^c})$ is defined in (4.8), the inequality holds.

**When $\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J \in \rho\partial K_{\mathrm{model}} \cap r(\rho)\Sigma_J^{-1/2}B_2^p$.** As mentioned earlier, when $\boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J$ is 'well-spread', the regularization term $\mathcal{R}_{\boldsymbol{\beta}_J}$ will dominate the term coming from the noise, i.e. the multiplier process. It follows from (4.49) that

$$\mathcal{R}_{\boldsymbol{\beta}_J} \geq -q\Big|\Big|\Big|\boldsymbol{\beta}^*_J\odot|\boldsymbol{\beta}^*_J|^{\odot(q-2)}\Big|\Big|\Big|r(\rho) + \frac{q-1}{q2^q}\rho^q.$$

As a result, $P_N\mathcal{L}^{(\mathrm{Reg})}_{\boldsymbol{\beta}_J} = P_N\mathcal{L}_{\boldsymbol{\beta}_J} + \mathcal{R}_{\boldsymbol{\beta}_J} > 0$ when

$$\frac{q-1}{q2^q}\rho^q > qC_{40}r(\rho)\frac{(\sigma_\xi^{q-1}+\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}^*_{J^c}\|_2^{q-1})N^{\frac{q-1}{2}}\sqrt{|J|}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)} + qr(\rho)\Big|\Big|\Big|\boldsymbol{\beta}^*_J\odot|\boldsymbol{\beta}^*_J|^{\odot(q-2)}\Big|\Big|\Big|.$$

There exists an absolute constant $C_{29}$ such that, for $\rho$ equal to the $\rho_*$ defined in (4.31), the inequality holds.

**When $1 < q < 2$**

When $1 < q < 2$, the proof is almost identical to the case $q \geq 2$, using (4.37). The only two differences are that: 1.) in order to guarantee (4.37), we need to assume that $r(V_J, V_{J^c}) \geq r_Q(\rho_*)$, where $r_Q(\rho_*)$ is defined in (4.36). A rough estimate yields $r_Q(\rho_*) \lesssim \sigma_\xi^{\frac{1}{3}}(|J|/N)^{\frac{1}{6}}$, which is precisely the origin of this term in $r(V_J, V_{J^c})$ defined in (4.8); 2.) Since $r(V_J, V_{J^c}) < \sigma_\xi^{\frac{2-q}{2}}$ when $N$ is sufficiently large, the choice of $\rho_*$ in (4.31) guarantees that $\frac{q-1}{q2^q}\rho_*^q > \frac{1}{2}c_{23}\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}r^2(\rho_*)$, and therefore in the well-spread part we still automatically obtain $P_N\mathcal{L}^{(\mathrm{Reg})}_{\boldsymbol{\beta}_J} > 0$.

**When $q = 1$**

**Lower bound for $\mathcal{R}_{\boldsymbol{\beta}_J}$ over $\boldsymbol{\beta}_J \in \boldsymbol{\beta}^*_J + (\rho S_1^J \cap r(\rho)\Sigma_J^{-1/2}B_2^p)$.** When $q = 1$, the lower bound for $\|\boldsymbol{\beta}_J\|_1 - \|\boldsymbol{\beta}^*_J\|_1$ can only be handled using the sub-gradient. Fortunately, in this case, we are dealing with a classical $\ell_1$ norm regularization. We apply the sparsity equation developed in [LM17, LM18] to establish the lower bound. We denote by $S = \mathrm{supp}(\boldsymbol{\beta}^*) \cap J$. In that case, any sub-gradient $g$ of the $\ell_1^J$ norm in $\boldsymbol{\beta}^*_J$ is such that $g_S = \mathrm{sign}(\boldsymbol{\beta}^*_S)$ and $g_{S^c}$ can take any values in $[-1,1]^{J\setminus\mathrm{supp}(\boldsymbol{\beta}^*)}$. Let $\boldsymbol{v} \in C_\rho := \{\boldsymbol{v} \in V_J : \|\Sigma_J^{1/2}\boldsymbol{v}\|_2 \leq r(\rho), \|\boldsymbol{v}\|_1 = \rho\}$ and choose $g \in \partial\|\cdot\|_1(\boldsymbol{\beta}^*_J)$ such that $g_{S^c} = \mathrm{sgn}(\boldsymbol{v}_{S^c})$. We have

$$\langle g, \boldsymbol{v}\rangle = \langle g_S, \boldsymbol{v}_S\rangle + \langle g_{S^c}, \boldsymbol{v}_{S^c}\rangle \geq \|\boldsymbol{v}_{S^c}\|_1 - \|\boldsymbol{v}_S\|_1 = \|\boldsymbol{v}\|_1 - 2\|\boldsymbol{v}_S\|_1 = \rho - 2\|\boldsymbol{v}_S\|_1$$

and

$$\|\boldsymbol{v}_S\|_1 = \sum_{j\in S}\sqrt{\sigma_j}|v_j|\frac{1}{\sqrt{\sigma_j}} \leq \sqrt{\sum_{j\in S}\sigma_j v_j^2}\sqrt{\sum_{j\in S}\frac{1}{\sigma_j}} \leq r(\rho)\sqrt{\sum_{j\in S}\frac{1}{\sigma_j}}.$$

As a consequence the sparsity equation

$$\inf\left(\sup\left(\langle g, \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J\rangle : g \in \partial^-\|\cdot\|_1(\boldsymbol{\beta}^*_J)\right) : \boldsymbol{\beta}_J - \boldsymbol{\beta}^*_J \in \rho S_1^J \cap r(\rho)\Sigma_J^{-1/2}B_2^J\right) \geq \frac{4}{5}\rho$$

is fulfilled for any $\rho$ such that $2r(\rho)\sqrt{\sum_{j\in S}\sigma_j^{-1}} \leq \rho/5$.

**Lower bound for $P_N\mathcal{L}_{\boldsymbol{\beta}_J}$ over $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_1^J \cap r(\rho)\Sigma_J^{-1/2}B_2^p)$.** On the event $\Omega_{\mathrm{DM,reg}}(\varepsilon_1, q)$, for all $\boldsymbol{\beta}_J \in V_J$, we apply the isometric property of the norm $\|\mathcal{A}[\cdot]\|_1$ to both $\boldsymbol{\mu} = \boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J$ and $\boldsymbol{\mu} = \boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*$ in (1.22) to get

$$\begin{aligned}
P_N\mathcal{L}_{\boldsymbol{\beta}_J} &= \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_1 - \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_1 \geq \frac{\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J\|_2}{(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_1^p)} - \frac{\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*\|_2}{(1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_1^p)} \\
&= \frac{1}{(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_1^p)}\left(\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J\|_2 - \left(1 + \frac{2\varepsilon_1}{1-\varepsilon_1}\right)\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*\|_2\right) \\
&\geq \frac{1}{(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_1^p)}\left(\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J\|_2 - \|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*\|_2\right) - \frac{4\varepsilon_1}{(1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_1^p)}\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*\|_2,
\end{aligned}$$

where we have used that $1 + \frac{2\varepsilon_1}{1-\varepsilon_1} \leq 1 + 4\varepsilon_1$ as long as $0 < \varepsilon_1 < 1/2$.

1. Suppose $\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^* \in \rho B_1^J \cap r(\rho)\Sigma_J^{-1/2}S_2^{p-1}$, that is, $\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*$ is a almost-sparse vector. By the same analysis as in the $1 < q \leq 2$ situation, but with
$$\mathcal{R}_{\boldsymbol{\beta}_J} = \|\boldsymbol{\beta}_J\|_1 - \|\boldsymbol{\beta}_J^*\|_1 = \|\boldsymbol{\beta}_J^* - (\boldsymbol{\beta}_J^* - \boldsymbol{\beta}_J)\|_1 - \|\boldsymbol{\beta}_J^*\|_1 \geq -\|\boldsymbol{\beta}_J^* - \boldsymbol{\beta}_J\|_1 \geq -\rho,$$
we need $P_N^{(\mathrm{Reg})}\mathcal{L}_{\boldsymbol{\beta}_J} = P_N\mathcal{L}_{\boldsymbol{\beta}_J} + \mathcal{R}_{\boldsymbol{\beta}_J} > 0$, which is implied by
$$\begin{aligned}
&\left(\frac{1}{4}c_{19}^2 r^2(\rho) - 8\delta_5^2\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\|_2^2 + (1-9\delta_5)\sigma_\xi^2\right)^{1/2} \\
&- \frac{(1+\varepsilon_1)}{(1-\varepsilon_1)}\sqrt{(1+\delta_5)}\left(\sigma_\xi + \left\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\right\|_2\right) - \rho(1+\varepsilon_1)\frac{\ell^*(\Sigma_{J^c}^{1/2}B_1^p)}{\sqrt{N}} > 0.
\end{aligned} \tag{4.52}$$
2. Suppose $\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^* \in \rho S_1^p \cap r(\rho)\Sigma_J^{-1/2}B_2^p$, that is, $\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*$ is a well-spread vector. Then $\|\mathbb{X}_J(\boldsymbol{\beta}_J^* - \boldsymbol{\beta}_J)\|_2^2 \geq 0$, hence
$$\frac{\|\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J\|_2}{(1+\varepsilon_1)(\ell^*(\Sigma_{J^c}^{1/2}B_1^p))} \geq \left[\left(\frac{(1-9\delta_5)N\sigma_\xi^2}{(1+\varepsilon_1)^2(\ell^*(\Sigma_{J^c}^{1/2}B_1^p))^2} - 2\frac{\bar{\kappa}_{RIP}Nr^2(\rho) + 8\delta_5^2\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\|_2^2}{(1+\varepsilon_1)^2(\ell^*(\Sigma_{J^c}^{1/2}B_1^p))^2}\right) \vee 0\right]^{1/2}.$$
As mentioned earlier, when $\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*$ is well-spread, the regularization term $\mathcal{R}_{\boldsymbol{\beta}_J}$ will dominate the noise term. This is done by sparsity equation. By $2r(\rho)\sqrt{\sum_{j\in S}\sigma_j^{-1}} \leq \rho/5$, to have $P_N\mathcal{L}_{\boldsymbol{\beta}_J}^{(\mathrm{Reg})} > 0$, we need
$$\left[\left(\frac{(1-3\delta_5)N\sigma_\xi^2}{(1+\varepsilon_1)^2(\ell^*(\Sigma_{J^c}^{1/2}B_1^p))^2} - 2\frac{\bar{\kappa}_{RIP}Nr^2(\rho)}{(1+\varepsilon_1)^2(\ell^*(\Sigma_{J^c}^{1/2}B_1^p))^2}\right) \vee 0\right]^{1/2} - \left[\frac{\sqrt{(1+\delta_5)N}\sigma_\xi}{(1-\varepsilon_1)\ell^*(\Sigma_{J^c}^{1/2}B_1^p)}\right] + \frac{7}{10}\rho > 0,$$
which is equivalent to
$$\begin{aligned}
&\left[\left((1-9\delta_5)\sigma_\xi^2 - 2\bar{\kappa}_{RIP}r^2(\rho)\right.\right. \\
&\left.\left. - 8\delta_5^2\|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\|_2^2\right) \vee 0\right]^{1/2} - \left[\frac{(1+\varepsilon_1)}{(1-\varepsilon_1)}\sqrt{(1+\delta_5)}(\sigma_\xi + \|\Sigma_{J^c}^{1/2}\boldsymbol{\beta}_{J^c}^*\|_2)\right] + \frac{7}{10}\rho(1+\varepsilon_1)\frac{\ell^*(\Sigma_{J^c}^{1/2}B_1^p)}{\sqrt{N}} > 0.
\end{aligned} \tag{4.53}$$

One may check that $r(V_J, V_{J^c})$ and $\rho_*$ defined in (4.8) and (4.31) solve (4.52) and (4.53).

**Proposition 31.** *Grant the same assumptions as in Theorem 10. There exists an absolute constant $C_{30} = C_{30}(q)$ such that on the event $\Omega_{reg,q\geq 2}(\varepsilon_1, \delta_5)$, $\Omega_{reg,1<q<2}(\varepsilon_1, \delta_5)$ or $\Omega_{reg,q=1}(\varepsilon_1, \delta_5)$ respectively, $\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\|_2 \leq C_{30}r(V_J, V_{J^c})$ for $r(V_J, V_{J^c})$ defined in (4.8).*

### 4.7.3 Price for overfitting in the regression model

We utilize the $\ell_q \to \ell_2$ operator norm of $\Sigma_{J^c}^{1/2}$ (which, due to the diagonal structure of $\Sigma_{J^c}$, coincides with $\mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)$) and then apply the Dvoretzky-Milman theorem to obtain that on $\Omega_{\mathrm{reg},q>1}(\varepsilon_1, \delta_5)$ or $\Omega_{\mathrm{reg},q=1}(\varepsilon_1, \delta_5)$, there holds
$$\left\|\Sigma_{J^c}^{1/2}\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\hat{\boldsymbol{\beta}}_J]\right\|_2 \lesssim \mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)\frac{\sqrt{N}(r_*(\rho_*) + \sigma_\xi)}{\ell_*(\Sigma_{J^c}^{1/2}B_q^p)} \lesssim \sqrt{\frac{N}{d_*(\Sigma_{J^c}^{-1/2}B_{q'}^p)}}(r_*(\rho_*) + \sigma_\xi). \tag{4.54}$$

### 4.7.4 The end of the proof of Theorem 10

Combining Proposition 30, Proposition 31, and equation (4.54), we complete the proof of Theorem 10.

## 4.8 Proof of Theorem 11: Benign overfitting of the minimum $\ell_2$-norm interpolant classifier

### 4.8.1 Stochastic Arguments for classification problem

**Dvoretzky-Milman theorem.** We apply Theorem 4 to $\phi_{J^c}(X) = YX_{J^c}$. Since $\|YX_{J^c}\|_2 = \|X_{J^c}\|_2$, $\|\langle U, \boldsymbol{v}\rangle\|_{L^{2+\epsilon}} = \|\langle X_{J^c}, \boldsymbol{v}\rangle\|_{L^{2+\epsilon}(\mu_X)}$, and $\|\langle U, \boldsymbol{v}\rangle\|_{L^2} = \|\langle X_{J^c}, \boldsymbol{v}\rangle\|_{L^2(\mu_X)}$. We know that Assumption 2 is satisfied by $U$, once it is satisfied by $X_{J^c}$. Moreover, $\mathbb{E}[(YX_{J^c}) \otimes (YX_{J^c})] = \mathbb{E}[X_{J^c} \otimes X_{J^c}]$. Finally, we observe that when $XY$ are centered sub-Gaussian random vectors, condition (1.26) can be verified through the Hanson-Wright inequality, where $\delta$ may be taken as $\sqrt{\operatorname{Tr}(\Sigma_{J^c}^2)}/\operatorname{Tr}(\Sigma_{J^c})$, see, for instance Appendix F of [P4]. Under the Dvoretzky-Milman condition, this can be further amplified to $N^{-\gamma}$, where $0 < \gamma < 1$. Therefore, when $N$ is sufficiently large, we should regard $\delta_{\text{distortion}}$ as being dominated by its principal term $\bar{\delta}$.

The main corollary of Theorem 4 is the self-regularization property of the minimum $\ell_2$-norm interpolant classifier, as stated in Proposition 26. Its proof has already been completed in the discussion following the Proposition 26.

**Risk decomposition for the empirical excess risk of squared hinge loss.** In this section, $\ell$ denotes the squared hinge loss, namely $\ell_{\boldsymbol{\beta}_J} : (\boldsymbol{x}, y) \in (\mathbb{R}^p \times \{-1, 1\}) \mapsto (1 - y\langle \boldsymbol{x}, \boldsymbol{\beta}_J\rangle)_+^2$. Consequently, the excess loss $\mathcal{L}_{\boldsymbol{\beta}_J} = \ell_{\boldsymbol{\beta}_J} - \ell_{\boldsymbol{\beta}_J^*}$, where $\boldsymbol{\beta}_J^*$ is the oracle defined in (4.5) for the squared hinge loss.

We now address the relationship between the empirical excess risk $P_N\mathcal{L}$ and the population excess risk $P\mathcal{L}$ for the squared hinge loss.

Using the identity $\|\boldsymbol{a}\|_2^2 - \|\boldsymbol{b}\|_2^2 = \|\boldsymbol{a} - \boldsymbol{b}\|_2^2 + 2\langle \boldsymbol{b}, \boldsymbol{a} - \boldsymbol{b}\rangle$, applied to $\boldsymbol{a} = [\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]_+$, $\boldsymbol{b} = [\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J^*]_+$, and the definition $P_N\mathcal{L}_{\boldsymbol{\beta}_J} = \frac{1}{N}\sum_{i=1}^N \left((1 - Y_i\langle X_i, \boldsymbol{\beta}_J\rangle)_+^2 - (1 - Y_i\langle X_i, \boldsymbol{\beta}_J^*\rangle)_+^2\right) = \frac{1}{N}\left(\|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]_+\|_2^2 - \|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J^*]_+\|_2^2\right)$, we obtain:

$$
\begin{aligned}
P_N\mathcal{L}_{\boldsymbol{\beta}_J} &= \frac{1}{N}\left(\|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]_+ - [\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J^*]_+\|_2^2 + 2\left\langle [\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J^*]_+,\, [\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]_+ - [\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J^*]_+\right\rangle\right) \\
&= \frac{1}{N}\sum_{i=1}^N \left((1 - Y_i\langle X_i, \boldsymbol{\beta}_J\rangle)_+ - (1 - Y_i\langle X_i, \boldsymbol{\beta}_J^*\rangle)_+\right)^2 \\
&+ \frac{2}{N}\sum_{i=1}^N (1 - Y_i\langle X_i, \boldsymbol{\beta}_J^*\rangle)_+ \left((1 - Y_i\langle X_i, \boldsymbol{\beta}_J\rangle)_+ - (1 - Y_i\langle X_i, \boldsymbol{\beta}_J^*\rangle)_+\right).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
P\mathcal{L}_{\boldsymbol{\beta}_J} &= \mathbb{E}\left[(1 - Y\langle X, \boldsymbol{\beta}_J\rangle)_+^2 - (1 - Y\langle X, \boldsymbol{\beta}_J^*\rangle)_+^2\right] \\
&= \mathbb{E}\left[\left((1 - Y\langle X, \boldsymbol{\beta}_J\rangle)_+ - (1 - Y\langle X, \boldsymbol{\beta}_J^*\rangle)_+\right)^2\right] \\
&+ 2\mathbb{E}\left[(1 - Y\langle X, \boldsymbol{\beta}_J^*\rangle)_+ \left((1 - Y\langle X, \boldsymbol{\beta}_J\rangle)_+ - (1 - Y\langle X, \boldsymbol{\beta}_J^*\rangle)_+\right)\right].
\end{aligned}
$$

Therefore, for any $\boldsymbol{\beta}_J \in V_J$, $\left|(P - P_N)(\mathcal{L}_{\boldsymbol{\beta}_J})\right| \le \mathcal{Q}_{\boldsymbol{\beta}_J} + 2\mathcal{M}_{\boldsymbol{\beta}_J}$, where

$$
\mathcal{Q}_{\boldsymbol{\beta}_J} = \left|\frac{1}{N}\sum_{i=1}^N \left((1 - Y_i\langle X_i, \boldsymbol{\beta}_J\rangle)_+ - (1 - Y_i\langle X_i, \boldsymbol{\beta}_J^*\rangle)_+\right)^2 - \mathbb{E}\left[\left((1 - Y\langle X, \boldsymbol{\beta}_J\rangle)_+ - (1 - Y\langle X, \boldsymbol{\beta}_J^*\rangle)_+\right)^2\right]\right| \tag{4.55}
$$

is called the quadratic process and

$$
\begin{aligned}
\mathcal{M}_{\boldsymbol{\beta}_J} = \Bigg|&\frac{1}{N}\sum_{i=1}^N (1 - Y_i\langle X_i, \boldsymbol{\beta}_J^*\rangle)_+ \left((1 - Y_i\langle X_i, \boldsymbol{\beta}_J\rangle)_+ - (1 - Y_i\langle X_i, \boldsymbol{\beta}_J^*\rangle)_+\right) \\
&- \mathbb{E}\left[(1 - Y\langle X, \boldsymbol{\beta}_J^*\rangle)_+ \left((1 - Y\langle X, \boldsymbol{\beta}_J\rangle)_+ - (1 - Y\langle X, \boldsymbol{\beta}_J^*\rangle)_+\right)\right]\Bigg|
\end{aligned} \tag{4.56}
$$

is called the multiplier process.

**Quadratic process** We first make some notations. Let $(F, \|\cdot\|_{L^2(\mu)})$ be a normed space. Define

$$\gamma_2(F, \|\cdot\|_{L^2(\mu)}) = \inf_{(\mathcal{A}_n)_{n\geq 1}} \sup_{f\in F} \sum_{n=1}^{\infty} 2^{\frac{n}{2}} \min\left(\|f - \pi_n f\|_{L^2(\mu)} : \pi_n f \in \mathcal{A}_n\right),$$

where $(\mathcal{A}_n)_{n\geq 1}$ is called an admissible sequence, [Tal21, Definition 2.7.1], be the Talagrand's $\gamma_2$ functional, [Tal21, Definition 2.7.3]. We make use of the following result from [Dir15]. According to Theorem 5.5 in [Dir15], there is an absolute constant $C_{48} \geq 1$ such that for all $t \geq 1$, with probability at least $1 - \exp(-t)$,

$$\begin{aligned}\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^{N} f^2(U_i) - \mathbb{E}f^2(U)\right| \leq C_{48}\Bigg(&\frac{\operatorname{diam}(F, L^2(\mu))\gamma_2(F, \|\cdot\|_{L^2(\mu)})}{\sqrt{N}} \\ &+ \frac{\gamma_2(F, \|\cdot\|_{L^2(\mu)})^2}{N} + \operatorname{diam}(F, L^2(\mu))^2\left(\sqrt{\frac{t}{N}} + \frac{t}{N}\right)\Bigg)\end{aligned} \tag{4.57}$$

where $\operatorname{diam}(F, L^2(\mu)) := \sup(\|f\|_{L^2(\mu)} : f \in F)$.

We now derive an upper bound for the quadratic process. Since we have assumed that Assumption 7 holds. Let $F = \{f_{\boldsymbol{\beta}_J}(\boldsymbol{u}) = (1 - \langle \boldsymbol{u}, \boldsymbol{\beta}_J\rangle)_+ - (1 - \langle \boldsymbol{u}, \boldsymbol{\beta}_J^*\rangle)_+ : \boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho)\Sigma_J^{-1/2} S_2^J)\}$. Then $\mathcal{Q}_{\boldsymbol{\beta}_J} = |\frac{1}{N}\sum_{i=1}^N f^2_{\boldsymbol{\beta}_J}(Y_i X_i) - \mathbb{E}f^2_{\boldsymbol{\beta}_J}(YX)|$. We compute the diameter.

$$\begin{aligned}&\operatorname{diam}(F, \|\cdot\|_{L^2}) = \sup\left(\|(1 - \langle XY, \boldsymbol{\beta}_J\rangle)_+ - (1 - \langle XY, \boldsymbol{\beta}_J^*\rangle)_+\|_{L^2} : \boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho)\Sigma_J^{-1/2} S_2^J)\right) \\ &\leq \sup\left(\|(1 - \langle XY, \boldsymbol{\beta}_J\rangle - 1 + \langle XY, \boldsymbol{\beta}_J^*\rangle)\|_{L^2} : \boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho)\Sigma_J^{-1/2} S_2^J)\right) \\ &= \sup\left(\|\langle XY, \boldsymbol{\beta}_J^* - \boldsymbol{\beta}_J\rangle\|_{L^2} : \boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho)\Sigma_J^{-1/2} S_2^J)\right) \\ &\leq \sup\left(\|\langle X, \boldsymbol{\beta}_J^* - \boldsymbol{\beta}_J\rangle\|_{L^2} : \boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho)\Sigma_J^{-1/2} S_2^J)\right) = r(\rho).\end{aligned}$$

Moreover, by Talagrand's majorizing measure theorem, [Tal21, Theorem 2.10.1],

$$\begin{aligned}&\gamma_2(F, \|\cdot\|_{L^2}) = \inf_{(\mathcal{A}_n)_{n\geq 1}} \sup\left(\sum_{n=1}^{\infty} 2^{\frac{n}{2}} \min\left(\left\|\Sigma_J^{1/2}(\boldsymbol{\beta}_J - \pi_n\boldsymbol{\beta}_J)\right\|_2 : \pi_n\boldsymbol{\beta}_J \in \mathcal{A}_n\right) : \boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho)\Sigma_J^{-1/2} S_2^J)\right) \\ &= \gamma_2(\rho\Sigma_J^{1/2} B_2^J \cap r(\rho) S_2^J) \lesssim r(\rho)\sqrt{\dim(V_J)}.\end{aligned}$$

Applying (4.57) we obtain that with probability at least $1 - 2\exp(-(1/(400c_{25}^2))\dim(V_J))$, we have

$$\sup(\mathcal{Q}_{\boldsymbol{\beta}_J} : \boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho)\Sigma_J^{1/2} B_2^J)) \leq c_{25} r^2(\rho)\sqrt{\frac{\dim(V_J)}{N}},$$

where $c_{25} < 1$ depending on $\|X\|_{\psi_2}$ is an absolute constant.

**Multiplier process.**

**Lemma 24** ([Men16]). *For $q > 2$, there are constants $C_{49}$, $c_{26}$, $c_{27}$, and $C_{50}$ that depend only on $q$, for which the following holds. Let $\zeta \in L^q$ and $\zeta_1, \cdots, \zeta_N$ to be independent copies of $\zeta$. Let $w, u > C_{49}$. Suppose $Z$ is a random vector and $Z_1, \cdots, Z_N$ be independent copies of $Z$. Here $Z$ is not necessarily independent with $\zeta$. Let $F$ be a class of real-valued functions. Suppose $F$ is a sub-Gaussian class, that is, there exists an absolute constant $C$ and a metric $d_{\psi_2}$ such that for any $f, g \in F$ and $t > 0$, $\mathbb{P}(|f - g|(X) \geq t) \leq C\exp(-t^2 d_{\psi_2}^{-2}(f, g))$. Then with probability at least*

$$1 - c_{26}\frac{\log^q(N)}{w^q N^{\frac{q}{2}-1}} - 2\exp\left(-c_{27}u^2\left(\frac{\gamma_2(F, d_{\psi_2})}{\operatorname{diam}(F, d_{\psi_2})}\right)^2\right),$$

$$\sup_{f\in F}\left|\sum_{i=1}^{N}(\zeta_i f(Z_i) - \mathbb{E}[\zeta f(Z)])\right| \leq C_{50} wu \|\zeta\|_{L^q}\sqrt{N}\gamma_2(F, d_{\psi_2}).$$

We take the same choice of $F$ as in the quadratic process, and let $\zeta = (1 - Y\langle X, \boldsymbol{\beta}_J^*\rangle)_+$ in Lemma 24. Notice that for any $f_{\boldsymbol{\beta}_J}, g_{\tilde{\boldsymbol{\beta}}_J} \in F$, that is, for any $\boldsymbol{\beta}_J, \tilde{\boldsymbol{\beta}}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho)\Sigma_J^{1/2} B_2^J)$,

$$\begin{aligned}\left\|f_{\boldsymbol{\beta}_J} - g_{\tilde{\boldsymbol{\beta}}_J}\right\|_{\psi_2} &= \left\|((1 - \langle U, \boldsymbol{\beta}_J\rangle)_+ - (1 - \langle U, \boldsymbol{\beta}_J^*\rangle)_+) - \left((1 - \langle U, \tilde{\boldsymbol{\beta}}_J\rangle)_+ - (1 - \langle U, \boldsymbol{\beta}_J^*\rangle)_+\right)\right\|_{\psi_2} \\ &= \left\|(1 - \langle U, \boldsymbol{\beta}_J\rangle)_+ - (1 - \langle U, \tilde{\boldsymbol{\beta}}_J\rangle)_+\right\|_{\psi_2} \le \left\|\langle U, \boldsymbol{\beta}_J - \tilde{\boldsymbol{\beta}}_J\rangle\right\|_{\psi_2} \le \left\|\Sigma_J^{1/2}\left(\boldsymbol{\beta}_J - \tilde{\boldsymbol{\beta}}_J\right)\right\|_2,\end{aligned}$$

hence we may take $d_{\psi_2}(\bullet_1, \bullet_2) = \|\Sigma_J^{1/2}(\bullet_1 - \bullet_2)\|_2$. Applying Lemma 24 to $q = 4$, $w = u = 2C_{49}$, there exists an absolute constant $C_{51} > 1$ depending on $\|X\|_{\psi_2}$ such that with probability at least $1 - c\frac{\log^4(N)}{N} - \exp(-C\dim(V_J))$,

$$2\sup(\mathcal{M}_{\boldsymbol{\beta}_J} : \boldsymbol{\beta}_J \in F) \le C_{51} P\ell_{\boldsymbol{\beta}_J^*} r(\rho)\sqrt{\frac{\dim(V_J)}{N}}. \tag{4.58}$$

As a result, with probability at least $1 - c\frac{\log^4(N)}{N} - \exp(-C\dim(V_J))$, the following random event holds

$$\begin{aligned}\Omega_{\text{profile}} := \Big\{ &\sup\left(\left|(P - P_N)\left(\mathcal{L}_{\boldsymbol{\beta}_J}\right)\right| : \boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho)\Sigma_J^{1/2} B_2^J)\right) \\ &\le c_{25} r^2(\rho)\sqrt{\frac{\dim(V_J)}{N}} + C_{51} P\ell_{\boldsymbol{\beta}_J^*} r(\rho)\sqrt{\frac{\dim(V_J)}{N}}\Big\}.\end{aligned} \tag{4.59}$$

**Concentration of oracle risk** Let $C_{35} > 1$ be an absolute constant. Denote

$$\Omega_{\text{oracle}} := \left\{P_N \ell_{\boldsymbol{\beta}_J^*} \le C_{35} P\ell_{\boldsymbol{\beta}_J^*}\right\}. \tag{4.60}$$

Recall that $P_N\ell_{\boldsymbol{\beta}_J^*} = \frac{1}{N}\sum_{i=1}^N (1 - \langle Y_i X_i, \boldsymbol{\beta}_J^*\rangle)_+^2$, and $\{Y_i X_i\}_{i=1}^N$ are i.i.d. sub-Gaussian random vectors. We use the fact that $(1-x)_+^2 \le (1-x)^2$ for any $x \in \mathbb{R}$ to conclude that $\|(1 - \langle Y_i X_i, \boldsymbol{\beta}_J^*\rangle)_+^2\|_{\psi_1} \le \|(1 - \langle Y_i X_i, \boldsymbol{\beta}_J^*\rangle)^2\|_{\psi_1} \lesssim \|1 - \langle Y_i X_i, \boldsymbol{\beta}_J^*\rangle\|_{\psi_2}^2 < \infty$. Therefore, $\{(1 - \langle Y_i X_i, \boldsymbol{\beta}_J^*\rangle)_+^2\}_{i=1}^N$ are i.i.d. sub-exponential random variables. By Bernstein's inequality, see, for instance, [Ver18, Theorem 2.8.1], there exists an absolute constant $c_{18} < 1$ depending on $C_{35}$ such that

$$\mathbb{P}\left(\Omega_{\text{oracle}}\right) \ge 1 - \exp\left(-c_{18}\frac{N}{\max\{\|1 - \langle Y_i X_i, \boldsymbol{\beta}_J^*\rangle\|_{\psi_2}^2, \|1 - \langle Y_i X_i, \boldsymbol{\beta}_J^*\rangle\|_{\psi_2}^4\}}\right). \tag{4.61}$$

We summarize the above stochastic argument in the following proposition.

**Proposition 32** ($q = 2$)**.** *There exist some absolute constants $0 < \kappa_{RIP}, \kappa_{DM} < 1$ such that the following holds. Suppose $X_{J^c}$ is a sub-Gaussian random vector. Recall $\delta_4$ defined in (1.29). Suppose the choice of $V_J$ satisfies that*

$$\kappa_{RIP}^{-1}\dim(V_J) \le N \le \kappa_{DM}\bar{\delta}^2\frac{\mathrm{Tr}(\Sigma_{J^c})}{\|\Sigma_{J^c}\|_{op}}.$$

*Then there exist absolute constants $c_{17} < 1$ and $C_{32} > 1$ such that for any $0 < \delta_4 < 1$, the random event*

$$\Omega_{\text{class}}(\delta_4) = \Omega_{oracle} \cap \Omega_{DM,class}(\delta_4) \cap \Omega_{profile}$$

*holds with probability at least*

$$1 - \bar{p}_{DM}(\delta_4) - c_{17}\frac{\log^4(N)}{N} - \exp(-C_{32}\dim(V_J)) - \mathbb{P}(\Omega_{oracle}^c),$$

*where $\mathbb{P}(\Omega_{oracle}^c)$ is provided by (4.61).*

### 4.8.2 Deterministic Arguments for classification problem

In this section, we place ourselves on the random event $\Omega_{\text{class}}(\delta_4)$.

Recall that when $q = 2$, (4.7) reduces to

$$|||\boldsymbol{\beta}_J||| := \sup\left( \langle \boldsymbol{\beta}, \boldsymbol{u} \rangle : \boldsymbol{u} \in \frac{\rho}{r(\rho)} B_2^J \cap \Sigma_J^{-1/2} B_2^J \right). \tag{4.62}$$

By (4.21),

$$\begin{aligned}
&(1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c}) \left(L(\boldsymbol{\beta}_J) - L(\boldsymbol{\beta}_J^*)\right) \\
&\geq \|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}} \boldsymbol{\beta}_J]_+\|_2^2 - \frac{(1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c})}{(1-\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c})} \|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}} \boldsymbol{\beta}_J^*]_+\|_2^2 + (1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c}) \left( \|\boldsymbol{\beta}_J\|_2^2 - \|\boldsymbol{\beta}_J^*\|^2 \right) \\
&\geq \|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}} \boldsymbol{\beta}_J]_+\|_2^2 - \left(1 + \frac{4\delta_4}{(1-\delta_4)^2}\right) \|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}} \boldsymbol{\beta}_J^*]_+\|_2^2 + (1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c}) \left( \|\boldsymbol{\beta}_J\|_2^2 - \|\boldsymbol{\beta}_J^*\|^2 \right) \\
&= \|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}} \boldsymbol{\beta}_J]_+\|_2^2 - \|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}} \boldsymbol{\beta}_J^*]_+\|_2^2 + (1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c}) \left( \|\boldsymbol{\beta}_J\|_2^2 - \|\boldsymbol{\beta}_J^*\|^2 \right) - \frac{4\delta_4}{(1-\delta_4)^2} \|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}} \boldsymbol{\beta}_J^*]_+\|_2^2 .
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{(1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c})}{N} \left(L(\boldsymbol{\beta}_J) - L(\boldsymbol{\beta}_J^*)\right) \\
&\geq P_N \mathcal{L}_{\boldsymbol{\beta}_J} + \frac{(1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c})}{N} \|\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\|_2^2 + 2 \frac{(1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c})}{N} \langle \boldsymbol{\beta}_J^*, \boldsymbol{\beta}_J - \boldsymbol{\beta}_J^* \rangle - \frac{4\delta_4}{(1-\delta_4)^2} P_N \ell_{\boldsymbol{\beta}_J^*},
\end{aligned}$$

where we recall that

$$\begin{aligned}
P_N \mathcal{L}_{\boldsymbol{\beta}_J} &= \frac{1}{N} \sum_{i=1}^N \left( (1 - Y_i \langle X_i, \boldsymbol{\beta}_J \rangle)_+ - (1 - Y_i \langle X_i, \boldsymbol{\beta}_J^* \rangle)_+ \right)^2 \\
&+ \frac{2}{N} \sum_{i=1}^N (1 - Y_i \langle X_i, \boldsymbol{\beta}_J^* \rangle)_+ \left( (1 - Y_i \langle X_i, \boldsymbol{\beta}_J \rangle)_+ - (1 - Y_i \langle X_i, \boldsymbol{\beta}_J^* \rangle)_+ \right).
\end{aligned}$$

In this paragraph, we analyze an upper bound on the excess risk under the local Bernstein's condition, that is, Assumption 7.

- In this item, we study almost sparse vectors, i.e., when $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho B_2^J \cap r(\rho) \Sigma_J^{-1/2} S_2^J)$. In this case, we obtain a non-trivial uniform lower bound $c_{19} r^2(\rho)$ for $P_N \mathcal{L}_{\boldsymbol{\beta}_J}$. By the local Bernstein condition (4.10), together with (4.59), for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + \rho B_2^J \cap r(\rho) \Sigma_J^{-1/2} S_2^J$, we have

$$\begin{aligned}
P_N \mathcal{L}_{\boldsymbol{\beta}_J} &\geq L_1 r^{2\kappa_2}(\rho) - \left( c_{25} r^2(\rho) \sqrt{\frac{\dim(V_J)}{N}} + C_{51} P \ell_{\boldsymbol{\beta}_J^*} r(\rho) \sqrt{\frac{\dim(V_J)}{N}} \right) \\
&\gtrsim L_1 r^{2\kappa_2}(\rho) - C_{51} \sqrt{\frac{\dim(V_J)}{N}} (P \ell_{\boldsymbol{\beta}_J^*} \vee r(\rho)) r(\rho).
\end{aligned}$$

  Therefore,

$$\begin{aligned}
&\frac{(1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c})}{N} \left(L(\boldsymbol{\beta}_J) - L(\boldsymbol{\beta}_J^*)\right) \\
&\gtrsim L_1 r^{2\kappa_2}(\rho) - C_{51} \sqrt{\frac{\dim(V_J)}{N}} (P \ell_{\boldsymbol{\beta}_J^*} \vee r(\rho)) r(\rho) - 4 \frac{\operatorname{Tr}(\Sigma_{J^c})}{N} r(\rho) |||\boldsymbol{\beta}_J^*||| - \frac{4\delta_4}{(1-\delta_4)^2} P_N \ell_{\boldsymbol{\beta}_J^*} > 0
\end{aligned}$$

  if

$$r(\rho) \gtrsim \left( \left( L_1^{-1} P \ell_{\boldsymbol{\beta}_J^*} \sqrt{\frac{\dim(V_J)}{N}} \right)^{\frac{1}{2\kappa_2 - 1}} + \left( L_1^{-1} \frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}_J^*||| \right)^{\frac{1}{2\kappa_2 - 1}} + \left( L_1^{-1} \delta_4 P \ell_{\boldsymbol{\beta}_J^*} \right)^{\frac{1}{2\kappa_2}} \right). \tag{4.63}$$

- In this item, we examine well-spread vectors, i.e., when $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + (\rho S_2^J \cap r(\rho) \Sigma_J^{1/2} B_2^J)$. In this case, we will dominate other terms in the empirical excess risk by establishing a non-trivial lower bound for $\|\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\|_2^2$.

First, we establish that the response $Y$ is in a "well-behaved" position relative to our function class $\{\langle\cdot,\boldsymbol{\beta}_J\rangle : \boldsymbol{\beta}_J \in V_J\}$—thanks to the convexity of our function class. We begin by proving the following claim:

$$\langle (1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+,\, (1-Y\langle X,\boldsymbol{\beta}_J\rangle)_+ - (1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+\rangle_{L^2(\mu)} \geq 0, \quad \forall \boldsymbol{\beta}_J \in V_J. \tag{4.64}$$

Such claims frequently appear in statistical learning theory to describe the geometric properties of a function class relative to the response, specifically whether the oracle is in a well-positioned region, as discussed in [Men18]. Let $r : \boldsymbol{\beta}_J \in V_J \mapsto (1-\langle\cdot,\boldsymbol{\beta}_J\rangle)_+ \in L^2(\mu)$. Consider $F : r \in L^2(\mu) \mapsto \frac{1}{2}\mathbb{E}[r^2]$. Then $(\mathbf{D}F)(r^*_{\boldsymbol{\beta}_J^*}) : r \in L^2(\mu) \mapsto \langle r_{\boldsymbol{\beta}_J^*}, r\rangle_{L^2(\mu)}$, inducing a bounded linear functional on $L^2(\mu)$, see [BC11] for convex analysis over Hilbert space. By definition, $\boldsymbol{\beta}_J^*$ is the minimizer of $\boldsymbol{\beta}_J \in V_J \mapsto P\ell_{\boldsymbol{\beta}_J} = \mathbb{E}[(1-Y\langle X,\boldsymbol{\beta}_J\rangle)_+^2] = 2F(r_{\boldsymbol{\beta}_J})$, which is a convex functional, by [BC11, Proposition 26.5], for any $r_{\boldsymbol{\beta}} \in L^2(\mu)$, that is, for any $\boldsymbol{\beta}_J \in V_J$, we have $\langle r_{\boldsymbol{\beta}_J^*}, r - r_{\boldsymbol{\beta}_J^*}\rangle_{L^2(\mu)} \geq 0$, that is, $\mathbb{E}[(1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+((1-Y\langle X,\boldsymbol{\beta}_J\rangle)_+ - (1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+)] \geq 0$. As a consequence of (4.64), we have:

$$\begin{aligned}
&\frac{(1+\delta_4)^2\operatorname{Tr}(\Sigma_{J^c})}{N}\left(L(\boldsymbol{\beta}_J) - L(\boldsymbol{\beta}_J^*)\right)\\
&= \frac{1}{N}\sum_{i=1}^N \left((1-Y_i\langle X_i,\boldsymbol{\beta}_J\rangle)_+ - (1-Y_i\langle X_i,\boldsymbol{\beta}_J^*\rangle)_+\right)^2 - \mathbb{E}\left[\left((1-Y\langle X,\boldsymbol{\beta}_J\rangle)_+ - (1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+\right)^2\right]\\
&+ \frac{2}{N}\sum_{i=1}^N (1-Y_i\langle X_i,\boldsymbol{\beta}_J^*\rangle)_+\left((1-Y_i\langle X_i,\boldsymbol{\beta}_J\rangle)_+ - (1-Y_i\langle X_i,\boldsymbol{\beta}_J^*\rangle)_+\right)\\
&- 2\mathbb{E}[(1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+((1-Y\langle X,\boldsymbol{\beta}_J\rangle)_+ - (1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+)]\\
&+ 2\mathbb{E}[(1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+((1-Y\langle X,\boldsymbol{\beta}_J\rangle)_+ - (1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+)] + \mathbb{E}\left[\left((1-Y\langle X,\boldsymbol{\beta}_J\rangle)_+ - (1-Y\langle X,\boldsymbol{\beta}_J^*\rangle)_+\right)^2\right]\\
&+ \frac{(1+\delta_4)^2\operatorname{Tr}(\Sigma_{J^c})}{N}\|\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\|_2^2 + 2\frac{(1+\delta_4)^2\operatorname{Tr}(\Sigma_{J^c})}{N}\langle\boldsymbol{\beta}_J^*, \boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\rangle - \frac{4\delta_4}{(1-\delta_4)^2}P_N\ell_{\boldsymbol{\beta}_J^*}\\
&\geq \frac{(1+\delta_4)^2\operatorname{Tr}(\Sigma_{J^c})}{N}\|\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\|_2^2 - r(\rho)\sqrt{\frac{\dim(V_J)}{N}}\left(c_{25}r(\rho)\vee P\ell_{\boldsymbol{\beta}_J^*}\right)\\
&- 4\frac{\operatorname{Tr}(\Sigma_{J^c})}{N}r(\rho)|||\boldsymbol{\beta}_J^*||| - 8C_{35}\delta_4 P\ell_{\boldsymbol{\beta}_J^*}.
\end{aligned} \tag{4.65}$$

By (4.63), we have $\frac{(1+\delta_4)^2\operatorname{Tr}(\Sigma_{J^c})}{N}\left(L(\boldsymbol{\beta}_J) - L(\boldsymbol{\beta}_J^*)\right) > 0$, if

$$\rho \sim r^{\kappa_2}(\rho)\frac{\sqrt{N}}{\sqrt{\operatorname{Tr}(\Sigma_{J^c})}}. \tag{4.66}$$

In summary, we have

$$\begin{aligned}
&r_*(\rho_*) \sim \left(\left(L_1^{-1}P\ell_{\boldsymbol{\beta}_J^*}\sqrt{\frac{\dim(V_J)}{N}}\right)^{\frac{1}{2\kappa_2-1}} + \left(L_1^{-1}\frac{\operatorname{Tr}(\Sigma_{J^c})}{N}|||\boldsymbol{\beta}_J^*|||\right)^{\frac{1}{2\kappa_2-1}} + \left(L_1^{-1}\delta_4 P\ell_{\boldsymbol{\beta}_J^*}\right)^{\frac{1}{2\kappa_2}}\right),\\
&\text{and, } \rho_* \sim \frac{\sqrt{N}}{\sqrt{\operatorname{Tr}(\Sigma_{J^c})}}r_*^{\kappa_2}(\rho_*).
\end{aligned} \tag{4.67}$$

Now we find an upper bound for the excess risk. There exists an absolute constant $C_{52} > 1$ such that

$$\begin{aligned}
&\frac{(1+\delta_4)^2\operatorname{Tr}(\Sigma_{J^c})}{N}\left(L(\hat{\boldsymbol{\beta}}_J) - L(\boldsymbol{\beta}_J^*)\right)\\
&\geq (P_N - P)\mathcal{L}_{\hat{\boldsymbol{\beta}}_J} + P\mathcal{L}_{\hat{\boldsymbol{\beta}}_J} + \frac{(1+\delta_4)^2\operatorname{Tr}(\Sigma_{J^c})}{N}\left\|\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*\right\|_2^2\\
&- 2\frac{(1+\delta_4)^2\operatorname{Tr}(\Sigma_{J^c})}{N}r(\rho_*)|||\boldsymbol{\beta}_J^*||| - \frac{4C_{35}\delta_4}{(1-\delta_4)^2}P\ell_{\boldsymbol{\beta}_J^*}\\
&\geq P\mathcal{L}_{\hat{\boldsymbol{\beta}}_J} - C_{52}r_*(\rho_*)\sqrt{\frac{\dim(V_J)}{N}}(P\ell_{\boldsymbol{\beta}_J^*}\vee r_*(\rho_*)) - 2\frac{(1+\delta_4)^2\operatorname{Tr}(\Sigma_{J^c})}{N}r(\rho_*)|||\boldsymbol{\beta}_J^*||| - \frac{4C_{35}\delta_4}{(1-\delta_4)^2}P\ell_{\boldsymbol{\beta}_J^*} > 0,
\end{aligned}$$

if

$$P\mathcal{L}_{\hat{\boldsymbol{\beta}}_J} > C_{52} r_*(\rho_*) \sqrt{\frac{\dim(V_J)}{N}} (P\ell_{\boldsymbol{\beta}^*_J} \vee r_*(\rho_*)) + 2\frac{(1+\delta_4)^2 \operatorname{Tr}(\Sigma_{J^c})}{N} r(\rho_*) |||\boldsymbol{\beta}^*_J||| + 8C_{35}\delta_4 P\ell_{\boldsymbol{\beta}^*_J}.$$

By Young's inequality, for any $1 < \gamma < \infty$ and $\gamma'$ be its conjugate index, $ab \le a^\gamma/\gamma + b^{\gamma'}/\gamma'$ we obtain

- $\left(L_1^{-1} P\ell_{\boldsymbol{\beta}^*_J} \sqrt{\frac{\dim(V_J)}{N}}\right)^{\frac{1}{2\kappa_2 - 1}} \sqrt{\frac{\dim(V_J)}{N}} P\ell_{\boldsymbol{\beta}^*_J} = L_1^{\frac{1}{1-2\kappa_2}} (P\ell_{\boldsymbol{\beta}^*_J})^{\frac{2\kappa_2}{2\kappa_2-1}} (\frac{\dim(V_J)}{N})^{\frac{\kappa_2}{2\kappa_2-1}}.$
- $\left(L_1^{-1} \frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{1}{2\kappa_2-1}} \sqrt{\frac{\dim(V_J)}{N}} P\ell_{\boldsymbol{\beta}^*_J} \lesssim_{\kappa_2} L_1^{\frac{2\kappa_2}{1-2\kappa_2}} \left(\frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{2\kappa_2}{2\kappa_2-1}} + (\sqrt{\frac{\dim(V_J)}{N}} P\ell_{\boldsymbol{\beta}^*_J})^{\frac{2\kappa_2}{2\kappa_2-1}}.$
- $\left(L_1^{-1}\delta_4 P\ell_{\boldsymbol{\beta}^*_J}\right)^{\frac{1}{2\kappa_2}} \sqrt{\frac{\dim(V_J)}{N}} P\ell_{\boldsymbol{\beta}^*_J} \lesssim_{\kappa_2} L_1^{-1}\delta_4 P\ell_{\boldsymbol{\beta}^*_J} + (\sqrt{\frac{\dim(V_J)}{N}} P\ell_{\boldsymbol{\beta}^*_J})^{\frac{2\kappa_2}{2\kappa_2-1}}.$
- $\left(L_1^{-1} P\ell_{\boldsymbol{\beta}^*_J} \sqrt{\frac{\dim(V_J)}{N}}\right)^{\frac{1}{2\kappa_2-1}} \frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J||| \lesssim_{\kappa_2} L_1^{\frac{2\kappa_2}{1-2\kappa_2}} (\sqrt{\frac{\dim(V_J)}{N}} P\ell_{\boldsymbol{\beta}^*_J})^{\frac{2\kappa_2}{2\kappa_2-1}} + \left(\frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{2\kappa_2}{2\kappa_2-1}}.$
- $\left(L_1^{-1} \frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{1}{2\kappa_2-1}} \frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J||| = L_1^{\frac{1}{1-2\kappa_2}} \left(\frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{2\kappa_2}{2\kappa_2-1}}.$
- $\left(L_1^{-1}\delta_4 P\ell_{\boldsymbol{\beta}^*_J}\right)^{\frac{1}{2\kappa_2}} \frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J||| \lesssim_{\kappa_2} L_1^{-1}\delta_4 P\ell_{\boldsymbol{\beta}^*_J} + \left(\frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{2\kappa_2}{2\kappa_2-1}}.$

In particular, since $L_1 < 1$, we obtain that $L_1^{\frac{2\kappa_2}{1-2\kappa_2}} \vee L_1^{\frac{1}{1-2\kappa_2}} = L_1^{\frac{2\kappa_2}{1-2\kappa_2}}$ In summary,

$$P\mathcal{L}_{\hat{\boldsymbol{\beta}}_J} \lesssim L_1^{\frac{2\kappa_2}{2\kappa_2-1}} (\sqrt{\frac{\dim(V_J)}{N}} P\ell_{\boldsymbol{\beta}^*_J})^{\frac{2\kappa_2}{2\kappa_2-1}} + L_1^{\frac{2\kappa_2}{2\kappa_2-1}} \left(\frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{2\kappa_2}{2\kappa_2-1}} + L_1^{-1}\delta_4 P\ell_{\boldsymbol{\beta}^*_J}.$$

In summary,

**Proposition 33.** *Grant Assumption 7 with constants $L_1 < 1$ and $\kappa_2 \ge 1$. On the random event $\Omega_{\text{class}}(\delta_4)$, we have:*

$$P\mathcal{L}_{\hat{\boldsymbol{\beta}}_J} \lesssim_{L_1} \left(\frac{\dim(V_J)}{N}\right)^{\frac{\kappa_2}{2\kappa_2-1}} (P\ell_{\boldsymbol{\beta}^*_J})^{\frac{2\kappa_2}{2\kappa_2-1}} + \left(\frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{2\kappa_2}{2\kappa_2-1}} + \delta_4 P\ell_{\boldsymbol{\beta}^*_J}, \text{ and}$$
$$\left\|\Sigma_J^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}^*_J)\right\|_2 \lesssim r_*(\rho_*),$$

*where*

$$r_*(\rho_*) = \left(\left(L_1^{-1} P\ell_{\boldsymbol{\beta}^*_J} \sqrt{\frac{\dim(V_J)}{N}}\right)^{\frac{1}{2\kappa_2-1}} + \left(L_1^{-1} \frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{1}{2\kappa_2-1}} + \left(L_1^{-1}\delta_4 P\ell_{\boldsymbol{\beta}^*_J}\right)^{\frac{1}{2\kappa_2}}\right). \tag{4.68}$$

*In particular, we also establish a non-exact oracle inequality, that is, there exists an absolute constant $C > 1$ such that*

$$P\ell_{\hat{\boldsymbol{\beta}}_J} - (1+\delta')P\ell_{\boldsymbol{\beta}^*_J} \lesssim L_1^{\frac{2\kappa_2}{2\kappa_2-1}} (\sqrt{\frac{\dim(V_J)}{N}} P\ell_{\boldsymbol{\beta}^*_J})^{\frac{2\kappa_2}{2\kappa_2-1}} + L_1^{\frac{2\kappa_2}{2\kappa_2-1}} \left(\frac{\operatorname{Tr}(\Sigma_{J^c})}{N} |||\boldsymbol{\beta}^*_J|||\right)^{\frac{2\kappa_2}{2\kappa_2-1}},$$

*where $\delta' = CL_1^{-1}\delta_4$.*

**From the squared hinge excess risk to the 0-1 excess risk.** The relationship between the 0-1 risk and the risk associated with the squared hinge loss is given by the following Zhang-type inequality [Zha04]. Here, we make use of the results from [BJM03, Theorem 3 and Table 1], $P\mathcal{L}^{\{0,1\}}_{\hat{\boldsymbol{\beta}}_J} \leq (P\mathcal{L}_{\hat{\boldsymbol{\beta}}_J})^{1/2}$. The Zhang-type inequality has undergone numerous improvements. For instance, [BJM06] demonstrated that under the Mammen-Tsybakov noise/margin condition [MT99, Tsy04], the Zhang-type inequality can be refined.

According to [Cha25, Section 7.1, pp. 165], for the Gaussian Mixture classification Model, we have $P\mathcal{L}^{\{0,1\}}_{\hat{\boldsymbol{\beta}}_J} \lesssim (P\mathcal{L}_{\hat{\boldsymbol{\beta}}_J})^{\frac{2}{3}}$. However, in logistic models, if one only employs Zhang's inequality, one can obtain at most a 2/3-rate of convergence. This result was recently improved to the optimal rate through a breakthrough by [CLM24]. We refer readers to H. Chardon's PhD thesis [Cha25] for further details.

**Lemma 25** ([Cha25], Proposition 7.1, Proposition 7.2 and Lemma 7.2)**.** *Let $\boldsymbol{\theta}^* \in \mathbb{R}^p$ be a deterministic vector. Let $B = \max\{e, \|\boldsymbol{\theta}^*\|_2\}$. Let $\gamma$ be a binary random variable such that $\mathbb{P}(\gamma = 1|\boldsymbol{\zeta}) = \sigma(\langle\boldsymbol{\theta}^*, \boldsymbol{\zeta}\rangle)$ where $\sigma$ is the sigmoid function.*

1. *Let $\boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, I_p)$ be a standard Gaussian random vector. Then for any $\boldsymbol{u} \in S_2^{p-1}$, there holds $\mathbb{P}(\gamma\langle\boldsymbol{\zeta}, \boldsymbol{u}\rangle < 0) - \mathbb{P}(\gamma\langle\boldsymbol{\zeta}, \boldsymbol{e}_*\rangle < 0) \leq 3.2\|\boldsymbol{\theta}^*\|_2\|\boldsymbol{u} - \boldsymbol{e}_*\|_2^2$, where $\boldsymbol{e}_* = \|\boldsymbol{\theta}^*\|_2^{-1}\boldsymbol{\theta}^*$.*
2. *More generally, if $\boldsymbol{\zeta}$ is isotropic and sub-exponential with sub-exponential norm $K$. Suppose there exists an absolute constant $c > 0$ such that for all $t > 0$, $\mathbb{P}(|\langle\boldsymbol{e}_*, X\rangle| \leq t) \leq ct$. Then for all $\boldsymbol{u} \in S_2^{p-1}$, such that $\|\boldsymbol{u} - \boldsymbol{e}_*\|_2 \leq \frac{1}{2}$, there holds $\mathbb{P}(\gamma\langle\boldsymbol{\zeta}, \boldsymbol{u}\rangle < 0) - \mathbb{P}(\gamma\langle\boldsymbol{\zeta}, \boldsymbol{u}^*\rangle < 0) \leq 10cK^2\|\boldsymbol{\theta}^*\|_2\|\boldsymbol{u} - \boldsymbol{e}_*\|_2^2 \log^2(\|\boldsymbol{u} - \boldsymbol{e}_*\|_2^{-1})$.*

*Moreover, let $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^p$ by a deterministic vector, suppose $\|\tilde{\boldsymbol{\theta}}\|_2 > 1$. Let $H = \|\tilde{\boldsymbol{\theta}}\|_2^{-3}\boldsymbol{e}_* \otimes \boldsymbol{e}_* + \|\tilde{\boldsymbol{\theta}}\|_2^{-1}(I_p - \boldsymbol{e}_* \otimes \boldsymbol{e}_*)$, and let $0 < r < 1$. Then for any $\boldsymbol{\theta} \in \mathbb{R}^p$ such that $\|H^{1/2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\|_2 \leq \frac{1}{\sqrt{B}}r$, there holds $\|\|\boldsymbol{\theta}\|_2^{-1}\boldsymbol{\theta} - \|\tilde{\boldsymbol{\theta}}\|_2^{-1}\tilde{\boldsymbol{\theta}}\|_2 \leq \frac{\sqrt{2}}{(1-r)B}r$.*

The connection between Lemma 25 and logistic model is as follows: we correspond $\boldsymbol{\zeta}$ to $\Sigma^{-1/2}X \sim \mathcal{N}(\mathbf{0}, I_p)$, $\gamma$ to $Y$, $\boldsymbol{\theta}$ to $\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J$, $\boldsymbol{\theta}^*$ to $2\Sigma^{-1/2}\boldsymbol{\mu}$, $\tilde{\boldsymbol{\theta}}$ to $\Sigma^{1/2}\boldsymbol{\beta}_J^*$, and $\boldsymbol{u}$ to $\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J/\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2$. Recall that $\boldsymbol{\beta}_J^*$ is parallel to $\Lambda^{-1}\boldsymbol{\mu}$ and $\Sigma = \Lambda$, then $\tilde{\boldsymbol{\theta}} = \Sigma^{-1/2}\boldsymbol{\mu}$ is parallel to $\boldsymbol{\theta}^*$. Since we have chosen $V_J = \mathrm{span}(\boldsymbol{e}_*)$, it follows that $H = \|\tilde{\boldsymbol{\theta}}\|_2^{-3}I_{V_J}$. Consequently, $\|H^{1/2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\|_2 = \|\Sigma_J^{1/2}\boldsymbol{\beta}_J^*\|_2^{-3/2}\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}_J - \boldsymbol{\beta}_J^*)\|_2 \leq r\|\Sigma_J^{1/2}\boldsymbol{\beta}_J^*\|_2^{-1/2}$, provided we take $r = C_{33}\|\Sigma_J^{1/2}\boldsymbol{\beta}_J^*\|_2^{-1}r_*(\rho_*)$. From Lemma 25 we obtain:

$$\begin{aligned}
&(1.17) = \mathbb{P}\left(Y\langle X, \hat{\boldsymbol{\beta}}_J\rangle < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y\langle X, \boldsymbol{\beta}_J^*\rangle < 0\right) = \mathbb{P}(\gamma\langle\boldsymbol{\zeta}, \boldsymbol{u}\rangle < 0) - \mathbb{P}(\gamma\langle\boldsymbol{\zeta}, \boldsymbol{e}_*\rangle < 0) \\
&\leq 6.4\left\|\Sigma^{-1/2}\boldsymbol{\mu}\right\|_2\left\|\frac{\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J}{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2} - \frac{\Sigma^{1/2}\boldsymbol{\beta}_J^*}{\|\Sigma^{1/2}\boldsymbol{\beta}_J^*\|_2}\right\|_2^2 \lesssim \left\|\Lambda^{-1/2}\boldsymbol{\mu}\right\|_2^{-3} r_*^2(\rho_*),
\end{aligned} \tag{4.69}$$

where we have used the fact that when $V_J = \mathrm{span}(\boldsymbol{e}_*)$, (1.18) vanishes and where $r_*(\rho_*)$ is defined in (4.68). For the case where $\Sigma^{-1/2}X$ follows a sub-exponential distribution, Lemma 25 shows that the estimation error of $\hat{\boldsymbol{\beta}}_J$ differs from the Gaussian case (4.69) by at most a logarithmic factor.

### 4.8.3 Price for Overfitting for classification problem

**Proof of Proposition 15**

In this subsection, the probability should be viewed as the condition probability on $(X_i, Y_i)_{i=1}^N$. We first prove the following.

$$\begin{aligned}
&\mathbb{P}\left(Y\hat{f}(X) < 0\right) - \mathbb{P}\left(Y\hat{f}_J(X) < 0\right) \\
&= 2\mathbb{P}\left(\mathrm{sign}(\hat{f}(X)) \neq \mathrm{sign}(\hat{f}_J(X)) \text{ and } Y = \mathrm{sign}(\hat{f}_J(X))\right) - \mathbb{P}(\mathrm{sign}(\hat{f}(X)) \neq \mathrm{sign}(\hat{f}_J(X)))
\end{aligned} \tag{4.70}$$

We first consider the random event $\Omega_{\mathrm{margin}} := \{\mathrm{sign}(\hat{f}(X)) = \mathrm{sign}(\hat{f}_J(X))\}$, which indicates that the margin of $\hat{f}_{J^c}$ is not large compared to the margin of $\hat{f}_J$ and thus does not alter the sign of $\hat{f}(X)$. Consequently, we obtain

$$\begin{aligned}
&\mathbb{P}\left(Y\hat{f}(X) < 0\right) - \mathbb{P}\left(Y\hat{f}_J(X) < 0\right) \\
&= \mathbb{P}(\Omega_{\mathrm{margin}})\big(\mathbb{P}(\mathrm{sign}(\hat{f}(X)) \neq Y \mid \Omega_{\mathrm{margin}}) - \mathbb{P}(\mathrm{sign}(\hat{f}_J(X)) \neq Y \mid \Omega_{\mathrm{margin}})\big) \\
&+ \mathbb{P}(\Omega_{\mathrm{margin}}^c)\big(\mathbb{P}(\mathrm{sign}(\hat{f}(X)) \neq Y \mid \Omega_{\mathrm{margin}}^c) - \mathbb{P}(\mathrm{sign}(\hat{f}_J(X)) \neq Y \mid \Omega_{\mathrm{margin}}^c)\big) \\
&= \mathbb{P}(\Omega_{\mathrm{margin}}^c)\big(\mathbb{P}(\mathrm{sign}(\hat{f}(X)) \neq Y \mid \Omega_{\mathrm{margin}}^c) - \mathbb{P}(\mathrm{sign}(\hat{f}_J(X)) \neq Y \mid \Omega_{\mathrm{margin}}^c)\big).
\end{aligned}$$

On $\Omega^c_{\text{margin}}$, there exists exactly one among $\text{sign}(\hat{f}(X))$ and $\text{sign}(\hat{f}_J(X))$ that equals $Y$. Consequently, we have $\mathbb{P}(\text{sign}(\hat{f}(X)) = Y \mid \Omega^c_{\text{margin}}) = \mathbb{P}(\text{sign}(\hat{f}_J(X)) \neq Y \mid \Omega^c_{\text{margin}})$, and similarly, $\mathbb{P}(\text{sign}(\hat{f}(X)) \neq Y \mid \Omega^c_{\text{margin}}) = \mathbb{P}(\text{sign}(\hat{f}_J(X)) = Y \mid \Omega^c_{\text{margin}})$. Since these two events are mutually exclusive, their conditional probabilities sum to 1, leading to the conclusion that

$$\begin{aligned}&\mathbb{P}(\text{sign}(\hat{f}(X)) \neq Y \mid \Omega^c_{\text{margin}}) - \mathbb{P}(\text{sign}(\hat{f}_J(X)) \neq Y \mid \Omega^c_{\text{margin}})\\ &= \mathbb{P}(\text{sign}(\hat{f}_J(X)) = Y \big| \Omega^c_{\text{margin}}) - \mathbb{P}(\text{sign}(\hat{f}(X)) = Y \big| \Omega^c_{\text{margin}}) = 2\mathbb{P}(\text{sign}(\hat{f}_J(X)) = Y \big| \Omega^c_{\text{margin}}) - 1.\end{aligned}$$

Hence (4.70) follows.

We provide an intuitive interpretation of the first term on the right-hand side of (4.70): the random event involves in this term indicates that the prediction made by $\hat{f}_J$ on a new data point $(X, Y)$ is itself correct (i.e., $Y = \text{sign}(\hat{f}_J(X))$), but the component $\hat{f}_{J^c}$, which is supposed to be used for noise interpolation, has a large margin, causing the overall predictor $\hat{f}$ to reverse the prediction and output an incorrect result. Therefore, the probability of this random event quantifies the impact of noise interpolation by $\hat{f}_{J^c}$ on the 0-1 classification risk.

Since analyzing the random event $\{\text{sign}(\hat{f}(X)) \neq \text{sign}(\hat{f}_J(X)) \text{ and } Y = \text{sign}(\hat{f}_J(X))\}$ is relatively intricate, we directly consider an upper bound on the probability of the event $\{\text{sign}(\hat{f}(X)) \neq \text{sign}(\hat{f}_J(X))\}$, which will serve as an upper bound for (4.70). Since $\text{sign}(\hat{f}(X)) \neq \text{sign}(\hat{f}_J(X))$ implies that $|\hat{f}_{J^c}(X)| > |\hat{f}_J(X)|$, it follows that we only need to analyze an upper bound on the probability of the random event $\{|\hat{f}_{J^c}(X)| > |\hat{f}_J(X)|\}$. This is precisely the conclusion claimed by Proposition 15, completing the proof.

**Proof of Proposition 27**

The correlation between $\langle X, \hat{\boldsymbol{\beta}}_J\rangle$ and $\langle X, \hat{\boldsymbol{\beta}}_{J^c}\rangle$ is given by $\frac{\langle \hat{\boldsymbol{\beta}}_J, \Sigma\hat{\boldsymbol{\beta}}_{J^c}\rangle}{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2}$. Therefore, the ratio $\frac{\langle X, \hat{\boldsymbol{\beta}}_{J^c}\rangle}{\langle X, \hat{\boldsymbol{\beta}}_J\rangle}$ follows the Cauchy distribution whose probability density function is given by

$$p(z) = \frac{1}{\pi}\frac{t}{(z-s)^2 + t^2},$$

where

$$s = \frac{\langle \Sigma^{1/2}\hat{\boldsymbol{\beta}}_J, \Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\rangle}{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2}\frac{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2}{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2}, \text{ and } t = \frac{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2}{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2}\sqrt{1 - \frac{\langle \Sigma^{1/2}\hat{\boldsymbol{\beta}}_J, \Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\rangle^2}{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2^2\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2^2}}.$$

Therefore, if we denote $F$ as the cumulant distribution function of $p(z)$, then

$$\mathbb{P}(|\langle X, \hat{\boldsymbol{\beta}}_{J^c}\rangle| > |\langle X, \hat{\boldsymbol{\beta}}_J\rangle|) = 1 - (F(1) - F(-1)) = 1 - \frac{1}{\pi}\left(\arctan\left(\frac{1-s}{t}\right) + \arctan\left(\frac{1+s}{t}\right)\right).$$

We may also obtain an upper bound for it in terms of $\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2$ and $\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2$. In that, by Cauchy's inequality, $s \le \|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2/\|\Sigma^{1/2}\boldsymbol{\beta}_J\|_2$. When $|z| \le \|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J_c}\|_2/\|\Sigma^{1/2}\boldsymbol{\beta}_J\|_2 < 1$, we have $(z-s)^2 + t^2 \ge (z-s)^2$ and hence $p(z) \le \frac{1}{\pi}\frac{t}{(z-s)^2}$, which indicates that $\int_1^\infty p(z)dz \le \frac{1}{\pi}\frac{t}{1-s}$. Similarly we obtain that $\int_{-\infty}^1 p(z)dz \le \frac{1}{\pi}\frac{t}{1+s}$. As a result, $1 - (F(1) - F(-1)) \le \frac{2t}{\pi(1-s^2)} \le \frac{2}{\pi}\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2\frac{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2}{\|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_J\|_2^2 - \|\Sigma^{1/2}\hat{\boldsymbol{\beta}}_{J^c}\|_2^2}$.

**Proof of Proposition 28: noise absorption in classification**

Condition on $\Omega_{\text{DM,class}}(\delta_4) \cap \Omega_{\text{oracle}}$, where we recall that $\Omega_{\text{DM,class}}(\delta_4)$ is defined in (1.22) and $\Omega_{\text{oracle}}$ is defined in (4.60). Then

$$\left\|\Sigma_{J^c}^{1/2}\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\hat{\boldsymbol{\beta}}_J]\right\|_2 \le \|\Sigma_{J^c}\|_{\text{op}}^{1/2}\left\|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\hat{\boldsymbol{\beta}}_J]\right\|_2 \le \frac{\|\Sigma_{J^c}\|_{\text{op}}^{1/2}}{(1-\delta_4)\sqrt{\text{Tr}(\Sigma_{J^c})}}\left\|[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\hat{\boldsymbol{\beta}}_J]_+\right\|_2 \le \frac{\sqrt{C_{35}\|\Sigma_{J^c}\|_{\text{op}}N}}{(1-\delta_4)\sqrt{\text{Tr}(\Sigma_{J^c})}}P\ell_{\boldsymbol{\beta}^*_J}.$$

Now we use

$$\frac{\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 + C_{33}r(V_J, V_{J^c})}{(\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 - C_{33}r(V_J, V_{J^c}))^2 - \frac{N\|\Sigma_{J^c}\|_{\text{op}}}{\text{Tr}(\Sigma_{J^c})}\frac{C_{35}}{(1-\delta_4)^2}P\ell_{\boldsymbol{\beta}^*_J}} \le 2,$$

provided by

$$\frac{N\|\Sigma_{J^c}\|_{\mathrm{op}}}{\mathrm{Tr}(\Sigma_{J^c})} \leq \frac{(\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 - C_{33}r(V_J, V_{J^c}))^2 - \frac{1}{2}(\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 + C_{33}r(V_J, V_{J^c}))}{\frac{C_{35}}{(1-\delta_4)^2}P\ell_{\boldsymbol{\beta}^*_J}}$$

to conclude the proof. Here, the assumption $r(V_J, V_{J^c}) < \frac{1}{4C_{33}}(4\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 + 1 - \sqrt{16\|\Sigma^{1/2}\boldsymbol{\beta}^*_J\|_2 + 1})$ ensures that such $\frac{N\|\Sigma_{J^c}\|_{\mathrm{op}}}{\mathrm{Tr}(\Sigma_{J^c})} > 0$ exists.

### 4.8.4 End of the proof of Theorem 11

The first part of Theorem 11 follows from the combination of Proposition 32 and Proposition 33. The second part of Theorem 11 is obtained as follows: Lemma 28 verifies Assumption 7; Proposition 28, through (4.26), provides an upper bound for (1.16); Lemma 27 shows that (1.18) equals zero; and (4.69), together with the first part of Theorem 11, yields an upper bound on (1.17) for the logistic model. Finally, the classical Zhang's inequality (see [Cha25, Section 7.1, pp. 165]), combined with the first part of Theorem 11, gives the upper bound on (1.17) for the Gaussian mixture classification model.

## 4.9 Auxiliary lemmas

### 4.9.1 Proof of Lemma 23

**Proof.** For any $f \in F$, we let $Z_f = \sum_{i=1}^N w_i(f(X_i) - \mathbb{E}[f(X)])$, then $\mathbb{E}[Z_f] = 0$. For any $f, g \in F$, since $f - g$ is sub-Gaussian, there exists an absolute constant $C_{53}$ depending only on $\theta_2$ such that, for any $\lambda \in \mathbb{R}$ (see, for instance, [Ver18, Proposition 2.5.2]),

$$\begin{aligned}
&\mathbb{E}\exp\left(\lambda(Z_f - Z_g)\right) = \mathbb{E}\exp\left[\sum_{i=1}^N \lambda w_i\left((f-g)(X_i) - \mathbb{E}[(f-g)(X)]\right)\right] \\
&= \prod_{i=1}^N \mathbb{E}\exp\left[\lambda w_i\left((f-g)(X_i) - \mathbb{E}[(f-g)(X)]\right)\right] \leq \prod_{i=1}^N \exp\left[C_{53}\lambda^2 w_i^2 \|f-g\|^2_{L^2(\mu_X)}\right] \\
&= \exp\left(C_{53}\lambda^2 \|\boldsymbol{w}\|_2^2 \|f-g\|^2_{L^2(\mu_X)}\right).
\end{aligned}$$

This implies that $\|Z_f - Z_g\|_{\psi_2} \lesssim \|\boldsymbol{w}\|_2 \|f-g\|_{L^2(\mu_X)}$. By generic chaining, see, for instance, [Ver18, Theorem 8.5.5] or [Tal21, Section 2.4], there exists an absolute constant $C_{42}$ depending only on $\theta_2$ such that for any $t > 0$, with probability at least $1 - 2\exp(-t^2)$,

$$\sup(|Z_f| : f \in F) \leq C_{42}\|\boldsymbol{w}\|_2\left(\gamma_2(F, d_{L^2(\mu_X)}) + t\,\mathrm{diam}(F, \|\cdot\|_{L^2(\mu_X)})\right).$$

■

### 4.9.2 A lemma on the $\ell_q$ norm

**Lemma 26.** *Let $1 < q < \infty$. Then for any $x$ and $\Delta$,*

$$|x+\Delta|^q \geq |x|^q + q|x|^{q-2}x\Delta + \frac{q-1}{q2^q}\alpha_q(|x|, \Delta).$$

*This bound is sharp up to multiplicative constant.*

Applying Lemma 26 with $x = s$ and $\Delta = t - s$ for any $s, t \in \mathbb{R}$ we have

$$|t|^q - |s|^q = |s + (t-s)|^q - |s|^q \geq qs\,|s|^{q-2}(t-s) + \frac{q-1}{q2^q}\alpha_q(|s|, t-s). \tag{4.71}$$

### 4.9.3 Property of squared hinges loss

**Lemma 27.** *Let $f^{**} \in \operatorname{argmin}(P\ell_f : f \in L^2(\mu))$, where $\ell_f : (\boldsymbol{x}, y) \in \mathbb{R}^p \times \{-1, 1\} \mapsto (1 - yf(\boldsymbol{x}))_+^2$ is the squared hinge loss. Then $f^{**}(\cdot) = \operatorname{sign}(2\eta(\cdot) - 1)$.*

**Proof.** Recall that the Bayes rule defined over the class of all *measurable* functions is given by $\operatorname{sign}(2\eta(\boldsymbol{x}) - 1)$. Denote $f^{**}(\boldsymbol{x}) = \operatorname{argmin}_{t\in\mathbb{R}}(\mathbb{E}[(1 - Yt)_+^2 | X = \boldsymbol{x}])$ be the oracle of squared hinge loss. Let $R(t) = \eta(\boldsymbol{x})(1 - t)_+^2 + (1 - \eta(\boldsymbol{x}))(1 + t)_+^2$. When $t \le -1$, then $R(t) = \eta(\boldsymbol{x})(1 - t)^2$ has minimum $4\eta$ when $t = -1$; when $t \ge 1$, $R(t) = (1-\eta(\boldsymbol{x}))(1+t)^2$ has minimum $4(1-\eta)$ at $t = 1$; when $-1 < t < 1$, $R(t) = \eta(\boldsymbol{x})(1-t)^2 + (1-\eta(\boldsymbol{x}))(1+t)^2 = \eta(\boldsymbol{x})(t^2 + 1 - 2t) + (1 - \eta(\boldsymbol{x}))(1 + t^2 + 2t)$, which has minimum at $t = 2\eta(\boldsymbol{x}) - 1 \in \{t : -1 < t < 1\}$. Therefore,

$$f^{**}(\boldsymbol{x}) = \operatorname{sign}(2\eta(\boldsymbol{x}) - 1).$$

Therefore, minimizing the squared hinge loss (in expectation) yields the Bayes classifier. ∎

### 4.9.4 Verification of the Local Bernstein Condition for Squared Hinge Loss

We verify the local Bernstein assumption, that is, the validity of Assumption 7. Although this condition is highly general, we provide the verification only for the two examples (Gaussian Mixture Model and logistic model) mentioned earlier.

**Lemma 28.** *Let $\boldsymbol{e}^* = \Lambda^{-1}\boldsymbol{\mu}/\|\Lambda^{-1}\boldsymbol{\mu}\|_2$ in gaussian mixture classification model and logistic model. There exists an absolute constant $\delta > 0$ which depends only on the probability distribution function of $\langle XY, \boldsymbol{e}^*\rangle$ such that for $r(\rho) < \delta$, and $V_J = \operatorname{span}(\boldsymbol{e}^*)$, then for the Gaussian Mixture Model and the logistic model, Assumption 7 holds with parameters $\kappa_2 = 1$ and some $L_1$.*

**Proof.** Since $P\ell_{\boldsymbol{\beta}_J}$ is a convex function of $\boldsymbol{\beta}_J$, there exists at least one $\alpha_* \in \mathbb{R}_+$ such that $\boldsymbol{\beta}_J^* = \alpha_*\boldsymbol{e}^*$, where $\boldsymbol{e}^* = \Lambda^{-1}\boldsymbol{\mu}/\|\Lambda^{-1}\boldsymbol{\mu}\|_2$. Define $R : \alpha \in \mathbb{R} \mapsto P\ell_{\alpha\boldsymbol{e}^*}$, and let $M = \langle XY, \boldsymbol{e}^*\rangle$. Then $R(\alpha) = \mathbb{E}(1 - \alpha M)_+^2$.

Since $M$ has a continuous probability distribution, both $R'(\alpha)$ and $R''(\alpha)$ exist for any $\alpha \in \mathbb{R}$. By the monotone convergence theorem and the first-order optimality condition, we have $R'(\alpha_*) = \mathbb{E}\left[\frac{d}{d\alpha}(1 - \alpha M)_+^2\right] = \mathbb{E}\left[-2M(1 - \alpha M)\mathbb{1}_{\{1-\alpha M>0\}}\right] = 0$. Applying the monotone convergence theorem again, we obtain $R''(\alpha_*) = 2\mathbb{E}\left[M^2\mathbb{1}_{\{1-\alpha M>0\}}\right]$. Now, since $M = \langle XY, \boldsymbol{e}^*\rangle$ has unbounded support, there exists a measurable set $\{m \neq 0 : 1 - \alpha_* m > 0\}$ with strictly positive probability measure (where the probability measure is $\mu_X \otimes \mu_Y$). Therefore, $R''(\alpha_*) > 0$ and $\mathbb{P}(1 - \alpha_* M > 0) > 0$. By the continuity of the probability distribution function of $M$, we know that there exists $\varepsilon > 0$ such that $\mathbb{P}(1 - \alpha_* M > 2\varepsilon) > \frac{1}{2}\mathbb{P}(1 - \alpha_* M > 0)$. Moreover, since $M < \infty$ almost surely, there exists $B > 0$ such that $\mathbb{P}(|M| \le B) \ge 1 - \frac{1}{4}\mathbb{P}(1 - \alpha_* M > 0)$. Therefore, the probability of the event $E = \{m \in \mathbb{R} : 1 - \alpha_* m > 2\varepsilon\} \cap \{m \in \mathbb{R} : |m| \le B\}$ is at least $\frac{1}{4}\mathbb{P}(1 - \alpha_* M > 0)$. Take $\delta = \frac{\varepsilon}{2B}$. Then for all $|\alpha - \alpha_*| < \delta$, for every $m \in E$, $1 - \alpha m = 1 - \alpha_* m - (\alpha - \alpha_*)m > 2\varepsilon - \delta|m| \ge 2\varepsilon - \frac{\varepsilon}{2B}B > \varepsilon > 0$. Therefore, for any $|\alpha - \alpha_*| < \delta$, we have $R''(\alpha) = \mathbb{E}[M^2\mathbb{1}_{\{1-\alpha M>0\}}] \ge \mathbb{E}[M^2\mathbb{1}_E] > 0$. By the mean value theorem for Taylor expansion, there exists $\bar{\alpha} \in [\alpha, \alpha_*]$ (or $\bar{\alpha} \in [\alpha_*, \alpha]$) such that $P\mathcal{L}_{\boldsymbol{\beta}_J} = R(\alpha) - R(\alpha_*) \ge R''(\bar{\alpha})(\alpha - \alpha_*)^2 \ge \mathbb{E}[M^2\mathbb{1}_E](\alpha - \alpha_*)^2$. On the other hand, we have $\|\Sigma_J^{1/2}(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)\|_2^2 = (\alpha - \alpha_*)^2\|\Sigma_J^{1/2}\boldsymbol{e}^*\|_2^2$. Therefore, if $r(\rho)/\|\Sigma_J^{1/2}\boldsymbol{e}^*\|_2 < \delta$, then for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + r(\rho)\Sigma_J^{1/2}S_2^J$, we have $P\mathcal{L}_{\boldsymbol{\beta}_J} \ge L_1\|\Sigma_J^{1/2}(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)\|_2^2$, where $L_1 = \mathbb{E}[M^2\mathbb{1}_E]/\|\Sigma_J^{1/2}\boldsymbol{e}^*\|_2^2$. Let $\boldsymbol{w} = \Lambda^{-1}\boldsymbol{\mu} = \|\Lambda^{-1}\boldsymbol{\mu}\|_2\boldsymbol{e}^*$, then $X_J = \frac{1}{\|\boldsymbol{w}\|_2^2}\langle \boldsymbol{w}, X\rangle\boldsymbol{w}$, and $\Sigma_J = \mathbb{E}[X_J \otimes X_J] = \|\langle \boldsymbol{e}^*, X\rangle\|_{L^2}^2\boldsymbol{e}^* \otimes \boldsymbol{e}^*$, so $\|\Sigma_J^{1/2}\boldsymbol{e}^*\|_2 = \|\langle \boldsymbol{e}^*, X\rangle\|_{L^2(\mu_X)}$. ∎

### 4.9.5 Proof of Lemma 19

We only treat the case $\boldsymbol{\beta}_{J^c}^* = \mathbf{0}$. The case $X_{J^c} \sim \mathcal{N}(\mathbf{0}, \Sigma_{J^c})$ is analogous, only the constants differ. A direct consequence of $\Omega_{\mathrm{DM,reg}}(\varepsilon_1)$ is the establishment of the local Bernstein condition, Assumption 9. For any $q \ge 1$, there exists an absolute constant $C_{41} > 1$ depending only on $q$ and $\varepsilon_1$ such that

$$\mathbb{E}_{\boldsymbol{\xi}}[\|\mathcal{A}[\boldsymbol{\xi}]\|_q^q | \mathbb{X}_{J^c}] \ge \frac{\mathbb{E}\|\boldsymbol{\xi}\|_2^q}{(1 + \varepsilon_1)^q \ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)} \ge \frac{N^{\frac{q}{2}}\sigma_\xi^q}{2C_{41}\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}. \tag{4.72}$$

**Proof.** Recall that for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + r(\rho)\Sigma_J^{-1/2}S_2^J \cap \rho K_{\mathrm{model}}$, $\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*) - \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, (\sigma_\xi^2 + r^2(\rho))I_N)$. As a result, condition on $\Omega_{\mathrm{DM,reg}}(\varepsilon_1)$, there exists an absolute constant $c_{28} < 1$ depending only on $q$ such that

$$\mathbb{E}_{\mathbb{X}_J,\boldsymbol{\xi}}\|\mathcal{A}[\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*) - \boldsymbol{\xi}]\|_q^q - \mathbb{E}_{\boldsymbol{\xi}}\|\mathcal{A}[\boldsymbol{\xi}]\|_q^q = \mathbb{E}_{\boldsymbol{\xi}}\|\mathcal{A}[\boldsymbol{\xi}]\|_q^q\left(\left(1 + \frac{r^2(\rho)}{\sigma_\xi^2}\right)^{q/2} - 1\right) \ge c_{28}\mathbb{E}_{\boldsymbol{\xi}}\|\mathcal{A}[\boldsymbol{\xi}]\|_q^q\frac{r^2(\rho)}{\sigma_\xi^2}.$$

By (4.72), we obtain that for any $\boldsymbol{\beta}_J \in \boldsymbol{\beta}_J^* + r(\rho)\Sigma_J^{-1/2} S_2^J \cap \rho K_{\text{model}}$,

$$P\mathcal{L}_{\boldsymbol{\beta}_J} = \mathbb{E}_{\mathbb{X}_J, \boldsymbol{\xi}} \|\mathcal{A}[\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*) - \boldsymbol{\xi}]\|_q^q - \mathbb{E}_{\boldsymbol{\xi}} \|\mathcal{A}[\boldsymbol{\xi}]\|_q^q \geq \frac{c_{28} N^{\frac{q}{2}} \sigma_\xi^{q-2}}{2C_{41} \ell_*^q(\Sigma_{J^c}^{1/2} B_q^p)} r^2(\rho).$$

■

### 4.9.6 Proof of Lemma 17

Let $X_{J^c} \sim \mathcal{N}(\mathbf{0}, \Sigma_{J^c})$, and $X_{J^c} = (x_j)_{j \in J^c}$. Then $\ell_*(\Sigma_{J^c}^{1/2} B_q^p) = \mathbb{E}\|X_{J^c}\|_{q'}$. Moreover, the function $t \mapsto t^{1/q'}$ is concave on $\{t : t \geq 0\}$. By Jensen's inequality for concave functions, $\mathbb{E}\|X_{J^c}\|_{q'} = \mathbb{E}[S^{1/q'}] \leq (\mathbb{E}[S])^{1/q'}$ with $S = \sum_{j \in J^c} |x_j|^{q'}$. Now, since $x_j \sim \mathcal{N}(0, \sigma_j)$, there exists an absolute constant $C_{31}$ depending only on $q$ such that $\mathbb{E}[S] = \sum_{j \in J^c} \mathbb{E}|x_j|^{q'} \leq C_{31}^{q'} \sum_{j \in J^c} \sigma_j^{q'/2}$. For the lower side, letting $|X_{J^c}| = (|x_j|)_{j \in J^c}$. Then $\|X_{J^c}\|_{q'} = \||X_{J^c}|\|_{q'}$. By Jensen's inequality for convex function applied for $\|\cdot\|_{q'}$, we have $\|\mathbb{E}|X_{J^c}|\|_{q'} \leq \mathbb{E}\||X_{J^c}|\|_{q'} = \mathbb{E}\|X_{J^c}\|_{q'}$. Moreover, $\|\mathbb{E}|X_{J^c}|\|_{q'} = (\sum_{j \in J^c} (\mathbb{E}|x_j|)^{q'})^{1/q'} = \sqrt{\frac{2}{\pi}} (\sum_{j \in J^c} \sigma_j^{q'/2})^{1/q'}$. Therefore, the lower side of Lemma 17 holds with $c_{16} = \sqrt{\frac{2}{\pi}}$.

# Chapter 5

# Generalizations of the Dvoretzky-Milman Theorem for the $\|\cdot\|_q$-Norm under a General Probability Measure

## 5.1 Dvoretzky-Milman theorem for $\|\cdot\|_q$ norm under general probability measure

Let $N < p$ be two positive integers, $Y \in \mathbb{R}^p$ be a random vector, $Y_1, \cdots, Y_N$ be i.i.d. copies of $Y$. Let $A$ be a $p \times p$ deterministic matrix and set $X_j = AY_j$. Let $\mathbb{X} = \mathbb{X}(A)$ (resp. $\mathbb{Y} = \mathbb{Y}(A)$) be the $N \times p$ matrix with rows $X_j$ (resp. $Y_j$), $\mathbb{X} = (X_j(i))_{j\le N, i\le p}$, where $X_j(i)$ is the $i$'th coordinate of $X_j$. By $\|\boldsymbol{x}\|_\rho = (\sum_i |x_i|^\rho)^{1/\rho}$ we denote the $\ell_\rho$-norm of a vector $\boldsymbol{x}$. For $q \ge 1$, define $\ell_q(A) = \mathbb{E}\|AG\|_q$, where $G$ is a standard Gaussian vector in $\mathbb{R}^p$ and $d_q(A) = \sup(\|A\boldsymbol{x}\|_q : \boldsymbol{x} \in B_2^p)$. Set $d_*(A) = (\ell_q(A)/d_q(A))^2$. Let $\mathrm{Log}(x) = \max\{1, \ln(x)\}$.

Let us now assume that $A = \mathrm{diag}(\sigma_1, \ldots, \sigma_p)$, so $X_j(i) = \sigma_i Y_j(i)$, $q \ge 1$ is fixed. In this case

$$\ell_q(A) = \mathbb{E}(\sum_{i=1}^p |\sigma_i|^q |g_i|^q)^{1/q} \ge c_0 \|\sigma\|_q, \tag{5.1}$$

where $c_0$ depends only on $q$, and

$$\ell_q(A) \le (\mathbb{E}\sum_{i=1}^p |\sigma_i|^q |g_i|^q)^{1/q} = \|\sigma\|_q. \tag{5.2}$$

Moreover,

$$d_q(A) = \begin{cases} \|\sigma\|_\infty & \text{if } q \ge 2 \\ \|\sigma\|_{\frac{2q}{2-q}} & \text{if } q < 2 \end{cases} \qquad \text{and so} \quad d_*(A) \underset{q}{\sim} \begin{cases} \left(\frac{\|\sigma\|_q}{\|\sigma\|_\infty}\right)^2 & \text{if } q \ge 2 \\ \left(\frac{\|\sigma\|_q}{\|\sigma\|_{\frac{2q}{2-q}}}\right)^2 & \text{if } q < 2. \end{cases} \tag{5.3}$$

Assume also that the entries of $Y$ are independent with mean zero and variance one, and satisfy $\mathbb{E}|Y_{ij}|^r \le \kappa^r$, for some $r > \max\{2, q\}$. The constants $C_i, c_i$ in this subsection will depend only on $q, r, \kappa$.

In this section, we prove the following theorem. Note: This $q$ is conjugate to the $q$ in Theorem 3 from Chapter 1. That is, the $q$ in Theorem 12 proved in this section should be regarded as the $q'$ in Theorem 3.

**Theorem 12** ([P1]). *Suppose $Y$ is mean zero and is isotropic. Moreover, the entries of $Y$ are independent with mean 0 and variance 1. By $Z_1, \cdots, Z_p$ we denote the rows of $\mathbb{Y}^\top$ (i.e., the matrix with columns $Y_1, \cdots, Y_N$), so $\mathbb{X}^\top \boldsymbol{\lambda} = (\sigma_i \langle Z_i, \boldsymbol{\lambda}\rangle)_i$. There exist absolute constants $0 < \kappa \le 1$ and $0 < \theta < 1$ such that*

$$\mathbb{P}(|\langle Z_i, \boldsymbol{\lambda}\rangle| \ge \theta) \ge \kappa \qquad \text{for all } \boldsymbol{\lambda} \in S_2^{N-1} \text{ and } i \le p.$$

*Then there exists an absolute constant $c > 0$ such that the following holds.*

1. *If $q \ge 2$, then there exists a constant $c > 0$, such that with probability at least $1 - \exp(-cd_*(A))$, for any $\boldsymbol{\lambda} \in S_2^{N-1}$, $\|\mathbb{X}^\top \boldsymbol{\lambda}\|_q \ge c\ell_q(A)$, provided that $N \le cd_*(A)$.*

*2. If $q < 2$, then there exists a constant $c > 0$, such that with probability at least*

$$1 - C\operatorname{Log}(p^{1/q}/d_*(A))e^{-\alpha c d_*(A)^\theta/\operatorname{Log}^{2\theta}(p^{1/q}/d_*(A))},$$

*for any $\boldsymbol{\lambda} \in S_2^{N-1}$, $\|\mathbb{X}^\top \boldsymbol{\lambda}\|_q \geq c\ell_q(A)$, provided that $N \leq \frac{c\gamma d_*(A)}{\operatorname{Log}^2(p^{1/q}/d_*(A))}$.*

*Suppose $A = \operatorname{diag}(\sigma_1, \cdots, \sigma_p)$, and the entries of $Y$ satisfy $\mathbb{E}|Y_j(i)|^{\max\{4,2q\}} \leq \kappa^{\max\{4,2q\}}$. Then the following hold. If $q > 2$ and $N \leq d_*^2(A)$, then there exist constants $c$ and $C$ such that*

$$\sup_{\lambda \in S_2^{N-1}} \|\mathbb{X}^\top \lambda\|_q \leq C\ell_q(A)$$

*with probability at least $1 - 2e^{-cd_*(A)} - Cd_*(A)^{-(q-2)/4}$.*

*If $q \leq 2$, $N \leq d_*^2(A)$ and we additionally assume that $\|Y_j(i)\|_{2q+\epsilon} \leq \kappa$ for some $\epsilon > 0$, then there exist constants $\epsilon_0$, $c$ and $C$ such that*

$$\sup_{\lambda \in S_2^{N-1}} \|\mathbb{X}^\top \lambda\|_q \leq C\operatorname{Log} p \cdot \ell_q(A)$$

*with probability at least $1 - 2e^{-cd_*(A)} - Cd_*(A)^{-c\min\{\epsilon,\epsilon_0\}}$.*

### 5.1.1 Upper bounds

Let us now assume that $A = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$, $q \geq 1$ is fixed, and the entries of $Y$ are independent with mean zero and variance one, and satisfy $\mathbb{E}|Y_j(i)|^{\max\{4,2q\}} \leq \kappa^{\max\{4,2q\}}$. The constants $C, C_i, c_i$ in this subsection will depend only on $q$ and $\kappa$.

Let us first provide upper estimates for the expectation of the operator norms of $\mathbb{X}^\top$ from $\ell_2^N$ to $\ell_q^p$. We do it separately for the cases $q \geq 2$ and $q < 2$.

**Proposition 34.** *If $q \geq 2$, then there exists a constant $C > 0$ such that for every $N \leq d_*(A)$,*

$$\mathbb{E}\|\mathbb{X}^\top \colon \ell_2^N \to \ell_q^p\| \leq C\ell_q(A).$$

**Proof.** Since $q \geq 2$ and $A = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$, $d_q(A) = \|\sigma\|_\infty$ Let $Y_j(i)$, for $i \leq p$ and $j \leq N$, be the $i$-th coordinate of $Y_j$ (the $j$-th copy of $Y$) and let $X_j(i) = \sigma_i Y_j(i)$, so $\mathbb{X}^\top = (\sigma_i Y_j(i))_{i \leq p, j \leq N}$. By our assumptions

$$\max\big\{(\mathbb{E}|Y_j(i)|^4)^{1/4}, (\mathbb{E}|Y_j(i)|^{2q})^{1/2q}\big\} \leq (\mathbb{E}|Y_j(i)|^{2q})^{1/(2q)} \leq \kappa.$$

Thus, if $N \leq d_* \leq \big(\frac{\|\sigma\|_q}{\|\sigma\|_\infty}\big)^2$, then [LS25, Corollary 13] and (5.1) yield

$$\begin{aligned}\mathbb{E}\|\mathbb{X}^\top \colon \ell_2^N \to \ell_q^p\| &\leq C_1\big(\|\sigma\|_\infty N^{1/2} + \kappa\|\sigma\|_q + \kappa\|\sigma\|_{2q}N^{1/4} + \kappa\|\sigma\|_{2q}N^{1/(2q)}\big)\\ &\leq C_2\ell_q(A) + 2C_1\kappa\|\sigma\|_{2q}N^{1/4}.\end{aligned}$$

Moreover, $\|\sigma\|_{2q}^{2q} \leq \|\sigma\|_\infty^q \|\sigma\|_q^q$, so the AM-GM inequality implies that

$$2\|\sigma\|_{2q}N^{1/4} \leq \|\sigma\|_q + \|\sigma\|_\infty N^{1/2} \leq 2c_0^{-1}\ell_q(A),$$

and the assertion follows. ∎

**Proposition 35.** *If $q < 2$, then there exists a constant $C > 0$ such that for every $N \leq d_*(A)$,*

$$\mathbb{E}\|\mathbb{X}^\top \colon \ell_2^N \to \ell_q^p\| \leq C\big(\operatorname{Log} N + (2-q)\operatorname{Log} p\big)\ell_q(A).$$

**Proof.** Assume without the loss of generality that $\|\sigma\|_\infty = 1$. Let $k_0$ be an integer to be fixed later,

$$I_k = \{i \leq p \colon 2^{-k-1} < |\sigma_i| \leq 2^{-k}\} \quad \text{for } k = 0, 1, \ldots, k_0 - 1,$$

and

$$I_{k_0} = \{i \leq p \colon |\sigma_i| \leq 2^{-k_0}\}.$$

Define $A_k = \operatorname{diag}\big((\sigma_i \mathbf{1}_{\{i\in I_k\}}\big)_{i\le p})$ and i.i.d. copies $X_1^{(k)},\dots X_N^{(k)}$ of $X^{(k)} = A_k Y$, and let $\mathbb{X}_k$ be the $N\times p$ matrix with rows $X_j^{(k)}$. Then $\mathbb{X}^\top = \sum_{k=0}^{k_0} \mathbb{X}_k^\top$, and the triangle inequality implies that

$$\mathbb{E}\|\mathbb{X}^\top\colon \ell_2^N \to \ell_q^p\| \le (k_0+1) \max_{0\le k\le k_0} \mathbb{E}\|\mathbb{X}_k^\top\colon \ell_2^N \to \ell_q^p\|. \tag{5.4}$$

Fix $0\le k\le k_0$ such that $I_k\neq\emptyset$. Let $(2/q)^* = 2/(2-q)$ denote the Hölder's conjugate to $2/q\in(1,2]$. Then, similarly as in the proof of Proposition 34, [LS25, Corollary 13] yields that whenever $N^{1/2}\le \frac{\|\sigma\|_q}{\|\sigma\|_{\frac{2q}{2-q}}}$, then

$$\begin{aligned}\mathbb{E}\|\mathbb{X}_k^\top\colon \ell_2^N\to\ell_q^p\| &\le \big\|\operatorname{diag}\big((|\sigma_i|^{1-q/2}\operatorname{sgn}(\sigma_i))_{i\in I_k}\big)\colon \ell_2^{I_k}\to\ell_q^{I_k}\big\| \cdot \mathbb{E}\big\|(|\sigma_i|^{q/2}X_j(i))_{\substack{i\in I_k\\ j\le N}}\colon \ell_2^N\to\ell_2^{I_k}\big\| \\ &\le C_3\|(|\sigma_i|^{(1-q/2)})_{i\in I_k}\|_{2q/(2-q)}\Big(\sup_{i\in I_k}|\sigma_i|^{q/2}N^{1/2} + \|(|\sigma|^{q/2})_{i\in I_k}\|_2\Big) \qquad (5.5)\\ &\le C_3 2^{-k}|I_k|^{\frac{2-q}{2q}}\frac{\|\sigma\|_q}{\|\sigma\|_{\frac{2q}{2-q}}} + C_3\|\sigma\|_q. \qquad (5.6)\end{aligned}$$

If $k\le k_0-1$, then

$$\|\sigma\|_{\frac{2q}{2-q}} \ge \Big(\sum_{i\in I_k}|\sigma_i|^{\frac{2q}{2-q}}\Big)^{\frac{2-q}{2q}} \ge 2^{-(k+1)}|I_k|^{\frac{2-q}{2q}}, \tag{5.7}$$

so it follows by (5.6) and (5.1) that

$$\mathbb{E}\|\mathbb{X}_k^\top\colon \ell_2^N\to\ell_q^p\| \le 3C_3\|\sigma\|_q \le C_4\ell_q(A), \tag{5.8}$$

provided that $N\le d_*(A)$.

If $k=k_0 := \lceil \log_2(\sqrt{N}p^{\frac1q-\frac12})\rceil$, then (5.5) implies for every $N$,

$$\mathbb{E}\|\mathbb{X}_{k_0}^\top\colon \ell_2^N\to\ell_q^p\| \le C_3 2^{-k_0}p^{\frac{2-q}{2q}}N^{1/2} + C_3\|\sigma\|_q \le C_3(\|\sigma\|_\infty + \|\sigma\|_q) \le C_4\ell_q(A). \tag{5.9}$$

Inequalities (5.4), (5.8), and (5.9) yield the assertion. ∎

**Proposition 36.** *If $q>2$ and $N\le d_*^2(A)$, then there exist constants $c$ and $C$ such that*

$$\sup_{\lambda\in S_2^{N-1}}\|\mathbb{X}^\top\lambda\|_q \le C\ell_q(A)$$

*with probability at least $1-2e^{-cd_*(A)} - Cd_*(A)^{-(q-2)/4}$.*

*If $q\le 2$, $N\le d_*^2(A)$ and we additionally assume that $\|Y_j(i)\|_{2q+\epsilon}\le\kappa$ for some $\epsilon>0$, then there exist constants $\epsilon_0$, $c$ and $C$ such that*

$$\sup_{\lambda\in S_2^{N-1}}\|\mathbb{X}^\top\lambda\|_q \le C\operatorname{Log} p\cdot\ell_q(A)$$

*with probability at least $1-2e^{-cd_*(A)} - Cd_*(A)^{-c\min\{\epsilon,\epsilon_0\}}$.*

**Proof.** Note that $\sup_{\lambda\in S_2^{N-1}}\|\mathbb{X}^\top\lambda\|_q = \|\mathbb{X}^\top\colon \ell_2^N\to\ell_q^p\|$.

Let us first consider the case $q>2$. By [Ada10, Theorem 2] (applied with $p:=q$, $\|\cdot\|$ being the operator norm from $\ell_2^N$ to $\ell_q^p$, $X_{(i,j)} := \sigma_i Y_j(i)e_i\otimes e_j$, $\eta=1$, and $\sigma:=\Sigma$ defined as below) and Propoistion 34 we have for every $t\ge 0$,

$$\mathbb{P}\big(\|\mathbb{X}^\top\colon \ell_2^N\to\ell_q^p\| \ge C_5\ell_q(A)+t\big) \le e^{-t^2/(3\Sigma^2)} + C_5\frac{\mathbb{E}\max_{i,j}|\sigma_i Y_j(i)|^q}{t^q}, \tag{5.10}$$

where

$$\Sigma^2 = \sup_{s\in B_{q^*}^p}\sup_{t\in B_2^N}\sum_{i,j}\sigma_i^2 s_i^2 t_j^2 = \max_{j\le N}\sup_{x\in B_{q^*/2}^p}\sum_i \sigma_i^2 x_i = \|\sigma\|_\infty^2 = d_q(A)^2.$$

Note that

$$\begin{aligned}\mathbb{E}\max_{i,j}|\sigma_i Y_j(i)|^q &\le \Big(\mathbb{E}\sum_{i,j}|\sigma_i|^{2q}|Y_j(i)|^{2q}\Big)^2 \le \kappa^q N^{1/2}\Big(\sum_i|\sigma_i|^{2q}\Big)^{1/2} \\ &\le \kappa^q d_*(A)^{1/2}\Big(\sum_i|\sigma_i|^q\Big)^{1/2}\|\sigma\|_\infty^{q/2} = \kappa^q d_*(A)^{1/2}\|\sigma\|_q^{q/2}\|\sigma\|_\infty^{q/2}.\end{aligned}$$

Thus, by taking $t = \ell_q(A) \sim_q \|\sigma\|_q$ (recall that $\sim$ follows by (5.3)) we get the assertion in the case $q > 2$.

In the case $q \le 2$ we proceed similarly, exploiting again [Ada10, Theorem 2] and Proposition 35 (instead of Proposition 34) to get for every $t \ge 0$,

$$\mathbb{P}\big(\|\mathbb{X}^\top \colon \ell_2^N \to \ell_q^p\| \ge C_5 \operatorname{Log} p \cdot \ell_q(A) + t\big) \le e^{-t^2/(3\Sigma^2)} + C_5 \frac{\mathbb{E}\max_{i,j} |\sigma_i Y_j(i)|^q}{t^q},$$

Since $(q^*/2)^* = q/(2-q)$,

$$\Sigma^2 = \max_{j\le N} \sup_{x \in B^p_{q^*/2}} \sum_i \sigma_i^2 x_i = \|\sigma\|^2_{2q/(q-2)} = d_q(A)^2.$$

If $q < 2$, we may assume without loss of generality that $\epsilon > 0$ is small enough to provide that $2q + \epsilon < \frac{2q}{2-q}$. Then $2q + \epsilon = \theta \frac{2q}{2-q} + (1-\theta)q$, where $\theta = (q+\epsilon)(2-q)/q^2 \in (0,1)$. Hence, we may use Hölder's inequality to obtain that

$$\begin{aligned}
\mathbb{E}\max_{i,j} |\sigma_i Y_j(i)|^q &\le \Big(\mathbb{E}\sum_{i,j} |\sigma_i|^{2q+\epsilon} |Y_j(i)|^{2q+\epsilon}\Big)^{q/(2q+\epsilon)} \le \kappa^q N^{q/(2q+\epsilon)} \Big(\sum_i |\sigma_i|^{2q+\epsilon}\Big)^{q/(2q+\epsilon)} \\
&\le \kappa^q d_*(A)^{q/(2q+\epsilon)} \|\sigma\|_q^{q - \frac{2q+2\epsilon}{2q+\epsilon}} \|\sigma\|_{2q/(2-q)}^{\frac{2q+2\epsilon}{2q+\epsilon}} \\
&= \kappa^q d_*(A)^{q/(2q+\epsilon)} \|\sigma\|_q^{q - \frac{2q+2\epsilon}{2q+\epsilon}} d_q(A)^{\frac{2q+2\epsilon}{2q+\epsilon}},
\end{aligned}$$

whereas for $q = 2$ we have

$$\begin{aligned}
\mathbb{E}\max_{i,j} |\sigma_i Y_j(i)|^q &\le \kappa^2 N^{2/(4+\epsilon)} \Big(\sum_i |\sigma_i|^{4+\epsilon}\Big)^{2/(4+\epsilon)} \le \kappa^2 d_*(A)^{2/(4+\epsilon)} \|\sigma\|_2^{\frac{4}{4+\epsilon}} \|\sigma\|_\infty^{\frac{4+2\epsilon}{4+\epsilon}} \\
&= \kappa^q d_*(A)^{q/(2q+\epsilon)} \|\sigma\|_q^{q - \frac{2q+2\epsilon}{2q+\epsilon}} d_q(A)^{\frac{2q+2\epsilon}{2q+\epsilon}}.
\end{aligned}$$

We finish by taking $t = \ell_q(A)$ as before. ∎

### 5.1.2 Reduction from the identity to an arbitrary diagonal covariance matrix

The next proposition shows that in order to have a lower bound up to logarithmic terms it suffices to consider the case when $A = \operatorname{diag}(\sigma_1, \dots, \sigma_p)$ with $\sigma_i \in \{-1, 0, 1\}$; this corresponds to $X$ being a projection of an isotropic vector into some coordinate subspace. The constants $c, C, c_i, C_i$ in this section depend only on $q$.

**Proposition 37.** *Let $p \ge N$ be integers, $q \ge 1$, $Y$ be an isotropic random vector in $\mathbb{R}^p$ and $\alpha, \beta, \gamma, \theta > 0$. Assume that for any nonempty $I \subset [p]$ and any $\eta_i \in \{-1, 1\}$, $i = 1, \dots, p$, with probability at least $1 - 2e^{-\alpha d_*(A(\eta,I))^\theta}$ for every $\lambda \in S_2^{N-1}$,*

$$\|\mathbb{X}(A(\eta, I))^\top \lambda\|_q \ge \beta \ell_q(A(\eta, I)),$$

*provided that $N \le \gamma d_*(A(\eta, I))$, where $A(\eta, I) = \operatorname{diag}\big((\eta_i \mathbf{1}_{\{i\in I\}})_{i\le p}\big)$. Then there exist constants $c, C > 0$, such that for every $\sigma \in \mathbb{R}^p \setminus \{0\}$ with probability at least $1 - C \operatorname{Log}(p^{1/q}/d_*(A)) e^{-\alpha c d_*(A)^\theta / \operatorname{Log}^{2\theta}(p^{1/q}/d_*(A))}$ for every $\lambda \in S_2^{N-1}$,*

$$\|\mathbb{X}(A)^\top \lambda\|_q \ge c\beta \ell_q(A),$$

*provided that $N \le \frac{c\gamma d_*(A)}{\operatorname{Log}^2(p^{1/q}/d_*(A))}$, where $A = \operatorname{diag}(\sigma_1, \dots, \sigma_p)$.*

**Proof.** Assume without the loss of generality that $\|\sigma\|_\infty = 1$ if $q \ge 2$ and $\|\sigma\|_{\frac{2q}{q-2}} = 1$ if $q < 2$. In both cases $\|\sigma\|_\infty \le 1$. Let $D = \|\sigma\|_q \sim \ell_q(A) = \sqrt{d_*(A)}$, $k_0 = \lceil \log_2(4p^{1/q}/D) \rceil$ and, as in the proof of Proposition 35, let

$$I_k = \{i \le p \colon 2^{-k-1} < |\sigma_i| \le 2^{-k}\} \quad \text{for } k = 0, 1, \dots, k_0 - 1.$$

and

$$I_{k_0} = \{i \le p \colon |\sigma_i| \le 2^{-k_0}\},$$

We consider two cases.

**Case I:** $q < 2$

Let

$$K = \Big\{k \in \{0, 1, \dots, k_0 - 1\} \colon \sum_{i \in I_k} \sigma_i^2 \geq \frac{D^2 2^{-2k}}{64 k_0^{2/q}}\Big\}.$$

Note that

$$\sum_{k \in [k_0-1] \setminus K} \sum_{i \in I_k} |\sigma_i|^q \leq \sum_{k \in [k_0-1] \setminus K} \Big(\sum_{i \in I_k} \sigma_i^2\Big)^{q/2} |I_k|^{\frac{2-q}{2}} \leq \sum_{k=0}^{k_0-1} \frac{D^q 2^{-kq}}{8^q k_0} |I_k|^{\frac{2-q}{2}}.$$

Moreover, if $k \leq k_0 - 1$, then inequality (5.7) shows that $|I_k|^{\frac{2-q}{2}} \leq \|\sigma\|_{\frac{2q}{2-q}}^q 2^{q(k+1)}$, so

$$\Big(\sum_{k \in [k_0-1] \setminus K} \sum_{i \in I_k} |\sigma_i|^q\Big)^{1/q} \leq \frac{D}{4}. \tag{5.11}$$

Since $k_0 = \lceil \log_2(4p^{1/q}/D) \rceil$,

$$\sum_{i \in I_{k_0}} |\sigma_i|^q \leq |I_{k_0}| 2^{-k_0 q} \leq p 2^{-k_0 q} \leq \frac{D^q}{4^q},$$

which together with (5.11) shows that

$$\|(\sigma_i)_{i \in \bigcup_{k \in K} I_k}\|_q \geq D/2,$$

so it suffices to prove that for every $k \in K$ with probability at least $1 - 2e^{-c_1 \alpha D^{2\theta}/k_0^{2\theta}}$ for every $\lambda \in S_2^{N-1}$,

$$\|\operatorname{Proj}_{I_k}(\mathbb{X}(A)^\top \lambda)\|_q^q \geq \beta^q \|(\sigma_i)_{i \in I_k}\|_q^q, \tag{5.12}$$

provided that $N \leq c_1 \gamma D^2/k_0^2$. Since every entry of the vector $(\sigma_i)_{i \in I_k}$ is comparable to its $\ell_\infty$ norm,

$$d_*(\operatorname{diag}((\sigma_i)_{i \in I_k})) \sim \Big(\frac{\|(\sigma_i)_{i \in I_k}\|_q}{\|(\sigma_i)_{i \in I_k}\|_{\frac{2q}{2-q}}}\Big)^2 \sim \Big(\frac{|I_k|^{1/q}}{|I_k|^{\frac{2-q}{2q}}}\Big)^2 = |I_k|.$$

Thus, the inequality (5.12) follows by the assumption, with probability at least $1 - 2e^{-c_2 \alpha |I_k|^\theta}$ and provided that $N \leq c_2 \gamma |I_k|$. Therefore, it suffices to show that $|I_k| \geq c_3 D^2/k_0^2$ whenever $k \in K$.

By the definition of $K$ and $I_k$ we have for every $k \in K$,

$$|I_k| 2^{-2k} \geq \sum_{i \in I_k} \sigma_i^2 \geq \frac{D^2 2^{-2k}}{64 k_0^{2/q}},$$

so indeed $|I_k| \geq c_3 D^2/k_0^2$.

**Case II:** $q \geq 2$

This case is similar to the previous one, so we omit some details. Let

$$K = \Big\{k \in \{0, 1, \dots, k_0 - 1\} \colon \sum_{i \in I_k} |\sigma_i|^q \geq \frac{D^q 2^{-kq}}{2 \cdot 4^q}\Big\}.$$

Then

$$\sum_{k \in [k_0-1] \setminus K} \sum_{i \in I_k} |\sigma_i|^q \leq \sum_{k=0}^{k_0} \frac{D^q 2^{-kq}}{2 \cdot 4^q} \leq \frac{D^q}{4^q}$$

and since $k_0 = \lceil \log_2(4p^{1/q}/D) \rceil$,

$$\sum_{i \in I_{k_0}} |\sigma_i|^q \leq p 2^{-kq} \leq \frac{D^q}{4^q},$$

so as in the Case I it suffices to provide (5.12) whenever $k \in K$, with appropriate probability and the upper bound for $N$. This again boils down to proving that for any $k \in K$,

$$c_4 D \leq \sqrt{d_*(\operatorname{diag}((\sigma_i)_{i \in I_k}))} \sim \frac{\|(\sigma_i)_{i \in I_k}\|_q}{\|(\sigma_i)_{i \in I_k}\|_\infty} \sim |I_k|^{1/q}.$$

The inequality $c_5 D \leq |I_k|^{1/q}$ follows by the definitions of $K$ and $I_k$, as before in Case I. ∎

### 5.1.3 Lower bounds in the independent case using selector processes

Let us now assume that $A = \operatorname{diag}(\sigma_1, \dots, \sigma_p)$ and the entries of $Y$ are independent with mean 0 and variance 1. By $Z_1, \dots, Z_p$ we denote the rows of $\mathbb{Y}^\top$ (i.e., the matrix with columns $Y_1, \dots Y_N$), so $\mathbb{X}^\top \boldsymbol{\lambda} = (\sigma_i \langle Z_i, \boldsymbol{\lambda}\rangle)_i$, and $\tilde{d}_* = \left(\frac{\|\boldsymbol{\sigma}\|_q}{\|\boldsymbol{\sigma}\|_\infty}\right)^2$. By Paley-Zygmund's inequality, c.f., [Kal21, Lemma 5.1] together with the norm equivalent assumption, there exist absolute constants $0 < \kappa \le 1$ and $0 < \theta < 1$ such that

$$\mathbb{P}(|\langle Z_i, \boldsymbol{\lambda}\rangle| \ge \theta) \ge \kappa \qquad \text{for all } \boldsymbol{\lambda} \in S_2^{N-1} \text{ and } i \le p. \tag{5.13}$$

The constants below depend only on $\kappa$ and $\theta$.

**Proposition 38.** *There exists a constant $c > 0$, such that when $N \le c\tilde{d}_*^{q/2}$, then with probability at least $1 - 2\exp(-c\tilde{d}_*^{q/2})$, for any $\boldsymbol{\lambda} \in S_2^{N-1}$, $\|\mathbb{X}^\top \boldsymbol{\lambda}\|_q \ge c\|\boldsymbol{\sigma}\|_q$.*

The assumption (5.13) is referred to as the small-ball assumption. The small-ball method is similar to the $\varepsilon$-net argument in that both require a single-scale approximation on an index set, and then extend this approximation to the whole index set via a union bound. The difference from the $\varepsilon$-net argument is that the small-ball method does not require an upper bound on the $\ell_2 \to \ell_q$ operator norm of $\mathbb{X}^\top$, but instead relies on upper bounds for the corresponding selector process, which in many cases only require mild moment assumptions.

**Proof.** We aim to prove the following fact: there exist a constant $0 < c < 1$ and a high-probability event $\Omega_0$ (more precisely, such that $\mathbb{P}(\Omega_0) \ge 1 - 2e^{-c\tilde{d}_*^{q/2}}$) such that, on this event, for every $\boldsymbol{\lambda} \in S_2^{N-1}$ there exists a set of indices $J_{\boldsymbol{\lambda}} \subset \{1, \dots, p\}$ satisfying

$$\|\mathbb{X}^\top \boldsymbol{\lambda}\|_q^q = \sum_{i=1}^p |\sigma_i|^q |\langle Z_i, \boldsymbol{\lambda}\rangle|^q \ge \sum_{i \in J_{\boldsymbol{\lambda}}} |\sigma_i|^q |\langle Z_i, \boldsymbol{\lambda}\rangle|^q \ge \sum_{i \in J_{\boldsymbol{\lambda}}} |\sigma_i|^q \ge c^q \|\boldsymbol{\sigma}\|_q^q.$$

This requires a single-scale approximation of the index set $S_2^{N-1}$. We separate the rest of the proof into three steps.

**Step 1. A union bound.**

Let $\varepsilon \in (0, 1)$ to be determined later (in Step 3). Take an $\varepsilon$-net $V_\varepsilon \subset S_2^{N-1}$ of cardinality not exceeding $(1 + 2/\epsilon)^N$. For every $\boldsymbol{\lambda} \in S_2^{N-1}$ let $\pi\boldsymbol{\lambda} \in V_\varepsilon$ be such that $\|\pi\boldsymbol{\lambda} - \boldsymbol{\lambda}\|_2 \le \varepsilon$. Fix $\boldsymbol{\lambda} \in S_2^{N-1}$ and define $\eta_i = \mathbb{1}_{\{|\langle Z_i, \boldsymbol{\lambda}\rangle| \ge \theta\}}$, that is, $\eta_i = 1$ if $|\langle Z_i, \boldsymbol{\lambda}\rangle| \ge \theta$ and $\eta_i = 0$ otherwise. Let $I_{\boldsymbol{\lambda}} = \{i \le p : \eta_i = 1\} = \{i \le p \colon |\langle Z_i, \boldsymbol{\lambda}\rangle| \ge \theta\}$. By assumption (5.13) $\mathbb{E}\eta_i \ge \kappa$. Moreover, $|\sigma_i|^q |\eta_i - \mathbb{E}\eta_i| \le |\sigma_i|^q \le \|\boldsymbol{\sigma}\|_\infty^q$ and $\mathbb{E}[|\sigma_i|^{2q}(\eta_i - \mathbb{E}\eta_i)^2] = |\sigma_i|^{2q} \operatorname{Var}(\eta_i) \le |\sigma_i|^{2q}$.

Applying Bernstein's inequality [Ver18, Theorem 2.8.2] for the sum of independent variables $\sigma_i^q(\eta_i - \mathbb{E}\eta_i)$ we obtain that

$$\mathbb{P}\left(\left|\sum_{i=1}^p |\sigma_i|^q (\eta_i - \mathbb{E}\eta_i)\right| \ge \frac{1}{2}\kappa \|\boldsymbol{\sigma}\|_q^q\right) \le 2\exp\left(-\frac{\frac{1}{4}\kappa^2 \|\boldsymbol{\sigma}\|_q^{2q}}{2\|\boldsymbol{\sigma}\|_{2q}^{2q} + \frac{1}{3}\kappa \|\boldsymbol{\sigma}\|_\infty^q \|\boldsymbol{\sigma}\|_q^q}\right).$$

Moreover, $\sum_{i \in I_\lambda} |\sigma_i|^q = \sum_{i=1}^p |\sigma_i|^q \eta_i \ge \kappa \|\boldsymbol{\sigma}\|_q^q + \sum_{i=1}^p |\sigma_i|^q (\eta_i - \mathbb{E}\eta_i)$, $\|\boldsymbol{\sigma}\|_{2q}^{2q} \le \|\boldsymbol{\sigma}\|_\infty^q \|\boldsymbol{\sigma}\|_q^q$, and $\kappa \le 1$, so for every $\boldsymbol{\lambda} \in S_2^{N-1}$,

$$\mathbb{P}\left(\sum_{i \in I_{\boldsymbol{\lambda}}} |\sigma_i|^q \ge \frac{1}{2}\kappa \|\boldsymbol{\sigma}\|_q^q\right) \ge 1 - 2\exp\left(-\frac{\kappa^2 \|\boldsymbol{\sigma}\|_q^q}{10 \|\boldsymbol{\sigma}\|_\infty^q}\right) = 1 - 2\exp\left(-\frac{\kappa^2}{10}\tilde{d}_*^{q/2}\right). \tag{5.14}$$

Let

$$\Omega_1 = \left\{\forall \boldsymbol{\lambda} \in V_\varepsilon \quad \sum_{i \in I_{\boldsymbol{\lambda}}} |\sigma_i|^q \ge \frac{1}{2}\kappa \|\boldsymbol{\sigma}\|_q^q\right\}.$$

Recall that $|V_\varepsilon| \le (1 + 2/\varepsilon)^N \le e^{2N/\epsilon}$. Therefore, under the assumption $\tilde{d}_*^{q/2} \frac{\kappa^2}{20} \ge \frac{2}{\varepsilon} N$, we get by the union bound and (5.14) that $\mathbb{P}(\Omega_1) \ge 1 - 2\exp\left(-\frac{\kappa^2}{20}\tilde{d}_*^{q/2}\right)$.

**Step 2. Using selector process.**

Let $\varepsilon_1,\cdots,\varepsilon_p$ be i.i.d. Rademacher random variables independent of other variables, and for $\boldsymbol{\lambda}\in S_2^{N-1}$, let $u_{\boldsymbol{\lambda}}(\cdot)=\mathbb{1}_{\{|\langle \boldsymbol{\lambda}-\pi\boldsymbol{\lambda},\cdot\rangle|>\theta/2\}}$. Then $u_{\boldsymbol{\lambda}}(\cdot)\le\frac{2}{\theta}|\langle\boldsymbol{\lambda}-\pi\boldsymbol{\lambda},\cdot\rangle|$. Let $\phi=\phi(Z_1,\cdots,Z_p)=\frac{1}{\|\boldsymbol{\sigma}\|_q^q}\sup\{\sum_{i=1}^p|\sigma_i|^q u_{\boldsymbol{\lambda}}(Z_i)\colon \boldsymbol{\lambda}\in S_2^{N-1}\}$. The symmetrization together with contraction principle [Ver18, Theorem 6.7.1] (applied conditionally) implies

$$\begin{aligned}
\mathbb{E}\phi&=\frac{1}{\|\boldsymbol{\sigma}\|_q^q}\mathbb{E}\sup_{\lambda\in S_2^{N-1}}\sum_{i=1}^p|\sigma_i|^q\mathbb{1}_{\{|\langle\boldsymbol{\lambda}-\pi\boldsymbol{\lambda},Z_i\rangle|>\frac{\theta}{2}\}}\\
&\le\frac{1}{\|\boldsymbol{\sigma}\|_q^q}\mathbb{E}\sup_{\boldsymbol{x}\in S_2^{N-1}}\sum_{i=1}^p|\sigma_i|^q\Big(\mathbb{1}_{\{|\langle\boldsymbol{x},Z_i\rangle|>\frac{\theta}{2}\}}-\mathbb{E}\mathbb{1}_{\{|\langle\boldsymbol{x},Z_i\rangle|>\frac{\theta}{2}\}}\Big)\\
&\quad+\frac{1}{\|\boldsymbol{\sigma}\|_q^q}\sup_{\boldsymbol{x}\in\epsilon S_2^{N-1}}\sum_{i=1}^p|\sigma_i|^q\mathbb{E}\mathbb{1}_{\{|\langle\boldsymbol{x},Z_i\rangle|>\frac{\theta}{2}\}}\\
&\le\frac{2}{\|\boldsymbol{\sigma}\|_q^q}\mathbb{E}\sup_{\boldsymbol{x}\in\epsilon S_2^{N-1}}\sum_{j=1}^p|\sigma_i|^q\varepsilon_i\mathbb{1}_{\{|\langle\boldsymbol{x},Z_i\rangle|>\frac{\theta}{2}\}}+\frac{2}{\theta\|\boldsymbol{\sigma}\|_q^q}\sup_{\boldsymbol{x}\in\epsilon S_2^{N-1}}\sum_{i=1}^p|\sigma_i|^q\mathbb{E}|\langle\boldsymbol{x},Z_i\rangle|\\
&\le\frac{4}{\theta\|\boldsymbol{\sigma}\|_q^q}\mathbb{E}\sup_{\boldsymbol{x}\in\epsilon S_2^{N-1}}\sum_{i=1}^p|\sigma_i|^q\varepsilon_i\langle\boldsymbol{x},Z_i\rangle+\frac{2\epsilon}{\theta\|\boldsymbol{\sigma}\|_q^q}\sup_{\boldsymbol{x}\in S_2^{N-1}}\sum_{i=1}^p|\sigma_i|^q(\mathbb{E}|\langle\boldsymbol{x},Z_i\rangle|^2)^{1/2}\\
&=\frac{4\epsilon}{\theta\|\boldsymbol{\sigma}\|_q^q}\mathbb{E}\Big\|\sum_{i=1}^p\varepsilon_i|\sigma_i|^qZ_i\Big\|_2+\frac{2\varepsilon}{\theta}.
\end{aligned}$$

Since $Z_i$ are independent and have independent entries with second moment equal to 1, $\mathbb{E}\|\sum_{i=1}^p\varepsilon_i|\sigma_i|^qZ_i\|_2^2=\sum_{i=1}^p|\sigma_i|^{2q}\mathbb{E}\|Z_i\|_2^2=N\|\boldsymbol{\sigma}\|_{2q}^{2q}$. Thus, $\mathbb{E}\|\sum_{i=1}^p\varepsilon_i|\sigma_i|^qZ_i\|_2\le\sqrt{N}\|\boldsymbol{\sigma}\|_\infty^{q/2}\|\boldsymbol{\sigma}\|_q^{q/2}$. As a result, $\mathbb{E}[\phi]\le\frac{4\varepsilon}{\theta}\sqrt{\frac{N}{\tilde{d}_*^{q/2}}}+\frac{2\varepsilon}{\theta}\le\frac{6\varepsilon}{\theta}$, provided that $\tilde{d}_*^{q/2}\ge N$.

Notice that for any $\boldsymbol{\lambda}\in S_2^{N-1}$, any $\boldsymbol{z}_i,\boldsymbol{z}_i'\in\mathbb{R}^N$ and $|u_{\boldsymbol{\lambda}}(\boldsymbol{z}_i)-u_{\boldsymbol{\lambda}}(\boldsymbol{z}_i')|\le1$, so for any $\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{i-1},\boldsymbol{z}_{i+1},\cdots,\boldsymbol{z}_p$,

$$\big|\phi(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{i-1},\boldsymbol{z}_i,\boldsymbol{z}_{i+1},\cdots,\boldsymbol{z}_p)-\phi(\boldsymbol{z}_1,\cdots,\boldsymbol{z}_{i-1},\boldsymbol{z}_i',\boldsymbol{z}_{i+1},\cdots,\boldsymbol{z}_p)\big|\le\frac{|\sigma_i|^q}{\|\boldsymbol{\sigma}\|_q^q}.$$

Therefore, by McDiarmid's inequality [Ver18, Theorem 2.9.1], for any $t>0$,

$$\mathbb{P}\left(\phi-\mathbb{E}[\phi]>t\right)\le\exp\left(-\frac{2t^2}{\sum_{i=1}^p\frac{|\sigma_i|^{2q}}{\|\boldsymbol{\sigma}\|_q^{2q}}}\right)\le\exp\left(-\frac{2t^2}{\frac{\|\boldsymbol{\sigma}\|_\infty^q\|\boldsymbol{\sigma}\|_q^q}{\|\boldsymbol{\sigma}\|_q^{2q}}}\right)=\exp\left(-2t^2\tilde{d}_*^{q/2}\right).$$

By putting $t=\frac{1}{\theta}\varepsilon$ we get that $\phi\le\frac{7\varepsilon}{\theta}$ with probability at least $1-\exp(-\frac{2\varepsilon^2}{\theta^2}\tilde{d}_*^{q/2})$. Hence, if we denote $B_{\boldsymbol{\lambda}}=\{1\le i\le p:\ |\langle\boldsymbol{\lambda}-\pi\boldsymbol{\lambda},Z_i\rangle|\le\frac{1}{2}\theta\}$, $B_{\boldsymbol{\lambda}}^c=\{1,2,\dots,p\}\setminus B_{\boldsymbol{\lambda}}$, and

$$\Omega_2=\Big\{\forall\boldsymbol{\lambda}\in S_2^{N-1}\quad\frac{1}{\|\boldsymbol{\sigma}\|_q^q}\sum_{i\in B_{\boldsymbol{\lambda}}^c}\sigma_i^q\le\frac{7\varepsilon}{\theta}\Big\},$$

then we get $\mathbb{P}(\Omega_2)\ge1-\exp(-\frac{2\varepsilon^2}{\theta^2}\tilde{d}_*^{q/2})$ provided that $\tilde{d}_*^{q/2}\ge N$.

**Step 3. The sets $J_{\boldsymbol{\lambda}}$**

Let $\epsilon=\frac{\theta}{28}\kappa$. We work on $\Omega_0=\Omega_1\cap\Omega_2$, whose probability is not smaller than $1-2e^{-c_1\tilde{d}_*^{q/2}}$ provided that $N\le c_2\tilde{d}_*^{q/2}$ (where $c_1$ and $c_2$ depend only on $\theta$ and $\kappa$). For every $\boldsymbol{\lambda}\in S_2^{N-1}$ let $J_{\boldsymbol{\lambda}}=I_{\pi\boldsymbol{\lambda}}\cap B_{\boldsymbol{\lambda}}$. Then $I_{\pi\boldsymbol{\lambda}}=(I_{\pi\boldsymbol{\lambda}}\cap B_{\boldsymbol{\lambda}})\sqcup(I_{\pi\boldsymbol{\lambda}}\cap B_{\boldsymbol{\lambda}}^c)=J_{\boldsymbol{\lambda}}\sqcup(I_{\pi\boldsymbol{\lambda}}\cap B_{\boldsymbol{\lambda}}^c)$, so

$$\sum_{i\in J_{\boldsymbol{\lambda}}}|\sigma_i|^q\ge\sum_{i\in I_{\pi\boldsymbol{\lambda}}}|\sigma_i|^q-\sum_{i\in B_{\boldsymbol{\lambda}}^c}|\sigma_i|^q\ge\left(\frac{\kappa}{2}-\frac{7\varepsilon}{\theta}\right)\|\boldsymbol{\sigma}\|_q^q=\frac{\kappa}{4}\|\boldsymbol{\sigma}\|_q^q.$$

Thus, by the definitions of the sets $I_{\pi\boldsymbol{\lambda}}$ and $B_{\boldsymbol{\lambda}}$ we know that on the set $\Omega_1 \cap \Omega_2$, for every $\boldsymbol{\lambda} \in S_2^{N-1}$,

$$\|\mathbb{X}^\top \boldsymbol{\lambda}\|_q \geq \Big(\sum_{i\in J_{\boldsymbol{\lambda}}} |\sigma_i|^q |\langle Z_i, \boldsymbol{\lambda}\rangle|^q\Big)^{1/q} \geq \Big(\sum_{i\in J_{\boldsymbol{\lambda}}} |\sigma_i|^q |\langle Z_i, \pi\boldsymbol{\lambda}\rangle|^q\Big)^{1/q} - \Big(\sum_{i\in J_{\boldsymbol{\lambda}}} |\sigma_i|^q |\langle Z_i, \pi\boldsymbol{\lambda} - \boldsymbol{\lambda}\rangle|^q\Big)^{1/q}$$
$$\geq \theta\Big(\sum_{i\in J_{\boldsymbol{\lambda}}} |\sigma_i|^q\Big)^{1/q} - \frac{\theta}{2}\Big(\sum_{i\in J_{\boldsymbol{\lambda}}} |\sigma_i|^q\Big)^{1/q} \geq \frac{\theta}{2}\Big(\sum_{i\in J_{\boldsymbol{\lambda}}} |\sigma_i|^q\Big)^{1/q} \geq \frac{\kappa^{1/q}\theta}{4}\|\boldsymbol{\sigma}\|_q.$$

As a result,

$$\mathbb{P}\Big(\forall \boldsymbol{\lambda} \in S_2^{N-1} \quad \|\mathbb{X}^\top \boldsymbol{\lambda}\|_q^q \geq \frac{\kappa\theta^q}{4^q}\|\boldsymbol{\sigma}\|_q^q\Big) \geq \mathbb{P}(\Omega_1 \cap \Omega_2) \geq 1 - 2e^{-c_1 \tilde{d}_*^{q/2}},$$

provided that $N \leq c_2 d_*^{q/2}$. ∎

If $q \geq 2$, then (5.3) implies that $\tilde{d}_*^{q/2} \geq d_*(A)$ for some constant $c$ depending only on $q$. Unfortunately, if $q < 2$, then $\tilde{d}_*^{q/2} \leq C d_*(A)$, since

$$\sum_{i=1}^p |\sigma_i|^{2q/(2-q)} \leq \|\boldsymbol{\sigma}\|_\infty^{q^2/(2-q)} \sum_{i=1}^p |\sigma_i|^q,$$

and $\tilde{d}_*^{q/2}$ may be of smaller order than $d_*(A)$ (for example if $\sigma_i = i^{(q-2)/2q}$). However, if $A = \mathrm{Id}_p$, then $\tilde{d}_*^{q/2} = p \sim d_*(A)$. Hence, Propositions 38 and 37, yield the following.

**Corollary 7.** *If $q \geq 2$, then there exists a constant $c > 0$, such that with probability at least $1 - \exp(-c d_*(A))$, for any $\boldsymbol{\lambda} \in S_2^{N-1}$, $\|\mathbb{X}^\top \boldsymbol{\lambda}\|_q \geq c\ell_q(A)$, provided that $N \leq c d_*(A)$.*

*If $q < 2$, then there exists a constant $c > 0$, such that with probability at least $1 - C\,\mathrm{Log}(p^{1/q}/d_*(A)) e^{-\alpha c d_*(A)^\theta / \mathrm{Log}^{2\theta}(p^{1/q}/d_*(A))}$, for any $\boldsymbol{\lambda} \in S_2^{N-1}$, $\|\mathbb{X}^\top \boldsymbol{\lambda}\|_q \geq c\ell_q(A)$, provided that $N \leq \frac{c\gamma d_*(A)}{\mathrm{Log}^2(p^{1/q}/d_*(A))}$.*

## 5.2 Proof of Theorem 4

**Proof.** We prove Theorem 4, and for the sake of generality, we replace the covariance matrix by $\Sigma$, that is, $\Sigma = \mathbb{E}[\phi(X) \otimes \phi(X)]$. For some $2 < p \leq 2 + \epsilon$, let

$$B = \sup\left(\mathbb{E}\,|\langle \phi(X), f\rangle_{\mathcal{H}}|^p : \|f\|_{\mathcal{H}} = 1\right). \tag{5.15}$$

Since we have $L_{2+\epsilon} - L_2$ norm equivalence of marginals (recall (1.27)), $B \leq \kappa^p \|\Sigma\|_{\mathrm{op}}^{p/2}$. To obtain a high-probability upper bound for Equation (1.30), we first use a one-scale net argument. Set $V_{\epsilon_0} = \{\pi\lambda : \lambda \in S_2^{N-1}\}$ to be an $\epsilon_0$-net of $S_2^{N-1}$ so that for all $\lambda \in S_2^{N-1}$, $\|\pi\lambda - \lambda\|_2 \leq \epsilon_0$ and $|V_{\epsilon_0}| \leq (5/\epsilon_0)^N$. Unlike the situation in [P4], the choice of $\epsilon_0$ here does not affect the probability deviation because we will use a different discretization procedure to control $\mathbb{E}_\eta V_{I_{\eta,\boldsymbol{u}}}$(which will be defined later) uniformly on $S_2^{N-1}$, which does not depend on $\epsilon_0$. Therefore, one may choose an $\epsilon_0$ as small as possible(say, 1/10) to compensate for absolute constants at the end of the proof.

It follows from the triangle inequality that for every $\boldsymbol{\lambda} \in S_2^{N-1}$,

$$\left|\frac{\|A\boldsymbol{\lambda}\|_{\mathcal{H}}^2}{(\ell^*)^2} - 1\right| = \left|\frac{\|A(\boldsymbol{\lambda} - \pi\boldsymbol{\lambda})\|_{\mathcal{H}}^2}{(\ell^*)^2} + \frac{\|A(\pi\boldsymbol{\lambda})\|_{\mathcal{H}}^2}{(\ell^*)^2} + \frac{2\langle A(\boldsymbol{\lambda} - \pi\boldsymbol{\lambda}), A(\pi\boldsymbol{\lambda})\rangle_{\mathcal{H}}}{(\ell^*)^2} - 1\right|$$
$$\leq \frac{\|A(\boldsymbol{\lambda} - \pi\boldsymbol{\lambda})\|_{\mathcal{H}}^2}{(\ell^*)^2} + \left|\frac{\|A(\pi\boldsymbol{\lambda})\|_{\mathcal{H}}^2}{(\ell^*)^2} - 1\right| + \left|\frac{2\langle A(\boldsymbol{\lambda} - \pi\boldsymbol{\lambda}), A(\pi\boldsymbol{\lambda})\rangle_{\mathcal{H}}}{(\ell^*)^2}\right|$$

and so, from the Cauchy-Schwarz inequality,

$$\sup_{\boldsymbol{\lambda}\in S_2^{N-1}} \left|\frac{\|A\boldsymbol{\lambda}\|_{\mathcal{H}}^2}{(\ell^*)^2} - 1\right| \leq \Phi^2 + \Psi^2 + 2\Phi\sqrt{\Psi^2 + 1}$$

where

$$\Phi^2 := \sup_{\boldsymbol{\lambda}\in S_2^{N-1}} \frac{\|A(\boldsymbol{\lambda} - \pi\boldsymbol{\lambda})\|_{\mathcal{H}}^2}{(\ell^*)^2} \text{ and } \Psi^2 := \sup_{\boldsymbol{\lambda}\in S_2^{N-1}} \left|\frac{\|A(\pi\boldsymbol{\lambda})\|_{\mathcal{H}}^2}{(\ell^*)^2} - 1\right|.$$

Thus, we only need to bound $\Phi$ and $\Psi$ from above. To that end, we start with a decoupling argument to deal with the cross terms.

**A decoupling argument** Let $(\eta_i)_{1\le i\le N}$ be $N$ i.i.d. selectors (i.e. $\mathbb{P}(\eta_i=0)=\mathbb{P}(\eta_i=1)=1/2$) and define $I_\eta := \{1\le i\le N : \eta_i = 1\}$. For all $\boldsymbol{u}\in\mathbb{R}^N$ and $I\subset[N]$, we also define

$$V_{I,\boldsymbol{u}} := \frac{1}{(\ell^*)^2}\left\langle \left(\sum_{i\in I} u_i\phi(X_i)\right), \left(\sum_{j\in I^c} u_j\phi(X_j)\right)\right\rangle_{\mathcal{H}}.$$

For every $\boldsymbol{u}\in B_2^N$, a decoupling technique (see for instance Chapter 6 of [Ver18]) leads to the following result

$$\begin{aligned}
\frac{\|A\boldsymbol{u}\|_{\mathcal{H}}^2}{(\ell^*)^2} &= \sum_{i=1}^N \frac{\|\phi(X_i)\|_{\mathcal{H}}^2 u_i^2}{(\ell^*)^2} + \sum_{i\neq j} u_iu_j \frac{\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}}{(\ell^*)^2} \\
&= \sum_{i=1}^N \frac{\|\phi(X_i)\|_{\mathcal{H}}^2 u_i^2}{(\ell^*)^2} + \sum_{i,j=1}^N \mathbb{E}_\eta 4\eta_i(1-\eta_j)u_iu_j \frac{\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}}{(\ell^*)^2} \\
&= \sum_{i=1}^N \frac{\|\phi(X_i)\|_{\mathcal{H}}^2 u_i^2}{(\ell^*)^2} + \frac{4}{(\ell^*)^2}\mathbb{E}_\eta\left\langle \left(\sum_{i\in I_\eta} u_i\phi(X_i)\right), \left(\sum_{j\in I_\eta^c} u_j\phi(X_j)\right)\right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^N \frac{\|\phi(X_i)\|_{\mathcal{H}}^2 u_i^2}{(\ell^*)^2} + 4\mathbb{E}_\eta V_{I_\eta,\boldsymbol{u}},
\end{aligned} \tag{5.16}$$

where $\mathbb{E}_\eta$ is the expectation with respect to $(\eta_i)_{i\in[N]}$ conditionally on all other random variables. With the decoupling technique, we are going to estimate $\Phi^2$ and $\Psi^2$ from above. We start with $\Psi^2$ and get

$$\begin{aligned}
\Psi^2 = \sup_{\boldsymbol{\lambda}\in S_2^{N-1}}\left|\frac{\|A(\pi\boldsymbol{\lambda})\|_{\mathcal{H}}^2}{(\ell^*)^2}-1\right| &\le \sup_{\lambda\in S_2^{N-1}}\left|\sum_{i=1}^N \frac{\|\phi(X_i)\|_{\mathcal{H}}^2}{(\ell^*)^2}(\pi\lambda)_i^2 - 1\right| + 4\sup_{\boldsymbol{\lambda}\in S_2^{N-1}}\left|\mathbb{E}_\eta V_{I_\eta,\pi\boldsymbol{\lambda}}\right| \\
&\le \max_{1\le i\le N}\left|\frac{\|\phi(X_i)\|_{\mathcal{H}}^2}{(\ell^*)^2}-1\right| + 4\sup_{\boldsymbol{\lambda}\in S_2^{N-1}}\left|\mathbb{E}_\eta V_{I_\eta,\pi\boldsymbol{\lambda}}\right|.
\end{aligned} \tag{5.17}$$

We next have $\Phi^2 = \sup\left(\|A(\boldsymbol{\lambda}-\pi\boldsymbol{\lambda})\|_{\mathcal{H}}^2/(\ell^*)^2 : \boldsymbol{\lambda}\in S_2^{N-1}\right) \le (\epsilon_0/\ell^*)^2\|A\|_{\text{op}}^2$, where $\|A\|_{\text{op}}$ is the operator norm of $A:(\mathbb{R}^N,\ell_2)\to(\mathcal{H},\|\cdot\|_{\mathcal{H}})$, thus it remains to prove a high probability upper bound on $\|A\|_{\text{op}}$. From (5.16) we know that

$$\frac{\|A\|_{\text{op}}^2}{(\ell^*)^2} \le \max_{i\in[N]}\frac{\|\phi(X_i)\|_{\mathcal{H}}^2}{(\ell^*)^2} + 4\sup_{\|\boldsymbol{\mu}\|_2=1}\mathbb{E}_\eta V_{I_\eta,\boldsymbol{\mu}}. \tag{5.18}$$

As a result,

$$\Phi^2 \le \frac{4\epsilon_0^2}{(\ell^*)^2}\sup_{\pi\boldsymbol{\mu}\in V_{1/2}}\|A(\pi\boldsymbol{\mu})\|_{\mathcal{H}}^2 \le 4\epsilon_0^2\left(\max_{1\le i\le N}\frac{\|\phi(X_i)\|_{\mathcal{H}}^2}{(\ell^*)^2} + 4\sup_{\pi\boldsymbol{\mu}\in V_{1/2}}\mathbb{E}_\eta V_{I_\eta,\pi\boldsymbol{\mu}}\right). \tag{5.19}$$

Therefore, we only need to find a high probability upper bound on $\left|\mathbb{E}_\eta V_{I_\eta,\boldsymbol{u}}\right|$ uniformly for all $\boldsymbol{u}$ in $V_{1/2}$ and $V_{\epsilon_0}$. In contrast to the method employed in [P4], we immediately derive this upper bound uniformly over $B_2^N$. This will result in $\epsilon_0$ being a free parameter, as we will show at the end of the proof ((5.27) holds on $S_2^{N-1}$, instead of $V_{\epsilon_0}$).

To achieve this, we adapt [Tik18]'s argument. We recall that Theorem 4 contains two aspects:

1. when $\text{Tr}(\Sigma)$ is the dominating term in $\lambda+\text{Tr}(\Sigma)$,
2. when $\lambda$ is dominating.

In case [1], we need not only the upper bound of $\sup\left(\left\|\mathbb{X}^\top\boldsymbol{\lambda}\right\|_{\mathcal{H}} : \boldsymbol{\lambda}\in S_2^{N-1}\right)$, but also the lower bound for

$$\inf\left(\left\|\mathbb{X}^\top\boldsymbol{\lambda}\right\|_{\mathcal{H}} : \boldsymbol{\lambda}\in S_2^{N-1}\right).$$

However, in case [2], we only need the upper bound for $\sup\left(\left\|\mathbb{X}^\top\boldsymbol{\lambda}\right\|_{\mathcal{H}} : \boldsymbol{\lambda}\in S_2^{N-1}\right)$. This will be clear in Section 5.2.

We first introduce some notation.

For all $I, J \subset [N]$ and $\ell \in [N]$, let $S_2^J$ be the unit sphere in Euclidean norm of $\mathbb{R}^J$(see as a subspace of $\mathbb{R}^N$ endowed by the canonical vectors indexed by $J$), and let

$$S_I^J := \left\{\lambda \in S_2^J : \lambda_i = 0, \forall i \in I^c\right\} \text{ and } S_{I,\ell}^J := \left\{\lambda \in S_I^J : |\{i \in I : \lambda_i \neq 0\}| \leq \ell\right\}.$$

For the sake of simplicity, we let $S_I^N := S_I^{[N]}$ and $S_{I,\ell}^N := S_{I,\ell}^{[N]}$. For each $k \leq N$ and $I, \mathcal{C} \subset [N]$, we denote

$$\begin{aligned} g(k, \mathcal{C}, I) &:= \sup\left(\left|\left\langle \sum_{i \in I \cap \mathcal{C}} v_i \phi(X_i), \sum_{j \in I^c \cap \mathcal{C}} u_j \phi(X_j)\right\rangle_{\mathcal{H}}\right|, \boldsymbol{u}, \boldsymbol{v} \in S_{[N],k}^N\right), \text{ and} \\ g(N, I) &:= g(k, [N], I). \end{aligned} \tag{5.20}$$

Further, for any vector $\boldsymbol{u} \in \mathbb{R}^N$, and any $i \leq N$, we set

$$W_{\boldsymbol{u},i} := \left\langle \phi(X_i), \sum_{j=1}^N u_j \phi(X_j)\right\rangle_{\mathcal{H}}, \tag{5.21}$$

and for some $J \subset [N]$, we denote $(|W_{\boldsymbol{u},J}|)_j^*$ as the $j$-th largest absolute value of the coordinates of $(W_{\boldsymbol{u},j})_{j \in J}$. Given $1 \leq k \leq N$ and $I \subset [N]$, we define

$$\mathcal{M}_{I,k} = \max\left(\left\|\sum_{i \in I} \lambda_i \phi(X_i)\right\|_{\mathcal{H}} : \boldsymbol{\lambda} \in S_{I,k}^N\right), \quad \text{and } \mathcal{M}_k = \mathcal{M}_{[N],k}.$$

In particular, $\|A\|_{\text{op}} = \mathcal{M}_N$, and by the decoupling argument again, see (5.18)

$$\mathcal{M}_N^2 \leq \max\left(\|\phi(X_i)\|_{\mathcal{H}}^2 : 1 \leq i \leq N\right) + 4(\ell^*)^2 \sup\left(\mathbb{E}_\eta V_{I_\eta, \boldsymbol{u}} : \boldsymbol{u} \in S_2^{N-1}\right). \tag{5.22}$$

Once (5.22) has been obtained, the next step is to derive an upper bound for the supremum of $\left|\mathbb{E}_\eta V_{I_\eta,\boldsymbol{u}}\right|$ over all $\boldsymbol{u} \in S_2^{N-1}$ in relation to $\mathcal{M}_N$. The purpose of the remaining part of the proof will be to establish this point. In order to accomplish this, we begin with

$$\begin{aligned} &4\left(\ell^*\right)^2 \sup\left(\left|\mathbb{E}_\eta V_{I_\eta,\boldsymbol{u}}\right| : \boldsymbol{u} \in S_2^{N-1}\right) \\ &= 4 \sup\left(\mathbb{E}_\eta\left|\left\langle\left(\sum_{i \in I_\eta} u_i \phi(X_i)\right), \left(\sum_{j \in I_\eta^c} u_j \phi(X_j)\right)\right\rangle_{\mathcal{H}}\right| : \boldsymbol{u} \in S_2^{N-1}\right) \\ &= \frac{4}{2^N} \sup\left(\sum_{I \subset [N]}\left|\left\langle \sum_{i \in I} u_i \phi(X_i), \sum_{j \in I^c} u_j \phi(X_j)\right\rangle_{\mathcal{H}}\right| : \boldsymbol{u} \in S_2^{N-1}\right) \\ &\leq \frac{4}{2^N} \sum_{I \subset [N]} g\left(N, I\right), \end{aligned} \tag{5.23}$$

where the last step is via Jensen's inequality and by the fact that optimizing over $(\boldsymbol{v}, \boldsymbol{u}) \in S_{[N],k}^N \times S_{[N],k}^N$ leads to a larger supreme.

**A sparsifying argument** We begin by the sparsifying lemma for $g\left(N, \mathcal{C}, I\right)$.

**Lemma 29.** *Let $0 < \varepsilon \leq 1$, $k \geq 12/\varepsilon^2$, $4 \leq m \leq k$, and let $I, \mathcal{C} \subset [N]$. There then exists an absolute constant $C_{54} > 0$ such that*

$$\begin{aligned} g(k, \mathcal{C}, I) \leq\ & g(m, \mathcal{C}, I) + 2C_{54}\varepsilon^{-2} \max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}\right) \\ &+ C_{54}\sqrt{k}\varepsilon^{-2}\Bigg(\sup\left(|W_{\boldsymbol{y}, I^c \cap \mathcal{C}}|_{\lfloor m/4 \rfloor}^* : \boldsymbol{y} \in S_{I \cap \mathcal{C}, \varepsilon k}^N\right) \\ &+ \sup\left(|W_{\boldsymbol{z}, I \cap \mathcal{C}}|_{\lfloor m/4 \rfloor}^* : \boldsymbol{z} \in S_{I^c \cap \mathcal{C}, \varepsilon k}^N\right)\Bigg). \end{aligned} \tag{5.24}$$

As mentioned in [Tik18], when compared to the findings of [BM22a], this method of sparsifying yields a lower cardinality when combined with the subsequent discretization argument. This gives rise to a broad condition about the tail of kernel features (from $L_{\log N} - L_2$ to $L_{2+\epsilon} - L_2$).

The next lemma is a combination of [Tik18, Lemma 13] and [Tik18, Lemma 4]:

**Lemma 30.** *Let $0 < \rho \leq 1$, $r, h, p, q \in \mathbb{N}$ and $r \geq 2$, let $V_\rho$ be a support-preserving Euclidean $\rho$-net of $S^q_{[q],h}$. Further, let $T$ be a $p \times q$ matrix. Then*

$$\sup\left(|T\boldsymbol{u}|^*_r : \boldsymbol{u} \in S^q_{[q],h}\right) \leq 2\sup\left(|T\boldsymbol{v}|^*_{\lfloor r/2 \rfloor} : \boldsymbol{v} \in V_\rho\right) + \frac{4\rho}{\sqrt{r}} \sup\left(\sqrt{\sum_{i=1}^{r} \left(|T\boldsymbol{u}|^*_i\right)^2} : \boldsymbol{u} \in S^q_{[q],h}\right).$$

*Here, the support-preserving Euclidean $\rho$-net means that $V_\rho$ is a $\rho$-net, such that for all $\boldsymbol{x} \in S^q_{[q],k}$, there exists $\boldsymbol{y} \in V_\rho$ with $\operatorname{supp}(\boldsymbol{y}) \subset \operatorname{supp}(\boldsymbol{x})$ and $\|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq \rho$. By [Tik18, Lemma 4], there exists $C_{55} > 0$ and $V_\rho \subset S^q_{[q],k}$ such that $|V_\rho| \leq (C_{55} q/\rho h)^h$, and $V_\rho$ is a support-preserving Euclidean $\rho$-Net of $S^q_{[q],h}$.*

**A dimension reduction argument** Combining Lemma 29 and Lemma 30 with an induction argument, we obtain the following proposition, as a deterministic argument. Compared to [Tik18, Proposition 14], we remove the condition $N \geq 128 C_{54} \varepsilon^{-2} k$. This assumption $N \geq 128 C_{54} \varepsilon^{-2} k$ prevents us from being in the Dvoretzky-Milman regime, because when $N = k$, we need $\varepsilon$ sufficiently large, however, Proposition 40 needs $\varepsilon$ to be sufficiently small. Removing this assumption is possible because we are going to choose $k = N$ in Proposition 40, which keeps $128 C_{54} k/(\varepsilon^2 N)$ as a constant in the proof.

**Proposition 39.** *Let $I, \mathcal{C} \subset [N]$, and let $0 < \varepsilon < 1/3$ and $k \geq 24/\varepsilon^2 \vee N$. Denote $t := \lfloor \log_2\left(\varepsilon^2 k/24\right) \rfloor$ and define $k_j := \lfloor k/2^j \rfloor$, $0 \leq j \leq t$. There are then subsets $V_j \subset S^I_{I,\varepsilon k_j}$ and $V'_j \subset S^{I^c}_{I^c,\varepsilon k_j}$ for all $0 \leq j \leq t-1$ such that for the absolute constants from Lemma 29 and Lemma 30,*

- $|V_j|, |V'_j| \leq (C_{55} N/\varepsilon k_j)^{2\varepsilon k_j}$ *for all* $0 \leq j \leq t-1$.
- *We have*

$$\begin{aligned} g(k, I) \quad &\leq \quad \exp\left(\frac{128 C_{54} k}{\varepsilon^2 N}\right)(48 + 2C_{54}\log_2(k))\,\varepsilon^{-2} \max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}\right) \\ &+ \quad 2c_{54} \exp\left(\frac{128 C_1 k}{\varepsilon^2 N}\right)\varepsilon^{-2} \sum_{j=0}^{t-1} \sqrt{k_j}\left(\sup\left(|W_{\boldsymbol{u}, I^c}|^*_{\lfloor k_{j+1}/16 \rfloor} : \boldsymbol{u} \in V_j\right)\right. \\ &+ \quad \left.\sup\left(|W_{\boldsymbol{v}, I}|^*_{\lfloor k_{j+1}/16 \rfloor} : \boldsymbol{v} \in V'_j\right)\right). \end{aligned}$$

**Proof.** First, we let $j < t$ and consider the quantity $g(k_j, \mathcal{C}, I)$. By Lemma 29, for $k = k_j$ and $m = k_{j+1}$ (where we have $k_j \geq 12/\varepsilon^2$ and $k_{j+1} \geq 4$),

$$\begin{aligned} g(k_j, \mathcal{C}, I) \quad &\leq \quad g(k_{j+1}, \mathcal{C}, I) + 2C_{54}\varepsilon^{-2} \max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}\right) \\ &+ \quad C_{54}\sqrt{k_j}\varepsilon^{-2}\left(\sup\left(|W_{\boldsymbol{y}, I^c \cap \mathcal{C}}|^*_{\lfloor k_{j+1}/4 \rfloor} : \boldsymbol{y} \in S^N_{I \cap \mathcal{C}, \varepsilon k_j}\right)\right. \\ &+ \quad \left.\sup\left(|W_{\boldsymbol{z}, I \cap \mathcal{C}}|^*_{\lfloor k_{j+1}/4 \rfloor} : \boldsymbol{z} \in S^N_{I^c \cap \mathcal{C}, \varepsilon k_j}\right)\right). \end{aligned}$$

We first find upper bounds for

$$\begin{aligned} &\sup\left(|W_{\boldsymbol{y}, I^c \cap \mathcal{C}}|^*_{\lfloor k_{j+1}/4 \rfloor} : \boldsymbol{y} \in S^N_{I \cap \mathcal{C}, \varepsilon k_j}\right), \text{ and} \\ &\sup\left(|W_{\boldsymbol{z}, I \cap \mathcal{C}}|^*_{\lfloor k_{j+1}/4 \rfloor} : \boldsymbol{y} \in S^N_{I^c \cap \mathcal{C}, \varepsilon k_j}\right). \end{aligned}$$

We only obtain the former, as the latter follows from the same ideas.

From the definition of $\boldsymbol{y} \in S^N_{I \cap \mathcal{C}, \varepsilon k_j}$ we know that $\boldsymbol{y}$ is supported on $I \cap \mathcal{C}$. By the definition of $W$ (see (5.21)), $W_{\boldsymbol{y}, I^c \cap \mathcal{C}} = \left(\left\langle \phi(X_i), \sum_{j \in I \cap \mathcal{C}} y_j \phi(X_j)\right\rangle_{\mathcal{H}}\right)_{i \in I^c \cap \mathcal{C}}$. Therefore, taking supreme over $S^N_{I \cap \mathcal{C}, \varepsilon k_j}$ is equivalent to taking it over $\boldsymbol{y} \in S^{I \cap \mathcal{C}}_{I \cap \mathcal{C}, \varepsilon k_j}$.

Apply Lemma 30 with $r = \lfloor k_{j+1}/4 \rfloor$, $\rho = k_j/N$, $h = \lfloor \varepsilon k_j \rfloor$, and matrix $T = \left(\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right)_{i,j}$ for $(i,j) \in ((I^c \cap \mathcal{C}) \times (I \cap \mathcal{C}))$, then $T\boldsymbol{y} = \left(\left\langle \phi(X_i), \sum_{j \in I \cap \mathcal{C}} y_j \phi(X_j) \right\rangle_{\mathcal{H}}\right)_{i \in I^c \cap \mathcal{C}} = W_{\boldsymbol{y}, I^c \cap \mathcal{C}}$ and thus there exists $V_j \subset S^{I \cap \mathcal{C}}_{I \cap \mathcal{C}, \varepsilon k_j} \subset S^N_{I, \varepsilon k_j}$, which is a support-preserving $k_j/N$-net of cardinality at most

$$\left(\frac{C_{55}\,|I \cap \mathcal{C}|}{((k_j/N)\varepsilon k_j)}\right)^{\varepsilon k_j} \leq \left(\frac{C_{55}}{\varepsilon}\right)^{\varepsilon k_j} \left(\frac{N}{k_j}\right)^{2\varepsilon k_j} \leq \left(\frac{C_{55} N}{\varepsilon k_j}\right)^{2\varepsilon k_j}$$

since we can choose $C_{55} > 1 \vee \varepsilon$, such that

$$\begin{aligned} & \sup\left(|W_{\boldsymbol{y}, I^c \cap \mathcal{C}}|^*_{\lfloor k_{j+1}/4 \rfloor} : \boldsymbol{y} \in S^N_{I \cap \mathcal{C}, \varepsilon k_j}\right) \leq 2 \sup\left(|W_{\boldsymbol{u}, I^c \cap \mathcal{C}}|^*_{\lfloor \lfloor k_{j+1}/4 \rfloor / 2 \rfloor} : \boldsymbol{u} \in V_j\right) \\ + \quad & \frac{4k_j/N}{\sqrt{\lfloor k_{j+1}/4 \rfloor}} \sup\left(\sqrt{\sum_{i=1}^{\lfloor k_{j+1}/4 \rfloor} \left(|W_{\boldsymbol{y}, I^c \cap \mathcal{C}}|^*_i\right)^2} : \boldsymbol{y} \in S^N_{I \cap \mathcal{C}, \varepsilon k_j}\right) \\ \leq \quad & 2 \sup\left(|W_{\boldsymbol{u}, I^c}|^*_{\lfloor \lfloor k_{j+1}/4 \rfloor / 2 \rfloor} : \boldsymbol{u} \in V_j\right) + \frac{4k_j/N}{\sqrt{\lfloor k_{j+1}/4 \rfloor}} g(k_{j+1}, \mathcal{C}, I), \end{aligned}$$

where we have used $k_{j+1}/4 \leq k_{j+1}$ and the fact that for any $\boldsymbol{u} \in \mathbb{R}^N$ and $\ell \in [N]$,

$$\sup\left(\sum_{i \in I} u_i v_i : \boldsymbol{v} \in S^I_{I,\ell}\right) = \sqrt{\sum_{i=1}^{\ell} ((\boldsymbol{u}_I)^*_i)^2}.$$

The analysis for $\sup\left(|W_{\boldsymbol{z}, I \cap \mathcal{C}}|^*_{\lfloor m/4 \rfloor} : \boldsymbol{z} \in S^N_{I^c \cap \mathcal{C}, \varepsilon k}\right)$ is similar. We obtain

$$\begin{aligned} g(k_j, \mathcal{C}, I) \quad \leq \quad & \left(1 + \frac{32 C_{54} k_j}{\varepsilon^2 N}\right) g(k_{j+1}, \mathcal{C}, I) + 2C_{54}\varepsilon^{-2} \max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}\right) \\ + \quad & 2C_{54}\sqrt{k_j}\varepsilon^{-2}\left(\sup\left(|W_{\boldsymbol{u}, I^c}|^*_{\lfloor k_{j+1}/16 \rfloor} : \boldsymbol{u} \in V_j\right) + \sup\left(|W_{\boldsymbol{v}, I}|^*_{\lfloor k_{j+1}/16 \rfloor} : \boldsymbol{v} \in V'_j\right)\right). \end{aligned}$$

Notice that for any $t' < t$,

$$\prod_{j=0}^{t'} \left(1 + \frac{32 C_{54} k_j}{\varepsilon^2 N}\right) \leq \prod_{j=0}^{t} \left(1 + \frac{32 C_{54} k_j}{\varepsilon^2 N}\right) \leq \exp\left(\sum_{j=0}^{t} \frac{32 C_{54} k_j}{\varepsilon^2 N}\right) \leq \exp\left(\frac{128 C_{54} k}{\varepsilon^2 N}\right).$$

By induction over $0 \leq j < t$, we obtain

$$\begin{aligned} & \exp\left(-\frac{128 C_{54} k}{\varepsilon^2 N}\right) g(k, \mathcal{C}, I) \leq g(k_t, \mathcal{C}, I) + 2C_{54}\varepsilon^{-2} \log_2(k) \max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}\right) \\ + \quad & 2C_{54}\varepsilon^{-2} \sum_{j=0}^{t-1} \sqrt{k_j}\left(\sup\left(|W_{\boldsymbol{u}, I^c}|^*_{\lfloor k_{j+1}/16 \rfloor} : \boldsymbol{u} \in V_j\right) + \sup\left(|W_{\boldsymbol{v}, I}|^*_{\lfloor k_{j+1}/16 \rfloor} : \boldsymbol{v} \in V'_j\right)\right). \end{aligned}$$

Finally, we bound $g(k_t, \mathcal{C}, I)$ from above:

$$\begin{aligned} g(k_t, \mathcal{C}, I) \quad \leq \quad & \sup\left(\sum_{i,j=1}^{N} \left|y_i z_j \langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : \boldsymbol{y} \in S^N_{I \cap \mathcal{C}, k_t}, \boldsymbol{z} \in S^N_{I^c \cap \mathcal{C}, k_t}\right) \\ \leq \quad & \max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}\right) \sup\left(\sum_{i,j=1}^{N} |y_i z_j| : \boldsymbol{y} \in S^N_{I \cap \mathcal{C}, k_t}, \boldsymbol{z} \in S^N_{I^c \cap \mathcal{C}, k_t}\right) \\ = \quad & \max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}\right) \cdot k_t \\ \leq \quad & 48\varepsilon^{-2} \max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}\right). \end{aligned}$$

■

**Stochastic arguments** We next introduce the randomness of $\phi(X)$. We will see in the next proposition that choosing $k = N$ does not destroy the proof, but leads to a larger constant compared with [Tik18, Proposition 15].

**Proposition 40.** *For the absolute constant $C_{54}$ from Lemma 29 and $C_{55}$ from Lemma 30, there are sufficiently large universal constants $C_{56} = (12C_{55}/\varepsilon^3)^{36/\varepsilon}$ for $\varepsilon < 1/256$ , $C_{57}$ and $C_{58}$ depending on $p$ such that: Suppose $N \geq (12C_{55}/\varepsilon^3)^6 C_{56}^{-\varepsilon/6} \vee \exp(48/C_{54})$, let $I \subset [N]$. With probability at least $1 - C_{57}/N^3$, for all $\mathcal{C} \subset [N]$,*

$$g(N, \mathcal{C}, I) \leq C_{58} \log_2(N) \max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}\right) + C_{58} B^{1/p} \sqrt{N} \mathcal{M}_N.$$

**Proof.** Let $t = \lfloor \log_2\left(\varepsilon^2 N/24\right) \rfloor$ and $k_j = \lfloor N/2^j \rfloor$ for each $j \in \{0, 1, \cdots, t\}$. Fix $j \in \{0, 1, \cdots, t-1\}$, we study the term $\sup\left(|W_{\boldsymbol{u}, I^c}|^*_{\lfloor k_{j+1}/16 \rfloor} : \boldsymbol{u} \in V_j\right)$ appearing by Proposition 39 for $k = N$. In the whole proof below $k = N$, in particular, $k_j = \lfloor N/2^j \rfloor$, $j = 0, \cdots, t$ and $t = \lfloor \log_2(\varepsilon^2 N/24) \rfloor$. At this time, $48 < C_{54} \log_2(N)$. For each $\boldsymbol{u} \in V_j$, condition on $(X_i)_{i \in I}$, $\left(\left\langle \phi(X_j), \sum_{i \in I} u_i \phi(X_i)\right\rangle_{\mathcal{H}}\right)_{j \in I^c} = (W_{\boldsymbol{u}, j})_{j \in I^c}$ are i.i.d. random variables. Recalling the definition of $B$, see (5.15), the conditional expectation of $|W_{\boldsymbol{u},j}|^p$ given $(X_i)_{i \in I}$ satisfies

$$\mathbb{E}\left[|W_{\boldsymbol{u},j}|^p \left| (X_i)_{i \in I}\right.\right] \leq B \left\| \sum_{i \in I} u_i \phi(X_i) \right\|^p_{\mathcal{H}} \leq B \mathcal{M}_N^p.$$

Let $\boldsymbol{u} \in V_j$(in particular, $\mathrm{supp}(\boldsymbol{u}) \subset I$). Condition on $(X_i)_{i \in I}$, $\mathbb{E}\left[|W_{\boldsymbol{u},j}|^p / \left\|\sum_{i \in I} u_i \phi(X_i)\right\|^p_{\mathcal{H}} \left| (X_i)_{i \in I}\right.\right] \leq B$. For $\tau_j^p = 32eB\left(N/k_{j+1}\right)^{1+256\varepsilon}$,

$$\begin{aligned}
&\mathbb{P}\left(\frac{|W_{\boldsymbol{u}, I^c}|^*_{\lfloor k_{j+1}/16 \rfloor}}{\left\|\sum_{i \in I} u_i \phi(X_i)\right\|_{\mathcal{H}}} \geq \tau_j \middle| (X_i)_{i \in I}\right) \\
&= \mathbb{P}\left(\exists J \subset I^c, |J| \geq \lfloor k_{j+1}/16 \rfloor, \text{ such that } \forall j \in J, \frac{|W_{\boldsymbol{u},j}|}{\left\|\sum_{i \in I} u_i \phi(X_i)\right\|_{\mathcal{H}}} \geq \tau_j \middle| (X_i)_{i \in I}\right) \\
&\leq \binom{|I^c|}{\lfloor k_{j+1}/16 \rfloor} \left(\mathbb{P}\left(\frac{|W_{\boldsymbol{u},j}|}{\left\|\sum_{i \in I} u_i \phi(X_i)\right\|_{\mathcal{H}}} \geq \tau_j \middle| (X_i)_{i \in I}\right)\right)^{\lfloor k_{j+1}/16 \rfloor} \\
&\leq \left(\frac{e\,|I^c|}{\lfloor k_{j+1}/16 \rfloor}\right)^{\lfloor k_{j+1}/16 \rfloor} \left(\frac{B}{\tau_j^p}\right)^{\lfloor k_{j+1}/16 \rfloor} \\
&\leq \left(\frac{|I^c|}{N} \left(\frac{k_{j+1}}{N}\right)^{256\varepsilon}\right)^{\lfloor k_{j+1}/16 \rfloor} \leq \left(\frac{k_{j+1}}{N}\right)^{4\varepsilon k_j}.
\end{aligned}$$

Hence, conditionally on $(X_i)_{i \in I}$, with probability at least $1 - (k_{j+1}/N)^{4\varepsilon k_j}$,

$$|W_{\boldsymbol{u}, I^c}|^*_{\lfloor k_{j+1}/16 \rfloor} \leq \tau_j \left\| \sum_{i \in I} u_i \phi(X_i) \right\|_{\mathcal{H}} \leq \tau_j \mathcal{M}_N.$$

Therefore, by Fubini's theorem, with probability at least $1 - (k_{j+1}/N)^{4\varepsilon k_j}$, $|W_{\boldsymbol{u}, I^c}|^*_{\lfloor k_{j+1}/16 \rfloor} \leq \tau_j \mathcal{M}_N$. Taking the union bound over all $\boldsymbol{u} \in V_j$(note that the cardinality of $V_j$ is given in Proposition 39) and using $k_{j+1} \leq k_j$,

$$\mathbb{P}\left(\sup\left(|W_{\boldsymbol{u}, I^c}|^*_{\lfloor k_{j+1}/16 \rfloor} : \boldsymbol{u} \in V_j\right) \geq \tau_j\right) \leq \left(\frac{k_{j+1}^2}{N^2} \cdot \frac{C_{55} N}{\varepsilon k_j}\right)^{2\varepsilon k_j} \leq \left(\frac{k_{j+1}}{N} \cdot \frac{C_{55}}{\varepsilon}\right)^{2\varepsilon k_j}.$$

Since $t \mapsto t \log(eN/t)$ is increasing on $\{t : 0 < t \leq N\}$ and $k_t \geq 12/\varepsilon^2$, we have for all $j = 0, \cdots, t-1$,

$$\begin{aligned}
2\varepsilon k_j \log\left(\frac{\varepsilon N}{C_{55} k_{j+1}}\right) &\geq 2\varepsilon k_j \log\left(\frac{\varepsilon N}{C_{55} k_j}\right) \geq 2\varepsilon k_t \log\left(\frac{\varepsilon N}{C_{55} k_t}\right) \geq 2\varepsilon \frac{12}{\varepsilon^2} \log\left(\frac{\varepsilon N}{C_{55} \frac{12}{\varepsilon^2}}\right), \\
&= \frac{24}{\varepsilon} \log\left(\frac{\varepsilon^3 N}{12 C_{55}}\right) \geq \log\left(\frac{N^4}{C_{56}}\right),
\end{aligned}$$

when $\left(\varepsilon^3 N/12C_{55}\right)^{24/\varepsilon} \geq N^4/C_{56}$, i.e., $N \geq (12C_{55}/\varepsilon^3)^6 \cdot C_{56}^{-\varepsilon/6}$. The polynomial rate can therefore balance the union bound over $j \in \{0, 1, \cdots, t-1\}$, which is a logarithmic rate with respect to $N$. With probability at least $1 - C_{57}/N^3$, for all $j \in \{0, 1, \cdots, t-1\}$,

$$\sup\left(|W_{\boldsymbol{u},I^c}|^*_{\lfloor k_{j+1}/16\rfloor} : \boldsymbol{u} \in V_j\right) < (32eB)^{1/p}\mathcal{M}_N\left(\frac{N}{k_{j+1}}\right)^{p^{-1}(1+256\varepsilon)}.$$

Finally, notice that

$$\sum_{j=0}^{t-1}\sqrt{k_j}\sup\left(|W_{\boldsymbol{u},I^c}|^*_{\lfloor k_{j+1}/16\rfloor} : \boldsymbol{u} \in V_j\right) \leq (32eB)^{1/p}\mathcal{M}_N\sqrt{N}\sum_{j=0}^{t-1}2^{j/p+256\varepsilon j/p-j/2}.$$

There exists an absolute constant $C_{58}$ such that $\sum_{j=0}^{t-1}2^{\left(\frac{1+256\varepsilon}{p}-\frac{1}{2}\right)j} \leq C_{58}$ for $1/2 > (1+256\varepsilon)/p$. We finish the proof of Proposition 40 by applying Proposition 39. ∎

**Case [1]: when $\mathrm{Tr}(\Sigma)$ is dominating.**

In this case, we apply all the aforementioned results (Proposition 40, Proposition 39 and Lemma 29) to $\mathcal{C} = [N]$.

Let us now apply Proposition 40,

$$\mathbb{E}\left|\left\{I \subset [N] : g(N,I) > C_{58}\log_2(N)\max\left(\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in [N]\right) + C_{58}B^{1/p}\sqrt{N}\mathcal{M}_N\right\}\right| \leq \frac{2^N}{N^3},$$

so with probability at least $1 - 1/N^2$, there are at most $2^N/N$ subsets $I \subset [N]$, such that

$$g(N,I) > C_{58}\log_2(N)\max\left(\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in [N]\right) + C_{58}B^{1/p}\sqrt{N}\mathcal{M}_N. \tag{5.25}$$

For these $2^N/N$ "spiky" subsets, we simply use a deterministic argument: for all $I \subset [N]$, we have

$$g(N,I) \leq N\max\left(\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in [N]\right).$$

As a consequence, if we denote by $\mathcal{I}$ the set of all subsets $I \subset [N]$ satisfying (5.25), with probability at least $1 - 1/N^2$,

$$\begin{aligned}\sum_{I\subset[N]} g(N,I) &= \sum_{I\in\mathcal{I}} g(N,I) + \sum_{I\notin\mathcal{I}} g(N,I)\\ &\leq 2C_{58}2^N\log_2(N)\max\left(\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in [N]\right) + 2^N C_{58}B^{1/p}\sqrt{N}\mathcal{M}_N,\end{aligned} \tag{5.26}$$

We are left with an upper bound that has a high probability of $\max\left(\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in [N]\right)$. We emphasize again that we do not use the sample coloring technique developed by [Tik18], but instead, we use the strong concentration of $\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right|$ to absorb this $\log N$ factor, because of (1.28).

**Upper bound for $\max\left(\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in [N]\right)$** Let $p = 2 + \epsilon$.

For any $i \neq j \in [N]$,

$$\begin{aligned}&\mathbb{E}\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right|^{2+\epsilon} = \mathbb{E}\left[\mathbb{E}\left[\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right|^p\middle|X_i\right]\right] \leq \kappa^p\left(\mathbb{E}\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right|^2\right)^{p/2}\\ &= \kappa^p\left(\mathbb{E}K_{k+1:\infty}(X_i,X_j)\right)^{p/2} = \kappa^p\left(\mathrm{Tr}\left(\Sigma^2\right)\right)^{p/2},\end{aligned}$$

where we used (1.27) to obtain the inequality. By union bound, for any $\tau > 0$,

$$\begin{aligned}\mathbb{P}\left(\max\left(\left|\langle\phi(X_i),\phi(X_j)\rangle\right|_{\mathcal{H}} : i \neq j \in [N]\right) > \tau\right) &\leq N^2\frac{\mathbb{E}\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right|^p}{\tau^p}\\ &\leq N^2\frac{\kappa^p\left(\mathrm{Tr}\left(\Sigma^2\right)\right)^{p/2}}{\tau^p}.\end{aligned}$$

Let $\tau = \bar{\delta}\operatorname{Tr}(\Sigma)/\log N$, with probability at least

$$1 - N^2 \frac{\kappa^p \left(\operatorname{Tr}\left(\Sigma^2\right)\right)^{p/2}}{\left(\frac{\bar{\delta}\operatorname{Tr}(\Sigma)}{\log N}\right)^p} = 1 - \left(\frac{\kappa}{\bar{\delta}}\right)^p (\log N)^p N^2 \left(\frac{\sqrt{\operatorname{Tr}\left(\Sigma^2\right)}}{\operatorname{Tr}(\Sigma)}\right)^p =: 1 - \bar{p},$$

we have

$$\max\left(\left|\langle \phi(X_i), \phi(X_j)\rangle\right|_{\mathcal{H}} : i \neq j \in [N]\right) \leq \frac{\bar{\delta}\operatorname{Tr}(\Sigma)}{\log N},$$

where

$$\bar{p} = \left(\frac{\kappa}{\bar{\delta}}\right)^{2+\epsilon} (\log N)^{2+\epsilon} N^{1-\frac{\epsilon}{2}} \left(\frac{\sqrt{N\operatorname{Tr}\left(\Sigma^2\right)}}{\operatorname{Tr}(\Sigma)}\right)^{2+\epsilon}.$$

Together with (5.26), with probability at least $1 - \bar{p} - N^{-2}$,

$$\sum_{I\subset[N]} g(N, I) \leq 2C_{58}2^N \bar{\delta}\operatorname{Tr}(\Sigma) + 2^N C_{58} B^{1/(2+\epsilon)}\sqrt{N}\mathcal{M}_N.$$

By (5.23), we obtain that

$$4(\ell^*)^2 \sup\left(\left|\mathbb{E}_\eta V_{I_\eta,\boldsymbol{u}}\right| : \boldsymbol{u} \in S_2^{N-1}\right) \lesssim \bar{\delta}\operatorname{Tr}(\Sigma) + \sqrt{N}B^{1/(2+\epsilon)}\mathcal{M}_N. \tag{5.27}$$

Combining Eq.(5.22), Eq.(5.27) and $B \leq \kappa^{2+\epsilon}\|\Sigma\|_{\mathrm{op}}^{\frac{2+\epsilon}{2}}$, there exists an absolute constant $C_{59} > 1$ such that with probability at least $1 - \gamma - \bar{p} - N^{-2}$,

$$\begin{aligned}
&\sup\left(\left|\mathbb{E}_\eta V_{I_\eta,\boldsymbol{u}}\right| : \boldsymbol{u} \in S^N_{[N],N}\right) \leq C_{59}\bar{\delta} \vee C_{59}\kappa^2 \frac{N\|\Sigma\|_{\mathrm{op}}}{\operatorname{Tr}(\Sigma)} \vee C_{59}\kappa\sqrt{\frac{N\|\Sigma\|_{\mathrm{op}}}{\operatorname{Tr}(\Sigma)}} \cdot \frac{\max\left(\|\phi(X_i)\|_{\mathcal{H}} : i \in [N]\right)}{\sqrt{\operatorname{Tr}(\Sigma)}} \\
&\vee C_{59}\kappa\sqrt{\bar{\delta}}\sqrt{\frac{N\|\Sigma\|_{\mathrm{op}}}{\operatorname{Tr}(\Sigma)}} \leq C_{59}\bar{\delta}\left(1 + \kappa^2\kappa_{DM}\frac{\operatorname{Tr}(\Sigma)+\lambda}{\operatorname{Tr}(\Sigma)} + \kappa\sqrt{\kappa_{DM}\frac{\operatorname{Tr}(\Sigma)+\lambda}{\operatorname{Tr}(\Sigma)}}\left(1 + \delta + \sqrt{\bar{\delta}}\right)\right),
\end{aligned} \tag{5.28}$$

where we used that $N \leq \kappa_{DM}\bar{\delta}^2 d^*_\lambda\left(\Sigma^{-1/2}_{k+1:\infty}B_{\mathcal{H}}\right)$, $\bar{\delta} < 1$ and (1.26) from Assumption 2.

Note that $S_2^{N-1} = S^N_{[N],N}$, and we plug (5.28) into (5.17) and (5.19) and take $\epsilon_0 = \sqrt{\delta}$ in (5.19),

$$\begin{aligned}
\Psi^2 &\leq \delta + 4C_{59}\bar{\delta}\left(1 + \kappa^2\kappa_{DM}\frac{\operatorname{Tr}(\Sigma)+\lambda}{\operatorname{Tr}(\Sigma)} + \kappa\sqrt{\kappa_{DM}\frac{\operatorname{Tr}(\Sigma)+\lambda}{\operatorname{Tr}(\Sigma)}}\left(1 + \delta + \sqrt{\bar{\delta}}\right)\right), \\
\Phi^2 &\leq \delta\left(1 + \delta + 4C_{59}\bar{\delta}\left(1 + \kappa^2\kappa_{DM}\frac{\operatorname{Tr}(\Sigma)+\lambda}{\operatorname{Tr}(\Sigma)} + \kappa\sqrt{\kappa_{DM}\frac{\operatorname{Tr}(\Sigma)+\lambda}{\operatorname{Tr}(\Sigma)}}\left(1 + \delta + \sqrt{\bar{\delta}}\right)\right)\right)
\end{aligned}$$

Since we have the right to choose sufficiently small $\bar{\delta}$ and $\kappa_{DM}$ as long as (1.28) holds, we can set

$$\kappa_{DM} \leq \left(\frac{1}{12C_{59}\kappa}\right)^2 < \frac{1}{4C_{59}\kappa^2}. \tag{5.29}$$

Because $\delta, \bar{\delta} < 1$,

$$\begin{aligned}
\Psi^2 &< \delta + 4C_{59}\bar{\delta} + 2\bar{\delta}\frac{\operatorname{Tr}(\Sigma)+\lambda}{\operatorname{Tr}(\Sigma)} \leq \delta + (4C_{59}+2)\bar{\delta}\frac{\operatorname{Tr}(\Sigma)+\lambda}{\operatorname{Tr}(\Sigma)}, \\
\Phi^2 &< 4\left(\delta + \delta + \delta\left(4C_{59}+2\right)\bar{\delta}\frac{\operatorname{Tr}(\Sigma)+\lambda}{\operatorname{Tr}(\Sigma)}\right).
\end{aligned}$$

Recall that in this subsection, we assume that there exists an absolute constant $C_3$ such that $\lambda \le C_3 \operatorname{Tr}(\Sigma)$. In this case, there exist absolute constants $C_2$, $C_4$, $C_5$, $C_6$ and $C_{60}$ such that

$$\Phi^2 + \Psi^2 + 2\Phi\sqrt{\Psi^2+1} < C_2\delta^2 + C_4\bar{\delta}^2 + 4\sqrt{\left(3\delta + C_5\bar{\delta}\right)\left(1+\delta+C_6\bar{\delta}\right)} =: \tilde{\delta} < 1,$$

provided that $\delta < 1/(100\sqrt{C_2})$ and $\bar{\delta} < 1/C_{60}$ (thus we can take $C_1$ in Assumption 2 as $C_{60}$), and where

$$\tilde{\delta} = C_2\delta^2 + C_4\bar{\delta}^2 + 4\sqrt{\left(3\delta + C_5\bar{\delta}\right)\left(1+\delta+C_6\bar{\delta}\right)}.$$

This proves that with probability at least $1-\gamma-\bar{p}-N^{-2}$, for all $\boldsymbol{\lambda} \in \mathbb{R}^N$,

$$\left(1-\tilde{\delta}\right)\ell^*\left\|\boldsymbol{\lambda}\right\|_2 \le \left\|\mathbb{X}_{\phi,k+1:\infty}^\top \boldsymbol{\lambda}\right\|_{\mathcal{H}} \le \left(1+\tilde{\delta}\right)\ell^*\left\|\boldsymbol{\lambda}\right\|_2,$$

provided that $N \le \kappa_{DM}\bar{\delta}^2 d_\lambda^*\left(\Sigma_{k+1:\infty}^{-1/2} B_{\mathcal{H}}\right)$.

**Case [2]: when $\lambda$ is dominating.**

When $\lambda > C_3 \operatorname{Tr}(\Sigma)$. In this case, we only make use of the fact that $\mathbb{X}_\phi \mathbb{X}_\phi^\top$ is of rank-$N$ and is positive semi-definite, hence $\sigma_N\left(\mathbb{X}_\phi\mathbb{X}_\phi^\top + \lambda I_N\right) \ge \lambda + \sigma_N\left(\mathbb{X}_\phi\mathbb{X}_\phi^\top\right) \ge \lambda \ge c_2\lambda + (1-c_2)C_3\operatorname{Tr}(\Sigma)$, where we recall that $0<c_2<1$ is some absolute constant. Hence our objective is to prove that there exists an absolute constant $C_{61}$ such that with high probability, we have $\left\|\mathbb{X}_\phi\mathbb{X}_\phi^\top + \lambda I_N\right\|_{\mathrm{op}} \le C_{61}\lambda$. We prove this by proving that there exists an absolute constant $C_{62}$ such that $C_{62}^2 \le C_{61}-1$, and with high probability we have $\left\|\mathbb{X}_\phi^\top\right\|_{\mathrm{op}} \le C_{62}\sqrt{\lambda}$.

Let $\{\mathcal{C}_m\}_{m\le\chi}$ for some $\chi \in \mathbb{N}_+$ be a partition of $[N]$, by Jensen's inequality, for any $\boldsymbol{\lambda} \in S_2^{N-1}$,

$$\left\|\mathbb{X}^\top\boldsymbol{\lambda}\right\|_{\mathcal{H}}^2 \le \chi\sum_{m=1}^{\chi}\left\|\sum_{i\in\mathcal{C}_m}\lambda_i\phi(X_i)\right\|_{\mathcal{H}}^2.$$

Applying (5.18) and (5.23) but with $A$ replaced by its restriction onto $\mathcal{C}_m$ for each $m \le \chi$, we obtain that for any $\boldsymbol{\lambda} \in S_2^{N-1}$,

$$\frac{\left\|\mathbb{X}^\top\boldsymbol{\lambda}\right\|_{\mathcal{H}}^2}{(\ell^*)^2} \le \chi^2 \max_{i\in[N]}\frac{\|\phi(X_i)\|_{\mathcal{H}}^2}{(\ell^*)^2} + \frac{\chi}{(\ell^*)^2}\sum_{m=1}^{\chi}\frac{4}{2^N}\sum_{I\subset[N]} g(N,\mathcal{C}_m,I). \tag{5.30}$$

At this time, we can make use of the sample coloring technique in [Tik18]. It is a technique used to truncate the inner products $\left(\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right)_{i\neq j\in\mathcal{C}}$. Given i.i.d. random vectors $(\phi(X_i))_{i\le N}$ and $H>0$, there exists an undirected graph $\mathcal{G}_H$ whose vertex set is $[N]$, and its edge set is:

$$\left\{(i,j):\, 1\le i<j\le N,\, \left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| > H\max\left(\|\phi(X_h)\|_{\mathcal{H}} :\, h\le N\right)\right\}.$$

The coloring of $\mathcal{G}_H$ is an assignment of "colors" to all vertices such that no adjacent vertices share the same color. The smallest possible number of colors sufficient to assign such a coloring is called the chromatic number of $\mathcal{G}_H$, denoted as $\chi(\mathcal{G}_H)$, and the collection $\{\mathcal{C}_m^H\}_{m\le\chi(\mathcal{G}_H)}$ is the associated partition by colors of $[N]$. That is to say, for any $m \le \chi(\mathcal{G}_H)$, and $i\neq j\in\mathcal{C}_m^H$, the vertices $i,j$ are not adjacent, so $\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| \le H\max\left(\|\phi(X_h)\|_{\mathcal{H}} :\, h\le N\right)$. Since $(\phi(X_i))_{i\le N}$ are random, $\mathcal{G}_H$ is a random graph, and the following lemma is a high probability estimate of $\chi(\mathcal{G}_H)$. The following lemma is a weaker version of [Tik18, Proposition 10], which is sufficient for our purpose.

**Lemma 31.** *Assume that for some $p>2$ we have $\sup\left(\mathbb{E}\left|\langle\phi(X),f\rangle_{\mathcal{H}}\right|^p :\, \|f\|_{\mathcal{H}}=1\right) = B$. Then for any $H>0$ and any integer $m>1$, the chromatic number of $\mathcal{G}_H$ satisfies $\chi(\mathcal{G}_H) \le m$ with probability at least $1-(BNH^{-p})^{m-1}N$.*

**Proof.** Let us introduce an auxiliary random process $(Y_i)_{i\in[N]}$ with values in $\mathbb{N}$, where $Y_1 := 1$ as a constant, and for all $i=2,\cdots,N$,

$$Y_i := \min\left(r\in\mathbb{N}_+ :\, \forall j<i,\, j\in\mathbb{N}_+,\ \text{with } Y_j = r,\ \text{we have } \left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| \le H\|\phi(X_j)\|_{\mathcal{H}}\right).$$

The process $(Y_i)_{i\in[N]}$ is "classifying" each $\phi(X_i)$ is if $|\langle\phi(X_1),\phi(X_2)\rangle_{\mathcal{H}}| > H\|\phi(X_1)\|_{\mathcal{H}}$, $|\langle\phi(X_1),\phi(X_3)\rangle_{\mathcal{H}}| \le H\|\phi(X_1)\|_{\mathcal{H}}$, $|\langle\phi(X_2),\phi(X_3)\rangle_{\mathcal{H}}| \le H\|\phi(X_2)\|_{\mathcal{H}}$, then $Y_2 = 2$(because $(1,2)$ is adjacent in $\mathcal{G}_H$), $Y_3 = 1$, because either $(1,3)$ or $(2,3)$ is not adjacent in $\mathcal{G}_H$. Such a $\mathcal{G}_H$ has chromatic number 2.

By the definition of $Y_i$, we have that any two numbers $i \neq j \in [N]$ such that $Y_i = Y_j$ are not adjacent in $\mathcal{G}_H$, and $Y_i = Y_j$ is a sufficient but not necessary condition for adjacency of $(i,j)$. In particular, $\chi(\mathcal{G}_H) \le \max(Y_i : i \in [N])$. Next for each $i > 1$ and $m \ge 1$, we have

$$\begin{aligned}\mathbb{P}(Y_i = m+1) &\le \mathbb{P}\left(\exists j \le i-1 \text{ s.t. } \left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| > H\|\phi(X_j)\|_{\mathcal{H}}, \text{ and } Y_j = m\right)\\ &\le \sum_{j=1}^{i-1}\mathbb{P}\left(\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| > H\|\phi(X_j)\|_{\mathcal{H}}, \text{ and } Y_j = m\right).\end{aligned}$$

For all $j = 0,\cdots,i-1$, since $Y_j$ is $\sigma(X_1,\cdots,X_j)$-measurable, it is independent of $X_i$, hence

$$\begin{aligned}&\mathbb{P}\left(Y_j = m, \left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| > H\|\phi(X_j)\|_{\mathcal{H}} \,\middle|\, (X_\ell)_{\ell=1}^{i-1}\right)\\ &= \mathbb{E}\left[\mathbb{1}_{\{Y_j=m\}}\mathbb{1}_{\left\{\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| > H\|\phi(X_j)\|_{\mathcal{H}}\right\}} \,\middle|\, (X_\ell)_{\ell=1}^{i-1}\right]\\ &= \mathbb{1}_{\{Y_j=m\}}\mathbb{P}\left(\left\{\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| > H\|\phi(X_j)\|_{\mathcal{H}}\right\} \,\middle|\, (X_\ell)_{\ell=1}^{i-1}\right)\\ &\le \mathbb{1}_{\{Y_j=m\}}\frac{\mathbb{E}\left[\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right|^p \,\middle|\, (X_\ell)_{\ell=1}^{i-1}\right]}{H^p\|\phi(X_j)\|_{\mathcal{H}}^p} \le \mathbb{1}_{\{Y_j=m\}}\frac{B}{H^p},\end{aligned}$$

where we used Markov's inequality to obtain the first inequality. Hence

$$\mathbb{P}\left(\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| > H\|\phi(X_j)\|_{\mathcal{H}}, \text{ and } Y_j = m\right) \le \mathbb{P}(Y_j = m)\frac{B}{H^p}.$$

Further, by $\mathbb{E}|\{j \le N : Y_j = m\}| = \sum_{j\le N}\mathbb{E}\mathbb{1}_{\{Y_j=m\}} = \sum_{j\le N}\mathbb{P}(Y_j = m) \ge \sum_{j\le i-1}\mathbb{P}(Y_j = m)$,

$$\sum_{j=1}^{i-1}\mathbb{P}\left(\left|\langle\phi(X_i),\phi(X_j)\rangle_{\mathcal{H}}\right| > H\|\phi(X_j)\|_{\mathcal{H}}, \text{ and } Y_j = m\right) \le BH^{-p}\mathbb{E}|\{j \le N : Y_j = m\}| \tag{5.31}$$

It follows from Equation (5.31) that

$$\sum_{i=1}^{N}\mathbb{P}(Y_i = m+1) = \mathbb{E}|\{j \le N : Y_j = m+1\}| \le BNH^{-p}\mathbb{E}|\{j \le N : Y_j = m\}|.$$

We next deal with $\mathbb{E}|\{j \le N : Y_j = 2\}|$. We simply upper bound $\mathbb{E}|\{j \le N : Y_j = 2\}|$ by $N$. Therefore,

$$\mathbb{E}|\{j \le N : Y_j = m+1\}| \le (BNH^{-p})^{m-1}N.$$

Note that the set of values $\{Y_j : j \le N\}$ is an interval in $\mathbb{N}$, hence

$$\mathbb{P}(\chi(\mathcal{G}_H) \ge m+1) \le \mathbb{P}(\exists j \le N : Y_j = m+1) \le \mathbb{E}|\{j \le N : Y_j = m+1\}| \le (BNH^{-p})^{m-1}N.$$

■

Combining Proposition 40 and the sample coloring technique from Lemma 31, we obtain an upper bound for $\mathbb{E}_\eta V_{I_\eta,\boldsymbol{u},\boldsymbol{v}}$ uniformly over all $\boldsymbol{u},\boldsymbol{v} \in S^N_{[N],N} = S_2^{N-1}$. We state this result in the following Proposition.

**Proposition 41.** *There are absolute constants $C_{63}$ and $C_{64}$ depending only on $p$, such that the following holds. If $N \ge C_{63}$, then for any $\lambda > -\operatorname{Tr}(\Sigma)$, with probability at least $1 - \bar{p} - N^{-2}$, where*

$$\bar{p} := N\left(\left(\frac{4\kappa^2\log^2(N)\|\Sigma\|_{op}}{\operatorname{Tr}(\Sigma)+\lambda}\right)^{p/2}N\right)^{\lceil(8+2p)/(p-2)\rceil-1}, \tag{5.32}$$

$$\mathcal{M}_N \le C_{65}\sqrt{\operatorname{Tr}(\Sigma)+\lambda} + C_{65}\sqrt{N}B^{1/p}. \tag{5.33}$$

**Proof.** Let $H > 0$ which will be determined later, and let $0 \le m \le \chi$, we apply Proposition 40 to $\mathcal{C} = \mathcal{C}_m^H$:

$$\mathbb{E}\left|\left\{I \subset [N] : g(N, \mathcal{C}_m^H, I) > C_{58}\log_2(N)\max\left(\left|\langle\phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}_m^H\right) + C_{58}B^{1/p}\sqrt{N}\mathcal{M}_N\right\}\right| \le \frac{2^N}{N^3},$$

then with probability at least $1 - 1/N^2$, there are at most $2^N/N$ subsets $I \subset [N]$, such that

$$g(N, \mathcal{C}_m^H, I) > C_{58}\log_2(N)\max\left(\left|\langle\phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}_m^H\right) + C_{58}B^{1/p}\sqrt{N}\mathcal{M}_N. \tag{5.34}$$

For these $2^N/N$ "spiky" subsets, we simply use a deterministic argument: for all $I \subset [N]$, we have

$$g(N, \mathcal{C}_m^H, I) \le N\max\left(\left|\langle\phi(X_i), \phi(X_j)\rangle_{\mathcal{H}}\right| : i \neq j \in \mathcal{C}_m^H\right) \le NH\max\left(\|\phi(X_i)\|_{\mathcal{H}} : i \le N\right),$$

where we use that $i, j \in \mathcal{C}_m^H$ have the same color and therefore are not adjacent in $\mathcal{G}_m^H$. As a consequence, if we denote by $\mathcal{I}$ the set of all subsets $I \subset [N]$ satisfying Equation (5.34), with probability at least $1 - 1/N^2$, for any $(\mathcal{C}_m^H)_{m\le\chi}$, we have

$$\begin{aligned}\sum_{I\subset[N]} g(N, \mathcal{C}_m^H, I) &= \sum_{I\in\mathcal{I}} g(N, \mathcal{C}_m^H, I) + \sum_{I\notin\mathcal{I}} g(N, \mathcal{C}_m^H, I)\\ &\le C_{58}2^N\log_2(N)H\max\left(\|\phi(X_i)\|_{\mathcal{H}} : i \le N\right) + 2^N C_{58}B^{1/p}\sqrt{N}\mathcal{M}_N,\end{aligned}$$

where we again used that $i$ and $j$ are not adjacent in $\mathcal{G}_m^H$.

Let $H = \bar{\delta}\sqrt{\mathrm{Tr}(\Sigma) + \lambda}/\log N$ with $\bar{\delta} = 1/2$ (unlike the case in Section 5.2, we only need an *isomorphic* upper bound, we can choose $\bar{\delta}$ to be an arbitrary absolute constant), then

$$BNH^{-p} \le \left(\frac{4\kappa^2\log^2(N)\|\Sigma\|_{op}}{(\mathrm{Tr}(\Sigma) + \lambda)}\right)^{p/2} N.$$

Let $\chi = \lceil(8 + 2p)/(p - 2)\rceil$ and apply Lemma 31 for $m = \chi$, with probability at least $1 - \bar{p}$, $\chi(\mathcal{G}_H) \le \chi$.

On the other hand, by (5.30) and the fact that (5.34) is valid uniformly over all $(\mathcal{C}_m^H)_{m\le\chi}$ (thanks to Proposition 40), there exists an absolute constant $C_{64}$ such that with probability $1 - \bar{p} - N^{-2}$, for all $\boldsymbol{\lambda} \in S_2^{N-1}$, we have that

$$\begin{aligned}\frac{\left\|\mathbb{X}^\top\boldsymbol{\lambda}\right\|_{\mathcal{H}}^2}{(\ell^*)^2} &\le \left(\frac{8 + 2p}{p - 2}\right)^2 \max_{i\in[N]}\frac{\|\phi(X_i)\|_{\mathcal{H}}^2}{(\ell^*)^2}\\ &\quad + \frac{\left(\frac{8+2p}{p-2}\right)^2}{(\ell^*)^2}\left(C_{64}\log_2(N)H\max\left(\|\phi(X_i)\|_{\mathcal{H}} : i \le N\right) + C_{64}B^{1/p}\sqrt{N}\mathcal{M}_N\right)\\ &= \left(\frac{8 + 2p}{p - 2}\right)^2 \max_{i\in[N]}\frac{\|\phi(X_i)\|_{\mathcal{H}}^2}{(\ell^*)^2}\\ &\quad + \frac{\left(\frac{8+2p}{p-2}\right)^2}{(\ell^*)^2}\left(C_{64}\bar{\delta}\sqrt{\mathrm{Tr}(\Sigma) + \lambda}\max\left(\|\phi(X_i)\|_{\mathcal{H}} : i \le N\right) + C_{64}B^{1/p}\sqrt{N}\mathcal{M}_N\right).\end{aligned} \tag{5.35}$$

Solving (5.35) gives that there exists an absolute constant $C_{65}$ depending only on $p$ such that with probability at least $1 - \bar{p} - N^{-2}$,

$$\mathcal{M}_N \le C_{65}\sqrt{\mathrm{Tr}\left(\Sigma\right) + \lambda} + C_{65}\sqrt{N}B^{1/p}.$$

■

Recall that we have assumed that $C_3\,\mathrm{Tr}\left(\Sigma\right) < \lambda$ in this case, and $\mathrm{Tr}\left(\Sigma\right) + \lambda \ge (\kappa_{DM}/4)^{-2}N\|\Sigma\|_{\mathrm{op}}$. Moreover, since we have $B \le \kappa^p\|\Sigma\|_{\mathrm{op}}^{p/2}$ for any $2 < p \le 2 + \epsilon$, we have

$$\begin{aligned}\mathcal{M}_N = \sup\left(\left\|\mathbb{X}^\top\boldsymbol{\lambda}\right\|_{\mathcal{H}} : \boldsymbol{\lambda} \in S_2^{N-1}\right) &\le C_{65}\sqrt{1 + C_3^{-1}\lambda} + C_{65}\kappa\sqrt{N\|\Sigma\|_{\mathrm{op}}}\\ &\le C_{65}\sqrt{1 + C_3^{-1}\lambda} + \frac{C_{65}}{4}\kappa\kappa_{DM}\sqrt{\mathrm{Tr}\left(\Sigma\right) + \lambda}\\ &\le \left(C_{65}\sqrt{1 + C_3^{-1}} + \frac{C_{65}}{4}\kappa\kappa_{DM}\sqrt{1 + C_3^{-1}}\right)\sqrt{\lambda}.\end{aligned}$$

Letting $C_{62} = C_{65}\sqrt{1 + C_3^{-1}} + C_{65}\kappa\kappa_{DM}\sqrt{1 + C_3^{-1}}/4$, this is precisely our initial objective. As a result, we may let $C_{61} = C_{62}^2 + 1$, and $C_7 = C_{61}$.

# Chapter 6

# Décomposition de l'Espace des Caractéristiques

*L'étude approfondie de la nature est la source la plus fertile des découvertes mathématiques.*

— *Joseph Fourier,* Théorie analytique de la chaleur, *Ch. 1, p. 7 (1822)*

Dans cette section, nous introduisons la principale contribution méthodologique de cette thèse : la méthode de Décomposition de l'Espace des Caractéristiques (FSD, pour *Feature Space Decomposition*). La méthode FSD a été développée dans une série de travaux par [P4, P2, P3, P1]. La méthode de Décomposition de l'Espace des Caractéristiques est avant tout un outil conçu pour aider les théoriciens à analyser l'excès de risque en population ; simultanément, elle pourrait également servir de nouveau cadre théorique potentiel pour la théorie de l'apprentissage statistique et la statistique mathématique, offrant aux théoriciens une perspective nouvelle pour comprendre les propriétés statistiques d'un estimateur.

Dans la Section 1.5.1, nous présentons le cadre de base de la méthode FSD pour les problèmes de régression supervisée à valeurs réelles et les problèmes de classification binaire. Dans la Section 1.5.2 et la Section 1.5.3, nous discutons respectivement des rôles des deux sous-espaces produits par la méthode FSD, et nous les illustrons par des exemples issus de divers problèmes d'apprentissage supervisé. Enfin, dans la Section 1.5.4, nous montrons comment la méthode FSD peut servir de nouveau cadre théorique potentiel. Tout au long de cette section, nous supposons toujours que $\mathcal{F}$ est un espace vectoriel, ou du moins qu'il peut être plongé dans un espace vectoriel. Suivant la tradition de la théorie de l'apprentissage statistique, nous désignerons alors $\mathcal{F}$ comme l'espace des caractéristiques (*feature space*), [VC68].

## 6.1 La méthode de Décomposition de l'Espace des Caractéristiques

Dans cette section, nous présentons la méthode FSD adaptée aux problèmes de régression et de classification supervisées. Nous commençons par les problèmes de régression supervisée à valeurs réelles.

**Problème de régression supervisée à valeurs réelles.** Nous rappelons de la Section 1.2 que l'objectif d'un théoricien est le suivant : étant donné un problème de régression supervisée à valeurs réelles $(\mu_X, f^*, \xi)$ et l'une de ses solutions $(\mathcal{F}, \hat{f}_N)$, il s'agit de caractériser l'erreur d'estimation $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$.

Pour l'erreur d'estimation, il existe deux manières fondamentalement différentes de la majorer :

1. Obtenir une borne supérieure pour $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$ via des annulations entre $\hat{f}_N$ et $f^*_{\mathcal{F}}$, c'est-à-dire en montrant que $\hat{f}_N$ et $f^*_{\mathcal{F}}$ sont proches sous la métrique $L^2(\mu_X)$ ;

2. Utiliser la petitesse de $\|\hat{f}_N\|^2_{L^2(\mu_X)}$ et de $\|f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$, c'est-à-dire appliquer l'inégalité triangulaire pour obtenir $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)} \leq 2(\|\hat{f}_N\|^2_{L^2(\mu_X)} + \|f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)})$.

Pour les problèmes de régression supervisée à valeurs réelles, la méthode FSD peut être vue formellement comme une interpolation entre ces deux approches. Pour le voir, nous définissons tout d'abord la FSD.

**Définition 8.** *Toute décomposition en somme directe $\mathcal{F} = V_J \oplus V_{J^c}$ de $\mathcal{F}$ est appelée une Décomposition de l'Espace des Caractéristiques (FSD) de $\mathcal{F}$. Notons par $P_{V_J}$ l'opérateur de projection sur le sous-espace vectoriel $V_J$, et par $P_{V_{J^c}}$ la projection sur $V_{J^c}$ ; de manière équivalente, l'opérateur identité $I_{\mathcal{F}} = P_{V_J} + P_{V_{J^c}}$ sur l'espace des caractéristiques $\mathcal{F}$ est décomposé. En particulier, si une FSD satisfait que $V_J$ et $V_{J^c}$ sont orthogonaux par rapport au produit scalaire de $L^2(\mu_X)$, nous l'appelons une FSD orthogonale, et nous la notons $\mathcal{F} = V_J \oplus^{\perp} V_{J^c}$.*

Pour tout $f \in \mathcal{F}$, nous écrivons $f_J = P_{V_J} f$ et $f_{J^c} = P_{V_{J^c}} f$. Nous abrégeons $P_{V_J} \hat{f}_N$ par $\hat{f}_J$, $P_{V_{J^c}} \hat{f}_N$ par $\hat{f}_{J^c}$, $P_{V_J} f^*_{\mathcal{F}}$ par $f^*_J$, et $P_{V_{J^c}} f^*_{\mathcal{F}}$ par $f^*_{J^c}$. Notez que nous ne confondrons pas $f^*_{\mathcal{F}}$ avec $f^*$, car nous pouvons toujours incorporer l'erreur d'approximation dans le bruit ; voir le Lemme 3. Étant donné une FSD quelconque $\mathcal{F} = V_J \oplus V_{J^c}$, l'erreur d'estimation admet la décomposition suivante :

$$\left\|\hat{f}_N - f^*_{\mathcal{F}}\right\|^2_{L^2(\mu_X)} \begin{cases} = \left\|\hat{f}_J - f^*_J\right\|^2_{L^2(\mu_X)} + \left\|\hat{f}_{J^c} - f^*_{J^c}\right\|^2_{L^2(\mu_X)}, & \text{si } V_J \perp V_{J^c} \text{ dans } L^2(\mu_X), \\ \leq 2\left\|\hat{f}_J - f^*_J\right\|^2_{L^2(\mu_X)} + 2\left\|\hat{f}_{J^c} - f^*_{J^c}\right\|^2_{L^2(\mu_X)}, & \text{sinon.} \end{cases} \tag{1.10}$$

L'interpolation entre le point *1* et le point *2* peut s'exprimer sous la forme de l'inégalité suivante :

$$\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)} \leq \min\left(2\|\hat{f}_J - f^*_J\|^2_{L^2(\mu_X)} + 4\|\hat{f}_{J^c}\|^2_{L^2(\mu_X)} + 4\|f^*_{J^c}\|^2_{L^2(\mu_X)} : \mathcal{F} = V_J \oplus V_{J^c}\right). \tag{1.11}$$

La méthode FSD consiste à rechercher des fonctions à valeurs réelles $r : (V_J, V_{J^c}) \mapsto r(V_J, V_{J^c}) \in \mathbb{R}_+$ et $\delta : (V_J, V_{J^c}) \mapsto \delta(V_J, V_{J^c}) \in [0, 1]$, telles que pour chaque (ou du moins pour une certaine) FSD, l'inégalité suivante soit vérifiée avec une probabilité d'au moins $1 - \delta(V_J, V_{J^c})$ (ou en espérance, si l'on souhaite une borne supérieure sur l'erreur d'estimation en espérance),

$$2\|\hat{f}_J - f^*_J\|^2_{L^2(\mu_X)} + 4\|\hat{f}_{J^c}\|^2_{L^2(\mu_X)} + 4\|f^*_{J^c}\|^2_{L^2(\mu_X)} \leq r^2(V_J, V_{J^c}). \tag{1.12}$$

Nous appelons un tel $r$ la fonction de taux de $(\mu_X, f^*, \xi)$ et de $(\mathcal{F}, \hat{f}_N)$. Ici, dire que nous cherchons une fonction de taux signifie chercher une fonction qui soit aussi petite que possible ; sinon, on pourrait trivialement prendre $r(V_J, V_{J^c}) = \infty$.

En tant que stratégie de preuve mathématique, l'idée centrale de la méthode FSD repose sur la conviction suivante :

1. Sur le sous-espace $V_J$, appelé sous-espace d'estimation, la statistique classique opère, c'est-à-dire que $\hat{f}_N$ estime $f^*_{\mathcal{F}}$ sur $V_J$ ; par conséquent, la distance entre $\hat{f}_J$ et $f^*_J$ sous la métrique $L^2(\mu_X)$ est petite, contribuant à l'erreur d'estimation via des annulations $\|\hat{f}_J - f^*_J\|^2_{L^2(\mu_X)}$.

2. D'un autre côté, nous pensons que $\hat{f}_N$ sur $V_{J^c}$ n'estime pas $f^*_{\mathcal{F}}$. Par conséquent, nous appelons $V_{J^c}$ le sous-espace libre. Sur ce sous-espace, $\hat{f}_{J^c}$ accomplit certaines tâches déterminées par la définition de $\hat{f}_N$, mais en général pas d'estimation ; par conséquent, nous nous attendons à ce que la distance entre $\hat{f}_{J^c}$ et $f^*_{J^c}$ sous la métrique $L^2(\mu_X)$ ne soit pas nécessairement petite par rapport à la somme de leurs normes $L^2(\mu_X)$, de sorte que l'application de l'inégalité triangulaire ne conduit pas nécessairement à une surestimation de $\|\hat{f}_{J^c} - f^*_{J^c}\|^2_{L^2(\mu_X)}$. Dans ce cas, l'erreur d'estimation reçoit des contributions sous la forme de la petitesse de $\|\hat{f}_{J^c}\|^2_{L^2(\mu_X)}$ et de $\|f^*_{J^c}\|^2_{L^2(\mu_X)}$.

En examinant (1.12), nous voyons qu'une FSD divise la borne supérieure de $\|\hat{f}_N - f^*_{\mathcal{F}}\|^2_{L^2(\mu_X)}$ en trois composantes. Chaque composante porte sa propre signification statistique : $\|\hat{f}_J - f^*_J\|_{L^2(\mu_X)}$ est l'erreur d'estimation encourue parce que $\hat{f}_J$ estime $f^*_J$ ; $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$ est l'« énergie » de la partie libre $\hat{f}_{J^c}$ ; et $\|f^*_{J^c}\|_{L^2(\mu_X)}$ est l'erreur d'approximation résultant du fait que $\hat{f}_J$ n'estime pas $f^*_{J^c}$.

**Proposition 4.** *Pour toute FSD $\mathcal{F} = V_J \oplus V_{J^c}$ et toute fonction de taux $r$, nous avons*

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

*Définissons*

$$(V^*_J, V^*_{J^c}) \in \operatorname{argmin}\left(r(V_J, V_{J^c}) : \mathcal{F} = V_J \oplus V_{J^c}\right). \tag{1.13}$$

*Nous appelons* $(V_J^*, V_{J^c}^*)$ *la FSD optimale pour la solution* $(\mathcal{F}, \hat{f}_N)$ *du problème de régression supervisée à valeurs réelles* $(\mu_X, f^*, \xi)$. *Alors en particulier,*

$$\mathbb{P}\left(\|\hat{f}_N - f_{\mathcal{F}}^*\|_{L^2(\mu_X)}^2 \leq r^2(V_J^*, V_{J^c}^*)\right) \geq 1 - \delta(V_J^*, V_{J^c}^*). \tag{1.14}$$

Dans ce qui suit, nous écrivons $P_{V_J^*} f$ comme $f_{J_*}$, $P_{V_{J^c}^*} f$ comme $f_{J_*^c}$ ; nous écrivons $P_{V_J^*} \hat{f}_N$ comme $\hat{f}_{J_*}$, $P_{V_{J^c}^*} \hat{f}_N$ comme $\hat{f}_{J_*^c}$ ; et nous écrivons $P_{V_J^*} f_{\mathcal{F}}^*$ comme $f_{J_*}^*$, $P_{V_{J^c}^*} f_{\mathcal{F}}^*$ comme $f_{J_*^c}^*$.

La théorie classique de l'apprentissage statistique introduite dans la Section 1.3 correspond au choix de la FSD triviale $V_J = \mathcal{F}$. Dans ce cas, la théorie classique de l'apprentissage statistique s'attend à ce que la statistique classique effectue l'estimation sur l'ensemble de l'espace des caractéristiques, obtenant ainsi une borne supérieure pour l'erreur d'estimation. Cette approche est intuitive étant donné que lorsqu'un estimateur $\hat{f}_N$ de $f_{\mathcal{F}}^*$ est consistant, nous nous attendons à ce que $\hat{f}_N$ estime $f_{\mathcal{F}}^*$ et pas seulement une partie de celui-ci. Une idée clé exposée par la méthode FSD est que cela peut ne pas être le cas, c'est-à-dire que cette FSD triviale n'est pas nécessairement optimale ; par conséquent, la borne supérieure qu'elle fournit pour l'erreur d'estimation n'est pas toujours fine. En fait, pour une large classe d'algorithmes spectraux—tels que la régression ridge, la descente de gradient, le flot de gradient, etc., voir l'Exemple 9, et pour tout problème de régression supervisée à valeurs réelles et pour tout espace de caractéristiques donné par un certain RKHS, avec une grande probabilité, nous pouvons inverser (1.14), c'est-à-dire que pour ces problèmes d'apprentissage supervisé et ces solutions, il existe une constante absolue $0 < c < 1$ et un nombre réel $0 < \delta < 1$ tels que l'inégalité suivante est vérifiée :

$$\mathbb{P}\left(\|\hat{f}_N - f_{\mathcal{F}}^*\|_{L^2(\mu_X)}^2 \geq c r^2(V_J^*, V_{J^c}^*)\right) \geq 1 - \delta. \tag{1.15}$$

Cela implique le phénomène remarquable suivant : pour cette classe de $(\mu_X, f^*, \xi)$ et de $(\mathcal{F}, \hat{f}_N)$, l'erreur d'estimation $\|\hat{f}_N - f_{\mathcal{F}}^*\|_{L^2(\mu_X)}^2$ est « caractérisée » par une interpolation entre ces deux approches distinctes. Ici, parce que $\|\hat{f}_N - f_{\mathcal{F}}^*\|_{L^2(\mu_X)}^2$ est avec une grande probabilité équivalente à $r(V_J^*, V_{J^c}^*)$, nous utilisons le terme « caractérisée ». De plus, il n'existe aucune autre manière de contrôler l'erreur d'estimation au-delà des deux voies décrites dans la Proposition 4.

**Problèmes de classification supervisée binaire.** Dans ce paragraphe, nous considérons l'excès de risque en population pour le problème de classification binaire $(\mu_X, \eta)$, dont nous rappelons qu'il est défini comme :

$$\begin{aligned} P\mathcal{L}_{\hat{f}_N}^{(0,1)} &= \mathbb{P}\left(Y\hat{f}_N(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y\left(\eta(X) - \frac{1}{2}\right) < 0\right), \text{ et} \\ P\mathcal{L}_{\hat{f}_N}^{(0,1),\mathcal{F}} &= \mathbb{P}\left(Y\hat{f}_N(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y f_{\mathcal{F}}^*(X) < 0\right), \end{aligned}$$

où $\eta : \boldsymbol{x} \in \Omega_X \mapsto \mathbb{P}(Y = 1 | X = \boldsymbol{x})$. Comme dans les problèmes de régression, $P\mathcal{L}_{\hat{f}_N}^{\mathcal{F}}$ ou $P\mathcal{L}_{\hat{f}_N}$ se compose de trois contributions. À savoir, étant donné une décomposition arbitraire $\mathcal{F} = V_J \oplus V_{J^c}$, soit $f_J^*$ une certaine fonction dans $V_J$ — nous la définirons plus tard. Nous décomposons le risque 0-1 de $\hat{f}_N$ comme suit :

$$\begin{aligned} P\mathcal{L}_{\hat{f}_N}^{(0,1)} &= \mathbb{P}\left(Y\hat{f}_N(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) && (1.16) \\ &+ \mathbb{P}\left(Y\hat{f}_J(X) < 0 \middle| (X_i, Y_i)_{i=1}^N\right) - \mathbb{P}\left(Y f_J^*(X) < 0\right) && (1.17) \\ &+ \mathbb{P}\left(Y f_J^*(X) < 0\right) - \mathbb{P}\left(Y\left(\eta(X) - \frac{1}{2}\right) < 0\right), && (1.18) \end{aligned}$$

où (1.16) est l'erreur causée par la partie libre $\hat{f}_{J^c}$ ; (1.17) est l'erreur de prédiction causée par $\hat{f}_J$ comparée à celle de $f_J^*$ ; et (1.18) est l'erreur de prédiction causée par $f_J^*$ comparée à celle de la règle de Bayes (ou, lorsque nous remplaçons $\eta(X) - 1/2$ par $f_{\mathcal{F}}^*(X)$, cela devient l'erreur d'approximation de $f_J^*$ par rapport à $f_{\mathcal{F}}^*$). Ces trois termes sont précisément les homologues de $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}^2$, $\|\hat{f}_J - f_J^*\|_{L^2(\mu_X)}^2$, et $\|f_{J^c}^*\|_{L^2(\mu_X)}^2$ dans (1.11).

De manière analogue au cas de la régression, la méthode FSD vise à trouver une fonction de taux non triviale $r : (V_J, V_{J^c}) \mapsto r(V_J, V_{J^c}) \in \mathbb{R}_+$ et une fonction de confiance $\delta : (V_J, V_{J^c}) \mapsto \delta(V_J, V_{J^c}) \in [0, 1]$ telles que pour toute FSD, l'inégalité suivante soit vérifiée avec une probabilité d'au moins $1 - \delta(V_J, V_{J^c})$ (ou en espérance) :

$$(1.16) + (1.17) + (1.18) \leq r(V_J, V_{J^c}).$$

De même, la proposition suivante est vérifiée.

**Proposition 5.** *Pour toute FSD $\mathcal{F} = V_J \oplus V_{J^c}$ et toute fonction de taux $r$, nous avons*

$$\mathbb{P}\left(P\mathcal{L}^{(0,1)}_{\hat{f}_N} \leq r^2(V_J, V_{J^c})\right) \geq 1 - \delta(V_J, V_{J^c}).$$

*Définissons*

$$(V_J^*, V_{J^c}^*) \in \operatorname{argmin}\left(r(V_J, V_{J^c}) : \mathcal{F} = V_J \oplus V_{J^c}\right). \tag{1.19}$$

*Nous appelons $(V_J^*, V_{J^c}^*)$ la FSD optimale pour la solution $(\mathcal{F}, \hat{f}_N)$ du problème de classification supervisée binaire $(\mu_X, \eta)$. Alors en particulier,*

$$\mathbb{P}\left(P\mathcal{L}^{(0,1)}_{\hat{f}_N} \leq r^2(V_J^*, V_{J^c}^*)\right) \geq 1 - \delta(V_J^*, V_{J^c}^*). \tag{1.20}$$

**La FSD en tant que méthode analytique.** Nous soulignons que la méthode FSD sert d'outil pour aider les théoriciens à analyser l'excès de risque de tout estimateur ainsi qu'à comprendre son comportement. C'est-à-dire que dans la construction des estimateurs $\hat{f}_N$, les praticiens n'ont aucun contrôle sur le choix de $V_J$ et $V_{J^c}$—parce que l'estimateur lui-même ne prend pas $V_J$ ou $V_{J^c}$ comme paramètres d'entrée. Par exemple, l'estimateur interpolant de norme minimale dans l'Exemple 10 n'a absolument aucun paramètre réglable. Par conséquent, nous affirmons que la décomposition de $\mathcal{F}$ en deux sous-espaces est effectuée implicitement par l'estimateur, et non par les praticiens. En conséquence, lorsque les praticiens exécutent cet algorithme statistique, cette décomposition se produit comme une opération en boîte noire. Pour les estimateurs avec des paramètres réglables, étant donné un paramètre défini par les praticiens, l'estimateur détermine automatiquement la FSD optimale $(V_J^*, V_{J^c}^*)$ en fonction à la fois de ce paramètre et du problème de régression lui-même. Certes, nous soulignons que les théoriciens peuvent tirer parti des nouvelles intuitions théoriques fournies par la méthode FSD pour aider à concevoir des méthodes pratiques. Par exemple, en utilisant la caractérisation précise de l'erreur d'estimation offerte par la méthode FSD pour concevoir un estimateur adaptatif via la méthode de Lepski, [Lep91], voir aussi la synthèse [Lep23] pour d'autres méthodes adaptatives.

Ci-dessous, dans la Section 1.5.2 et la Section 1.5.3, nous expliquons séparément les rôles de ces deux sous-espaces et comment ils assistent spécifiquement les théoriciens dans leur analyse.

## 6.2 $V_J$ définit un morphisme dans la catégorie des problèmes d'apprentissage supervisé

Pour des raisons de commodité, tout au long de cette section, nous supposons toujours que $f^* \in \mathcal{F}$, et par conséquent, nous ne distinguons pas $f^*$ de $f^*_{\mathcal{F}}$. Avant de commencer cette section, nous rappelons que pour obtenir une borne supérieure pour $\|\hat{f}_N - f^*\|^2_{L^2(\mu_X)}$ ou pour $P\mathcal{L}^{(0,1)}_{\hat{f}_N}$ dans les problèmes de classification binaire via la méthode FSD, sur $V_J$, nous avons besoin d'une borne supérieure pour $\|\hat{f}_J - f_J^*\|_{L^2(\mu_X)}$ ou pour $\mathbb{P}(Y\hat{f}_J(X) < 0 | (X_i, Y_i)_{i=1}^N) - \mathbb{P}(Yf_J^*(X) < 0)$. C'est précisément la tâche de la théorie classique de l'apprentissage statistique et de la statistique mathématique. Quel est alors le rôle de la méthode FSD sur $V_J$ ?

Pour tout quintuplet donné $(\mu_X, f^*, \xi, \mathcal{F}, \hat{f}_N)$ constitué d'un problème de régression supervisée à valeurs réelles et d'une solution, la FSD fournit, via $V_J$, la flèche suivante :

$$\bullet_J : (\mu_X, f^*, \xi, \mathcal{F}, \hat{f}_N) \longmapsto (\mu_X, f_J^*, \zeta, V_J, \hat{f}_J), \text{ où } \zeta = \xi + f_{J^c}^*,$$

à travers la relation suivante :

$$Y = f^*(X) + \xi = f_J^*(X) + \zeta.$$

En d'autres termes, la méthode FSD dote le théoricien du pouvoir de passer du traitement d'un problème de régression supervisée et de sa solution $(\mu_X, f^*, \xi, \mathcal{F}, \hat{f}_N)$ à un autre problème de régression supervisée et sa solution $(\mu_X, f_J^*, \zeta, V_J, \hat{f}_J)$. De plus, si l'on souhaite uniquement obtenir une borne supérieure pour $\|\hat{f}_N - f^*\|^2_{L^2(\mu_X)}$, alors le théoricien possède la liberté de choisir la flèche, c'est-à-dire en sélectionnant une FSD, choisissant ainsi librement le problème de régression supervisée cible et sa solution $(\mu_X, f_J^*, \zeta, V_J, \hat{f}_J)$. Cela peut souvent conférer au théoricien un pouvoir analytique supplémentaire au-delà de la théorie classique de l'apprentissage statistique introduite dans la Section 1.3 — car il suffit alors d'appliquer la théorie classique de l'apprentissage statistique sur le nouveau modèle $V_J$, et le nouveau signal $f_J^*$ peut être plus facile à analyser. Bien sûr, si l'on vise à obtenir une borne supérieure pour

$\|\hat{f}_N - f^*\|^2_{L^2(\mu_X)}$ qui soit aussi fine que possible, ou même une caractérisation précise de $\|\hat{f}_N - f^*\|^2_{L^2(\mu_X)}$ au sens de (1.15), il est alors nécessaire de choisir une bonne FSD $(V_J, V_{J^c})$, de sorte que la fonction de taux $r(V_J, V_{J^c})$ soit aussi petite que possible—voire la FSD optimale $(V_J^*, V_{J^c}^*)$.

Illustrons maintenant ce point par quelques exemples.

### 6.2.1 $\bullet_J$ définit le nouveau $\hat{f}_J$.

Bien que $\hat{f}_J$ soit par définition $P_{V_J}\hat{f}_N$, si $V_J$ est choisi de manière appropriée, $\hat{f}_J$ peut admettre une caractérisation équivalente autre que $P_{V_J}\hat{f}_N$, que le théoricien peut alors exploiter pour faciliter l'analyse. Trois exemples suivent. Leurs preuves sont aisées et donc omises, voir également la Proposition 20 plus loin au Chapitre 2 pour la preuve de la Proposition 8 ci-dessous.

**Proposition 6** (auto-régularisation de l'estimateur interpolant de norme $\|\cdot\|_q$ minimale)**.** *Soit $p \in \mathbb{N}_+$, $\mathcal{F} = \{\langle\cdot, \boldsymbol{\beta}\rangle : \boldsymbol{\beta} \in \mathbb{R}^p\}$. Soient $\boldsymbol{e}_1, \dots, \boldsymbol{e}_p$ une base de $\mathbb{R}^p$. Soit $1 \le q < \infty$ un nombre réel, et $\|\cdot\|_q$ la norme $\ell_q$ sur $\mathbb{R}^p$ par rapport à cette base. Considérons l'estimateur interpolant de norme $\|\cdot\|_q$ minimale défini dans l'Exemple 10, c'est-à-dire,*

$$\hat{\boldsymbol{\beta}} \in \operatorname{argmin}(\|\boldsymbol{\beta}\|_q : \mathbb{X}\boldsymbol{\beta} = \boldsymbol{y}), \text{ où } \mathbb{X} = [X_1|\cdots|X_N]^\top, \boldsymbol{y} = (Y_1, \cdots, Y_N).$$

*Prenons n'importe quelle FSD $\mathbb{R}^p = V_J \oplus V_{J^c}$, où $V_J = \operatorname{span}(\boldsymbol{e}_j : j \in J)$ pour un certain $J \subset \{1, \cdots, p\}$. Définissons $\mathcal{A} : \boldsymbol{\mu} \in \mathbb{R}^N \mapsto \mathcal{A}[\boldsymbol{\mu}] \in \arg\min(\|\boldsymbol{\nu}\|_q : \mathbb{X}\boldsymbol{\nu} = \boldsymbol{\mu},\ \boldsymbol{\nu} \in V_{J^c})$. Alors $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{A}[\boldsymbol{y} - \mathbb{X}\hat{\boldsymbol{\beta}}_J]$, et*

$$\hat{\boldsymbol{\beta}}_J \in \operatorname*{argmin}_{\boldsymbol{\beta}_J \in V_J} \left(L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) + \|\boldsymbol{\beta}_J\|_q^q\right), \text{ où } L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q.$$

La Proposition 6 nous indique que bien que $\hat{\boldsymbol{\beta}}_J$ soit par définition égale à $P_{V_J}\hat{\boldsymbol{\beta}}$, en tant que théoriciens, lorsque nous choisissons une FSD appropriée, nous pouvons lui conférer une nouvelle signification statistique—une minimisation du risque empirique régularisée (RERM) dont la fonction de perte $L_{\boldsymbol{\beta}_J}$ est en fait une fonction de perte stochastique et $\|\cdot\|_q^q$ est la fonctionnelle d'auto-régularisation. Parce que cette régularisation est imposée par $\hat{\boldsymbol{\beta}}$ sur lui-même, plutôt que d'être explicitement définie par le praticien, nous l'appelons auto-régularisation. Cette régularisation ne dépend pas de l'algorithme d'entraînement spécifique, et diffère donc de la régularisation implicite introduite dans la Section 1.4, voir [BMR21, pp. 92].

**Proposition 7** (auto-régularisation du classifieur interpolant de norme $\|\cdot\|_2$ minimale)**.** *Si $\mathcal{F}$ est identifié avec $\mathbb{R}^p$, et que $\hat{\boldsymbol{\beta}}$ est le classifieur interpolant de norme $\|\cdot\|_2$ minimale (Exemple 10). Prenons une FSD arbitraire $\mathbb{R}^p = V_J \oplus V_{J^c}$, notons $\mathbb{1} = (1, \dots, 1) \in \mathbb{R}^N$, et soit $\mathbb{X}_{\boldsymbol{y},J^c} = [Y_1 P_{V_{J^c}} X_1|\cdots|Y_N P_{V_{J^c}} X_N]^\top$. Définissons $\mathcal{B} : \boldsymbol{\mu} \in \mathbb{R}^N \mapsto \mathcal{B}[\boldsymbol{\mu}] \in \arg\min(\|\boldsymbol{\nu}\|_{\mathcal{H}} : \mathbb{X}_{\boldsymbol{y},J^c}\boldsymbol{\nu} \succeq \boldsymbol{\mu})$. Alors $\hat{\boldsymbol{\beta}}_{J^c} = \mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\hat{\boldsymbol{\beta}}_J]$, et*

$$\hat{\boldsymbol{\beta}}_J \in \operatorname{argmin}\left(L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) + \|\boldsymbol{\beta}_J\|_2^2 : f_J \in V_J\right), \text{ où } L_{f_J}((X_i, Y_i)_{i=1}^N) = \|\mathcal{B}[\mathbb{1} - \mathbb{X}_{\boldsymbol{y}}\boldsymbol{\beta}_J]\|_2^2.$$

*Ici, pour tout $\boldsymbol{a} = (a_i)_{i=1}^N$ et $\boldsymbol{b} = (b_i)_{i=1}^N$, nous écrivons $\boldsymbol{a} \succeq \boldsymbol{b}$, si $a_i \ge b_i$ pour tout $1 \le i \le N$.*

De manière similaire, ici $\hat{f}_J$ est identifié comme une RERM dont la fonction de perte est une fonction de perte stochastique.

Pour ces deux nouvelles fonctions de perte, parce qu'elles intègrent une régularisation, elles ne souffrent pas de surapprentissage. Par conséquent, l'application de la théorie classique de l'apprentissage statistique sur $V_J$ donne une inégalité d'oracle dont le terme résiduel peut tendre vers zéro. C'est précisément l'avantage apporté par le nouvel estimateur $\hat{f}_J$ via la FSD.

**Proposition 8** (régularisation effective)**.** *Si $\mathcal{F}$ est identifié avec un RKHS $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ avec la carte de caractéristiques $\phi$, et que $\hat{f}_N$ est la régression ridge sur $\mathcal{F}$ avec paramètre $t^{-1}$, c'est-à-dire, $\hat{f}_N = \frac{1}{N}\mathbb{X}^\top(\frac{1}{N}\mathbb{X}\mathbb{X}^\top + \frac{1}{t}I_N)^{-1}\boldsymbol{y}$, où $\boldsymbol{y} = (Y_1, \dots, Y_N)$ et $\mathbb{X} : f \in \mathcal{H} \mapsto (\langle\phi(X_i), f\rangle_{\mathcal{H}})_{i=1}^N$. Prenons une FSD arbitraire $\mathcal{H} = V_J \oplus V_{J^c}$, et notons $\mathbb{X}_{J^c} = \mathbb{X}P_{J^c}$. Alors*

$$\hat{f}_J \in \operatorname{argmin}\left(L_{f_J}((X_i, Y_i)_{i=1}^N) + \|f_J\|_{\mathcal{H}}^2\right), \text{ où } L_{f_J}((X_i, Y_i)_{i=1}^N) = \|Q(\boldsymbol{y} - \mathbb{X}f_J)\|_{\mathcal{H}}^2,$$

*et $Q : \mathbb{R}^N \to V_{J^c}$ est un opérateur linéaire borné tel que $Q^\top Q = (\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + t^{-1}I_N)^{-1}$.*

En d'autres termes, $\hat{f}_J$ est identifié comme une RERM dont la fonction de perte $L_{f_J}$ est également une fonction de perte stochastique. Ici, la régression ridge a un paramètre de réglage $t^{-1}$ ; ainsi, pour tout paramètre de réglage $t^{-1}$ donné par le praticien, $\hat{f}_N$ sélectionne lui-même une FSD, générant une nouvelle régularisation $(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + t^{-1}I_N)$, qui est appelée régularisation effective.

### 6.2.2 $\bullet_J$ définit le nouveau signal $f_J^*$.

En choisissant une FSD, le théoricien peut également sélectionner un nouveau signal approprié avec lequel travailler. Dans ce paragraphe, nous présentons deux exemples : la régression à facteurs latents et le classifieur interpolant de norme $\|\cdot\|_{\mathcal{H}}$ minimale.

**Régression à facteurs latents.** La régression à facteurs latents est une classe spéciale de problèmes de régression à valeurs réelles où la dépendance entre $(X, Y)$ est régie par un vecteur aléatoire latent $Z$, une matrice de plongement inconnue $A$, et deux types de bruit.

**Définition 9** (Problème de régression à facteurs latents)**.** *Soient $k < p$ deux entiers positifs, soit $\Omega_X = \mathbb{R}^p$, et soit $A \in \mathbb{R}^{p\times k}$ une matrice fixe mais inconnue. Soit $Z \in \mathbb{R}^k$ un vecteur aléatoire, appelé facteur latent. Soit $W \in \mathbb{R}^p$ un vecteur aléatoire centré, indépendant de $Z$, avec pour matrice de covariance $\Sigma_W = \mathbb{E}[W \otimes W]$. Soit $\xi \in \mathbb{R}$ une variable aléatoire centrée de variance $\sigma_\xi^2$, indépendante de $(Z, W)$. Le vecteur d'observation est défini par $X = AZ + W$. Ainsi, dans ce modèle, le vecteur d'observation observable $X$ provient d'un facteur latent $Z$ par une transformation linéaire non observée $A$, ainsi que d'une perturbation par un bruit non observé $W$, de sorte que $X = AZ + W$.*

*Soit $\Omega_Y = \mathbb{R}$, et soit $Y$ défini comme suit. Soit $\boldsymbol{\alpha}^* \in \mathbb{R}^k$ un vecteur de position, et la variable de réponse par $Y = \langle \boldsymbol{\alpha}^*, Z\rangle + \xi$. La variable de réponse $Y$ dépend uniquement du facteur latent $Z$, du signal inconnu $\boldsymbol{\alpha}^* \in \mathbb{R}^k$, et d'une perturbation par un bruit non observé $\xi$. Dans la régression à facteurs latents, la fonction de perte la plus courante est la perte quadratique $\ell : (y_1, y_2) \in \mathbb{R}\times\mathbb{R} \mapsto (y_1 - y_2)^2$. Voir, par exemple, [BBSMW21].*

Soit $\mathcal{F} = \{f_{\boldsymbol{\beta}}(\cdot) = \langle \boldsymbol{\beta}, \cdot\rangle : \boldsymbol{\beta} \in \mathbb{R}^p\}$. Le problème de régression à facteurs latents est mal spécifié à moins que $(Z, X)$ ne soit conjointement gaussien. En fait, la règle de Bayes est donnée par $f^* : \boldsymbol{x} \mapsto \langle \boldsymbol{\alpha}^*, \mathbb{E}[Z \mid X = \boldsymbol{x}]\rangle$. Cependant, le modèle statistique $\mathcal{F}$ est la classe des fonctionnelles linéaires. L'oracle dans $\mathcal{F}$ est donné par $f_{\mathcal{F}}^*$, identifié par un vecteur $\boldsymbol{\beta}^*$ à travers $f_{\mathcal{F}}^*(\cdot) = \langle \cdot, \boldsymbol{\beta}^*\rangle$, défini comme $\boldsymbol{\beta}^* \in \operatorname{argmin}(P\ell_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathbb{R}^p) = \operatorname{argmin}(\mathbb{E}[(\langle \boldsymbol{\beta}, X\rangle - Y)^2] : \boldsymbol{\beta} \in \mathbb{R}^p)$ . Soit $\Sigma = \mathbb{E}[X \otimes X]$ l'opérateur de covariance de $X$. Un calcul direct donne $\Sigma = A\Sigma_Z A^\top + \Sigma_W$, où $\Sigma_Z = \mathbb{E}[Z \otimes Z] : \mathbb{R}^k \to \mathbb{R}^k$. Puisque $\Sigma_W$ est définie positive, $\Sigma$ est également définie positive, et $\Sigma$ peut être vue comme la composante informative de rang $k$, $A\Sigma_Z A^\top$, perturbée par $\Sigma_W$. Il est calculé dans [BBSMW21, Équation 6] que $\boldsymbol{\beta}^* = \Sigma^{-1} A\Sigma_Z \boldsymbol{\alpha}^*$. Soit $\mathbb{Z} : \boldsymbol{\alpha} \in \mathbb{R}^k \mapsto (\langle Z_i, \boldsymbol{\alpha}\rangle)_{i=1}^N \in \mathbb{R}^N$. Dans le problème de régression à facteurs latents, le vecteur de réponse est $\boldsymbol{y} = \mathbb{Z}\boldsymbol{\alpha}^* + \boldsymbol{\xi}$, mais nous devons résoudre le problème dans $\mathbb{R}^p$, et l'oracle dans $\mathbb{R}^p$ est $\boldsymbol{\beta}^*$.

Ci-dessous, nous montrons comment, en choisissant une bonne FSD—c'est-à-dire un bon $V_J$—nous pouvons explorer la composante informative de rang $k$, $A\Sigma_Z A^\top$, cachée dans $\mathbb{R}^p$, ce qui est exactement le but du problème de régression à facteurs latents. Prenons $V_J = \operatorname{Im}(A\Sigma_Z A^\top) = \operatorname{Im}(A)$. Dans ce cas, $\boldsymbol{\beta}_J^* = P_{V_J}\boldsymbol{\beta}^* = \boldsymbol{\beta}^*$. Par conséquent, nous avons le problème de régression supervisée suivant $(\mu_X, \boldsymbol{\beta}_J^*, \zeta)$, où $\zeta = \xi + (\langle Z, \boldsymbol{\alpha}^*\rangle - \langle X, \boldsymbol{\beta}_J^*\rangle)$. Ici, le nouveau bruit se compose de deux parties : $\xi$ est le bruit original, tandis que $\langle Z, \boldsymbol{\alpha}^*\rangle - \langle X, \boldsymbol{\beta}_J^*\rangle = \langle Z, \boldsymbol{\alpha}^*\rangle - \langle X, \boldsymbol{\beta}^*\rangle$ correspond à l'erreur d'approximation de $\boldsymbol{\alpha}^*$ sur $\mathbb{R}^p$. Dans [BBSMW21], il est prouvé que ce terme est une composante irréductible de l'erreur d'estimation. Par conséquent, pour le problème de régression à facteurs latents, en choisissant un $V_J$ approprié, nous réduisons la dimension du problème à $k$, tout en garantissant que le signal dans cet espace satisfait $\boldsymbol{\beta}_J^* = \boldsymbol{\beta}^*$.

**Classifieur interpolant de norme $\|\cdot\|_2$ minimale.** Dans ce paragraphe, nous considérons le classifieur interpolant de norme $\|\cdot\|_2$ minimale défini dans l'Exemple 10, c'est-à-dire que nous supposons que $\mathcal{F}$ est identifié avec $\mathbb{R}^p$. Nous illustrons maintenant qu'en choisissant une FSD de manière appropriée, l'erreur d'approximation résultant de la restriction de l'estimation à $V_J$—à savoir, (1.18)—peut être éliminée. Nous examinons le modèle standard suivant pour les problèmes de classification supervisée binaire :

**Définition 10** (Problème de classification logistique)**.** *Soit $\boldsymbol{\mu} \in \mathbb{R}^p$ appelé le signal, et $\Lambda \in \mathbb{R}^{p\times p}$ un opérateur linéaire borné défini positif. Soit $X \sim \mathcal{N}(\mathbf{0}, \Lambda)$ un vecteur aléatoire gaussien de moyenne $\mathbf{0}$ et d'opérateur de covariance $\Lambda$. En définissant $\eta(\boldsymbol{x}) = \mathbb{P}(Y = 1 \mid X = \boldsymbol{x}) = 1/(1 + \exp(-2\langle \Lambda^{-1}\boldsymbol{\mu}, \boldsymbol{x}\rangle))$ et $\mathbb{P}(Y = -1 \mid X = \boldsymbol{x}) = 1 - \eta(\boldsymbol{x})$, nous spécifions la distribution de $Y$. Ce problème est appelé le modèle logistique, [Gir14, Section 11.1.3].*

Un calcul direct montre que le classifieur de Bayes pour le problème de classification logistique est $f^*(\cdot) = \operatorname{sign}(\langle \cdot, \Lambda^{-1}\boldsymbol{\mu}\rangle)$. Ainsi, le classifieur de Bayes peut être identifié avec $\Lambda^{-1}\boldsymbol{\mu}$. Par conséquent, tant que la FSD est choisie de telle sorte que $f_J^*$ et $\Lambda^{-1}\boldsymbol{\mu}$ soient bien alignés, (1.18) devient nulle. Plus tard, dans la Proposition 11, nous prouvons que si $\Lambda^{-1}\boldsymbol{\mu} \in V_J$, alors cela est effectivement vrai.

Le modèle logistique représente une classe de modèles de classification supervisée binaire ; le modèle de classification par mélange gaussien [WT21] et le modèle de classification à facteurs latents [BW23] partagent la même caractéristique—à savoir, il existe un classifieur linéaire optimal qui correspond à $f^*$.

### 6.2.3 $\bullet_J$ réduit les points fixes.

Parce que nous pensons que l'estimation se produit uniquement sur $V_J$, en conséquence le théoricien devrait appliquer la théorie classique de l'apprentissage statistique—c'est-à-dire les méthodes de la Section 1.3—uniquement sur $V_J$. Un résultat de cette démarche est que, puisque le problème d'apprentissage supervisé et sa solution ont tous deux changé, l'application de la théorie classique de l'apprentissage statistique sur $V_J$ peut produire un point fixe plus petit, et ainsi une borne plus petite sur $\|\hat{f}_J - f_J^*\|^2_{L^2(\mu_X)}$. La raison principale pour laquelle on s'attend à ce que les points fixes sur $V_J$ soient plus petits que celui sur $\mathcal{F}$ est parce que nous avons généralement $\dim(V_J) \ll \dim(\mathcal{F})$, non pas parce que $f_J^*$ et $\hat{f}_J$ ont changé. Dans cette section, nous illustrons comment la FSD réduit les points fixes définis dans la Section 1.3.1, en utilisant l'exemple du surapprentissage bénin pour l'estimateur interpolant de norme $\|\cdot\|_q$ minimale. Pour la régression ridge, la FSD peut également réduire le point fixe multiplicateur et le point fixe quadratique, mais la preuve est plus complexe et ne sera pas présentée ici (voir [P2]).

**La FSD réduit le point fixe multiplicateur.** La version formelle et la preuve de la Proposition 9 suivante peuvent être trouvées dans [P1], voir également la Section 4.3.4 ; nous ne les répétons pas ici. Prouver ces propriétés nécessite les outils géométriques sur $V_{J^c}$ introduits dans la Section 1.5.3.

**Proposition 9** (informel)**.** *En utilisant la notation de la Proposition 6.*

*Sous certaines hypothèses, il existe des constantes absolues* $0 < \delta_{\mathrm{M}} < \frac{1}{100}$, $c, c' < 1$, $\ell_* > 0$ *et* $c'' = c''(c, c', \delta_{\mathrm{M}}) > 1$ *telles que pour tout sous-ensemble de localisation* $\mathcal{G} \subset V_J$, $r_{\mathrm{M}}(\mathcal{G}, \delta_{\mathrm{M}}, \frac{2}{q}, 4c\frac{N^{\frac{q}{2}}}{\ell_*^q}) \leq c''\sigma_\xi(\frac{|J|}{N})^{\frac{1}{2(q-1)}}$ *lorsque* $q \geq 2$ *; et* $r_{\mathrm{M}}(\mathcal{G}, \delta_{\mathrm{M}}, 1, 4c'\sigma_\xi^{q-2}\frac{N^{\frac{q}{2}}}{\ell_*^q}) \leq c''\sigma_\xi^{q-1}(\frac{|J|}{N})^{\frac{1}{2}}$ *lorsque* $1 \leq q < 2$.

**La FSD réduit le point fixe quadratique.** Pour l'estimateur interpolant de norme $\|\cdot\|_q$ minimale, la FSD fournie par la Proposition 6 peut également réduire le point fixe quadratique. La version formelle et la preuve de la proposition suivante peuvent être trouvées dans [P1], voir également la Section 4.3.4 plus loin.

**Proposition 10** (informel)**.** *Sous les hypothèses de la Proposition 9, il existe une constante absolue* $0 < \delta_{\mathrm{Q}} < \frac{1}{100}$, $c = c(q)$, *et* $c' = c'(q)$, *telles que ce qui suit est vérifié.*

1. *Lorsque* $q \geq 2$. *Alors pour tout* $r > 0$, *et tout sous-ensemble de localisation* $\mathcal{G}$, *avec une probabilité d'au moins* $1 - \delta_{\mathrm{Q}}$, *pour tout* $\boldsymbol{\beta}_J \in \mathcal{G} \cap S_{L^2(P_{V_J}\mu_X)}(\boldsymbol{\beta}_J^*; r)$,

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

   *et* $\boldsymbol{g} = \nabla L_{\boldsymbol{\beta}_J^*}$.

2. *Lorsque* $1 \leq q < 2$. *Supposons que* $X_J$ *est un vecteur aléatoire gaussien centré, alors pour tout sous-ensemble de localisation* $\mathcal{G}$ *et tout* $0 < r < \sigma_\xi$, *avec une probabilité d'au moins* $1 - \delta_{\mathrm{Q}}$, $(\mathcal{G}, X_J, L_\bullet)$ *satisfait la condition de Bernstein locale à l'échelle* $r$, *avec les paramètres* $(\diamondsuit, 1)$, *où*

$$\diamondsuit = c'\frac{N^{\frac{q}{2}}\sigma_\xi^{q-2}}{\ell_*^q}.$$

Note : la condition de Bernstein locale qui est vérifiée avec une grande probabilité au point *2* est due au fait que dans la Proposition 6, la fonction de perte $L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) = \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q$ est une fonction de perte stochastique qui dépend de $\mathbb{X}P_{V_{J^c}}$ ; par conséquent, l'excès de risque en population $P\mathcal{L}_{\boldsymbol{\beta}_J}^{V_J}$ est une espérance conditionnelle $\mathbb{E}_{\mathbb{X}_J,\boldsymbol{\xi}}P_N\mathcal{L}_{\boldsymbol{\beta}_J}^{V_J}$, et la condition de Bernstein locale est vérifiée avec une grande probabilité. On peut prouver que $\boldsymbol{\beta}_J^* \in \arg\min(P\ell_{\boldsymbol{\beta}_J} : \boldsymbol{\beta}_J \in V_J)$ est vrai presque sûrement, où $P\ell_{\boldsymbol{\beta}_J} = \mathbb{E}_{\mathbb{X}_J,\boldsymbol{\xi}}L_{\boldsymbol{\beta}_J}$, voir le Lemme 21 plus loin.

La Proposition 10 nous dit que lorsque $q \geq 2$, si la FSD est convenablement choisie, il existe $\delta_{\mathrm{Q}} < \frac{1}{100}$ tel que pour chaque sous-ensemble de localisation $\mathcal{G} \subset V_J$, le point fixe quadratique $r_{\mathrm{Q}}(\mathcal{G}, \delta_{\mathrm{Q}}, \frac{2}{q}) = 0$ lorsque $q \geq 2$. En conséquence, dans cette situation, la FSD élimine complètement le point fixe quadratique.

**$\bullet_J$ comme une coquille enveloppant l'analyse mathématico-statistique classique**

Les estimateurs de régression ridge et interpolant de norme minimale étudiés précédemment peuvent tous deux être écrits comme une RERM (ou leurs limites). Les estimateurs de cette forme tombent généralement dans le champ d'application de la théorie de l'apprentissage statistique, [VC68]. Dans cette section, nous montrons que la méthode FSD est non seulement applicable aux estimateurs définis par l'ERM et la RERM, qui sont courants en théorie de l'apprentissage statistique, mais aussi aux estimateurs classiques qui appartiennent plus largement au domaine de la statistique mathématique : les méthodes spectrales (Exemple 9). Appliquer la méthode FSD à l'analyse de l'erreur d'estimation de tels estimateurs revient à envelopper d'une coquille l'analyse mathématico-statistique originelle—c'est-à-dire, confiner l'analyse de l'erreur d'estimation, qui couvrait initialement l'espace des caractéristiques tout entier, au sous-espace $V_J$. Même si cela ne crée pas nécessairement un nouvel estimateur ni ne réduit les points fixes comme cela le fait pour les estimateurs interpolants de norme minimale ou la régression ridge, cela fournit toujours un signal « correct » $f_J^*$ avec lequel travailler.

La théorie statistique classique pour les méthodes spectrales fournit des moyens d'obtenir une borne supérieure sur $\|\hat{f}_N - f^*\|_{L^2(\mu_X)}$, par exemple, [SZ07, YRC07, BPR07, LGRO+08, BM16, BM18, BMM19, ZLL23, LGSL24]. Cependant, si nous effectuons d'abord une FSD, alors nous n'avons besoin d'appliquer la théorie classique que pour obtenir une borne supérieure sur $\|\hat{f}_J - f_J^*\|_{L^2(\mu_X)}$. Cela signifie que nous sommes passés de l'estimation du signal complet $f^*$ à l'estimation du « signal effectif » $f_J^*$, et par conséquent nous pouvons obtenir une caractérisation de l'erreur d'estimation (au sens de (1.15))—quelque chose que l'approche classique ne peut pas accomplir.

### 6.2.4 $V_{J^c}$ : de nouveaux outils issus des Aspects Géométriques de l'Analyse Fonctionnelle

Puisqu'aucune estimation de $f_{J^c}^*$ par $\hat{f}_{J^c}$ n'a lieu dans le sous-espace libre, nous disons qu'aucune statistique ne se produit sur ce sous-espace. Par conséquent, les outils requis pour ce sous-espace n'appartiennent pas à la statistique mathématique classique, et pour cette raison nous en savons encore relativement peu à son sujet. Nos travaux constituent donc les premiers exemples d'analyse de certains estimateurs dans l'espace libre. Bien sûr, nous avons utilisé des outils existants issus des Aspects Géométriques de l'Analyse Fonctionnelle (GAFA) qui n'étaient pas utilisés auparavant en statistique et nous avons dû les étendre pour les adapter à notre cadre statistique.

En ce qui concerne le sous-espace libre et l'estimateur $\hat{f}_{J^c}$ sur celui-ci, nous nous concentrons principalement sur les deux problèmes suivants :

1. les propriétés stochastiques que le sous-espace libre fournit pour $\hat{f}_J$ ;
2. l'énergie $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$ de $\hat{f}_{J^c}$.

### 6.2.5 $V_{J^c}$ fournit des propriétés stochastiques de $\hat{f}_J$

Dans cette section, nous considérons l'estimateur interpolant de norme $\|\cdot\|_q$ minimale (Exemple 10) et la régression ridge.

1. Estimateur interpolant de norme $\|\cdot\|_q$ minimale.

   Dans la Proposition 6, la Proposition 9, et la Proposition 10, nous avons déjà vu que la FSD identifie $\hat{\boldsymbol{\beta}}_J$ de manière équivalente comme une RERM dont la fonction de perte est donnée par $L_{\boldsymbol{\beta}_J} : (X_i, Y_i)_{i=1}^N \in \Omega^N \mapsto \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta}_J]\|_q^q$. Ici, nous rappelons sa définition : soient $\mathbb{X}_J = \mathbb{X}P_{V_J}$ et $\mathbb{X}_{J^c} = \mathbb{X}P_{V_{J^c}}$ ; alors $\mathcal{A} : \boldsymbol{\mu} \in \mathbb{R}^N \mapsto \mathcal{A}[\boldsymbol{\mu}] \in \arg\min(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}\boldsymbol{\nu} = \boldsymbol{\mu})$. Ainsi $\mathcal{A} : (\mathbb{R}^N, \|\cdot\|_2) \to (V_{J^c}, \|\cdot\|_q)$ est un opérateur de plongement aléatoire, et par conséquent $L_\bullet$ est une fonction de perte stochastique.

2. Régression ridge.

   De même, la Proposition 8 nous indique que pour une régression ridge avec le paramètre $t^{-1}$, son $\hat{f}_J$ est également une RERM dont la fonction de perte est $L_{f_J}((X_i, Y_i)_{i=1}^N) = \|Q(\boldsymbol{y} - \mathbb{X}f_J)\|_{\mathcal{H}}^2$, où $Q^\top Q = (\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + t^{-1}I_N)^{-1}$.

Suivant le credo de la FSD—appliquer la statistique mathématique classique et la théorie de l'apprentissage statistique (voir la Section 1.3) sur $V_J$—nous devons étudier les propriétés de ces fonctions de perte stochastiques afin de compléter les preuves de la Proposition 9 et de la Proposition 10, ainsi que pour calculer les points fixes multiplicateur et quadratique pour la régression ridge. Les propriétés de ces fonctions de perte stochastiques nécessitent donc une analyse à l'aide d'outils géométriques spécialisés. Cet outil est fourni par le célèbre théorème de Dvoretzky-Milman, [Dvo59, Dvo61, Mil71].

**Le théorème de Dvoretzky-Milman et son rôle dans le surapprentissage bénin pour l'estimateur interpolant de norme $\|\cdot\|_q$ minimale.** Pour tout sous-ensemble compact $K \subset \mathbb{R}^p$, nous définissons $\ell_*(K) = \mathbb{E}(\sup\langle \boldsymbol{v}, G\rangle : \boldsymbol{v} \in K)$ comme la largeur moyenne gaussienne de $K$, où $G \in \mathbb{R}^p$ est un vecteur aléatoire gaussien standard. Nous posons $\mathrm{diam}(K) = \max(\|\boldsymbol{v}\|_2 : \boldsymbol{v} \in K)$ comme le diamètre $\ell_2$ de $K$. Nous notons $K^\circ = \{\boldsymbol{v} \in \mathbb{R}^p : \langle \boldsymbol{v}, \boldsymbol{u}\rangle \leq 1, \forall \boldsymbol{u} \in K\}$ comme le corps polaire de $K$. Notons $d_*(K) = (\ell_*(K^\circ)/\,\mathrm{diam}(K^\circ))^2$ la dimension de Dvoretzky de $K$. Nous notons $q'$ par $\frac{q}{q-1}$. Ci-dessous se trouve la version de Milman du théorème de Dvoretzky ; voir [Pis89].

**Théorème 2** (Dvoretzky-Milman)**.** *Il existe des constantes absolues $\kappa_{DM} \leq 1$ et $c_1$ telles que ce qui suit est vérifié. Soit $|||\cdot|||$ une norme sur $\mathbb{R}^p$ et désignons par $B$ sa boule unité. Notons par $\mathbb{G} := \mathbb{G}^{(N\times p)}$, la matrice gaussienne standard $N \times p$ avec des entrées i.i.d. $\mathcal{N}(0,1)$. Étant donné tout $0 < \varepsilon_1 \leq 1$. Supposons que $N \leq \kappa_{DM}\varepsilon_1^2 d_*(B)$. Alors, avec une probabilité d'au moins $1 - \exp(-c_1\varepsilon_1^2 d_*(B))$, pour tout $\boldsymbol{\lambda} \in \mathbb{R}^N$,*

$$(1-\varepsilon_1)\,\|\boldsymbol{\lambda}\|_2\,\ell_*(B^*) \leq |||\mathbb{G}^\top \boldsymbol{\lambda}||| \leq (1+\varepsilon_1)\,\|\boldsymbol{\lambda}\|_2\,\ell_*(B^*). \tag{1.21}$$

Pour tout $0 < \varepsilon_1 < 1$, nous définissons l'événement

$$\Omega_{\mathrm{DM,reg}}(\varepsilon_1) := \left\{\forall \boldsymbol{\lambda} \in \mathbb{R}^N : \|\boldsymbol{\lambda}\|_2\,(1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_q^p) \leq \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \leq \|\boldsymbol{\lambda}\|_2\,(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_q^p)\right\} \tag{1.22}$$

$$\subset \left\{\forall \boldsymbol{\mu} \in \mathbb{R}^N : \frac{\|\boldsymbol{\mu}\|_2}{(1+\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_q^p)} \leq \|\mathcal{A}[\boldsymbol{\mu}]\|_q \leq \frac{\|\boldsymbol{\mu}\|_2}{(1-\varepsilon_1)\ell_*(\Sigma_{J^c}^{1/2}B_q^p)}\right\}. \tag{1.23}$$

Il découle du Théorème 2 appliqué à la norme $|||\cdot||| = \|\Sigma_{J^c}^{1/2}\cdot\|_{q'}$ que, si $X_{J^c}$ est un vecteur aléatoire gaussien et $\kappa_{DM}\varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2}B_{q'}^p) \geq N$, alors $\mathbb{P}(\Omega_{\mathrm{DM,reg}}(\varepsilon_1)) \geq 1 - \exp(-c_1\varepsilon_1^2 d_*(\Sigma_{J^c}^{-1/2}B_{q'}^p))$. L'inclusion de (1.23) découle de la dualité forte : pour tout $\boldsymbol{\mu} \in \mathbb{R}^N$,

$$\|\mathcal{A}[\boldsymbol{\mu}]\|_q = \min\left(\|\boldsymbol{\nu}\|_q : \mathbb{X}_{J^c}^\top \boldsymbol{\nu} = \boldsymbol{\mu}\right) = \max\left(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle : \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{q'} \leq 1\right). \tag{1.24}$$

Même si $\mathcal{A} : (\mathbb{R}^N, \ell_2) \to (V_{J^c}, \ell_q)$ est un plongement métrique non linéaire (sauf lorsque $q = 2$), il satisfait un théorème DM hérité de $\mathbb{X}_{J^c}^\top$.

Ci-dessous, nous démontrons comment utiliser le théorème de Dvoretzky-Milman pour prouver la Proposition 10, point *1*.

**Proof.** (de la Proposition 10, point *1*) Par l'Exemple 12, nous avons

$$\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J]\|_q^q - \|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q^q \geq \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\rangle + \frac{q-1}{q2^q}\,\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q^q,$$

où $\boldsymbol{g}$ est défini dans la Proposition 9. À partir de la définition de $\mathcal{A}$, nous avons $\|\mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J] - \mathcal{A}[\boldsymbol{y} - \mathbb{X}_J\boldsymbol{\beta}_J^*]\|_q \geq \|\mathcal{A}[\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)]\|_q$. Ensuite, en utilisant (1.23), nous obtenons

$$P_N\mathcal{L}_{\boldsymbol{\beta}_J} \geq \langle \boldsymbol{g}, \boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*\rangle + \frac{q-1}{q2^q}\frac{\|\mathbb{X}_J(\boldsymbol{\beta}_J - \boldsymbol{\beta}_J^*)\|_2^q}{(1+\varepsilon_1)^q\ell_*^q(\Sigma_{J^c}^{1/2}B_q^p)}.$$

Enfin, à partir de l'hypothèse $\dim(V_J) \lesssim N$ et du fait que pour tout $\mathcal{G} \subset V_J$, nous avons $r_{\mathrm{RIP},-}(\mathcal{G}) = 0$ (voir l'Exemple 11), la preuve de la Proposition 10, point *1* est complétée. ∎

**Le théorème de Dvoretzky-Milman pour les normes $\|\cdot\|_{q'}$ sous des mesures de probabilité générales.** Le Théorème 2 fournit le théorème de Dvoretzky-Milman pour les mesures gaussiennes. Parce que nous devons étudier le cas où $X_{J^c}$ est distribué selon une mesure de probabilité générale, nous avons besoin d'une extension du théorème de Dvoretzky-Milman pour les normes $\|\cdot\|_{q'}$. Le théorème suivant, tiré de [P1], est une contribution aux GAFA qui a été motivée précisément par la méthode FSD. Sa preuve peut être trouvée dans la Section 5.1. Notons $\mathrm{Log}(x) = \max\{1, \ln(x)\}$.

**Hypothèse 1.** *$\boldsymbol{\zeta} = (\zeta_j)_{j=1}^p$ est un vecteur aléatoire isotrope centré dans $\mathbb{R}^p$ avec des coordonnées i.i.d., satisfaisant $\mathbb{E}[\zeta_1^2] = 1$, et il existe des constantes absolues $0 < \kappa \leq 1$ et $\varepsilon > 0$ telles que $\mathbb{E}|\zeta_1|^{\max\{4,2q+\varepsilon\}} \leq \kappa^{\max\{4,2q+\varepsilon\}}$.*

**Théorème 3.** *Soit $\boldsymbol{\zeta}$ un vecteur aléatoire satisfaisant l'Hypothèse 1, et soit $\Sigma$ une matrice diagonale définie positive sur $\mathbb{R}^p$, $\Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_p)$. Soit $X = \Sigma^{1/2}\boldsymbol{\zeta}$, et soient $X_1, \cdots, X_N$ des copies indépendantes de $X$, formant la matrice aléatoire $\mathbb{X} = [X_1|\cdots|X_N]^\top = [Z_1|\cdots|Z_p]$, où $(Z_j)_{j=1}^p$ sont les vecteurs colonnes de $\mathbb{X}$. Notons $\ell_* = \ell_*(\Sigma^{1/2}B_q^p)$ et $d_* = d_*(\Sigma^{-1/2}B_{q'}^p)$. Sans perte de généralité, supposons que $d_* \geq 1$. Il existe alors une constante absolue $0 < \theta < 1$ telle que pour tout $\boldsymbol{\lambda} \in S_2^{N-1}$, $\mathbb{P}(|\langle Z_j, \boldsymbol{\lambda}\rangle| \geq \theta) \geq \kappa$. De plus, il existe des constantes absolues $c, c', C, C', C'', \kappa_{DM}, \varepsilon_0 > 0$ telles que les faits suivants sont vérifiés.*

1. *Lorsque $q \geq 2$. Si $N \leq \kappa_{DM} d_* \mathrm{Log}^{-2}(p^{\frac{1}{q'}}/d_*)$, alors avec une probabilité d'au moins*

$$1 - C'\mathrm{Log}\left(\frac{p^{\frac{1}{q'}}}{d_*}\right)\exp\left(-C''\kappa_{DM}\frac{d_*^\theta}{\mathrm{Log}^{2\theta}\left(\frac{p^{\frac{1}{q'}}}{d_*}\right)}\right) - 2\exp\left(-C'd_*\right) - C'd_*^{-c\min\{\varepsilon,\varepsilon_0\}} =: 1 - \bar{p}_{DM},$$

*il est vérifié pour tout $\boldsymbol{\lambda} \in S_2^{N-1}$,*

$$c\ell_* \leq \left\|\mathbb{X}^\top\boldsymbol{\lambda}\right\|_{q'} \leq C\mathrm{Log}(p)\ell_*.$$

2. *Lorsque $q < 2$. Si $N \leq \kappa_{DM} d_*(\Sigma^{-1/2}B_{q'}^p)$, alors*

$$\mathbb{P}\left(\forall\boldsymbol{\lambda} \in S_2^{N-1},\, c\ell_* \leq \|\mathbb{X}^\top\boldsymbol{\lambda}\|_{q'} \leq C\ell_*\right) \geq 1 - 3\exp(-c'd_*) - C'd_*^{-\frac{q'-2}{4}} =: 1 - \bar{p}_{DM}.$$

**Le théorème de Dvoretzky-Milman pour les normes $\|\cdot\|_{\mathcal{H}}$ sous des mesures de probabilité générales.** Lorsque $q = 2$, le théorème de Dvoretzky-Milman peut être valable pour des mesures de probabilité plus générales, par exemple, pour une carte de caractéristiques générée par des RKHS dont les fonctions noyaux sont des polynômes de degré fini. Pour tout $\lambda \geq 0$, définissons

$$d_\lambda^*(\Sigma_{J^c}^{-1/2}B_{\mathcal{H}}) := \frac{\mathrm{Tr}(\Sigma_{J^c}) + \lambda}{\|\Sigma_{J^c}\|_{\mathrm{op}}}. \tag{1.25}$$

**Hypothèse 2.** *Il existe des constantes absolues $C_1 > 1$, $C_2 > 1$, $0 \leq \gamma < 1/16$, $0 \leq \delta < 1/(100\sqrt{C_2})$, $\bar{\delta} < C_1^{-1}$, $\epsilon > 0$ et $\kappa > 1$ telles que*

- *Avec une probabilité d'au moins $1 - \gamma$,*

$$\max_{1\leq i\leq N}\left|\frac{\|\phi_{J^c}(X_i)\|_{\mathcal{H}}^2}{(\ell^*)^2} - 1\right| \leq \delta, \tag{1.26}$$

  *où nous définissons $\ell^* = \sqrt{\mathbb{E}\,\|\phi_{J^c}(X)\|_{\mathcal{H}}^2} = \sqrt{\mathrm{Tr}\left(\Sigma_{J^c}\right)}$.*

- *Pour tout $f \in V_{J^c}$, nous avons*

$$\|f\|_{L^{2+\epsilon}(\mu_X)} \leq \kappa\,\|f\|_{L^2(\mu_X)}. \tag{1.27}$$

- *Selon le choix de $\epsilon$, il y a deux cas :*

  1. *si $\epsilon > 2$, alors aucune hypothèse supplémentaire n'est requise.*
  2. *si $0 < \epsilon \leq 2$, alors*

$$\kappa N^{\frac{2-\epsilon}{2\epsilon+\epsilon^2}}\log(N)\left(\sqrt{\frac{N\|\Sigma_{J^c}\|_{\mathrm{op}}}{\mathrm{Tr}\left(\Sigma_{J^c}\right)}}\right) < \bar{\delta}. \tag{1.28}$$

Un exemple typique satisfaisant l'Hypothèse 2 est lorsque $\phi_{J^c}$ est l'application identité et $X_i$ est un vecteur aléatoire sous-gaussien ; dans ce cas, le résultat découle de l'inégalité de Hanson-Wright.

**Théorème 4.** *Soit $X$ un vecteur aléatoire distribué selon $\mu_X$ dans un ensemble compact $\Omega_X \subset \mathbb{R}^d$, et soient $X_1, \cdots, X_N$ des copies i.i.d. de $X$. Soit $\phi : \boldsymbol{x} \in \Omega_X \mapsto K(\boldsymbol{x}, \cdot) \in \mathcal{H}$ la carte de caractéristiques du RKHS $\mathcal{H}$. Soient $C_3$, $C_4$, $C_5$ et $C_6$ des constantes absolues.*

1. *Supposons que* $\lambda \le C_3 \operatorname{Tr}(\Sigma_{J^c})$. *Considérons* $0 < \delta, \bar{\delta} < 1$ *de l'Hypothèse 2, définissons*

$$\tilde{\delta} = C_2\delta^2 + C_4\bar{\delta}^2 + 4\sqrt{(3\delta + C_5\bar{\delta})\left(1 + \delta + C_6\bar{\delta}\right)} \tag{1.29}$$

*Supposons que pour un certain* $\lambda \ge 0$, *nous ayons* $N \le \kappa_{DM}\bar{\delta}^2 d_\lambda^*\left(\Sigma_{J^c}^{-1/2} B_{\mathcal{H}}\right)$ *pour une constante suffisamment petite* $\kappa_{DM} < 1$ *qui ne dépend que de* $\kappa$. *Nous supposons que* $\phi_{J^c}$ *satisfait l'Hypothèse 2. Alors avec une probabilité d'au moins*

$$1 - \gamma - \frac{1}{N^2} - \left(\frac{\kappa}{\bar{\delta}}\right)^{2+\epsilon}\left(\sqrt{\frac{N\|\Sigma_{J^c}\|_{\mathrm{op}}}{\operatorname{Tr}(\Sigma_{J^c})}}\right)^{2+\epsilon}\frac{\log^{2+\epsilon}(N)}{N^{\frac{\epsilon}{2}-1}} =: 1 - \bar{p}_{DM},$$

*pour tout* $\boldsymbol{\lambda} \in \mathbb{R}^N$, *il est vérifié que*

$$\left(1 - \tilde{\delta}\right)\sqrt{\operatorname{Tr}(\Sigma_{J^c})}\,\|\boldsymbol{\lambda}\|_2 \le \left\|\mathbb{X}_{J^c}^\top \boldsymbol{\lambda}\right\|_{\mathcal{H}} \le \left(1 + \tilde{\delta}\right)\sqrt{\operatorname{Tr}(\Sigma_{J^c})}\,\|\boldsymbol{\lambda}\|_2 . \tag{1.30}$$

2. *Supposons que* $\lambda > C_3 \operatorname{Tr}(\Sigma_{J^c})$. *Supposons que* $\phi_{J^c}$ *satisfait les deux premiers points de l'Hypothèse 2. Supposons que pour un certain* $\lambda \ge 0$, *nous ayons* $N \le (\kappa_{DM}/4) d_\lambda^*(\Sigma_{J^c}^{-1/2} B_{\mathcal{H}})$. *Alors il existe des constantes absolues* $C_7$ *dépendant de* $\epsilon, \kappa, \kappa_{DM}$, *et* $0 < c_2 < 1$ *telles qu'avec une probabilité d'au moins*

$$1 - \gamma - N\left(\left(\frac{\kappa_{DM}\kappa^2\log^2(N)}{N}\right)^{1+\epsilon/2} N\right)^{\lceil (12+2\epsilon)/\epsilon\rceil - 1} - \frac{1}{N^2} =: 1 - \bar{p}_{DM},$$

*nous ayons* $\left\|\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + \lambda I_N\right\|_{\mathrm{op}} \le C_7\lambda + \operatorname{Tr}(\Sigma_{J^c})$ *et*

$$\sigma_N(\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top + \lambda I_N) \ge c_2\lambda + (1 - c_2)C_3 \operatorname{Tr}(\Sigma_{J^c}) .$$

**Proposition 11** (informel)**.** *Supposons que* $\mathcal{F}$ *est identifié avec des fonctionnelles linéaires sur* $\mathbb{R}^p$. *Si* $\mathbb{R}^p = V_J \oplus V_{J^c}$ *est une FSD satisfaisant les propriétés suivantes : 1.* $YX_{J^c}$ *est un vecteur aléatoire sous-gaussien centré ; 2.* $N \le \kappa_{\mathrm{DM}}\bar{\delta}^2 \frac{\operatorname{Tr}(\Sigma_{J^c})}{\|\Sigma_{J^c}\|_{op}}$, *où* $\Sigma_{J^c} = \mathbb{E}[X_{J^c} \otimes X_{J^c}]$. *Alors avec une grande probabilité la fonction de perte* $L_{\boldsymbol{\beta}_J}$ *définie dans la Proposition 7 possède la propriété suivante : pour tout* $\boldsymbol{\beta}_J \in V_J$,

$$\frac{N}{(1+\tilde{\delta})^2 \operatorname{Tr}(\Sigma_{J^c})} P_N \ell_{\boldsymbol{\beta}_J}^{(\mathrm{sh})} \le L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N) \le \frac{N}{(1-\tilde{\delta})^2 \operatorname{Tr}(\Sigma_{J^c})} P_N \ell_{\boldsymbol{\beta}_J}^{(\mathrm{sh})} \le L_{\boldsymbol{\beta}_J}((X_i, Y_i)_{i=1}^N),$$

*où* $\tilde{\delta}$ *vient du Théorème 4, et* $P_N\ell_{\boldsymbol{\beta}_J}^{(\mathrm{sh})} = \frac{1}{N}\sum_{i=1}^N \ell_{\boldsymbol{\beta}_J}^{(\mathrm{sh})}(X_i, Y_i)$, *et* $\ell_{\boldsymbol{\beta}_J}^{(\mathrm{sh})}(\boldsymbol{x}, y) = (1 - y\langle \boldsymbol{\beta}_J, \boldsymbol{x}\rangle)_+^2$ *est la perte charnière quadratique. De plus, pour tout problème de classification binaire* $(\mu_X, \eta)$, *si* $f^{**} = \arg\min(P\ell_f^{(\mathrm{sq})} : f$ *est mesurable*), *alors* $f^{**} = f^*$, *c'est-à-dire la règle de Bayes.*

Dans la proposition, puisque $YX_{J^c}$ est un vecteur aléatoire sous-gaussien, l'Hypothèse 2 est naturellement vérifiée, et le Théorème 4 s'applique donc.

**Proof.** En appliquant le Théorème 4 à $\phi_{J^c}(X) = YX_{J^c}$ et à $\mathcal{H} = \mathbb{R}^p$, nous avons seulement besoin de prouver que l'inclusion suivante est vraie

$$\Omega_{\mathrm{DM,class}}(\tilde{\delta}) := \left\{\forall \boldsymbol{\lambda} \in \mathbb{R}^N : \|\boldsymbol{\lambda}\|_2 (1 - \tilde{\delta})\sqrt{\operatorname{Tr}(\Sigma_{J^c})} \le \left\|\mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}\right\|_2 \le \|\boldsymbol{\lambda}\|_2 (1 + \tilde{\delta})\sqrt{\operatorname{Tr}(\Sigma_{J^c})}\right\} \tag{1.31}$$

$$\subseteq \left\{\forall \boldsymbol{\mu} \in \mathbb{R}^N : \frac{\|[\boldsymbol{\mu}]_+\|_2}{(1+\tilde{\delta})\sqrt{\operatorname{Tr}(\Sigma_{J^c})}} \le \|\mathcal{B}[\boldsymbol{\mu}]\|_2 \le \frac{\|[\boldsymbol{\mu}]_+\|_2}{(1-\tilde{\delta})\sqrt{\operatorname{Tr}(\Sigma_{J^c})}}\right\}, \tag{1.32}$$

où $[\boldsymbol{\mu}]_+ = (\max(\mu_i, 0))_{i=1}^N$. Par un argument de dualité standard, voir, par exemple, [BV14, Équation 5.11], nous obtenons que

$$\|\mathcal{B}[\boldsymbol{\mu}]\|_2 = \max\left(\langle \boldsymbol{\mu}, \boldsymbol{\lambda}\rangle : \boldsymbol{\lambda} \succeq \mathbf{0},\ \left\|\mathbb{X}_{\boldsymbol{y},J^c}^\top \boldsymbol{\lambda}\right\|_2 \le 1\right). \tag{1.33}$$

Conditionnellement à $\Omega_{\mathrm{DM,class}}(\tilde{\delta})$, voir (1.31), nous avons

$$\max_{\boldsymbol{\lambda}\succeq\mathbf{0}}\left(\langle\boldsymbol{\mu},\boldsymbol{\lambda}\rangle:\|\boldsymbol{\lambda}\|_2\leq\frac{1}{(1+\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}}\right)\leq\|\mathcal{B}[\boldsymbol{\mu}]\|_2\leq\max_{\boldsymbol{\lambda}\succeq\mathbf{0}}\left(\langle\boldsymbol{\mu},\boldsymbol{\lambda}\rangle:\|\boldsymbol{\lambda}\|_2\leq\frac{1}{(1-\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}}\right).$$

Soit $H(\boldsymbol{\mu}):=\{i\in[N]:\mu_i<0\}$ et soit $\boldsymbol{\lambda}^-$ le maximiseur du problème de maximisation du côté gauche et $\boldsymbol{\lambda}^+$ le maximiseur du problème de maximisation du côté droit. Nous prouvons que si $i\in H(\boldsymbol{\mu})$, alors $\lambda_i^-=0$. Nous le prouvons par contradiction. Supposons que $i\in H(\boldsymbol{\mu})$ mais $\lambda_i^->0$, alors en posant $\tilde{\boldsymbol{\lambda}}^-=(\lambda_1^-,\cdots,\lambda_{i-1}^-,0,\lambda_{i+1}^-,\cdots,\lambda_N^-)$, nous savons que $\|\tilde{\boldsymbol{\lambda}}^-\|_2<\|\boldsymbol{\lambda}^-\|_2\leq\frac{1}{(1+\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}}$. De plus, $\langle\boldsymbol{\mu},\tilde{\boldsymbol{\lambda}}^-\rangle=\sum_{i'\neq i}\mu_{i'}\lambda_{i'}^->\sum_{i'=1}^N\mu_{i'}\lambda_{i'}^-=\langle\boldsymbol{\mu},\boldsymbol{\lambda}^-\rangle$ puisque $\mu_i\lambda_i^-<0$. Cela implique que $\tilde{\boldsymbol{\lambda}}^-$ est une solution réalisable mais avec une plus grande valeur de fonction objectif, contredisant ainsi l'hypothèse que $\boldsymbol{\lambda}^-$ est le maximiseur. En rappelant la contrainte que $\boldsymbol{\lambda}\succeq\mathbf{0}$, nous avons : pour tout $i\in H(\boldsymbol{\mu})$, nous avons nécessairement $\lambda_i^-=0$. La même chose vaut également pour $\boldsymbol{\lambda}^+$. Maintenant, par Cauchy-Schwartz, nous avons $\boldsymbol{\lambda}^-=(\boldsymbol{\mu}/(\|\boldsymbol{\mu}\|_2(1+\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}))_+$, et $\boldsymbol{\lambda}^+=(\boldsymbol{\mu}/(\|\boldsymbol{\mu}\|_2(1-\tilde{\delta})\sqrt{\mathrm{Tr}(\Sigma_{J^c})}))_+$. Par conséquent, conditionnellement à $\Omega_{\mathrm{DM,class}}(\tilde{\delta})$, (1.32) s'ensuit. Pour les preuves des propriétés de la perte charnière quadratique, voir la Section 4.9.3. ■

### 6.2.6 Énergie du $\hat{f}_{J^c}$.

La proposition suivante montre que $\hat{f}_{J^c}$ est une RERM du résidu de $\hat{f}_J$.

**Proposition 12.** *Supposons qu'il existe une fonction différentiable $L:\mathbb{R}^N\to\mathbb{R}$ telle que pour tout $f\in\mathcal{F}$, $L_f:(X_i,Y_i)_{i=1}^N\in\Omega^N\mapsto L(\boldsymbol{y}-\mathbb{X}f)$, où $\mathbb{X}:f\in\mathcal{F}\mapsto(f(X_i))_{i=1}^N$ et $\boldsymbol{y}=(Y_1,\cdots,Y_N)$. Soit $\lambda\in\mathbb{R}$ un nombre réel quelconque, soit $\Psi$ une fonction différentiable et supposons que $\mathcal{F}$ est un espace linéaire. Définissons*

$$\hat{f}_N\in\operatorname*{argmin}_{f\in\mathcal{F}}\left(L(\boldsymbol{y}-\mathbb{X}f)+\lambda\Psi(f)\right)$$

*comme une RERM. Soit $\mathcal{F}=V_J\oplus V_{J^c}$ une FSD. Supposons que $\Psi:\mathcal{F}\to\mathbb{R}$ est décomposable par rapport à $V_J\oplus V_{J^c}$, au sens où pour tout $f\in\mathcal{F}$, $\Psi(f)=\Psi(f_J)+\Psi(f_{J^c})$. Alors*

$$\hat{f}_{J^c}\in\operatorname*{argmin}_{f_{J^c}\in V_{J^c}}\left(L(\boldsymbol{y}'-\mathbb{X}f_{J^c})+\lambda\Psi(f_{J^c})\right),\ \text{où}\ \boldsymbol{y}'=\boldsymbol{y}-\mathbb{X}\hat{f}_J.\tag{1.34}$$

Ici, $\hat{f}_{J^c}$ apprend le résidu de $\hat{f}_J$, en utilisant $\mathbb{X}_{J^c}$ qui ne sont pas nécessairement isomorphes à $\Sigma_{J^c}^{1/2}$, voir la Figure 1.1.

**Proof.** Par hypothèse, nous pouvons développer l'application $(f_J,f_{J^c})\in V_J\times V_{J^c}\mapsto L(\boldsymbol{y}-\mathbb{X}f)+\lambda\Psi(f)$ sous la forme $(f_J,f_{J^c})\in V_J\times V_{J^c}\mapsto L((\boldsymbol{y}-\mathbb{X}f_J)-\mathbb{X}f_{J^c})+\lambda\Psi(f_J)+\lambda\Psi(f_{J^c})$. Fixons maintenant $f_J=\hat{f}_J=P_{V_J}\hat{f}_N$. Nous savons que $\hat{f}$ doit satisfaire la condition d'optimalité du premier ordre, c'est-à-dire que le gradient de cette application est égal à $\mathbf{0}$,

$$\mathbf{0}=-2P_{V_{J^c}}\mathbb{X}^\top\nabla L((\boldsymbol{y}-\mathbb{X}\hat{f}_J)-\hat{f}_{J^c})+\lambda(\nabla\Psi)(\hat{f}_{J^c}).$$

C'est précisément la condition nécessaire du premier ordre du problème d'optimisation défini dans (1.34). ■

Dans le problème de régression à valeurs réelles (Exemple 1), $\boldsymbol{y}-\mathbb{X}_J\hat{f}_J=\boldsymbol{\xi}+\mathbb{X}_{J^c}f_{J^c}^*+\mathbb{X}_J(\hat{f}_J-f_J^*)$.

**Estimateurs linéaires et estimateurs interpolants de norme minimale.** Ensuite, nous considérons une borne supérieure pour $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$—un aspect de la méthode FSD que nous avons pu étudier dans 4 cas, mais qui nécessite une compréhension plus approfondie. En fait, à l'heure actuelle, nous avons une compréhension relativement complète uniquement lorsque $\hat{f}_{J^c}$ est un opérateur linéaire en $\boldsymbol{y}$ dans le problème de régression linéaire. Ci-dessous, nous l'illustrons en utilisant deux classes d'opérateurs linéaires—la régression ridge et les méthodes spectrales—et nous concluons en discutant du cas où $\hat{f}_{J^c}$ est un estimateur interpolant de norme minimale, un opérateur non linéaire comparativement simple.

Pour les opérateurs linéaires en régression linéaire, c'est-à-dire, lorsqu'il existe un opérateur linéaire aléatoire $A:\mathbb{R}^N\to V_{J^c}$ indépendant de $(Y_i)_{i=1}^N$ tel que $\hat{f}_{J^c}:(X_i,Y_i)_{i=1}^N\in\Omega^N\mapsto\langle\cdot,A\boldsymbol{y}\rangle\in V_{J^c}$, nous identifions $\hat{f}_{J^c}$ comme $\hat{\boldsymbol{\beta}}_{J^c}(\boldsymbol{y})=A\boldsymbol{y}$. Alors, dans le modèle de régression additif, $\boldsymbol{y}=\mathbb{X}\boldsymbol{\beta}^*+\boldsymbol{\xi}$, et par conséquent $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}\leq$

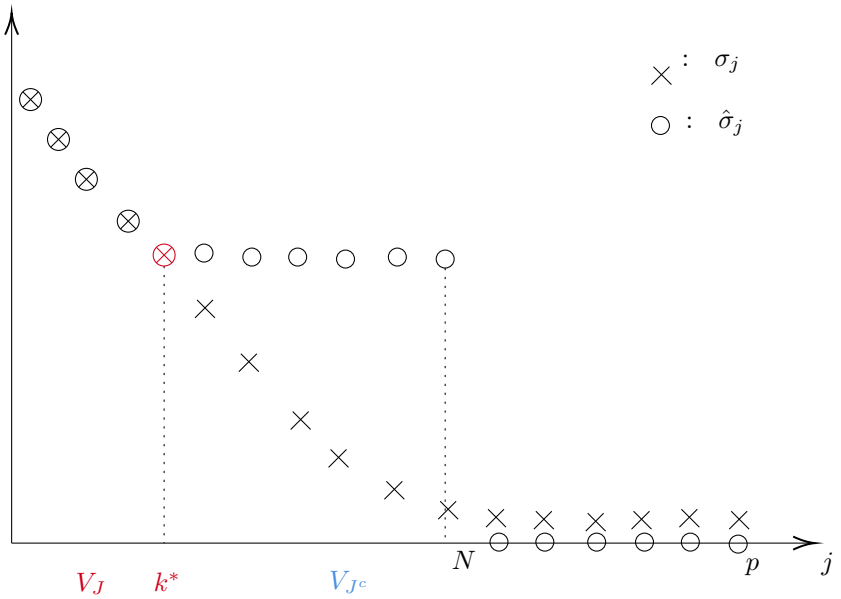


Figure 6.1: D'après le Théorème 4, le spectre de $\frac{1}{N}\mathbb{X}^\top\mathbb{X} + t^{-1}I$ a un plateau de hauteur $\frac{1}{N}\operatorname{Tr}(\Sigma_{J^c}) + t^{-1}$ qui est différent de $\operatorname{Spec}(\Sigma_{J^c}) = \{\sigma_j : j \in J^c\}$. C'est la raison pour laquelle aucune estimation n'est possible sur $V_{J^c}$.

$\|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\mathbb{X}\boldsymbol{\beta}^*_J)\rangle\|_{L^2(\mu_X)} + \|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\mathbb{X}\boldsymbol{\beta}^*_{J^c})\rangle\|_{L^2(\mu_X)} + \|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\boldsymbol{\xi})\rangle\|_{L^2(\mu_X)}$. Pour le terme $\|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\boldsymbol{\xi})\rangle\|_{L^2(\mu_X)}$, nous avons $\mathbb{E}_{\boldsymbol{\xi}}\|\langle X, \hat{\boldsymbol{\beta}}_{J^c}(\boldsymbol{\xi})\rangle\|^2_{L^2(\mu_X)} = \sigma_\xi^2 \operatorname{Tr}(A^\top \Sigma_{J^c} A)$, où $\Sigma_{J^c} = P_{J^c}\Sigma P_{J^c}$. C'est précisément la stratégie que nous adoptons lors de l'analyse des méthodes spectrales.

**Méthodes spectrales.** Rappelons les méthodes spectrales définies dans l'Exemple 9. Nous savons que les méthodes spectrales sont des estimateurs linéaires, et dans ce cas $A = \frac{1}{N}\varphi(\hat{\Sigma})\mathbb{X}^\top$. Par conséquent,

$$\|\hat{f}_{J^c}\|_{L^2(\mu_X)} \le \|P_{J^c}\varphi_t(\hat{\Sigma})\hat{\Sigma}f^*_J\|_{L^2(\mu_X)} + \|P_{J^c}\varphi_t(\hat{\Sigma})\hat{\Sigma}f^*_{J^c}\|_{L^2(\mu_X)} + \|P_{J^c}\varphi_t(\hat{\Sigma})[N^{-1}\mathbb{X}^\top]\boldsymbol{\xi}\|_{L^2(\mu_X)}.$$

Nous ne continuons pas à montrer le traitement ultérieur de ces termes ici ; voir [P5].

**Régression ridge.** Un cas encore plus spécial se produit lorsque $\hat{f}_N$ est la régression ridge, c'est-à-dire, lorsque $\mathcal{F}$ est identifié avec un certain RKHS $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$, $L : \boldsymbol{u} \in \mathbb{R}^N \mapsto \frac{1}{N}\|\boldsymbol{u}\|_2^2$, et $\Psi : f \in \mathcal{H} \mapsto t^{-1}\|f\|^2_{\mathcal{H}}$. Dans ce cas, non seulement nous savons que $\hat{f}_{J^c}$ est un opérateur linéaire, mais aussi que $\hat{f}_{J^c}$ est une régression ridge avec paramètre $t^{-1}$ appliquée à $\boldsymbol{y} - \mathbb{X}_J\hat{f}_J$. En fait, la Proposition 12 montre que pour toute FSD, la régression ridge $\hat{f}_N$ avec paramètre de réglage $t^{-1}$ satisfait

$$\hat{f}_{J^c} = \frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}(\boldsymbol{y} - \mathbb{X}_J\hat{f}_J). \tag{1.35}$$

À partir de (1.35), nous savons que $\hat{f}_{J^c}$ est un opérateur linéaire sur $\boldsymbol{y} - \mathbb{X}_J\hat{f}_J$, et puisque $\boldsymbol{y} = \mathbb{X}_J\boldsymbol{\beta}^*_J + \mathbb{X}_{J^c}\boldsymbol{\beta}^*_{J^c} + \boldsymbol{\xi}$, l'inégalité triangulaire donne

$$\begin{aligned}\|\hat{f}_{J^c}\|_{L^2(\mu_X)} \le{}& \left\|\frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}\mathbb{X}_J(f^*_J - \hat{f}_J)\right\|_{L^2(\mu_X)} \\ &+ \left\|\frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}\mathbb{X}_{J^c}f^*_{J^c}\right\|_{L^2(\mu_X)} + \left\|\frac{1}{N}\mathbb{X}^\top_{J^c}\left(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N\right)^{-1}\boldsymbol{\xi}\right\|_{L^2(\mu_X)}.\end{aligned} \tag{1.36}$$

Parmi ces trois termes, les deux premiers ne nécessitent que des normes d'opérateurs, tandis que le dernier terme $\|\frac{1}{N}\mathbb{X}^\top_{J^c}(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}^\top_{J^c} + t^{-1}I_N)^{-1}\boldsymbol{\xi}\|_{L^2(\mu_X)}$ est le plus particulier ; nous devons exploiter le caractère aléatoire de $\boldsymbol{\xi}$. La proposition suivante est appelée le « côté supérieur » du théorème de Dvoretzky-Milman ; sa preuve peut être trouvée dans [P2].

**Hypothèse 3.** *Il existe des constantes absolues $\gamma_1 \in \left(0, \frac{1}{16}\right)$, $\delta_1 \ge 0$, $\epsilon > 0$ et $\kappa' > 1$ telles que*

- $$\mathbb{P}\left(\max_{1\le i\le N}\frac{\left\|\Sigma_{J^c}^{1/2}\phi_{J^c}(X_i)\right\|_{\mathcal{H}}^2}{\mathrm{Tr}\left(\Sigma_{J^c}^2\right)}\le 1+\delta_1\right)\ge 1-\gamma_1, \tag{1.37}$$

- *pour tout* $f\in V_{J^c}$, $\|f\|_{L^{4+\epsilon}(\mu_X)}\le\kappa'\|f\|_{L^2(\mu_X)}$.

**Proposition 13.** *Supposons que l'Hypothèse 3 soit vérifiée. Il existe des constantes absolues* $c_3$ *et* $C_8>0$ *telles qu'avec une probabilité d'au moins* $1-\bar{p}_{DMU}$, *où* $\bar{p}_{DMU}=\frac{c_3}{N^\epsilon}+\gamma_1$, *on ait* $\left\|\Sigma_{J^c}^{1/2}\mathbb{X}_{J^c}^\top\right\|_{op}\le C\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)}+C\sqrt{N}\|\Sigma_{J^c}\|_{op}$.

Par conséquent, si la Proposition 13 et le Théorème 4 sont vérifiés, alors à partir de (1.35) nous savons qu'il existe une constante absolue $C$ telle que sur l'intersection des deux événements aléatoires ce qui suit est vérifié :

$$\begin{aligned}\left\|\langle X,\frac{1}{N}\mathbb{X}_{J^c}^\top(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top+t^{-1}I_N)^{-1}\mathbb{X}_J(f_J^*-\hat{f}_J)\rangle\right\|_{L^2(\mu_X)}&\le C\frac{\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)}+\sqrt{N}\|\Sigma_{J^c}\|_{\mathrm{op}}}{Nt^{-1}+\mathrm{Tr}(\Sigma_{J^c})}\|\mathbb{X}_J(\hat{\boldsymbol{\beta}}_J-\boldsymbol{\beta}_J^*)\|_2 \text{ et}\\ \left\|\langle X,\frac{1}{N}\mathbb{X}_{J^c}^\top(\frac{1}{N}\mathbb{X}_{J^c}\mathbb{X}_{J^c}^\top+t^{-1}I_N)^{-1}\mathbb{X}_{J^c}f_{J^c}^*\rangle\right\|_{L^2(\mu_X)}&\le C\frac{\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)}+\sqrt{N}\|\Sigma_{J^c}\|_{\mathrm{op}}}{Nt^{-1}+\mathrm{Tr}(\Sigma_{J^c})}\|\mathbb{X}_{J^c}f_{J^c}^*\|_2.\end{aligned} \tag{1.38}$$

**Estimateur interpolant de norme $\|\cdot\|_q$ minimale.** Rappelons que lorsque $\hat{\boldsymbol{\beta}}$ est l'estimateur interpolant de norme $\|\cdot\|_q$ minimale (Exemple 10), la Proposition 6 stipule : si $(V_J,V_{J^c})$ est une FSD de la forme définie dans la Proposition 6, alors $\hat{\boldsymbol{\beta}}_{J^c}=\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\hat{\boldsymbol{\beta}}_J]$. Cela confère à $\hat{\boldsymbol{\beta}}_{J^c}$ une signification statistique : $\hat{\boldsymbol{\beta}}_{J^c}$ est l'estimateur interpolant de norme $\|\cdot\|_q$ minimale du résidu de $\hat{\boldsymbol{\beta}}_J$. Combinée avec le théorème de Dvoretzky-Milman, cette interprétation statistique nous permet d'obtenir le contrôle suivant sur la borne supérieure pour $\|\langle X,\hat{\boldsymbol{\beta}}_{J^c}\rangle\|_{L^2(\mu_X)}$.

**Proposition 14.** *En utilisant la notation de la Proposition 6. Si* $N\le\kappa_{\mathrm{DM}}\varepsilon^2 d_*(\Sigma_{J^c}^{1/2}B_{q'}^p)$, *alors il existe une constante absolue* $C$ *telle que sur l'événement aléatoire* $\Omega_{DM,reg}(\varepsilon)$, *nous avons*

$$\|\langle X,\hat{\boldsymbol{\beta}}_{J^c}\rangle\|_{L^2(\mu_X)}\le C\frac{\mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)}{\ell_*(\Sigma_{J^c}^{1/2}B_{q'}^p)}\|\boldsymbol{y}-\mathbb{X}_J\hat{\boldsymbol{\beta}}_J\|_2.$$

**Proof.** À partir de la Proposition 6, nous avons $\hat{\boldsymbol{\beta}}_{J^c}=\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\hat{\boldsymbol{\beta}}_J]$, et donc sous la condition de Dvoretzky-Milman $N\le\kappa_{DM}\varepsilon^2 d_*(\Sigma_{J^c}^{1/2}B_{q'}^p)$ vérifiée, puisque $\|\Sigma_{J^c}^{1/2}\|_{\ell_q\to\ell_2}=\mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)$, sur $\Omega_{\mathrm{DM,reg}}(\varepsilon)$ par (1.23), on a $\|\Sigma_{J^c}^{1/2}\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\hat{\boldsymbol{\beta}}_J]\|_2\le\mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)\|\mathcal{A}[\boldsymbol{y}-\mathbb{X}_J\hat{\boldsymbol{\beta}}_J]\|_q\le C\frac{\mathrm{diam}(\Sigma_{J^c}^{1/2}B_q^p)}{\ell_*(\Sigma_{J^c}^{1/2}B_{q'}^p)}\|\boldsymbol{y}-\mathbb{X}_J\hat{\boldsymbol{\beta}}_J\|_2$. ∎

**Classifieur interpolant de norme $\|\cdot\|_2$ minimale.** Pour les problèmes de classification binaire, nous avons une compréhension plus unifiée de l'énergie de $\hat{f}_{J^c}$. La proposition suivante est prouvée dans [P1], voir également la Section 4.8.3.

**Proposition 15.** *Soit* $\mathcal{F}=V_J\oplus V_{J^c}$ *une FSD quelconque et* $\hat{f}_N$ *un estimateur quelconque. Alors,* $\mu^{\otimes N}$*-p.s.,*

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

**LASSO avec récupération du support.** Comme autre exemple d'estimateur non linéaire, nous considérons dans ce paragraphe le cas où $\hat{\boldsymbol{\beta}}$ est l'estimateur LASSO, c'est-à-dire, $\hat{\boldsymbol{\beta}}\in\mathrm{argmin}(\frac{1}{2N}\|\boldsymbol{y}-\mathbb{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1)$. Soit $\boldsymbol{\beta}^*\in\mathbb{R}^p$ un vecteur inconnu et notons son support par $S=\mathrm{supp}(\boldsymbol{\beta}^*)$, c'est-à-dire, $S=\{j\in[p]:\langle\boldsymbol{\beta}^*,\boldsymbol{e}_j\rangle\neq 0\}$, où nous rappelons que $\boldsymbol{e}_1,\dots,\boldsymbol{e}_p$ est la base canonique de $\mathbb{R}^p$. Soit $s=|S|$ et supposons que $s\le\lfloor c\frac{N}{\log(p/N)}\rfloor$ pour une certaine constante $c<1$. Pour mettre en évidence la FSD et éviter d'être distrait par des arguments stochastiques, nous travaillons sur l'événement stochastique suivant

$$\Omega_{\mathrm{LASSO}}=\left\{\mathrm{supp}(\hat{\boldsymbol{\beta}})=S,\ \left\|\frac{1}{N}\mathbb{X}_S\mathbb{X}_S^\top\right\|_{\mathrm{op}}\le 10,\ \left\|\frac{1}{N}\mathbb{X}_S^\top\boldsymbol{\xi}\right\|_2\le\sigma_\xi\sqrt{\frac{2s}{N}}\right\},$$

où $\mathbb{X}_S = [X_{1,S}|\cdots|X_{N,S}]^\top \in \mathbb{R}^{N\times s}$ est la restriction de $\mathbb{X}$ à $S$, et $X_{i,S}$ est la restriction de $X_i$ à $S$. Lorsque l'événement $\operatorname{supp}(\hat{\boldsymbol{\beta}}) = S$ se produit, nous disons que $\hat{\boldsymbol{\beta}}$ atteint la récupération du support. Des conditions suffisantes pour cet événement ont été largement étudiées dans la théorie du LASSO, par exemple dans [Gir14, Section 5.5.2]. L'avantage de travailler sur cet événement stochastique est que, si l'on pose $J = S$ et $V_J = \operatorname{span}(\{\boldsymbol{e}_j : j \in S\})$, nous avons $\hat{\boldsymbol{\beta}}_{J^c} = \boldsymbol{\beta}^*_{J^c} = \mathbf{0}$, ce qui élimine le besoin de considérer l'énergie de $\hat{\boldsymbol{\beta}}_{J^c}$. Le cas où la récupération du support ne se produit pas nécessairement reste une direction particulièrement intéressante pour les recherches futures. Dans ce cas, nous avons la proposition suivante.

**Proposition 16.** *Supposons que* $\lambda > \sigma_\xi\sqrt{\frac{\log(ep/s)}{N}}$ *et* $p > e^7 s$. *Alors, sur* $\Omega_{\text{LASSO}}$, *nous avons* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \geq \frac{1}{20}\sigma_\xi\sqrt{\frac{s\log(ep/s)}{N}}$.

**Proof.** Par la définition de $\hat{\boldsymbol{\beta}}$ et les conditions KKT, nous avons $\frac{1}{N}\mathbb{X}_S^\top(\mathbb{X}_S\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda\operatorname{sign}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. En substituant $\boldsymbol{y} = \mathbb{X}_S\boldsymbol{\beta}^*_S + \boldsymbol{\xi}$ et en utilisant le fait que $\operatorname{sign}(\hat{\boldsymbol{\beta}}) = \operatorname{sign}(\boldsymbol{\beta}^*_S)$, nous obtenons $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*_S = (\frac{1}{N}\mathbb{X}_S^\top\mathbb{X}_S)^{-1}[\lambda\operatorname{sign}(\boldsymbol{\beta}^*_S) - \frac{1}{N}\mathbb{X}_S^\top\boldsymbol{\xi}]$. En prenant la norme $\ell_2$ des deux côtés et en appliquant l'inégalité triangulaire, nous avons

$$\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\right\|_2 \geq \frac{1}{10}\left(\lambda\left\|\operatorname{sign}(\boldsymbol{\beta}^*_S)\right\|_2 - \sigma_\xi\sqrt{\frac{2s}{N}}\right). \tag{1.39}$$

Étant donné que $\|\operatorname{sign}(\boldsymbol{\beta}^*_S)\|_2 = \sqrt{s}$ et l'hypothèse $\lambda > \sigma_\xi\sqrt{\frac{\log(ep/s)}{N}}$, la preuve est complète. ∎

La Proposition 16 indique que lorsque la récupération du support (support recovery) a lieu, la borne d'erreur d'estimation du LASSO est optimale par instance pour *ce* $\boldsymbol{\beta}^*$ spécifique. Plus précisément, si $X$ est un vecteur aléatoire isotrope, alors $(V_J^*, V_{J^c}^*)$ est donné par $V_J^* = \operatorname{span}(\{\boldsymbol{e}_j : j \in S\})$, et dans ce cas $\|\langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle\|_{L^2(\mu_X)} \sim r(V_J^*, V_{J^c}^*)$, où $r(V_J^*, V_{J^c}^*) = \sigma_\xi\sqrt{\frac{s\log(ep/s)}{N}}$ avec une grande probabilité. Ce résultat est plus fort que l'optimalité minimax, car il n'affirme pas simplement l'existence d'un certain $\boldsymbol{\beta}^*$ pour lequel l'erreur d'estimation n'est pas inférieure à cette borne ; au lieu de cela, la borne inférieure est valable pour tout $\boldsymbol{\beta}^*$ satisfaisant la condition de récupération du support. En fait, à partir de (1.39), nous pouvons observer que $\lambda\|\operatorname{sign}(\boldsymbol{\beta}^*_S)\|_2$ et $\sigma_\xi\sqrt{\frac{2s}{N}}$ correspondent respectivement aux termes de biais et de variance (rappelons que $s = \dim(V_J^*)$) du sous-espace d'estimation dans la fonction de taux de la FSD (cf. (1.41) ci-dessous). Ici, puisque les projections de $\hat{\boldsymbol{\beta}}$ et de $\boldsymbol{\beta}^*$ sur le sous-espace libre sont toutes deux $\mathbf{0}$, il n'y a pas de termes de biais ou de variance provenant du sous-espace libre dans la fonction de taux.

**Direction de recherche.** Lorsque $\hat{f}_{J^c}$ est un estimateur non linéaire — comme dans le cas où $\hat{f}_{J^c}$ est l'estimateur d'interpolation de norme $\|\cdot\|_q$ minimale, le LASSO, ou un RERM général — comment pouvons-nous développer une boîte à outils mathématique systématique qui nous permette d'obtenir une borne supérieure fine avec grande probabilité pour $\|\hat{f}_{J^c}\|_{L^2(\mu_X)}$ ? Ici, une méthode systématique désigne une méthode qui fonctionne pour un $\mu_X$ général et un $\hat{f}_{J^c}$ général, et qui, lorsqu'elle est spécialisée à la régression ridge et aux méthodes spectrales, permet de retrouver les bornes supérieures fines qui ont été obtenues en utilisant le fait que $\hat{f}_{J^c}$ est un estimateur linéaire, [P4]. Pour la (R)ERM, cela constitue une classe de problèmes qui n'a jamais été explorée ; cela pourrait stimuler le développement de nouveaux outils géométriques.

## 6.3 La FSD comme cadre théorique

À travers la discussion de la Section 1.5.1, de la Section 1.5.2, et de la Section 1.5.3, nous avons expliqué le rôle de la FSD en tant que méthode analytique. Dans cette section, nous soutenons en outre que la méthode FSD sert également de cadre théorique pour comprendre comment une solution s'attaque à un problème d'apprentissage supervisé spécifique, offrant aux théoriciens une nouvelle perspective potentielle et une nouvelle façon de penser.

Dans cette section, nous nous concentrons sur les situations où une FSD optimale $(V_{J_*}, V_{J_*^c})$ peut être construite, de sorte que (1.15) puisse être prouvée. Dans de tels cas, la méthode FSD révèle comment un estimateur emploie réellement l'espace des caractéristiques pour l'estimation—quelles caractéristiques l'estimateur utilise lors de la résolution du problème—et comment il utilise $V_{J_*^c}$ pour traiter le signal et le bruit. Cela diffère de l'objectif classique de la théorie de l'apprentissage statistique consistant à « établir une inégalité d'oracle qui correspond à la borne inférieure minimax » comme dans la Section 1.1 ; au lieu de cela, il se concentre davantage sur la compréhension,

d'un point de vue mathématique, de la relation d'adéquation entre la solution et le problème lui-même. Ainsi, nous disons que la méthode FSD introduit un nouveau paradigme de recherche potentiel dans la théorie de l'apprentissage statistique.

La méthode FSD, en tant que cadre théorique, peut—comme tout cadre théorique réussi—fournir les « bonnes » définitions. Nous illustrons ce point en définissant les trois concepts suivants :

1. une définition d'un préordre sur les méthodes spectrales pour un problème de régression supervisée donné,
2. une définition de l'effet de saturation généralisé et ses conditions nécessaires et suffisantes, et
3. une définition mathématique de la propriété d'apprentissage de caractéristiques ainsi que de la propriété d'alignement signal-caractéristiques.

Les définitions de ces concepts doivent s'appuyer sur l'étude des méthodes spectrales via la méthode FSD. Dans [P5], nous appliquons la méthode FSD pour étudier les méthodes spectrales (Exemple 9) pour résoudre des problèmes de régression linéaire dans $\mathbb{R}^p$, c'est-à-dire que nous supposons que $\mathcal{H}$ est identifié avec $\mathbb{R}^p$ via $f(\cdot) = \langle \cdot, \boldsymbol{\beta} \rangle$ ; alors la méthode spectrale $\hat{f}_N$ est identifiée avec $\hat{\boldsymbol{\beta}}$, et le signal $f^*$ est identifié avec $\boldsymbol{\beta}^*$.

**Définition 11.** *Rappelons que $\sigma_1 \geq \cdots$ sont les valeurs propres de $\Sigma$. Soit $b > 0$ et $t \geq 1$. La dimension d'estimation de la méthode spectrale $\hat{\boldsymbol{\beta}}$ avec la fonction de filtre $\varphi_t$ est définie comme*

$$k^* = k^*_{t^{-1},b} = \min\Big\{ k \in [p] : \sigma_{k+1} \leq bt^{-1} \Big\}. \tag{1.40}$$

Soit $V_{J_*} = \mathrm{span}(\boldsymbol{e}_j : j \in J_*)$, $J_* = \{1, \cdots, k^*\}$, $(\boldsymbol{e}_j)_j$ sont les vecteurs propres de $\Sigma$ et $\psi_t$ est la fonction résiduelle définie par $\psi_t(x) = 1 - x\varphi_t(x)$. Définissons

$$r(V_{J_*}, V_{J_*^c}) = \left\| \Sigma^{1/2}_{J_*} \psi_t(\Sigma) \boldsymbol{\beta}^*_{J_*} \right\|_2 + \sigma_\xi \sqrt{\frac{|J_*|}{N}} + \left\| \Sigma^{1/2}_{J_*^c} \boldsymbol{\beta}^*_{J_*^c} \right\|_2 + \sigma_\xi t \sqrt{\frac{\mathrm{Tr}(\Sigma^2_{J_*^c})}{N}}, \tag{1.41}$$

où nous rappelons que nous notons $\Sigma_{J_*} = P_{V_{J_*}} \Sigma P_{V_{J_*}}$ et $\Sigma_{J_*^c} = P_{V_{J_*^c}} \Sigma P_{V_{J_*^c}}$. La conclusion principale de [P5], voir également le Chapitre 3, est que, sous des hypothèses générales, les méthodes spectrales $\hat{\boldsymbol{\beta}} = \frac{1}{N} \varphi_t(\hat{\Sigma}) \mathbb{X}^\top \boldsymbol{y}$ satisfont la propriété suivante avec une grande probabilité :

$$\| \langle X, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle \|_{L^2(\mu_X)} \sim r(V_{J_*}, V_{J_*^c}). \tag{1.42}$$

### 6.3.1 Préordre des Méthodes Spectrales

Grâce à la méthode FSD, (1.42) fournit des bornes supérieure et inférieure correspondantes pour un problème de régression linéaire arbitraire $\mathcal{R} = (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, plutôt que d'être restreinte à une décroissance spectrale spécifique ou à une classe particulière de $\boldsymbol{\beta}^*$. Par conséquent, pour toute $\mathcal{R}$, comparer l'excès de risque en population de deux méthodes spectrales se réduit à comparer deux nombres réels donnés par leur FSD optimale.

Puisqu'un algorithme spectral est déterminé de manière unique par sa fonction de filtre, nous considérons deux méthodes spectrales $\hat{\boldsymbol{\beta}}^{(A)}_{t_A}$ et $\hat{\boldsymbol{\beta}}^{(B)}_{t_B}$ avec paramètres $t_A, t_B$ , et avec les fonctions de filtre $\varphi^{(A)}_{t_A}$ et $\varphi^{(B)}_{t_B}$, respectivement. Par (1.42), il existe $r^{(A)}_{t_A}(V^{(A)}_{J_*}, V^{(A)}_{J_*^c})$ et $r^{(B)}_{t_B}(V^{(B)}_{J_*}, V^{(B)}_{J_*^c})$ caractérisant l'excès de risque en population pour la perte quadratique $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}^{(A)}_{t_A} - \boldsymbol{\beta}^*)\|_2$ et $\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}^{(B)}_{t_B} - \boldsymbol{\beta}^*)\|_2$ pour ces deux méthodes spectrales dans ce problème de régression linéaire. Étant donné tout $\mathcal{R} = (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi) \in \mathbb{R}^{p \times p} \times \mathbb{R}^p \times \mathbb{R}$, nous définissons le préordre suivant « $\preceq_{\mathcal{R}}$ » sur l'ensemble de toutes les méthodes spectrales.

**Définition 12** (Préordre des algorithmes spectraux dans les problèmes de régression linéaire)**.** *Pour le problème de régression linéaire $\mathcal{R} := (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$, nous écrivons*

$$\hat{\boldsymbol{\beta}}^{(A)}_{t_A} \preceq_{\mathcal{R}} \hat{\boldsymbol{\beta}}^{(B)}_{t_B} \quad \textit{si} \quad r^{(A)}_{t_A}(V^{(A)}_{J_*}, V^{(A)}_{J_*^c}) = O\Big( r^{(B)}_{t_B}(V^{(B)}_{J_*}, V^{(B)}_{J_*^c}) \Big)$$

*lorsque $N$ et $p$ tendent vers l'infini. En particulier, si $r^{(A)}_{t_A}(V^{(A)}_{J_*}, V^{(A)}_{J_*^c}) = \Theta\Big( r^{(B)}_{t_B}(V^{(B)}_{J_*}, V^{(B)}_{J_*^c}) \Big)$, nous écrivons $\hat{\boldsymbol{\beta}}^{(A)}_{t_A} \asymp_{\mathcal{R}} \hat{\boldsymbol{\beta}}^{(B)}_{t_B}$. Il est facile de vérifier que « $\asymp_{\mathcal{R}}$ » définit une relation d'équivalence sur l'ensemble de toutes les méthodes spectrales, tandis que $\preceq_{\mathcal{R}}$ définit un préordre.*

Dans ce qui suit, nous considérons le cas où $t_A = t_B$.

**Corollaire 2.** *Étant donné tout problème de régression linéaire $\mathcal{R} = (\Sigma, \boldsymbol{\beta}^*, \sigma_\xi)$. Pour tout $t \geq 1$, $\hat{\boldsymbol{\beta}}_t^{(A)} \preceq_{\mathcal{R}} \hat{\boldsymbol{\beta}}_t^{(B)}$ si et seulement si lorsque $N$ et $p$ tendent vers l'infini*

$$\left\| \Sigma_{J_*}^{\frac{1}{2}} \psi_t^{(A)}(\Sigma) \boldsymbol{\beta}_{J_*}^* \right\|_2 = O\left( \left\| \Sigma_{J_*}^{\frac{1}{2}} \psi_t^{(B)}(\Sigma) \boldsymbol{\beta}_{J_*}^* \right\|_2 \right).$$

**Corollaire 3** (Le Flot de Gradient surpasse Ridge)**.** *Pour tout problème de régression linéaire, $\varphi_t^{(\mathrm{GF})} \preceq_{\mathcal{R}} \varphi_t^{(\mathrm{Ridge})}$, où $\varphi_t^{(\mathrm{Ridge})}(x) = \frac{1}{x + t^{-1}}$ est la fonction de filtre de la régression ridge ; tandis que $\varphi_t^{(\mathrm{GF})}(x) = \frac{1 - \exp(-tx)}{x}$ est la fonction de filtre du flot de gradient.*

### 6.3.2 Effet de saturation généralisé

**Définition 13** (Effet de Saturation Généralisé)**.** *Pour tout problème de régression linéaire $\mathcal{R}$, tout intervalle $I \subset [1, +\infty)$ et toutes familles de fonctions de filtre $\{\varphi_t^{(A)}\}_{t \geq 1}$ et $\{\varphi_t^{(B)}\}_{t \geq 1}$, nous écrivons $\{\varphi_t^{(A)}\}_{t \in I} \preceq_{\mathcal{R}} \{\varphi_t^{(B)}\}_{t \in I}$ si lorsque $N$ et $p$ tendent vers l'infini*

$$\inf\left( r_{t_A}^{(A)}(V_{J_*}, V_{J_*^c}) : t_A \in I \right) = O\left( \inf\left( r_{t_B}^{(B)}(V_{J_*}, V_{J_*^c}) : t_B \in I \right) \right).$$

*Si $\{\varphi_t^{(A)}\}_{t \in I} \preceq_{\mathcal{R}} \{\varphi_t^{(B)}\}_{t \in I}$, nous disons que l'algorithme spectral $\hat{\boldsymbol{\beta}}^{(B)}$ défini par la famille de fonctions de filtre $\{\varphi_t^{(B)}\}_{t \geq 1}$ sature par rapport à la famille de fonctions de filtre $\{\varphi_t^{(A)}\}_{t \geq 1}$ dans $I$.*

**Corollaire 4** (Effet de Saturation dans l'Espace de Sobolev)**.** *Soient $\varphi_t^{(\mathrm{GF})} : x \mapsto (1 - \exp(-tx))/x$ et $\varphi_t^{(\mathrm{Ridge})} : x \mapsto (x + t^{-1})^{-1}$. Soit $\mathcal{R} \in \mathfrak{R}_{\mathrm{Sob}}(s, \alpha)$. Nous avons $\{\varphi^{(\mathrm{GF})}\}_{t \geq 1} \preceq_{\mathcal{R}} \{\varphi^{(\mathrm{Ridge})}\}_{t \geq 1}$. De plus, lorsque $t^{-1} \sim N^{-\frac{\alpha}{1+\tilde{s}\alpha}}$, où $\tilde{s} = s \wedge 2$ pour la régression ridge, et $\tilde{s} = s$ pour le flot de gradient, nous avons $(r_t^{(\mathrm{GF})}(V_{J_*}, V_{J_*^c}))^2 \sim N^{-\frac{\alpha s}{1+s\alpha}}$ et $(r_t^{(\mathrm{Ridge})}(V_{J_*}, V_{J_*^c}))^2 \sim N^{-\frac{\alpha \tilde{s}}{1+\tilde{s}\alpha}}$.*

**Corollaire 5** (Effet de saturation dans le modèle de covariance à pointes)**.** *Supposons qu'il existe certains $k \lesssim N \lesssim p - k$, $\sigma > \varepsilon > 0$ tels que $\sigma_1 = \cdots = \sigma_k = \sigma$, et $\sigma_{k+1} = \cdots = \sigma_p = \varepsilon$. Soit $J = \{1, \cdots, k\}$ et supposons qu'il existe un nombre réel $\alpha_* > 0$ tel que $|\langle \boldsymbol{\beta}^*, \boldsymbol{e}_j \rangle| = \alpha_*$ pour tout $j \in J$ alors que $\langle \boldsymbol{\beta}^*, \boldsymbol{e}_j \rangle = 0$ sinon. Posons*

$$\mathrm{SNR} = \frac{\|\Sigma^{1/2} \boldsymbol{\beta}^*\|_2}{\sigma_\xi} \frac{\sigma \sqrt{N}}{\sqrt{\mathrm{Tr}(\Sigma_{J^c}^2)}}.$$

*Supposons que $4 < \mathrm{SNR} \leq b\frac{\sigma}{\varepsilon}$, où $b$ provient de (1.40). Soit $I = \{t > 1 : b^{-1}\varepsilon \leq t^{-1} < \sigma\}$. Alors*

$$\min_{t \in I} r^{(\mathrm{GF})}(V_{J_*}, V_{J_*^c}) \leq \min_{t \in I} r^{(\mathrm{Ridge})}(V_{J_*}, V_{J_*^c}).$$

*De plus, lorsque $\mathrm{SNR} \to \infty$ et $\sigma = \Omega(\varepsilon)$, $\{\varphi_t^{(\mathrm{Ridge})}\}_{t \in I} \prec_{\mathcal{R}} \{\varphi_t^{(\mathrm{GF})}\}_{t \in I}$.*

### 6.3.3 La FSD pour définir la propriété d'apprentissage de caractéristiques

La définition suivante fournit une définition mathématique, dans le domaine de la théorie de l'apprentissage profond, de ce qui est appelé la propriété d'apprentissage de caractéristiques (feature learning property) (voir la Section 1.4.3). Dans la définition suivante, les $k$ premiers vecteurs propres impliqués dans la propriété d'alignement sont spécifiquement liés à la compréhension acquise en étudiant les méthodes spectrales via la méthode FSD — à savoir, que les méthodes spectrales apprennent en utilisant les vecteurs propres (caractéristiques) correspondant aux $k$ plus grandes valeurs propres, voir la Définition 11. Grosso modo, la FSD elle-même explique comment un estimateur utilise les caractéristiques dans l'espace des caractéristiques, tandis que la propriété d'apprentissage de caractéristiques se concentre sur la façon dont les caractéristiques bénéfiques pour résoudre un problème d'apprentissage supervisé spécifique sont implicitement construites et utilisées par les estimateurs, en particulier les réseaux de neurones. Par conséquent, le rôle de la FSD dans la définition de la propriété d'apprentissage de caractéristiques est de nous aider à comprendre comment ces caractéristiques construites par les réseaux de neurones sont liées au problème spécifique.

**Définition 14.** ***[Propriété d'alignement]*** *Soit $(\mu_X, f^*, \xi)$ un problème de régression supervisée, soit $\mathcal{H}$ un RKHS, et soit $\hat{g}_N$ un estimateur prenant ses valeurs dans $\mathcal{H}$. Soit $g^*_{\mathcal{H}} \in \operatorname{argmin}\{\|g - f^*\|_{L^2(\mu_X)} : g \in \mathcal{H}\}$ l'oracle dans $\mathcal{H}$. Étant donné une fonction croissante $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$, une suite de nombres réels positifs ou nuls $\{\gamma_j\}_{j=k+1}^{\infty}$ et un nombre réel $0 < \delta < 1$, nous disons que $\hat{g}_N$ satisfait la propriété d'alignement $(\Phi, k, \delta)$ par rapport à $\{\gamma_j\}_{j=k+1}^{\infty}$ si, avec une probabilité d'au moins $1-\delta$, on a $\|\hat{g}_N - g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)} \leq \Phi(\sum_{j>k}^{\infty} \gamma_j \langle g^*_{\mathcal{H}}, \boldsymbol{e}_j\rangle^2)$, où $\langle g^*_{\mathcal{H}}, \boldsymbol{e}_j\rangle$ est le produit scalaire dans $\mathcal{H}$ entre $g^*_{\mathcal{H}}$ et $\boldsymbol{e}_j$.*

Un estimateur $\hat{g}_N$ satisfaisant la propriété d'alignement présente la caractéristique suivante : lorsque l'oracle $g^*_{\mathcal{H}}$ est bien aligné avec les vecteurs propres (fonctions propres) correspondant aux $k$ plus grandes valeurs propres de l'opérateur de covariance $\Sigma$, l'estimateur peut exploiter cette structure et ainsi atteindre une erreur d'estimation plus faible $\|\hat{g}_N - g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)}$. Il est connu que de nombreux estimateurs satisfont cette propriété, y compris la régression ridge, le flot de gradient (gradient flow), la descente de gradient, la régression sur les composantes principales, et la RERM avec des termes de régularisation dont la divergence de Bregman est non triviale ; voir le Théorème 1. Pour une RERM générale, les poids $\gamma_j$ peuvent être choisis égaux à 1 ; pour la régression ridge, le flot de gradient, la descente de gradient et la régression sur les composantes principales, $\gamma_j$ peut être choisi égal à $\sigma_j$, voir (1.41).

La propriété d'alignement caractérise la capacité d'un estimateur à exploiter un alignement géométrique favorable existant ; cependant, un tel alignement n'est pas toujours intrinsèquement présent. Lorsque cette relation d'alignement est sous-optimale, certains estimateurs produisent de grandes erreurs de prédiction. Pour étudier l'impact de l'alignement défavorable entre l'oracle $g^*_{\mathcal{H}}$ et les fonctions propres de l'opérateur de covariance de $\mathcal{H}$ sur l'erreur d'estimation d'une large classe d'estimateurs, nous introduisons le concept suivant d'efficacité de l'alignement.

**Définition 15.** *Soit $(\mu_X, f^*, \xi)$ un problème de régression supervisée, $\sigma^2_\xi$ la variance de $\xi$, $\mathcal{H}$ un RKHS et $N$ la taille de l'échantillon. Soit $0 < \alpha < 1$ un seuil pré-spécifié. Nous définissons la dimension d'équilibre (balance dimension) $k^\circ(N) = \min\{k \in \mathbb{N} : \|P_{k+1:\infty} g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)} \leq \sigma^2_\xi \frac{k}{N}\}$, où $P_{k+1:\infty} = \sum_{j>k} \boldsymbol{e}_j \otimes \boldsymbol{e}_j$ est la projection sur $\operatorname{span}(\boldsymbol{e}_j : j > k)$. Si $k^\circ(N) \leq \alpha N$, nous disons que l'alignement est efficace ; si $k^\circ(N) > \alpha N$, nous disons que l'alignement est déficient.*

La Définition 15 a pour but de caractériser la relation entre la dimension d'équilibre et la taille de l'échantillon $N$. D'après la FSD, la projection $P_{k+1:\infty} g^*_{\mathcal{H}}$ n'est pas estimée et entre donc dans l'erreur d'estimation comme le prix de la non-estimation. Parallèlement, $\sigma^2_\xi \frac{k}{N}$ représente le terme de variance associé au sous-espace d'estimation. Puisque les deux termes sont indépendants du choix spécifique de l'estimateur, la dimension d'équilibre — en décrivant l'équilibre entre ces deux composantes — révèle l'impact sur l'erreur d'estimation qui dépend exclusivement de l'alignement de $g^*_{\mathcal{H}}$ avec les vecteurs propres de l'opérateur de covariance de $\mathcal{H}$, indépendamment de l'algorithme spécifique.

Dans la fonction de taux de la FSD, les termes $\|P_{k+1:\infty} g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)}$ et $\sigma^2_\xi \frac{k}{N}$ sont universels, ce qui signifie qu'ils sont indépendants du choix spécifique de l'estimateur et apparaissent toujours dans la fonction de taux. Par conséquent, l'équilibre entre ces deux termes caractérise, dans un sens indépendant de l'estimateur (estimator-free), l'impact de l'alignement entre $g^*_{\mathcal{H}}$ et les vecteurs propres de l'opérateur de covariance sur l'erreur d'estimation de n'importe quel estimateur. En particulier, par la définition de $k^\circ(N)$, pour tout $k \in \mathbb{N}$ (spécialement pour la dimension d'estimation $k^*$), nous avons toujours $\|P_{k^\circ+1:\infty} g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)} + \sigma^2_\xi \frac{k^\circ}{N} \leq 2(\|P_{k+1:\infty} g^*_{\mathcal{H}}\|^2_{L^2(\mu_X)} + \sigma^2_\xi \frac{k}{N})$. Cela implique que la fonction de taux optimale de la FSD est toujours soumise à une borne inférieure fournie par $\sigma^2_\xi \frac{k^\circ}{N}$, quel que soit l'estimateur sélectionné pour la FSD. Lorsque la dimension d'équilibre est excessivement grande, cette borne inférieure peut être assez substantielle, conduisant à une erreur d'estimation sous-optimale. Un élément clé de la propriété d'apprentissage de caractéristiques introduite dans ce qui suit est que l'apprentissage de caractéristiques peut construire automatiquement des alignements géométriques favorables grâce à l'apprentissage autonome des caractéristiques.

**Définition 16.** ***[Propriété d'alignement et propriété d'apprentissage de caractéristiques]*** *Considérons un problème de régression supervisée à valeurs réelles $(\mu_X, f^*, \xi)$. Soit $\hat{f}_N$ un estimateur. Nous disons que $\hat{f}_N$ accomplit la propriété d'apprentissage de caractéristiques (feature learning property) en résolvant $(\mu_X, f^*, \xi)$, s'il existe un RKHS $\mathcal{H}_{\text{fea}}$ avec sa carte de caractéristiques (feature map) canonique notée $\phi_{\text{fea}} : \Omega_X \to \mathcal{H}_{\text{fea}}$, et un élément $\hat{g}_N \in \mathcal{H}_{\text{fea}}$ tels que les conditions suivantes soient vérifiées à la limite quand $N \to \infty$ :*

1. *$\hat{g}_N$ satisfait la propriété d'alignement avec une dimension d'estimation $k$ pour un certain $k \in \mathbb{N}_+$ ;*
2. *$\hat{f}_N(\cdot) = \hat{g}_N(\phi_{\text{fea}}(\cdot))$ où $\hat{g}_N(\phi_{\text{fea}}(\cdot)) = \langle \hat{g}_N, \phi_{\text{fea}}(\cdot)\rangle_{\mathcal{H}_{\text{fea}}}$ ;*
3. *l'oracle $g^*_{\mathcal{H}_{\text{fea}}} \in \operatorname{argmin}\{\|g - f^*\|_{L^2(\mu_X)} : g \in \mathcal{H}_{\text{fea}}\}$ satisfait $\|f^* - g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)} = o_{\mathbb{P}}(1)$ ;*
4. *$k = O(d)$, et $\|P_{k+1:\infty} g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)} = o_{\mathbb{P}}(1)$ où $P_{k+1:\infty} = \sum_{j>k} \boldsymbol{e}_j \otimes \boldsymbol{e}_j$.*

*Nous appelons un tel $\mathcal{H}_{\text{fea}}$ le sous-espace de caractéristiques appris (learned feature subspace).*

La signification de la Définition 16 est la suivante : si $\hat{f}_N$ construit à partir des échantillons d'entraînement un sous-espace de caractéristiques $\mathcal{H}_{\text{fea}}$ (que nous supposons être un RKHS) tel que, pour le problème de régression à valeurs réelles $(\mu_X, f^*, \xi)$, l'erreur d'approximation de l'espace de caractéristiques construit $\mathcal{H}_{\text{fea}}$, $\|f^* - g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)}$, est petite, et qu'au sein de ce sous-espace de caractéristiques $\mathcal{H}_{\text{fea}}$, les caractéristiques qui sont utiles pour estimer $f^*$ sont effectivement utilisées pour estimer $f^*$. En d'autres termes : la capacité d'ingénierie des caractéristiques de $\hat{f}_N$ sur ce problème de régression se manifeste par le sous-espace de caractéristiques $\mathcal{H}_{\text{fea}}$ construit à partir des données, qui possède une petite erreur d'approximation, et les $k$ premières caractéristiques construites sont bénéfiques pour estimer $f^*$ dans le sens où, sur $\mathcal{H}_{\text{fea}}$, il existe un estimateur latent $\hat{g}_N$ capable d'utiliser les $k$ premiers vecteurs propres les plus importants pour obtenir une petite erreur d'estimation afin d'estimer l'oracle $g^*_{\mathcal{H}_{\text{fea}}}$. Enfin, $\hat{f}_N$ est proche de cet estimateur latent $\hat{g}_N$ en distance $L^2(\mu_X)$.

Par conséquent, le phénomène observé à la fois par les praticiens et les théoriciens lorsque $\hat{f}_N$ résout le problème $(\mu_X, f^*, \xi)$ est le suivant : sur la base des données d'entraînement $(X_i, Y_i)_{i=1}^N$, $\hat{f}_N$ semble construire automatiquement, dans $L^2(\mu_X)$, un espace de caractéristiques $\mathcal{H}_{\text{fea}}$ qui possède des propriétés statistiques favorables pour le problème, apprend « à l'intérieur » de cet espace, et atteint une petite erreur d'estimation — comme si $\hat{f}_N$ lui-même était un estimateur défini sur $\mathcal{H}_{\text{fea}}$. Ce phénomène est précisément ce que l'on appelle souvent la propriété d'apprentissage de caractéristiques dans la théorie de l'apprentissage profond.

Par conséquent, si $\hat{f}_N$ possède la propriété d'apprentissage de caractéristiques lors de la résolution de $(\mu_X, f^*, \xi)$, alors son erreur d'estimation (à un carré près) peut être majorée comme suit :

$$\|\hat{f}_N - f^*\|_{L^2(\mu_X)} \leq \|\hat{f}_N - \hat{g}_N\|_{L^2(\mu_X)} + \|\hat{g}_N - g^*_{\mathcal{H}_{\text{fea}}}\|_{L^2(\mu_X)} + \|g^*_{\mathcal{H}_{\text{fea}}} - f^*\|_{L^2(\mu_X)}.$$

L'exigence dans la Définition 16 que $\|\hat{f}_N - \hat{g}_N\|_{L^2(\mu_X)}$ soit petite ne peut pas être omise, car sous des conditions très larges, pour tout $f^* \in L^2(\mu_X)$, il existe toujours un RKHS déterministe $\mathcal{H}$ (dont la carte de caractéristiques est notée $\phi$) tel qu'il existe un $\hat{g}_N$ possédant la propriété d'alignement pour lequel $\|f^* - g^*_{\mathcal{H}}\|_{L^2(\mu_X)}$ est petit, où $g^*_{\mathcal{H}} : \boldsymbol{x} \in \Omega_X \mapsto \mathbb{E}[Y \mid \phi(\boldsymbol{x})]$ est la fonction de régression dans $\mathcal{H}$. En fait, les méthodes spectrales avec des fonctions de filtrage analytiques possèdent toujours la propriété d'alignement, et de nombreux RKHS $\mathcal{H}$ sont denses dans $L^2(\mu_X)$. Par conséquent, ce que la Définition 16 souligne, c'est qu'un tel RKHS doit dépendre de $(X_i, Y_i)_{i=1}^N$ ainsi que de $(\mathcal{F}, \hat{f}_N)$ (c'est pourquoi nous le notons $\mathcal{H}_{\text{fea}}$), et tel que $\hat{g}_N$ soit proche de $\hat{f}_N$ pour la métrique $L^2(\mu_X)$. Ce n'est qu'alors que ce sous-espace de caractéristiques latent et l'estimateur latent $\hat{g}_N$ défini sur celui-ci deviennent capables d'expliquer la propriété d'apprentissage de caractéristiques de $\hat{f}_N$ pour le problème donné.

# Publications


[P1] Radoslaw Adamczak, Guillaume Lecué, Zong Shang, and Marta Strzelecka. Benign overfitting property of the minimum norm interpolant estimator in regression and classification via feature space decomposition. In preparation, 2026.

[P2] Georgios Gavrilopoulos, Guillaume Lecué, and Zong Shang. A geometrical analysis of kernel ridge regression and its applications. *The Annals of Statistics*, 53(6):2592–2616, December 2025. Publisher: Institute of Mathematical Statistics.

[P3] Guillaume Lecué, Zhifan Li, and Zong Shang. Sharp convergence rates for Spectral methods via the feature space decomposition method, December 2025. arXiv:2512.14473 [math].

[P4] Guillaume Lecué and Zong Shang. A geometrical viewpoint on the benign overfitting property of the minimum $$\ell _2$$-norm interpolant estimator and its universality. *Probability Theory and Related Fields*, November 2024.

[P5] Guillaume Lecué, Zong Shang, Taiji Suzuki, and Tomoya Wakayama. On the generalization error of mean field shallow neural network, 2026. in preparation.

[P6] Zong Shang. Upper bounds for the L^q empirical process via generic chaining, November 2025. arXiv:2511.06338 [math].

# Bibliography


[ACL19] Pierre Alquier, Vincent Cottet, and Guillaume Lecué. Estimation bounds and sharp oracle inequalities of regularized procedures with Lipschitz loss functions. *The Annals of Statistics*, 47(4):2117–2144, August 2019.

[Ada10] Radosław Adamczak. A Few Remarks on the Operator Norm of Random Toeplitz Matrices. *Journal of Theoretical Probability*, 23(1):85–108, March 2010.

[AKT19] Alnur Ali, J. Zico Kolter, and Ryan J. Tibshirani. A Continuous-Time View of Early Stopping for Least Squares Regression. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1370–1378. PMLR, April 2019.

[ASH21] Navid Ardeshir, Clayton Sanford, and Daniel Hsu. Support vector machines and linear regression coincide with very high-dimensional features, October 2021. arXiv:2105.14084 [cs, math, stat].

[Bac24] Francis Bach. *Learning Theory from First Principles*. MIT Press, 2024.

[BB21] Anas Barakat and Pascal Bianchi. Convergence and Dynamical Behavior of the ADAM Algorithm for Nonconvex Stochastic Optimization. *SIAM Journal on Optimization*, 31(1):244–274, January 2021.

[BBL05] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of Classification: a Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9:323–375, November 2005.

[BBSMW21] Xin Bing, Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp. Prediction Under Latent Factor Regression: Adaptive PCR, Interpolating Predictors and Beyond. *Journal of Machine Learning Research*, 22, 2021.

[BC11] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, New York, NY, 2011.

[BCLL18] Sébastien Bubeck, Michael B. Cohen, Yin Tat Lee, and Yuanzhi Li. An homotopy method for $\ell_p$ regression provably beyond self-concordance and in input-sparsity time, June 2018. arXiv:1711.01328 [math].

[BDVRV23] Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Understanding neural networks with reproducing kernel Banach spaces. *Applied and Computational Harmonic Analysis*, 62:194–236, January 2023.

[BES+22] Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation, May 2022. arXiv:2205.01445 [cs, math, stat].

[BJM03] Peter Bartlett, Michael Jordan, and Jon Mcauliffe. Large Margin Classifiers: Convex Loss, Low Noise, and Convergence Rates. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.

[BJM06] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, Classification, and Risk Bounds. *Journal of the American Statistical Association*, 101(473):138–156, March 2006. _eprint: https://doi.org/10.1198/016214505000000907.

[BK20] Pierre C. Bellec and Arun K. Kuchibhotla. First order expansion of convex regularized estimators, March 2020. arXiv:1910.05480 [math].

[BM06] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334, July 2006.

[BM16] Gilles Blanchard and Nicole Mücke. Kernel regression, minimax rates and effective dimensionality: beyond the regular case, November 2016. arXiv:1611.03979 [stat].

[BM18] Gilles Blanchard and Nicole Mücke. Optimal Rates for Regularization of Statistical Inverse Learning Problems. *Foundations of Computational Mathematics*, 18(4):971–1013, August 2018.

[BM22a] Daniel Bartl and Shahar Mendelson. Random embeddings with an almost Gaussian distortion. *Advances in Mathematics*, 400:108261, May 2022.

[BM22b] Daniel Bartl and Shahar Mendelson. Structure preservation via the Wasserstein distance, September 2022. arXiv:2209.07058 [math, stat].

[BM24] Daniel Bartl and Shahar Mendelson. Optimal Non-Gaussian Dvoretzky–Milman Embeddings. *International Mathematics Research Notices*, 2024(10):8459–8480, May 2024.

[BMA24] Daniel Beaglehole, Ioannis Mitliagkas, and Atish Agarwala. Gradient descent induces alignment between weights and the empirical NTK for deep non-linear networks, February 2024. arXiv:2402.05271 [cs, stat].

[BMM18] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To Understand Deep Learning We Need to Understand Kernel Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 541–549. PMLR, July 2018.

[BMM19] Gilles Blanchard, Peter Mathé, and Nicole Mücke. Lepskii Principle in Supervised Learning, May 2019. arXiv:1905.10764 [math, stat].

[BMR21] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta Numerica*, 30:87–201, May 2021.

[Boy22] Claire Boyer. Living la vida loca: learning in interpolation regimes. Lecture notes, Sorbonne University, Paris Sacaly University, 2022.

[BP12] Viorel Barbu and Teodor Precupanu. *Convexity and Optimization in Banach Spaces*. Springer Monographs in Mathematics. Springer Science+Business Media, Bucuresti, Romania, 4nd edition, 2012.

[BPR07] Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, February 2007.

[BS24] Daniel Barzilai and Ohad Shamir. Generalization in Kernel Regression Under Realistic Assumptions, February 2024. arXiv:2312.15995 [cs, stat].

[BV14] Stephen P. Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2014.

[BvdG11] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, Berlin, Heidelberg, 2011.

[BVRV24] Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Neural reproducing kernel Banach spaces and representer theorems for deep networks, March 2024. arXiv:2403.08750 [stat].

[BW23] Xin Bing and Marten Wegkamp. Optimal discriminant analysis in high-dimensional latent factor models. *The Annals of Statistics*, 51(3):1232–1257, June 2023.

[CCLN21] Stéphane Chrétien, Mihai Cucuringu, Guillaume Lecué, and Lucie Neirac. Learning with semi-definite programming: statistical bounds based on fixed point analysis and excess risk curvature. *Journal of Machine Learning Research*, 22(230):1–64, 2021.

[CGB22] Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk Bounds for Over-parameterized Maximum Margin Classification on Sub-Gaussian Mixtures, January 2022. arXiv:2104.13628 [cs, math, stat].

[CGLP12] Djalil Chafai, Olivier Guédon, Guillaume Lecué, and Alain Pajor. *Interactions between compressed sensing, random matrices, and high dimensional geometry*. Société Mathématique de France, December 2012.

[Cha25] Hugo Chardon. *Finite-sample theory for maximum-likehood estimation in logistic regression*. PhD thesis, Institut Polytechnique de Paris, Palaiseau, June 2025.

[CL19] Sabyasachi Chatterjee and John Lafferty. Adaptive risk bounds in unimodal regression. *Bernoulli*, 25(1):1–25, February 2019.

[CLL20] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust statistical learning with Lipschitz and convex loss functions. *Probability Theory and Related Fields*, 176(3):897–940, April 2020.

[CLL21] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle. Robust high dimensional learning for Lipschitz and convex losses, January 2021. arXiv:1905.04281 [math, stat].

[CLM24] Hugo Chardon, Matthieu Lerasle, and Jaouad Mourtada. Finite-sample performance of the maximum likelihood estimator in logistic regression, December 2024. arXiv:2411.02137 [math].

[CLvdG22] 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.

[CM22] Chen Cheng and Andrea Montanari. Dimension free ridge regression, October 2022. arXiv:2210.08571 [math, stat].

[CT05] Emmanuel J. Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[CW21] 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.

[Dir15] Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20(none):1–29, January 2015.

[DRSY22] 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.

[Dvo59] Aryeh Dvoretzky. A THEOREM ON CONVEX BODIES AND APPLICATIONS TO BANACH SPACES*. *Proceedings of the National Academy of Sciences of the United States of America*, 45(2):223–226, February 1959.

[Dvo61] A P Dvoredsky. Some results on convex bodies and Banach spaces. In *Proc. Symp. on Linear Spaces*, Jerusalem, 1961.

[EHN00] Heinz Werner Engl, Martin Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Springer Science & Business Media, March 2000.

[FCB22] Spencer Frei, Niladri S. Chatterji, and Peter Bartlett. Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 2668–2703. PMLR, June 2022.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Springer, New York, NY, 2013.

[FVBS23] Spencer Frei, Gal Vardi, Peter L. Bartlett, and Nathan Srebro. Benign Overfitting in Linear Classifiers and Leaky ReLU Networks from KKT Conditions for Margin Maximization, March 2023. arXiv:2303.01462.

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[Gir14] Christophe Giraud. *Introduction to High-Dimensional Statistics*. Taylor & Francis, December 2014. Google-Books-ID: qRuVoAEACAAJ.

[GLPTJ07] Yehoram Gordon, Alexander Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Random epsilon-nets and embeddings in l(infinity)(N). *Studia Mathematica*, 178(1):91–98, 2007.

[GLSS18] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit Bias of Gradient Descent on Linear Convolutional Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[HI25] Qiyang Han and Masaaki Imaizumi. Precise gradient descent training dynamics for finite-width multi-layer neural networks, May 2025. arXiv:2505.04898 [cs].

[HMX21] Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 91–99. PMLR, March 2021.

[HW19] Qiyang Han and Jon A. Wellner. Convergence rates of least squares regression estimators with heavy-tailed errors. *The Annals of Statistics*, 47(4):2286–2319, August 2019.

[HW23] Laura Hucker and Martin Wahl. A note on the prediction error of principal component regression in high dimensions. *Theor. Probability and Math. Statist.*, 109:37–53, 2023. arXiv:2212.04959 [math, stat].

[JGH18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 8580–8589, Red Hook, NY, USA, 2018. Curran Associates Inc.

[JHZ+24] Jiarui Jiang, Wei Huang, Miao Zhang, Taiji Suzuki, and Liqiang Nie. Unveil Benign Overfitting for Transformer in Vision: Training Dynamics, Convergence, and Generalization. *Advances in Neural Information Processing Systems*, 37:135464–135625, December 2024.

[Kal21] Olav Kallenberg. *Foundations of Modern Probability*, volume 99 of *Probability Theory and Stochastic Modelling*. Springer International Publishing, Cham, 2021.

[Kat95] Tosio Kato. *Perturbation Theory for Linear Operators*, volume 132 of *Classics in Mathematics*. Springer, Berlin, Heidelberg, 1995.

[KCCG23] Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. Benign Overfitting in Two-layer ReLU Convolutional Neural Networks. In *Proceedings of the 40th International Conference on Machine Learning*, pages 17615–17659. PMLR, July 2023.

[KL16] Vladimir Koltchinskii and Karim Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(4):1976–2013, November 2016.

[KM15] Vladimir Koltchinskii and Shahar Mendelson. Bounding the Smallest Singular Value of a Random Matrix Without Concentration. *International Mathematics Research Notices*, 2015(23):12991–13008, January 2015.

[Kol09] Vladimir Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *The Annals of Statistics*, 37(3):1332–1359, June 2009.

[Kol11] 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.

[Kol18] Vladimir Koltchinskii. Asymptotic efficiency in high-dimensional covariance estimation. In *Proceedings of the International Congress of Mathematicians (ICM 2018)*, pages 2903–2923. WORLD SCIENTIFIC, June 2018.

[KRV24] Varun Kanade, Patrick Rebeschini, and Tomas Vaškevičius. Exponential Tail Local Rademacher Complexity Risk Bounds Without the Bernstein Condition. *Journal of Machine Learning Research*, 25:1 – 43, 2024.

[KYS23] Guy Kornowski, Gilad Yehudai, and Ohad Shamir. From Tempered to Benign Overfitting in ReLU Neural Networks. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023.

[KZSS21] Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds and Benign Overfitting. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, October 2021.

[Lax02] Peter D. Lax. *Functional Analysis*. Wiley–Blackwell, New York, April 2002.

[Lec11] Guillaume Lecué. *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis*. thesis, Université Paris-Est, December 2011.

[Led96] Michel Ledoux. Isoperimetry and Gaussian analysis. In Pierre Bernard, editor, *Lectures on Probability Theory and Statistics*, volume 1648, pages 165–294. Springer Berlin Heidelberg, Berlin, Heidelberg, 1996. Series Title: Lecture Notes in Mathematics.

[Led05] Michel Ledoux. *The Concentration of Measure Phenomenon*. American Mathematical Society, Providence, RI, February 2005.

[Lep91] O. V. Lepskii. On a Problem of Adaptive Estimation in Gaussian White Noise. *Theory of Probability & Its Applications*, 35(3):454–466, January 1991.

[Lep23] Oleg V. Lepski. Theory of adaptive estimation. In *International Congress of Mathematicians*, pages 5478–5498. European Mathematical Society - EMS - Publishing House GmbH, December 2023.

[LGRO+08] L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, July 2008.

[LGSL24] Yicheng Li, Weiye Gan, Zuoqiang Shi, and Qian Lin. Generalization Error Curves for Analytic Spectral Algorithms under Power-law Decay, July 2024. arXiv:2401.01599.

[LM12] Guillaume Lecué and Shahar Mendelson. General Nonexact Oracle Inequalities for Classes with a Subexponential Envelope. *The Annals of Statistics*, 40(2):832–860, 2012.

[LM16] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes : Upper and minimax bounds, September 2016. arXiv:1305.4825 [math, stat].

[LM17] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method II: complexity dependent error rates. *Journal of Machine Learning Research*, 18(146):1–48, 2017.

[LM18] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: sparse recovery. *The Annals of Statistics*, 46(2):611–641, 2018.

[LN24] Guillaume Lecué and Lucie Neirac. Learning with a linear loss function: excess risk and estimation bounds for ERM, minmax MOM and their regularized versions with applications to robustness in sparse PCA. *Journal of Machine Learning Research*, 25(399):1–90, 2024.

[LR21] Tengyuan Liang and Benjamin Recht. Interpolating Classifiers Make Few Mistakes, July 2021. arXiv:2101.11815 [cs, math, stat].

[LRJ23] Haochuan Li, Alexander Rakhlin, and Ali Jadbabaie. Convergence of Adam under Relaxed Assumptions, June 2023. arXiv:2304.13972 [cs, math, stat].

[LS24] Guillaume Lecué and Zong Shang. A geometrical viewpoint on the benign overfitting property of the minimum $$\ell _2$$-norm interpolant estimator and its universality. *Probability Theory and Related Fields*, November 2024.

[LS25] Rafał Latała and Marta Strzelecka. Operator $\ell_p\to\ell_q$ norms of Gaussian matrices, February 2025. arXiv:2502.02186 [math].

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1991.

[LW15] Po-Ling Loh and Martin J. Wainwright. Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *Journal of Machine Learning Research*, 16(19):559–616, 2015.

[LZL23] Yicheng Li, Haobo Zhang, and Qian Lin. On the Asymptotic Learning Curves of Kernel Ridge Regression under Power-law Decay, September 2023. arXiv:2309.13337 [cs, math, stat].

[Men15] Shahar Mendelson. Learning without Concentration. *Journal of the ACM*, 62(3):21:1–21:25, 2015.

[Men16] Shahar Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Processes and their Applications*, 126(12):3652–3680, December 2016.

[Men17] Shahar Mendelson. On Multiplier Processes Under Weak Moment Assumptions. In Bo'az Klartag and Emanuel Milman, editors, *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*, Lecture Notes in Mathematics, pages 301–318. Springer International Publishing, Cham, 2017.

[Men18] Shahar Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1):459–502, June 2018.

[Men19] Shahar Mendelson. An Unrestricted Learning Procedure. *J. ACM*, 66(6):42:1–42:42, November 2019.

[Men22] Shahar Mendelson. An isomorphic Dvoretzky-Milman Theorem using general random ensembles. *Journal of Functional Analysis*, 283(2):109473, July 2022.

[Mil71] V. D. Milman. New proof of the theorem of A. Dvoretzky on intersections of convex bodies. *Functional Analysis and Its Applications*, 5(4):288–295, October 1971.

[MMM22] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, July 2022.

[MRSS23] Andrea Montanari, Feng Ruan, Basil Saeed, and Youngtak Sohn. Universality of max-margin classifiers, September 2023. arXiv:2310.00176 [math, stat].

[MRSY25] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: Benign overfitting and high dimensional asymptotics in the overparametrized regime. *The Annals of Statistics*, 53(2):822–853, April 2025.

[MSA+22] Neil Rohit Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, Tempered, or Catastrophic: Toward a Refined Taxonomy of Overfitting. October 2022.

[MT99] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, December 1999.

[MTJ08] Shahar Mendelson and Nicole Tomczak-Jaegermann. A subgaussian embedding theorem. *Israel Journal of Mathematics*, 164(1):349–364, March 2008.

[MU25] Andrea Montanari and Pierfrancesco Urbani. Dynamical Decoupling of Generalization and Overfitting in Large Two-Layer Networks, February 2025. arXiv:2502.21269 [stat].

[MZC23] Xuran Meng, Difan Zou, and Yuan Cao. Benign Overfitting in Two-Layer ReLU Convolutional Neural Networks for XOR Data, October 2023. arXiv:2310.01975 [cs].

[Nes83] Yurii Nesterov. A method for solving the convex programming problem with convergence rate O(1/k^2). *Proceedings of the USSR Academy of Sciences*, 269:543–547, January 1983.

[NWS22] Atsushi Nitanda, Denny Wu, and Taiji Suzuki. Convex Analysis of the Mean Field Langevin Dynamics. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR, May 2022.

[PHD20] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663, October 2020.

[Pis89] Gilles Pisier. *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, 1989.

[Pis16] Gilles Pisier. *Martingales in Banach Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 2016.

[Pol87] Boris T. Polyak. *Introduction to optimization*. New York, Optimization Software, 1987.

[PR19] Nicolò Pagliana and Lorenzo Rosasco. Implicit Regularization of Accelerated Methods in Hilbert Spaces, December 2019. arXiv:1905.13000 [cs].

[PVRB18] Loucas Pillaud-Vivien, Alessandro Rudi, and Francis Bach. Statistical Optimality of Stochastic Gradient Descent on Hard Learning Problems through Multiple Passes, November 2018. arXiv:1805.10074 [cs, math, stat].

[PVRF22] Loucas Pillaud-Vivien, Julien Reygner, and Nicolas Flammarion. Label noise (stochastic) gradient descent implicitly solves the Lasso for quadratic parametrisation, June 2022. arXiv:2206.09841 [cs, math, stat].

[RBD25] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Dmitriy Drusvyatskiy. Linear Recursive Feature Machines provably recover low-rank matrices. *Proceedings of the National Academy of Sciences*, 122(13):e2411325122, April 2025.

[RBPB22] Adityanarayanan Radhakrishnan, Daniel Beaglehole, Parthe Pandit, and Mikhail Belkin. Feature learning in neural networks and kernel machines that recursively learn features, December 2022. arXiv:2212.13881 [cs].

[Roc70] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[Ros62] Frank Rosenblatt. *Principles of neurodinamics: Perceptron and theory of brain mechanisms*. Spartan Books, Washington D.C., 1962.

[RW05] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. November 2005.

[SC08] Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, NY, 2008.

[Sch13] Rolf Schneider. *Convex Bodies: The Brunn–Minkowski Theory*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 2 edition, 2013.

[Sha22] Ohad Shamir. The Implicit Bias of Benign Overfitting. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pages 448–478. PMLR, June 2022.

[SLS+20] Fanhua Shang, Yuanyuan Liu, Fanjie Shang, Hongying Liu, Lin Kong, Licheng Jiao, Fanhua Shang, Yuanyuan Liu, Fanjie Shang, Hongying Liu, Lin Kong, and Licheng Jiao. A Unified Scalable Equivalent Formulation for Schatten Quasi-Norms. *Mathematics*, 8(8), August 2020. Company: Multidisciplinary Digital Publishing Institute Distributor: Multidisciplinary Digital Publishing Institute Institution: Multidisciplinary Digital Publishing Institute Label: Multidisciplinary Digital Publishing Institute.

[SS16] Saburou Saitoh and Yoshihiro Sawano. *Theory of Reproducing Kernels and Applications*, volume 44 of *Developments in Mathematics*. Springer, Singapore, 2016.

[STC04] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.

[SZ07] Steve Smale and Ding-Xuan Zhou. Learning Theory Estimates via Integral Operators and Their Approximations. *Constructive Approximation*, 26(2):153–172, August 2007.

[Tal96] Michel Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24(3):1049–1103, July 1996.

[Tal14] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, Berlin, Heidelberg, 2014.

[Tal21] Michel Talagrand. *Upper and Lower Bounds for Stochastic Processes: Decomposition Theorems*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics*. Springer International Publishing, Cham, 2021.

[TB23] Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

[TCFB25] Alexander Tsigler, Luiz F. O. Chamon, Spencer Frei, and Peter L. Bartlett. Benign Overfitting and the Geometry of the Ridge Regression Solution in Binary Classification, March 2025. arXiv:2503.07966 [stat].

[Tik18] Konstantin Tikhomirov. Sample Covariance Matrices of Heavy-Tailed Distributions. *International Mathematics Research Notices*, 2018(20):6254–6289, October 2018.

[Tsy03] Alexandre B. Tsybakov. Optimal Rates of Aggregation. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 303–313, Berlin, Heidelberg, 2003. Springer.

[Tsy04] Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, February 2004.

[Tsy09] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009.

[Vap00] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, 2000.

[VC68] Vladimir N. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Doklady Akademii Nauk USSR*, 181(4), 1968.

[VDVW23] A. W. Van Der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer International Publishing, Cham, 2023.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018.

[vH18] Ramon van Handel. Chaining, interpolation, and convexity. *Journal of the European Mathematical Society*, 20(10):2413–2435, July 2018.

[VPY24] Maksim Velikanov, Maxim Panov, and Dmitry Yarotsky. Generalization error of spectral algorithms, March 2024. arXiv:2403.11696 [cs, stat].

[Wai19] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2019.

[WBL+25] Jingfeng Wu, Peter L. Bartlett, Jason D. Lee, Sham M. Kakade, and Bin Yu. Risk Comparisons in Linear Regression: Implicit Regularization Dominates Explicit Regularization, September 2025. arXiv:2509.17251 [stat].

[WDY22] 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.

[Wei25] Alexander S. Wein. Computational Complexity of Statistics: New Insights from Low-Degree Polynomials, June 2025. arXiv:2506.10748 [math].

[WMT21] Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign Overfitting in Multiclass Classification: All Roads Lead to Interpolation. In *NeurIPS*, 2021.

[WT21] Ke Wang and Christos Thrampoulidis. Binary Classification of Gaussian Mixtures: Abundance of Support Vectors, Benign Overfitting and Regularization, September 2021. arXiv:2011.09148 [cs, math, stat].

[XWF+23] Zhiwei Xu, Yutong Wang, Spencer Frei, Gal Vardi, and Wei Hu. Benign Overfitting and Grokking in ReLU Networks for XOR Cluster Data, October 2023. arXiv:2310.02541 [cs, stat].

[YH21] Greg Yang and Edward J. Hu. Tensor Programs IV: Feature Learning in Infinite-Width Neural Networks. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11727–11737. PMLR, 2021.

[YRC07] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, 26(2):289–315, August 2007.

[Zha04] Tong Zhang. Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. *The Annals of Statistics*, 32(1):56–85, 2004.

[ZKS+22] Lijia Zhou, Frederic Koehler, Pragya Sur, Danica J. Sutherland, and Nathan Srebro. A Non-Asymptotic Moreau Envelope Theory for High-Dimensional Generalized Linear Models, October 2022. arXiv:2210.12082 [stat].

[ZLL23] Haobo Zhang, Yicheng Li, and Qian Lin. On the Optimality of Misspecified Spectral Algorithms, August 2023. arXiv:2303.14942 [math, stat].

[ZLRB24] Libin Zhu, Chaoyue Liu, Adityanarayanan Radhakrishnan, and Mikhail Belkin. Catapults in SGD: spikes in the training loss and their impact on generalization through feature learning, March 2024. arXiv:2306.04815 [cs, math, stat].

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

**Titre :** Décomposition d'Espace-Caractéristiques

**Mots clés :** Argument de Convergence Uniforme, Théorème de Dvoretzky-Milman, Points Fixes, Surapprentissage Bénin, Régression Ridge, Descente/Flux de Gradient

**Résumé :** Le problème de la prédiction statistique constitue l'une des questions les plus centrales des domaines de la théorie de l'apprentissage statistique et de la statistique mathématique. L'argument de convergence uniforme est l'une des méthodes les plus fondamentales pour étudier ce problème. Ces dernières années, l'argument de convergence uniforme s'est révélé inadéquat pour étudier les estimateurs surajustés, en particulier en raison de son incapacité à expliquer le phénomène de surapprentissage bénin.
Motivé par ce constat, cette thèse propose un cadre mathématique systématique pour raffiner l'argument de convergence uniforme, nommé la méthode de Décomposition d'Espace de Caractéristiques. Elle vise à étudier le phénomène de surapprentissage bénin des estimateurs d'interpolation à norme $\ell_q$ minimale en régression linéaire, ainsi que de l'estimateur d'interpolation à norme $\ell_2$ minimale en classification linéaire. En outre, cette thèse applique la méthode à la régression ridge et, plus généralement, aux méthodes spectrales (qui incluent la descente de gradient et le flux de gradient), permettant d'obtenir des descriptions non asymptotiques et précises de leur erreur d'estimation dans presque tout problème de régression linéaire.
Ces applications démontrent que la méthode permet non seulement aux statisticiens théoriciens de caractériser le risque excédentaire en population d'un estimateur, mais offre également un nouveau cadre théorique potentiel, fournissant une perspective analytique novatrice. Pour illustrer cela, nous utilisons le cadre de la méthode de Décomposition d'Espace de Caractéristiques pour définir trois concepts nouveaux : un ordre partiel sur l'ensemble des méthodes spectrales pour un problème de régression linéaire donné, un effet de saturation généralisé, et une définition mathématique de la propriété d'apprentissage de caractéristiques.
Outre les implications statistiques mentionnées précédemment, cette thèse établit également une généralisation du théorème de Dvoretzky-Milman pour la norme $\ell_q$ sous une mesure de probabilité générale. Ce résultat, qui découle naturellement du développement de la méthode de Décomposition d'Espace de Caractéristiques, relève des aspects géométriques de l'analyse fonctionnelle.

**Title :** Feature Space Decomposition

**Keywords :** Uniform Convergence Argument, Dvoretzky-Milman theorem, Fixed Points, Benign Overfitting, Ridge Regression, and Gradient Descent/Flow

**Abstract :** The problem of statistical prediction stands as one of the most central issues in the fields of statistical learning theory and mathematical statistics. The uniform convergence argument is one of the most fundamental methods for studying this problem. In recent years, the uniform convergence argument has proven inadequate for investigating overfitting estimators, particularly in its inability to explain the phenomenon of benign overfitting.
Motivated by this observation, this thesis proposes a systematic mathematical framework to refine the uniform convergence argument, termed the Feature Space Decomposition method, to study the benign overfitting phenomenon of minimum $\ell_q$-norm interpolant estimators in linear regression, and of minimum $\ell_2$-norm interpolant estimator in linear classification. Furthermore, this thesis applies the method to ridge regression and, more generally, to spectral methods (which include gradient descent, and gradient flow), thereby obtaining non-asymptotic, sharp descriptions of their estimation error in nearly any linear regression problem.
These applications demonstrate that the method not only aids theoretical statisticians in characterizing the population excess risk of an estimator but also offers a potential new theoretical framework, providing a novel perspective for analysis. To illustrate this point, we employ the framework of the Feature Space Decomposition method to define three new concepts: a partial order on the set of spectral methods for a given linear regression problem, a generalized saturation effect, and a mathematical definition of the feature learning property.
In addition to the aforementioned statistical implications, this thesis also establishes a generalization of the Dvoretzky-Milman theorem for the $\ell_q$ norm under a general probability measure. This result, which naturally arose from the development of the Feature Space Decomposition method, constitutes a conclusion within the geometric aspects of functional analysis.

**Institut Polytechnique de Paris**
91120 Palaiseau, France