ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Statistics
Prof. Dr. S. van de Geer

Master's Thesis in Applied Mathematics

Spring Semester 2023

Jivan Waber

Interactions between Benign Overfitting and Regularization

Supervisors: Prof. Dr. S. van de Geer & Dr. G. Chinot

August 2023

Abstract

Until recent years, the common wisdom among statisticians and machine learning researchers was to consider overfitting as a universally bad phenomenon. Research had shown that overfitting on training data led to poor generalization performances in many general settings. One incarnation of this phenomenon is the so-called bias-variance trade-off which implies that by reducing the bias of a model, necessarily, the variance of the model increases. However, with the emergence of deep learning, researchers have come to notice that some deep learning models are able to perform very well despite heavily overfitting the data, seemingly contradicting the common wisdom. This led researchers to investigate the concept of benign overfitting, overfitting that does not hurt generalization capabilities. In particular, in order to analyze this surprising phenomenon, they tried to understand under which circumstances it would arise in classical models that are simpler than neural networks. In this paper, we review a variety of results concerning benign overfitting that have been recently uncovered in the context of minimum l_2-norm interpolators for linear regression and ridge regression. Along the way, certain tools that are not in the traditional toolbox of researchers in the field such as the Dvoretzky-Milman theorem are presented in order to prove some of the results. These tools might provide a novel perspective on the way to approach certain problems in statistical learning theory. Finally, we make an attempt to generalize the results for minimum l_p-norm interpolators by adapting the geometric approach based on Dvoretzky-Milman theorem.

Contents

Chapter 1

Introduction

1.1 The framework of function approximation

In this section, we set the framework of the general problem of function approximation. To this end, we draw inspiration from the second chapter of the book [HTF09].

We start by presenting the general framework. Let $\mathcal{X}$ be the space of all possible inputs and $\mathcal{Y}$ the space of all possible outputs. Assume that there is some unknown joint probability distribution P on the set $\mathcal{X} \times \mathcal{Y}$. Given an random vector $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ We are looking for a function $f : \mathcal{X} \to \mathcal{Y}$ from a given function class $\mathcal{F}$, that predicts outputs Y from inputs X. In order to discriminate functions in terms of their prediction performance, we must introduce some notion of metric allowing us to quantify how well a function predicts the output. This motivates the definition of a loss functional $l(Y, f(X))$ whose purpose is to penalize the errors in predictions incurred by the choice of a certain function. The risk of f is defined as $\mathcal{R}(f) \coloneqq \mathbb{E}_{(X,Y)\sim P}[l(Y, f(X))]$. It measures how far the prediction $f(X)$ is from Y with respect to the loss in expectation, for a randomly drawn pair $(X, Y) \sim P$. The goal is then to find a prediction function $f \in \mathcal{F}$ that minimizes the risk $\mathcal{R}(f)$:

$$f^* \in \arg\min_{f\in\mathcal{F}} \mathcal{R}(f). \tag{1.1}$$

This framework is defined with the following situation in mind: we have limited data $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathcal{X} \times \mathcal{Y}$ that are input-output pairs drawn from an underlying unknown joint distribution P. We wish to find a function f that is able to predict new outputs $y \in \mathcal{Y}$ from new inputs $x \in \mathcal{X}$. In this setting, we do not have access to the risk since we do not know the probability distribution P; hence, we define the empirical risk that is designed to estimate the true risk, $\hat{\mathcal{R}}(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i)$. In order to get as close as possible to solving the problem (1.1) with our limited data, we look for a solution

$$\hat{f} \in \arg\min_{f\in\mathcal{F}} \hat{\mathcal{R}}(f). \tag{1.2}$$

This is the problem of empirical risk minimization (ERM).

From now on we will restrict our attention to the setting where $\mathcal{X} = \mathbb{R}^p$, and $\mathcal{Y} = \mathbb{R}$ for simplicity. Now that the general setting has been introduced, let us present some classical examples that fit into this framework.

1.2 Linear regression

Linear regression consists in predicting real-valued outputs y from an affine function of inputs x in $\mathbb{R}^p$. Consider data $(x_i, y_i)_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ coming from an unknown probability distribution P. The loss that is commonly considered in linear regression is the squared error loss, given by $l(Y, f(X)) \coloneqq (Y - f(X))^2$. As announced, the function class we restrict to in the linear model is $\mathcal{F}_L \coloneqq \{f : \mathbb{R}^p \to \mathbb{R} | f \text{ is affine}\}$. $\mathcal{F}_L$ can equivalently be written as $\mathcal{F}_L = \{\langle \beta, \cdot \rangle + \beta_0 | \beta \in \mathbb{R}^p, \beta_0 \in \mathbb{R}\}$. Moreover, we assume that the data obeys the relationship

$$y_i = f^*(x_i) + \xi_i = \beta_0^* + \langle \beta^*, x_i \rangle + \xi_i, \quad \forall i \in \{1, \dots, n\}, \tag{1.3}$$

where $\xi \in \mathbb{R}^n$ describes an unobserved noise that we model to be random, from a known probability distribution. In words, we assume there is a linear relationship between the inputs and outputs, modulo some noise ξ. For the remainder of the text, unless stated otherwise, we assume centered $\sigma_\xi-$subgaussian noise.

Now, let us turn our attention to what the linear regression solution looks like. For the remainder of this exposition we will use the convention that the input vector $x_i = (1, x_i)^\top \in \mathbb{R}^p$ for all $i \in \{1, \dots, n\}$ to simplify the notations by accounting for an intercept term. First of all, let $\mathbb{X} \in \mathbb{R}^{n \times p}$ denote the matrix with $i-$th row being the observation x_i for all $i \in \{1, \dots, n\}$, and $y \in \mathbb{R}^n$ denote the output vector $(y_1, \dots, y_n)^\top$. Notice that $\mathcal{F}_L = \{\langle \beta, \cdot \rangle | \beta \in \mathbb{R}^p\}$ and thus,

$$\begin{aligned}\min_{f \in \mathcal{F}_L} \hat{\mathcal{R}}(f) &= \min_{f \in \mathcal{F}_L} \frac{1}{n} \sum_{i=1}^n l(f(x_i), y_i) = \min_{f \in \mathcal{F}_L} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 \\ &= \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n (y_i - \langle \beta, x_i \rangle)^2 = \frac{1}{n} \min_{\beta \in \mathbb{R}^p} \|y - \mathbb{X}\beta\|_2^2.\end{aligned}$$

Due to the convexity of the objective, the solution to this minimization problem is obtained by setting the gradient with respect to β to zero. This derivation gives the so-called normal equation which the minimizer $\hat{\beta}$ must fulfill:

$$\mathbb{X}^\top \mathbb{X} \beta = \mathbb{X} y.$$

The solution to the normal equation depends on the rank of the matrix $\mathbb{X}$, as well as on the relationship between the dimension of the feature space p and the number of observations n. We distinguish the cases as follows. If $p = n$ or in the underparametrized regime, that is when $p < n$, if $\mathbb{X}$ has full rank p, then the normal equation has a unique solution given by $\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X} y$. If $\mathbb{X}$ does not have full rank, then the normal equation has infinitely many solutions. In the overparametrized regime, when $n < p$, if $\mathbb{X}$ has full rank n, the normal equation does not have a unique solution and in this case we can use regularization techniques to find a suitable solution. If $\mathbb{X}$ does not have full rank, then the normal equation has no solution. In this paper, we will explore the behaviour of the linear model exclusively in the overparametrized regime, with full rank design matrix $\mathbb{X}$; we will discuss certain regularization techniques that will help us choose particular solutions.

1.2.1 Squared error loss and bias-variance trade-off

In the context of the squared error loss, suppose that the class function $\mathcal{F}$ is unrestricted. Then the solution to the minimization problem (1.1) is given by

$$f^*(x) = \mathbb{E}(Y|X = x), \tag{1.4}$$

that is, the solution is defined as the expected value of Y given X. However, we do not have access to this expected value since the conditional probability distribution of Y given X is unknown. The next best thing that we have at our disposal is arguably the $k-$nearest-neighbor prediction, which is given by

$$\hat{f}(x) \coloneqq \frac{1}{k} \sum_{x_i \in N_k(x)} y_i,$$

where $N_k(x)$ denotes the neighborhood of x given by the k closest points x_i with respect to the Euclidean norm, and where k is a parameter to be chosen a priori. This solution might seem satisfactory at first glance but it has a major drawback. In order to maintain a certain density of points in the feature space $\mathbb{R}^p$ as the dimension p increases we need the number of observations n to grow exponentially with p. This is known as the curse of dimensionality. The solution above provides us with the best local estimate but in high dimension, with a limited number of observations, this local estimate turns out to be downright bad. This motivates the seemingly strong structural assumption on the function space that is made in the linear model. On top of that, the solution to (1.2) with respect to this function class is easily interpretable, which also explains the ubiquity of the linear model in applications.
The situation above illustrates the necessity of choosing the right amount of structural assumptions in order to find meaningful and robust solutions. To make this choice optimally and resolve the tension between expressivity and restrictiveness of the model, one popular formalism is given by the bias-variance trade-off.
The squared error loss has the particular property that its associated risk $\mathcal{R}(f) = \mathbb{E}_{(X,Y)}(Y - f(X))^2$, also called the mean squared error (MSE), admits a useful decomposition into bias and variance terms. The bias of a one-dimensional statistical estimator $\hat{\theta}$ of a parameter $\theta \in \Theta$ is defined as $\text{Bias}_\theta(\hat{\theta}, \theta) = \mathbb{E}_\theta(\hat{\theta} - \theta)$ and the variance as $\text{Var}_\theta(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2$ where the θ in the subscript of the expectation is to indicate that the expectation is taken only over θ. The decomposition of the MSE into bias and variance terms in the one-dimensional case is the following.

$$\text{MSE}_\theta(\hat{\theta}) = \text{Bias}_\theta(\hat{\theta}, \theta)^2 + \text{Var}_\theta(\hat{\theta}). \tag{1.5}$$

In higher dimensions, say for $\theta \in \Theta = \mathbb{R}^p$, the bias is still defined the same way, but instead of the variance, the term that will appear in the bias-variance decomposition is some value depending on the covariance matrix of θ. The bias-variance decomposition is given in this case by the following.

$$\text{MSE}_\theta\hat{\theta} = \|\text{Bias}_\theta(\hat{\theta}, \theta)\|_2^2 + \text{tr}(\text{cov}(\hat{\theta})). \tag{1.6}$$

Proof.

$$\begin{aligned}\text{MSE}_\theta\hat{\theta} = \mathbb{E}\|\hat{\theta} - \theta\|^2 &= \sum_{i=1}^{p} \mathbb{E}[\hat{\theta}_i - \theta_i]^2 = \sum_{i=1}^{p} \Big(\text{Bias}_\theta(\hat{\theta}_i, \theta_i)^2 + \text{Var}_{\theta_i}(\hat{\theta}_i)\Big) \\ &= \|\text{Bias}_\theta(\hat{\theta}, \theta)\|_2^2 + \text{tr}(\text{cov}(\hat{\theta})).\end{aligned}$$

□

The softer the structural assumptions made on the function class $\mathcal{F}$ the lower the bias because the model is able to find functions that are closer to the true function. However, the variance tends to be larger in that case, because the model is more sensitive to noise, as it has more freedom to fit to the observations. Conversely, the stronger the structural assumptions, the larger the bias and the lower the variance become because the model finds a solution from a restricted function class, and is thus less sensitive to noise and more prone to find solutions further from the observations. This is known as the bias-variance trade-off.
In the case of the linear model, we restrict ourselves to affine functions and therefore the bias will increase when the linear relationship between X and Y that we assume in (1.3) does not reflect the reality.

Therefore, the bias-variance trade-off is telling us that we should expect an algorithm that perfectly fits the training data (and minimizes the bias) to perform poorly on new unseen data. However, it has been observed in practice, that there are algorithms that generalize well on new data, despite interpolating the training data. In this manuscript, we review some papers and try to understand when and why this peculiar phenomenon can be observed in linear regression.

1.3 Ridge regression and LASSO

Ridge regression is a learning algorithm akin to linear regression coming with a regularization parameter $\lambda \in \mathbb{R}$. The minimization problem of ridge regression is the following.

$$\min_{\beta \in \mathbb{R}^p} \|y - \mathbb{X}\beta\|_2^2 + \lambda \|\beta\|_2^2.$$

where λ is a regularization hyperparameter that must be chosen a priori. The parameter λ produces a penalty in term of the norm of the estimator β. As λ tends to $+\infty$, the estimator β must shrink to 0 and thus the minimization problem finds solutions that have small l_2-norm. Conversely, as λ tends to 0, the regularization constraint on the estimator β vanishes and the minimization problem's solution coincides with linear regression. The case when $\lambda = 0$ is called the ridgeless case. In the overparametrized regime $(p > n)$, the solution to the ridge regression minimization problem is the following.

$$\hat{\beta} = \mathbb{X}^\top (\mathbb{X}\mathbb{X}^\top + \lambda I)^{-1} y.$$

LASSO is another regularized learning algorithm, where this time the penalty is in the l_1-norm of the estimator. The minimization problem of LASSO is the following.

$$\min_{\beta \in \mathbb{R}^p} \|y - \mathbb{X}\beta\|_2^2 + \lambda \|\beta\|_1$$

LASSO favors sparse solutions, that is, estimators that have only a few non-zero entries. At a heuristic level, this is due to the fact that the l_1-norm can be thought of as a convex surrogate to the so-called l_0-norm (which is not a norm) which is defined as follows.
Given any vector $\beta \in \mathbb{R}^p$,

$$\|\beta\|_0 = \text{`number of non-zero entries of } \beta'$$

Introducing a penalty in term of the l_0-norm would make the minimization problem non-convex and therefore hard to solve, that is why we resort to an l_1-penalty. The fact that the l_1-norm can serve as a surrogate to the l_0-norm can be understood at the intuitive level thanks to the observation that the shape of the l_1-ball favors sparsity, when compared to the geometry of the l_2-ball.

1.4 Minimum norm interpolators

In 1.2, we mention the need for regularization techniques to differentiate between solutions to the normal equation in the overparametrized regime; in the present section, we discuss the regularization consisting in picking a solution which has minimal norm among all possible solutions to the normal equation. Given a norm $\|\cdot\|$, a minimum norm interpolator is an estimator $\hat{\beta}$ that is a solution to the minimization problem

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|. \tag{1.7}$$

$$\text{s.t. } \mathbb{X}\beta = y \tag{1.8}$$

There are multiple reasons for being interested in minimum norm interpolators. A motivation for studying the interpolating regime is the observation of certain deep neural networks that perfectly fit the data and are still capable of good generalization, as in [ZBH$^+$16], [ZBH$^+$21], and [BHMM19a]. Another motivation, this time for studying interpolating estimators that have minimal norm, is the example of basis pursuit in compressed sensing that we are going to develop in 1.4.2.

1.4.1 l_2-norm

In the case of the l_2-norm, this problem admits a closed form solution given by the Moore-Penrose pseudoinverse, that is,

$$\hat{\beta} = \mathbb{X}^\dagger y = \mathbb{X}^\top (\mathbb{X}\mathbb{X}^\top)^{-1} y. \tag{1.9}$$

Proof. Let $\bar{\beta} = \mathbb{X}^\dagger y$, and let $\beta \in \mathbb{R}^p$, then,

$$\mathbb{X}\beta - y = \mathbb{X}(\beta - \mathbb{X}^\dagger y) + (I - \mathbb{X}\mathbb{X}^\dagger)(-y).$$

Since $I - \mathbb{X}\mathbb{X}^\dagger$ is the orthogonal projector onto $\text{Ker}(\mathbb{X}^\top) = \text{Ker}(\mathbb{X}^\dagger)$, the two terms in the right-hand side of the equation above are orthogonal. Therefore, by the Pythagorean theorem,

$$\|\mathbb{X}\beta - y\|^2 = \|\mathbb{X}(\beta - \mathbb{X}^\dagger y)\|^2 + \|(I - \mathbb{X}\mathbb{X}^\dagger)(-y)\|^2 = \|\mathbb{X}(\beta - \bar{\beta})\|^2 + \|\mathbb{X}\bar{\beta} - y\|^2 \geq \|\mathbb{X}\bar{\beta} - y\|^2.$$

□

The minimum l_2-norm interpolator is relevant beyond the fact that it admits a closed-form solution. For example, as discussed in [BHMM19b], there are certain types of neural networks for which the stochastic gradient descent (SGD) algorithm, with weights initialized at 0, converges to the minimum l_2-norm interpolator.

In the case of other l_p-norms, the minimization problem (1.7) is harder to approach. Let us now take a look at the case of the l_1-norm, this is the so-called problem of basis pursuit (BP).

1.4.2 l_1-norm and basis pursuit

Historically, this problem arose in the field of compressed sensing, whose goal is to find recovery guarantees of compressed signals. Hence, this part is inspired by lectures notes on compressed sensing [Tao09] and the paper [CW08]. The setup of compressed sensing is the following. In order to relate this problem to our setting, we use the same notations as before, which are perhaps more traditional to statistics and machine learning.
Let $\mathbb{X} \in \mathbb{R}^{n\times p}$, with $n < p$, $\beta \in \mathbb{R}^p$, and $y \in \mathbb{R}^n$. Suppose we want to measure a high-dimensional vector β that can be thought of as an observation of a signal from the real world such as an image or a video. Now, if we restrain ourselves to linear measurements, how many measurements do we need to recover the observation? That is the question compressed sensing is interested in answering. In mathematical terms, viewing the matrix $\mathbb{X}$ as a matrix whose rows perform linear measurements of β, the question can be formulated as the following. For which n can we solve the system of equations

$$\mathbb{X}\beta = y \tag{1.10}$$

for β. The key assumptions made in compressed sensing are about the sparsity of the signal captured by β as well as the incoherence of the measurement matrix $\mathbb{X}$. We assume that β is $s-$sparse, for some positive number s, where $s-$sparsity of a vector is defined as the vector having at most s non-zero coordinates. If the support of the vector β was known, the problem would be trivial, as one could simply reduce the problem to the recovery of an $s-$dimensional signal. But if the support is unknown, apart from the fact that it is $s-$dimensional, then what can we say about the numbers of measurements n required to reconstruct exactly β from the linear measurements in (1.10)?

Before answering this question, let us ask ourselves another one. Why would we assume sparsity in the first place? We can argue that most of what we do in statistics and machine learning is trying to uncover the underlying low-dimensional structure of noisy data with different techniques. On a more philosophical level, if there was no low-dimensional structure to be found in high dimensional data, it would make the search for meaning in the data foolish. This presupposes that there exists a certain basis or representation in which the data would be low-dimensional, and hence sparse. Going back to our previous question, compressed sensing gives a surprising answer; it states that the number of measurements required to exactly recover the signal β needs only be proportional to the sparsity level of the signal, that is $n \gtrsim s$. Of course,

to obtain this conclusion, we need to make some assumptions. Let us state a result from [Tao09].

Proposition 1.4.1. *Suppose that any $2s$ columns of the $\mathbb{X}$ are linearly independent. Then any $s-$sparse signal $\beta \in \mathbb{R}^p$ can be reconstructed uniquely from $\mathbb{X}\beta$.*

Proof. For a contradiction, suppose it is not the case. Then there exist two $s-$sparse signals $\beta, \beta' \in \mathbb{R}^p$ that satisfy $\mathbb{X}\beta = \mathbb{X}\beta'$, which implies that $\mathbb{X}(\beta - \beta') = 0$. But $\beta - \beta'$ is a $2s-$sparse vector, and hence there are $2s$ columns of $\mathbb{X}$ that are dependent. This gives a contradiction. □

The assumption in the proposition above is a quantification of the incoherence of the measurement matrix, and it is reasonable once $n \geq 2s$. Moreover, the proof shows how the $s-$sparse signal β can be reconstructed. It is the unique solution to the minimization problem

$$\beta = \arg\min_{\beta \in \mathbb{R}^p} \{\|\beta\|_0 : \mathbb{X}\beta = y\}, \tag{1.11}$$

where $\|\beta\|_0$ denotes the l_0-norm of β that we define in 1.3. However, this optimization problem is not convex and is computationally intractable. Therefore, we must resort to another way of finding the signal β. This leads to basis pursuit, which consists in solving the surrogate convex optimization problem

$$\beta = \arg\min_{\beta \in \mathbb{R}^p} \{\|\beta\|_1 : \mathbb{X}\beta = y\}, \tag{1.12}$$

which can be solved by reformulating it as a linear program. However, basis pursuit does not always recover the solution, so let us expand a bit on one case in which it can recover a solution, without going into the details. This leads to the introduction of the restricted isometry property (RIP).

Definition 1.4.1. For any integer $s > 0$, define the isometry constant δ_s of a matrix M as the smallest number such that

$$(1 - \delta_s)\|v\|_2^2 \leq \|Mv\|_2^2 \leq (1 + \delta_s)\|v\|_2^2$$

holds for all $s-$sparse vectors $v \in \mathbb{R}^p$.

This constant basically quantifies how close to an isometry the matrix M is for $s-$sparse vectors. We say that a matrix M obeys the RIP of order s if δ_s is not too close to 1. When this property holds, we are able to recover an approximation of the original signal. This is captured in the following result from [CRT05].

Theorem 1.4.2. *Assume that $\delta_{2s} < \sqrt{2} - 1$, then the solution to (1.12) obeys*

$$\|\beta^* - \beta\|_2 \leq c\|\beta - \beta_s\|_1/\sqrt{s} \quad \textit{and} \quad \|\beta^* - \beta\|_1 \leq c\|\beta - \beta_s\|_1,$$

for some constant c, where β_s is the vector β with all but the largest s components set to 0, and β^* denotes the true signal.

In particular, if β is $s-$sparse, then $\beta = \beta_s$ and we achieve exact recovery. There are many interesting matrices that satisfy some RIP with value s close to n. Some notable such matrices are random matrices that satisfy that kind of RIP with high probability: for example, the matrix consisting of column vectors drawn uniformly from the unit sphere in $\mathbb{R}^n$, or the matrix whose entries are i.i.d standard Gaussian random variables. These obey the condition of the above theorem with high probability, as soon as

$$n \geq c_1 s \log(n),$$

for some constant $c_1 > 0$.

Chapter 2

Benign Overfitting

Benign overfitting refers to the phenomenon of learning algorithms that fit to the training data very closely but still manage to generalize well on new data. In this chapter, we look at several results that try to uncover what characteristics the problem must have in order for benign overfitting to happen. We restrict ourselves to the specific case of linear regression as well as ridge regression, in high dimensions, and to the situation where the estimator $\hat{\beta}$ of the true parameter β^* interpolates the data $(\mathbb{X}, y)$. We first make an exposition of the results of Bartlett et al. from [BLLT20] that focus on linear regression. Then we discuss the results of Tsigler et al. from [TB20] focusing on ridge regression and compare them with [BLLT20] in the absence of regularization. Next, we analyze the paper of Lecué and Shang [LS22], devoted to the linear regression model as well, and contrast their results with the previous two papers, and the differences in their approach.

2.1 Notations

This section regroups a few notations that we use throughout the paper. Every quantity denoted by c, C, c_i, C_i for some $i \in \mathbb{Z}$ is a positive constant if not specified otherwise. For two vectors $a, b \in \mathbb{R}^l$, for some $l > 0$, we interchangeably denote their inner product by $a^\top b$ or $\langle a, b \rangle$. We denote the operator norm of a matrix M by $\|M\|_{op}$, and if it is a square matrix, we denote its trace by $\mathrm{tr}(M)$. We use $\lesssim$ and $\gtrsim$ for inequality up to a constant and write a $\approx$ b if both $a \lesssim b$ and $a \gtrsim b$ hold. Given p the dimension of the feature space $\mathbb{R}^p$, we denote $f(p) \asymp g(p)$ as $p \to +\infty$ if there exist two positive constants C_1, C_2 and two positive integers p_1, p_2 such that

$$|f(p)| \geq C_1 g(p) \quad \forall p \geq p_1, \quad \text{and} \quad |f(p)| \leq C_2 g(p) \quad \forall p \geq p_2. \tag{2.1}$$

2.2 Benign Overfitting in linear regression

The paper [BLLT20] examines in which regime benign overfitting can or cannot occur in the linear regression model with subgaussian assumptions on the data (see A.1.1). The main tools it uses are concentration of subgaussian and subexponential random variables as well as classical generic chaining results for empirical processes to provide

sharp upper bounds on the excess risk, a measure of the generalization error. In order to do that, the excess risk is split into two terms that are analyzed separately. The results they obtain only depend on the covariance matrix Σ and not on the interplay between the true parameter β^* and Σ; as we will see later, this may not provide the full picture.

2.2.1 Setup of the paper of Bartlett et al.

In [BLLT20], they consider a general feature space $\mathbb{H}$ that is a separable Hilbert space. However, we will only present their results for $\mathbb{H} = \mathbb{R}^p$ because it is the most interesting case for our purpose. The linear regression setup is the following.
We consider an input-output pair $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$ coming from a centered distribution P. We assume that $Y = X^\top \beta^* + \epsilon$ where $\beta^* \in \mathbb{R}^p$ is defined as in the introduction, and ϵ is some centered random noise coming from a distribution that will be specified below. We let the covariance Σ have the spectral decomposition $\Sigma = U\Lambda U^\top$ with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ and corresponding eigenvectors $u_1, \ldots, u_p$. We assume that $X = U\Lambda^{1/2}Z$ where Z has components that are independent σ_X-subgaussian random variables, for some positive constant σ_X. Now let's specify the distribution of the noise. We assume that ϵ is $\sigma_\xi-$subgaussian conditionally on X, that is, for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\lambda\epsilon|X] \leq \exp\left(C\sigma_\xi^2\lambda^2\right),$$

and that the conditional noise variance is bounded from below by some constant σ^2,

$$\mathbb{E}[\epsilon^2|X] \geq \sigma^2.$$

Now let us consider a training sample $(x_i, y_i)_{i=1}^n$ of n independent random variables coming from P. Define $\mathbb{X} \in \mathbb{R}^p \times \mathbb{R}^n$ and $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ as in the introduction, and $\xi \coloneqq (\epsilon_i, \ldots, \epsilon_n) \in \mathbb{R}^n$. With these notations, we are in the setup of linear regression where $y = \mathbb{X}\beta^* + \xi$. The last assumption is that, almost surely, the projection of $\mathbb{X}$ on the space orthogonal to any eigenvector of Σ spans a space of dimension n.

Remark 2.2.1. Notice that in particular all the assumptions made in the setup are fulfilled for (X, Y) coming from a centered multivariate normal distribution and when $\text{rank}(\Sigma) > n$.

The situation they are interested in is the one where the parameter β interpolates the data and hence they consider the minimum l_2-norm interpolator that is introduced in 1.4 and which admits a closed form solution.

Notations

Let us fix some more notations that will become handy when we discuss the papers [TB20], [LS22]. Given $k \leq p$, let us denote the subspaces of the feature space $\mathbb{R}^p$ spanned by the first k eigenvectors $u_1, \ldots, u_k$ of Σ, respectively the last $p-k$ eigenvectors $u_{k+1}, \ldots, u_p$ by $V_{1:k}$, respectively $V_{k+1:p}$. Accordingly, let $P_{1:k}$ and $P_{k+1:p}$ denote the orthogonal projections onto $V_{1:k}$ and $V_{k+1:p}$ respectively, and let $\beta_{1:k} \coloneqq P_{1:k}\beta$ and $\beta_{k+1:p} \coloneqq P_{k+1:p}\beta$. Let us also denote $\Sigma_{1:k} = U\Lambda_{1:k}U^\top$ where $\Lambda_{1:k} \coloneqq \text{diag}(\lambda_1, \ldots, \lambda_k, 0, \ldots, 0)$

and $\Sigma_{k+1:p}$ analogously. Finally let $X_{1:k} := \mathbb{X}P_{1:k}$ and $X_{k+1:p} := \mathbb{X}P_{k+1:p}$ so that $\mathbb{X} = X_{1:k} + X_{k+1:p}$.

Definition 2.2.1. Given a covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$, and given k such that $\lambda_{k+1} > 0$, we define the effective ranks

$$r_k(\Sigma) := \frac{\Sigma_{i>k}\lambda_i}{\lambda_{k+1}} = \frac{\mathrm{tr}(\Sigma_{k+1:p})}{\|\Sigma_{k+1:p}\|_{op}}, \quad \text{and,} \tag{2.2}$$

$$R_k(\Sigma) := \frac{(\Sigma_{i>k}\lambda_i)^2}{\Sigma_{i>k}\lambda_i^2} = \frac{(\mathrm{tr}(\Sigma_{k+1:p}))^2}{\mathrm{tr}(\Sigma_{k+1:p}^2)}. \tag{2.3}$$

2.2.2 Excess risk and bias-variance decomposition

Definition 2.2.2. The excess risk of an estimator $\hat{\beta}$ is defined as the difference between the risk of $\hat{\beta}$ conditioned on the data $(\mathbb{X}, y)$ and the risk of the true parameter β^*.

$$\mathcal{R}(\hat{\beta}) := R(\hat{\beta}) - R(\beta^*) = \mathbb{E}[(Y - X^\top\hat{\beta})^2|(\mathbb{X}, y)] - \mathbb{E}[(Y - X^\top\beta^*)^2].$$

In our setting, the excess risk can be decomposed into a bias and variance term which allows us to nicely decouple the noise and the signal in the analysis of the excess risk. In the following lemma we show how the decomposition is made in the case of data $X \sim \mathcal{N}(0, \Sigma)$. Then in Lemma 2.3.2, we present a similar decomposition with a proof that is more technical due to the weaker assumptions. Nevertheless, we think Lemma 2.2.1 is as enlightening if not more for the understanding of the decomposition and the quantities involved.

Lemma 2.2.1. *For $X \sim \mathcal{N}(0, \Sigma), Y = X^\top\beta^* + \epsilon, \xi \sim \mathcal{N}(0, v_\xi^2 I)$, and $\epsilon \sim \mathcal{N}(0, v_\xi^2)$ independent of X,*

$$\mathcal{R}(\hat{\beta}) = \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 = \|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I)\beta^*\|_2^2 + \|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1})\xi\|_2^2,$$

and moreover,

$$\mathbb{E}_\xi\mathcal{R}(\hat{\beta}) = \|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I)\beta^*\|_2^2 + v_\xi^2\,\mathrm{tr}((\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X}\Sigma\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1})$$

Proof. Let us start by proving the first equality.

$$\begin{aligned}
\mathcal{R}(\hat{\beta}) &= \mathbb{E}[(Y - X^\top\hat{\beta})^2|(\mathbb{X}, y)] - \mathbb{E}[(Y - X^\top\beta^*)^2] = \mathbb{E}[(X^\top(\beta^* - \hat{\beta}) + \epsilon)^2|(\mathbb{X}, y)] - \mathbb{E}[\epsilon^2] \\
&= \mathbb{E}[(X^\top(\beta^* - \hat{\beta}))^2|(\mathbb{X}, y)] + 2\mathbb{E}[(X^\top(\beta^* - \hat{\beta})|(\mathbb{X}, y)] \cdot \mathbb{E}[\epsilon] = \mathbb{E}[(X^\top(\beta^* - \hat{\beta}))^2|(\mathbb{X}, y)] \\
&= \mathbb{E}[X^\top(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^\top X|(\mathbb{X}, y)] = \mathbb{E}[\mathrm{tr}(X^\top(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^\top X)|(\mathbb{X}, y)] \\
&= \mathbb{E}[\mathrm{tr}((\beta^* - \hat{\beta})^\top XX^\top(\beta^* - \hat{\beta}))|(\mathbb{X}, y)] = \mathrm{tr}((\beta^* - \hat{\beta})^\top\mathbb{E}[XX^\top](\beta^* - \hat{\beta})) \\
&= \mathrm{tr}((\beta^* - \hat{\beta})^\top\Sigma(\beta^* - \hat{\beta})) = (\beta^* - \hat{\beta})^\top\Sigma(\beta^* - \hat{\beta}) = \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2.
\end{aligned}$$

For the second equality, we plug in the value of $\hat{\beta}$, the minimum l_2-norm interpolator.

$$\begin{aligned}
\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 &= (\hat{\beta} - \beta^*)^\top\Sigma(\hat{\beta} - \beta^*) = \beta^{*\top}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I)^\top\Sigma(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I)\beta^* \\
&\quad + \xi^\top(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1})^\top\Sigma(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1})\xi \\
&= \|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I)\beta^*\|_2^2 + \|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1})\xi\|_2^2.
\end{aligned}$$

For the last equality, we use the same trick with the trace as for the first equality.

$$\begin{aligned}\mathbb{E}_\xi\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1})\xi\|_2^2 &= \operatorname{tr}(\Sigma(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1})\mathbb{E}_\xi[\xi\xi^\top](\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1})^\top)\\ &= v_\xi^2 \operatorname{tr}((\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X}\Sigma\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}).\end{aligned}$$

□

Definition 2.2.3. We define the bias of the excess risk as

$$B \coloneqq \|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I)\beta^*\|_2^2$$

and the variance of the excess risk as

$$V \coloneqq \operatorname{tr}((\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X}\Sigma\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1})/v_\xi^2.$$

Remark 2.2.2. Notice that the quantities in Definition 2.2.3 do not depend on the noise ξ. Furthermore, Lemma 2.2.1 and Definition 2.2.3 give us the nice expression

$$\mathbb{E}_\xi\mathcal{R}(\hat{\beta}) = B + v_\xi^2 V.$$

Remark 2.2.3. It turns out that the effective ranks of the covariance matrix are the right notions to control the excess risk.

2.3 First sharp bound on the excess risk

We have now all the necessary tools to state the main theorem of the paper of Bartlett et al. ([BLLT20, Theorem 4]).

Theorem 2.3.1. *For any σ_X, there are $b, c, c_1 > 1$ for which the following holds. Consider a linear regression problem satisfying the conditions listed above. Define*

$$k^* \coloneqq \min\{k > 0 : r_k(\Sigma) \geq bn\}, \tag{2.4}$$

where the minimum of the empty set is defined to be ∞. Suppose that $\delta < 1$ with $\log(1/\delta) < n/c$. If $k^* \geq n/c_1$, then $\mathbb{E}\mathcal{R}(\hat{\beta}) \geq \sigma^2/c$. Otherwise,

$$\begin{aligned}\mathcal{R}(\hat{\beta}) \leq & c\|\beta^*\|_2^2\|\Sigma\|_{op}\max\left\{\sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}}\right\}\\ &+ c\log(1/\delta)\sigma_\xi^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)\end{aligned}$$

with probability at least $1-\delta$, and

$$\mathbb{E}\mathcal{R}(\hat{\beta}) \geq \frac{\sigma^2}{c}\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right).$$

Overview of the findings

Bartlett et al. were able to show in [BLLT20] that under suitable assumptions on the covariance matrix, specifically on its effective ranks r_k and R_k, benign overfitting is observed with constant probability, that is, there are good sharp upper bounds on the excess risk $\mathcal{R}(\hat{\beta})$. Informally, the spectrum of the covariance matrix must be such that the eigenvalues that are more attributable to noise than signal do not decay too fast, in order for the estimator to spread the noise over the dimensions of these eigenvalues. In that setting, the estimator $\hat{\beta}$ will exhibit benign overfitting properties. The most important contributions of this work are probably the definition of k^* and the implicit feature space decomposition that happens in the proof of Lemma 2.3.3.

Theorem 2.3.1 motivates the following definition.

Definition 2.3.1. Let Σ_n denote the covariance matrix for data with n observations and let $k_n^* \coloneqq \min\{k > 0 : r_k(\Sigma_n) \geq bn\}$. We say that a sequence of covariance matrices Σ_n is benign if

$$\lim_{n\to\infty} \frac{r_0(\Sigma_n)}{n} = \lim_{n\to\infty} \frac{k_n^*}{n} = \lim_{n\to\infty} \frac{n}{R_{k_n^*}(\Sigma_n)} = 0.$$

Remark 2.3.1. When the sequence of covariance matrices is benign, the excess risk of the estimator $\hat{\beta}$ converges to 0 as n tends to $+\infty$.

The proof of Theorem 2.3.1 relies on a handful of lemmas, ranging from linear algebra and convergence results to concentration of probability arguments. We are going to examine parts of the proof more in detail to understand the phenomenons at play and get a sense of the proof technique. Although the proof of the lower bound on the excess risk is of interest, we focus our attention to the proof of the upper bound since the purpose of the lower bound is only to show we cannot expect a much better upper bound.
The proof starts by finding bounds on the excess risk, depending on two random matrices that depend on the data. It is some kind of bias-variance decomposition that is of the same nature as the one we perform in Lemma 2.2.1.

Lemma 2.3.2 (Lemma 7 in [BLLT20])**.** *The excess risk of the minimum norm estimator satisfies*

$$\mathcal{R}(\hat{\beta}) \leq 2{\beta^*}^\top D\beta^* + c\sigma^2 \log(1/\delta)\operatorname{tr}(C)$$

with probability at least $1-\delta$ *over* ϵ*, and*

$$\mathbb{E}_\epsilon \mathcal{R}(\hat{\beta}) \geq {\beta^*}^\top D\beta^* + \sigma^2 \operatorname{tr}(C),$$

where

$$D = (I - \mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X})\Sigma(I - \mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X}), \quad \textit{and} \quad C = (\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X}\Sigma\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}.$$

The proof of Lemma 2.3.2 mostly consists in algebraic manipulations that are not enlightening so we skip it; it must be noted however that it requires a high probability upper bound on some quadratic form of C by its trace that is outside the scope of this

exposition. Now that we have control over the excess risk with respect to quantities depending on B and C, the rest of the proof of Theorem 2.3.1 consists in bounding them efficiently.
The upper bound on ${\beta^*}^\top D\beta^*$ relies on standard arguments and directly uses a result from another paper. Since $I - \mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X}$ is a projection, it has operator norm bounded by 1. Then by rearranging the quadratic form, we can obtain

$$ {\beta^*}^\top D\beta^* \leq \left\| \Sigma - \frac{1}{n}\mathbb{X}^\top\mathbb{X} \right\|_{op} \|\beta^*\|_2^2. $$

From there, using [KL14, Theorem 9], we immediately recover the bound that makes Theorem 2.3.1 work. Note that [KL14, Theorem 9] is not trivial at all and relies on tools such as generic chaining for empirical processes. The same kind of tools are needed in some part of the proof of the main theorem of [LS22].
The crux of the paper of Bartlett et al., however, is bounding the trace of C. The upper bound they obtain is the following.

Lemma 2.3.3 (Lemma 11 in [BLLT20])**.** *There are constants $b, c \geq 1$ such that if $0 \leq k \leq n/c, r_k(\Sigma) \geq bn$, and $l \leq k$, then with probability at least $1 - 7e^{-n/c}$,*

$$ \mathrm{tr}(C) \leq c\left(\frac{l}{n} + n\frac{\sum_{i>l}\lambda_i^2}{(\sum_{i>k}\lambda_i)^2} \right). $$

Then to be able to piece everything together to obtain Theorem 2.3.1, we also need to find which l minimizes the upper bound of Lemma 2.3.3. This is elucidated by the next lemma whose proof relies solely on clever rearrangements and on the definition of k^*.

Lemma 2.3.4 (Lemma 17 in [BLLT20])**.** *For any $b \geq 1$ and k^* as in Theorem 2.3.1, if $k^* < \infty$, we have*

$$ \min_{l \leq k^*}\left(\frac{l}{bn} + bn\frac{\sum_{i>l}\lambda_i^2}{(\lambda_{k^*+1}r_{k^*}(\Sigma))^2} \right) = \frac{k^*}{bn} + bn\frac{\sum_{i>l}\lambda_i^2}{(\lambda_{k^*+1}r_{k^*}(\Sigma))^2} = \frac{k^*}{bn} + \frac{bn}{R_{k^*}(\Sigma)}. $$

By putting together the bound on ${\beta^*}^\top B\beta^*$ and the one on $\mathrm{tr}(C)$ from Lemma 2.3.4 into Lemma 2.3.2, we obtain the upper bound in Theorem 2.3.1. We now spend some time on the proof of Lemma 2.3.3.

2.3.1 Proof of Lemma 2.3.3

The first part of the proof consists in writing the trace of C as a function of independent subgaussian random variables, thanks to the assumptions on our model. The technique to perform this rephrasing of $\mathrm{tr}(C)$ is purely algebraic; it relies on the Sherman-Morrison-Woodbury formula, see [BLLT20, Lemma 20] for the proof.

Lemma 2.3.5 (Lemma 8 in [BLLT20])**.** *Consider the covariance matrix Σ with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p$ such that $\lambda_n > 0$ and spectral decomposition $\Sigma = \sum_j \lambda_j u_j u_j^\top$ where the*

orthonormal vectors $u_j \in \mathbb{R}^p$ are the eigenvectors corresponding to the eigenvalue λ_j, for all $j \in \{1, \dots, p\}$. For i with $\lambda_i > 0$, define $z_i = \mathbb{X}u_i/\sqrt{\lambda_i}$. Then,

$$\mathrm{tr}(C) = \sum_i \left[\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i \right],$$

and these z_i are independent σ_X-subgaussian with unit variance. Furthermore, for any i with $\lambda_i > 0$, we have

$$\lambda_i^2 z_i^\top \left(\sum_j \lambda_j z_j z_j^\top \right)^{-2} z_i = \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2},$$

where $A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top$.

With this new expression of $\mathrm{tr}(C)$, we can now make use of probabilistic arguments to find high probability bounds by exploiting the subgaussianity. Let us introduce the following quantities.

$$A = \sum_i \lambda_i z_i z_i^\top, \quad A_{-i} = \sum_{j \neq i} \lambda_j z_j z_j^\top, \quad A_k = \sum_{i > k} \lambda_i z_i z_i^\top,$$

and denote the $i-$th largest eigenvalue of a matrix M by $\mu_i(M)$. We arrive now at the heart of the proof which consists in controlling the spectrum of these matrices as well as the squared norms of multiple random vectors with subgaussian entries simultaneously, both with high probability. This is done in the two following lemmas that eventually conclude the proof of Lemma 2.3.3.

Lemma 2.3.6 (Lemma 9 in [BLLT20])**.** *There is a universal constant c such that with probability at least $1 - 2e^{-n/c}$,*

$$\frac{1}{c} \sum_i \lambda_i - c\lambda_1 n \leq \mu_n(A) \leq \mu_1(A) \leq c \left(\sum_i \lambda_i + \lambda_1 n \right).$$

Proof. We give now a proof sketch of Lemma 2.3.6. Let us first sketch the reasoning of an $\epsilon-$net argument. To obtain the result, we want to control the operator norm of A which corresponds to controlling the quadratic form $v^\top A v$ for any vector $v \in \mathbb{S}_2^{n-1}$, the l_2-sphere of dimension $n-1$. The strategy is to find a high probability upper bound on the quadratic form for a single vector v. Then we would like to simply use a union bound over all such vectors v but we cannot do that over all of $\mathbb{S}^{n-1}$. We are therefore using an $\epsilon-$net to control the operator norm of A with the maximum of the quadratic forms $v^\top A v$, only for v in the $\epsilon-$net $\mathcal{N}_\epsilon$. According to [BLLT20, Lemma 25], for all $\epsilon < 1/2$,

$$\mu_1(A) = \|A\|_{op} \leq (1 - \epsilon)^2 \max_{v \in \mathcal{N}_\epsilon} \|v^\top A v\|.$$

Note that an $\epsilon-$net of $\mathbb{S}_2^{n-1}$ has cardinality $|\mathcal{N}_\epsilon| = (C\epsilon)^n$, for some constant $C > 0$.
The second step consists in noticing that for any v and i, $v^\top z_i$ is a $c_1\|v\|\sigma_X-$subgaussian random variable (see Example A.1.2), for some constant c_1 and thus the quantity $v^\top A v = \sum_i \lambda_i (v^\top z_i)^2$ can be controlled via Bernstein-like bounds. Indeed notice that

since $v^\top z_i$ is $c_1\|v\|_2\sigma_X-$subgaussian, $(v^\top z_i)^2$ is $c_1^2\|v\|_2^2\sigma_X^2-$subexponential (see Lemma A.1.1), Now we are able to show that $v^\top Av$ concentrates with high probability around $\mathrm{tr}(\Sigma) = \sum_i \lambda_i$. In order to do that, we can use Corollary A.1.4 for $X_i = \lambda_i((v^\top z_i)^2 - 1)$ and $a_i = 1$, for all $0 \leq i \leq n$ and $(X_i, a_i) = (-\lambda_i, 1)$ otherwise to obtain that for all $t > 0$, with probability at least $1 - 2e^{-t}$,

$$\left| v^\top Av - \sum_i \lambda_i \right| \leq c_2\sigma_X^2 \max\left(\lambda_1 t, \sqrt{t\sum_i \lambda_i^2}\right),$$

for some constant c_2.
The last step is to union bound over the vectors in $\mathcal{N}_\epsilon$ to get the high probability bound on the eigenvalues of A that we need. Notice that a union bound is now possible since it is over a finite number of vectors. Without getting into the details, this yields the result of the lemma. □

Corollary 2.3.7 (Lemma 10 in [BLLT20]). *There are constants $b, c \geq 1$ such that for any $k \geq 0$, with probability at least $1 - 2e^{-n/c}$,*

1. *for all $i \geq 1$,*

$$\mu_{k+1}(A_{-i}) \leq \mu_{k+1}(A) \leq \mu_1(A_k) \leq c\left(\sum_{j>k} \lambda_j + \lambda_{k+1}n\right),$$

2. *for all $1 \leq i \leq k$,*

$$\mu_n(A) \geq \mu_n(A_{-i}) \geq \mu_n(A_k) \geq \frac{1}{c}\sum_{j>k} \lambda_j - c\lambda_{k+1}n,$$

3. *if $r_k(\Sigma) \geq bn$,*

$$\frac{1}{c}\lambda_{k+1}r_k(\Sigma) \leq \mu_n(A_k) \leq \mu_1(A_k) \leq c\lambda_{k+1}r_k(\Sigma).$$

Remark 2.3.2. The proof of the corollary relies on Courant-Fischer Theorem and does not provide any deep insight so we skip it.

The second lemma is the following.

Lemma 2.3.8 (Corollary 24 in [BLLT20]). *There is a universal constant c such that for any centered random vector $z \in \mathbb{R}^n$ with independent $\sigma-$subgaussian coordinates with unit variances, any random subspace $\mathcal{L}$ of $\mathbb{R}^n$ of codimension k that is independent of z, and any $t > 0$, with probability at least $1 - 3e^{-t}$,*

$$\|z\|_2^2 \leq n + c\sigma^2(t + \sqrt{tn}), \quad \textit{and} \quad \|\Pi_{\mathcal{L}} z\|_2^2 \geq n - c\sigma^2(k + t + \sqrt{nt}),$$

where $\Pi_{\mathcal{L}}$ is the orthogonal projection onto $\mathcal{L}$.

Remark 2.3.3. Lemma 2.3.8 shows how the the sum of squares of subgaussian random variables behave and concentrate, as well as how much information or energy of the vector is preserved with high probability by a random projection onto a lower-dimensional subspace.

Proof. Let us sketch the proof of Lemma 2.3.8. The first inequality relies on Corollary A.1.4, for $X_i = z_i^2 - 1$ and $a_i = 1$ for all $0 \leq i \leq n$ to obtain for an absolute constant c and for all $t > 0$, with probability at least $1 - 2e^{-t}$,

$$|\|z\|_2^2 - n| = \left| \sum_i^n (z_i^2 - 1) \right| \leq c\sigma^2 \max(t, \sqrt{nt}).$$

Then the second inequality relies on an upper bound on the the quadratic form $z^\top M z$ by the trace of M, where $M = \Pi_{\mathcal{L}^\perp}^\top \Pi_{\mathcal{L}^\perp}$, This upper bound is out of the scope of this exposition and is the same as the one we allude to when we present Lemma 2.3.2. It allows us to obtain, with a union bound to control $\|z\|_2^2$ simultaneously,

$$\|\Pi_{\mathcal{L}} z\|_2^2 = \|z\|_2^2 - \|\Pi_{\mathcal{L}^\perp} z\|_2^2 \geq \|z\|_2^2 - \sigma^2(2k + 4t) \geq n - \sigma^2(2k + 4t + c\max(t, \sqrt{nt}),$$

with probability at least $1 - 3e^{-t}$. □

Now we are ready to provide a proof sketch of Lemma 2.3.3.

Proof. The proof relies on rewriting the trace of C with the expression in Lemma 2.3.5 and by splitting the sum as follows

$$\operatorname{tr}(C) = \sum_i \lambda_i^2 z_i^\top A^{-2} z_i = \sum_i^l \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} + \sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i. \tag{2.5}$$

The terms on the right-hand side are handled separately by leveraging the low dimensionality of the first term and the small magnitude of the last $p - l$ eigenvalues of the second. For the first one, thanks to the assumption on the effective rank $r_k(\Sigma) \geq bn$, we can use Lemma 2.3.7 to bound the following quadratic forms with high probability for all $0 \leq i \leq l$.

$$z^\top A_{-i}^{-2} z \lesssim \frac{\|z\|_2^2}{(\lambda_{k+1} r_k(\Sigma))^2}, \quad \text{and} \quad z^\top A_{-i}^{-1} \gtrsim \frac{\|\Pi_{\mathcal{L}_i}\|_2^2}{\lambda_{k+1} r_k \Sigma},$$

where $\mathcal{L}_i$ is the span of the $n-k$ eigenvectors of A_{-i} corresponding to its smallest $n-k$ eigenvalues. All in all, we obtain

$$\frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \lesssim \frac{\|z\|_2^2}{\|\Pi_{\mathcal{L}_i} z_i\|_2^4}.$$

Then, by applying Lemma 2.3.8 l times, and taking the union bound over all z_i for $0 \leq i \leq l$, we obtain that with high probability,

$$\sum_i^l \frac{\lambda_i^2 z_i^\top A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^\top A_{-i}^{-1} z_i)^2} \lesssim \frac{l}{n}. \tag{2.6}$$

For the second term in Equation (2.5), we use again a combination of Lemma 2.3.6 and Corollary A.1.4 to obtain the high probability bound

$$\sum_{i>l} \lambda_i z_i^\top A^{-2} z_i \lesssim n \frac{\sum_{i>l} \lambda_i^2}{(\lambda_{k+1} r_k(\Sigma))^2}. \tag{2.7}$$

The first step is to notice that with high probability $\mu_n(A) \gtrsim \lambda_{k+1} r_k(\Sigma)$ and thus,

$$\sum_{i>l} \lambda_i^2 z_i^\top A^{-2} z_i \lesssim \frac{\sum_{i>l} \lambda_i^2 \|z_i\|_2^2}{(\lambda_{k+1} r_k(\Sigma))^2},$$

and then to realize that the $\lambda_i^2 \|z_i\|_2^2$ are subexponential random variables for all $i > l$ and use Corollary A.1.4 to upper bound their sum. Assembling the two bounds in Equations (2.6) and (2.7) together gives us the result. □

2.4 Benign Overfitting in ridge regression

In this section, we briefly review the work of Tsigler et al., in the paper [TB20]. We do not look at it in its full generality but only consider non-negative regularization. Let us introduce the setup. We want to upper bound the excess risk of the minimum l_2-norm estimator of ridge regression as introduced in 1.3 in the overparametrized regime, that is, when $p > n$. Its solution is given by $\hat{\beta} = \mathbb{X}^\top (\mathbb{X}\mathbb{X}^\top + \lambda I)^{-1} y$. The assumptions made on the model are the same as the ones in 2.2.1, except for the fact that we do not assume independence of the observations, but instead assume some regularity in the form of a bound on the condition number of the tail of the matrix $\mathbb{X}\mathbb{X}^\top + \lambda I$.

Definition 2.4.1. We define a new quantity, similar to the effective rank but that takes into account the regularization λ.

$$\rho_k \coloneqq \frac{\sum_{i>k} \lambda_i + \lambda}{n \lambda_{k+1}}$$

2.4.1 Main result

According to Tsigler et al., [TB20, Theorem 1],

Theorem 2.4.1. *Fix any constants $b > 0, L > 0$. Let $k^* = \min\{\kappa > 0 : \rho_\kappa \leq b\}$ and $A_k \coloneqq X_{k+1:p} X_{k+1:p}^\top + \lambda I$.*

There exists a constant c which only depends on σ_X, b, L such that the following holds. Suppose that for some $\bar{k} < n/c$ and $\delta \leq 1 - ce^{-n/c}$, with probability at least $1 - \delta$ the matrix $A_{\bar{k}}$ is positive-definite (PD) with condition number at most L. Take $k = \min\{\bar{k}, k^*\}$. Then with probability at least $1 - ce^{-n/c} - \delta$,

$$B \leq c \left(\|\Sigma_{k+1:p}^{1/2} \beta_{k+1:p}^*\|_2^2 + \|\Sigma_{1:k}^{-1/2} \beta_{1:k}^*\|_2^2 \left(\frac{\lambda + \sum_{i>k} \lambda_i}{n} \right)^2 \right), \quad \textit{and}$$

$$V \leq c \left(\frac{k}{n} + \frac{n \sum_{i>k} \lambda_i^2}{(\lambda + \sum_{i>k} \lambda_i)^2} \right).$$

where B and V are the quantities defined in 2.2.2.

Remark 2.4.1. Therefore, we obtain an upper bound on the excess risk by following the excess risk decomposition in 2.2.2.

Remark 2.4.2. In the ridgeless setting, the similarity between the results of [TB20] and [BLLT20] becomes more evident as the bounds become

$$B \leq c \left(\|\Sigma_{k+1:p}^{1/2} \beta_{k+1:p}^*\|_2^2 + \|\Sigma_{1:k}^{-1/2} \beta_{1:k}^*\|_2^2 \left(\frac{\lambda_{k+1} r_k(\Sigma)}{n} \right)^2 \right), \quad \text{and}$$

$$V \leq c \left(\frac{k}{n} + \frac{n}{R_k(\Sigma)} \right).$$

Let us describe some of the main differences between Theorem 2.3.1 and Theorem 2.4.1 in the ridgeless setting. First of all, Tsigler et al. relax the independence condition on the data and focus instead on the condition number of the tail of the matrix $\mathbb{X}\mathbb{X}^\top + \lambda I$, in an attempt to generalize the results of [BLLT20]. Then, they proceed to show that the conditions given by Bartlett et al. make the assumption on the condition number hold with high probability. However, they do not provide compelling evidence that their assumption on the condition number can be fulfilled in other interesting regimes. Secondly, contrary to [BLLT20], whose main feat is the sharp bound on the variance term, Tsigler et al. manage to additionally control the bias term in a sharper way. Their bound relies upon the realization that the implicit feature space decomposition into two components that takes place in the proof of Lemma 2.3.3 can also be used to control the bias. As can be seen in Theorem 2.4.1, the upper bound on the bias–which could potentially be dominated by the behavior of the norm of β^* in [BLLT20]–now depends on two norms in terms of $\beta_{1:k}^*$ and $\beta_{k+1:p}^*$. Furthermore, these norms exploit the entanglement of the pair (Σ, β^*). This relationship, that is absent in [BLLT20], gives a richer picture of the phenomena at play. Moreover, the l_2-norm of β^* that appears in [BLLT20] can be significant in the setting we examine, since the dimension of the feature space is assumed to be large. Concerning the variance term however, the result of [TB20] and [BLLT20] coincide. All things considered, from our perspective, the two major contributions of Theorem 2.4.1 are the exploitation of the feature space decomposition into a low-dimensional and a high-dimensional part for both the bias and the variance, and the appearance of the relationship between the covariance matrix and the true parameter.

Example 2.4.1. *Let us build an example where the bias term in [TB20] is much better than the bias term in [BLLT20] to illustrate the improvement made in [TB20]. Let* $\beta^* = (1, \dots, 1)^\top$, $\epsilon \in (0, 1)$, $\Sigma =$*diag*$(1, \dots, 1, \epsilon, \dots, \epsilon)$, *and let* $\tilde{k}$ *denote the number of eigenvalues of* Σ *that are equal to* 1. *Therefore,* $k^* = \tilde{k}$ *since* $r_{\tilde{k}}(\Sigma) = p - k \gtrsim n$ *but* $r_{\tilde{k}-1}(\Sigma) = (p-k)\epsilon$, *and thus, by taking* ϵ *small enough, we cannot have* $r_{\tilde{k}-1}(\Sigma) \gtrsim n$. *We have that* $r_0(\Sigma) = k^* + (p - k^*)\epsilon$, $\|\beta^*\|_2^2 = p$ *and* $\|\Sigma\|_{op} = 1$. *Therefore, we obtain the following bias term in Theorem 2.3.1, as* ϵ *goes to* 0.

$$\|\beta^*\|_2^2 \|\Sigma\|_{op} \max \left\{ \sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n}, \sqrt{\frac{\log(1/\delta)}{n}} \right\} = p \frac{k^* + (p - k^*)\epsilon}{n},$$

which tends to $p\frac{k^*}{n}$ *as* ϵ *goes to* 0. *However, in Theorem 2.4.1, we obtain the following bias*

term.

$$\left(\|\Sigma_{k+1:p}^{1/2}\beta_{k+1:p}^*\|_2^2 + \|\Sigma_{1:k}^{-1/2}\beta_{1:k}^*\|_2^2 \left(\frac{\lambda + \sum_{i>k}\lambda_i}{n} \right)^2 \right) = (p-k^*)\epsilon^2 + k^* \left(\frac{\epsilon}{n} \right)^2,$$

which tends to 0 *as* ϵ *goes to* 0*.*

2.4.2 Feature space decomposition

We decompose the feature space as $\mathbb{R}^p = V_{1:k} \oplus V_{k+1:p}$ according to the notations from 2.2.1 and analyze $\hat{\beta}_{1:k}$ and $\hat{\beta}_{k+1:p}$ separately. $\hat{\beta}_{1:k}$ corresponds to a prediction component whereas $\hat{\beta}_{k+1:p}$ corresponds to an overfitting component. Indeed, it turns out that the role of $\hat{\beta}_{1:k}$ is to estimate the essential part of the true parameter given by $\beta_{1:k}^*$, while the role of $\hat{\beta}_{k+1:p}$ is more or less to fit to the noise. This is shown in Proposition 2.5.1 in the next section. Moreover, we also go over the notion of implicit regularization that is discussed in [TB20] in the next section as well.

2.5 The geometric viewpoint of Benign Overfitting

In this section, we look at the techniques and results introduced in [LS22]. This paper restricts to the case of a design matrix $\mathbb{X}$ with independent centered Gaussian entries and anisotropic covariance matrix Σ. In this context, it improves the main theorems from [BLLT20], [TB20] (Theorems 2.3.1 and 2.4.1) and perhaps more importantly it provides a new geometric perspective on these results. The main result of [LS22] relies on very different tools, namely Dvoretzky-Milman theorem and some type of restricted isomorphy property.

2.5.1 Setup of the paper of Lecué et al.

We consider the now familiar linear regression problem in the overparametrized regime, but this time with Gaussian assumptions on the data instead of the subgaussianity assumed previously. That is, $y = \mathbb{X}\beta^* + \xi$, where $\mathbb{X} \in \mathbb{R}^{n\times p}$ is a Gaussian matrix with i.i.d. $\mathcal{N}(0,\Sigma)$ row vectors, $\xi \sim \mathcal{N}(0, v_\xi^2 I)$ is some independent Gaussian noise, $\beta^* \in \mathbb{R}^p$ is the unknown true parameter of the model, and $p > n$. We write

$$\mathbb{X} = \mathbb{G}^{(n\times p)}\Sigma^{1/2},$$

where $\mathbb{G}^{(n\times p)}$ is an $n \times p$ random matrix with i.i.d. $\mathcal{N}(0,1)$ entries.

In the context of [LS22], as discussed in Lemma 2.2.1, the excess risk has the form

$$\mathcal{R}(\hat{\beta}) = \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2.$$

2.5.2 Feature space decomposition and self-induced regularization

According to [LS22, Proposition 3], we can split the minimum l_2-norm interpolator as follows.

Proposition 2.5.1. *We have* $\hat{\beta} = \hat{\beta}_{1:k} + \hat{\beta}_{k+1:p}$ *where*

$$\hat{\beta}_{1:k} \in \underset{\beta_1 \in \mathbb{R}^p}{\arg\min} \left(\|X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} (y - X_{1:k}\beta_1)\|_2^2 + \|\beta_1\|_2^2 \right) \quad \textit{and} \tag{2.8}$$

$$\hat{\beta}_{k+1:p} = X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} (y - X_{1:k}\hat{\beta}_{1:k}). \tag{2.9}$$

Proof. The minimum l_2-norm interpolator is given by $\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} (\|\beta\|_2^2 : \mathbb{X}\beta = y)$. Since for all $\beta \in \mathbb{R}^p$, $\|\beta\|_2^2 = \|\beta_{1:k}\|_2^2 + \|\beta_{k+1:p}\|_2^2$, $\hat{\beta} = \hat{\beta}_{1:k} + \hat{\beta}_{k+1:p}$ and

$$\begin{aligned}(\hat{\beta}_{1:k}, \hat{\beta}_{k+1:p}) &\in \underset{(\beta_1,\beta_2) \in V_{1:k} \times V_{k+1:p}}{\arg\min} \left(\|\beta_1\|_2^2 + \|\beta_2\|_2^2 : X_{1:k}\beta_1 + X_{k+1:p}\beta_2 = y \right) \\ &= \underset{(\beta_1,\beta_2) \in \mathbb{R}^p \times \mathbb{R}^p}{\arg\min} \left(\|\beta_1\|_2^2 + \|\beta_2\|_2^2 : X_{1:k}\beta_1 + X_{k+1:p}\beta_2 = y \right).\end{aligned}$$

For the last equality, one inequality is trivial since we search over a larger space, and for the other inequality, suppose that there exists $(\hat{\beta}_1, \hat{\beta}_2) \notin V_{1:k} \times V_{k+1:p}$ for a contradiction. Without loss of generality, suppose that $\hat{\beta}_1 \notin V_{1:k}$, then

$$\hat{\beta}_1 = \hat{\beta}_1' + \hat{\beta}_1'', \quad \text{where} \quad \hat{\beta}_1' \in V_{1:k}, 0 \neq \hat{\beta}_1'' \in V_{k+1:p}.$$

$X_{1:k}\hat{\beta}_1 = G^{(n\times p)}\Sigma_{1:k}^{1/2}\hat{\beta}_1 = G^{(n\times p)}\Sigma_{1:k}^{1/2}\hat{\beta}_1' + G^{(n\times p)}\Sigma_{1:k}^{1/2}\hat{\beta}_1'' = G^{(n\times p)}\Sigma_{1:k}^{1/2}\hat{\beta}_1'$. Therefore, using $\hat{\beta}_1'$ instead of $\hat{\beta}_1$, we obtain a solution to $X_{1:k}\beta_1 + X_{k+1:p}\beta_2 = y$ that has strictly smaller l_2-norm and we get a contradiction. Now, optimizing separately, first in β_2 for a fixed β_1,

$$\hat{\beta}_{k+1:p}(\beta_1) \in \underset{\beta_2 \in \mathbb{R}^P}{\arg\min} (\|\beta_2\|_2^2 : X_{k+1:p}\beta_2 = y - X_{1:k}\beta_1),$$

and a solution is given by $X_{k+1:p}^\dagger (y - X_{1:k}\beta_1)$ where $X_{k+1:p}^\dagger$ denotes the Moore-Penrose pseudoinverse of $X_{k+1:p}$ which is given by $X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1}$. Thus,

$$\hat{\beta}_{k+1:p}(\beta_1) = X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} (y - X_{1:k}\beta_1).$$

Optimizing in β_1,

$$\begin{aligned}\hat{\beta}_{1:k} &\in \underset{\beta_1 \in \mathbb{R}^p}{\arg\min} \left(\|\beta_1\|_2^2 + \|\hat{\beta}_{k+1:p}(\beta_1)\|_2^2 \right) \\ &= \underset{\beta_1 \in \mathbb{R}^p}{\arg\min} \left(\|\beta_1\|_2^2 + \|X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} (y - X_{1:k}\beta_1)\|_2^2 \right),\end{aligned}$$

and thus,

$$\hat{\beta}_{k+1:p} = \hat{\beta}_{k+1:p}(\hat{\beta}_{1:k}) = X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} (y - X_{1:k}\hat{\beta}_{1:k}).$$

□

Remark 2.5.1. Proposition 2.5.1 has a deep meaning. It demonstrates how the solution to the minimum l_2-norm interpolation problem decouples into two components whose interpretations are different.
First of all, $\hat{\beta}_{1:k}$ is the solution to a minimization problem that is reminiscent of ridge regression but that does not quite coincide with it. Later, we formally see that $\hat{\beta}_{1:k}$ approximately corresponds to a ridge estimator with regularization coefficient $\mathrm{tr}(\Sigma_{k+1:p})$, that is, $\hat{\beta}_{1:k}$ is approximately a solution to

$$\underset{\beta_1 \in \mathbb{R}^p}{\arg\min} \left(\|y - X_{1:k}\beta_1\|_2^2 + \mathrm{tr}(\Sigma_{k+1:p})\|\beta_1\|_2^2 \right)$$

This formalization is achieved through the use of Dvoretzky-Milman theorem, which in this case allows us to say that with high probability, under some conditions on the the number of observations n and the covariance matrix Σ, $\|X_{k+1:p}^\top (X_{k+1:p}X_{k+1:p}^\top)^{-1}\cdot\|_2$ is isomorphic (that is, equivalent up to absolute constants) to $\mathrm{tr}(\Sigma_{k+1:p})^{-1/2}\|\cdot\|_2$. This ridge regression appearing seemingly out of nowhere is an incarnation of the so-called self-regularization property of the estimator $\hat{\beta}$, (see [BMR21]). The fact that we can think of $\hat{\beta}_{1:k}$ as a ridge estimator means we have at our disposal many tools to find the solution, as the regularized least squares problem has been extensively studied and is solvable in a variety of settings. In 2.5.5, we use classical techniques to arrive at a solution.
Secondly, $\hat{\beta}_{k+1:p}$ corresponds to the solution to the ordinary linear regression problem, as an estimator of $y - X_{1:k}\hat{\beta}_{1:k}$. Notice that $y - X_{1:k}\hat{\beta}_{1:k}$ essentially corresponds to the noise, because $X_{1:k}\hat{\beta}_{1:k}$ is an estimator of y as k is chosen in a way that implicitly states that the data is essentially only $k-$dimensional, that is, most of the energy of the parameter resides in $V_{1:k}$. This reflection leads to the following interpretation: $\hat{\beta}_{k+1:p}$ perfectly fits the noise and hence corresponds to the overfitting component of the estimator $\hat{\beta}$. Therefore, we end up with a solution $\hat{\beta}$ to the linear regression problem which decouples into two parts; the first one behaves like a ridge estimator–and acts as a prediction of the true parameter β^*–while the second one behaves like an overfitting component.

Excess risk decomposition

Following the feature space decomposition, we make the excess risk decomposition

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 = \|\Sigma_{1:k}^{1/2}(\hat{\beta}_{1:k} - \beta_{1:k}^*)\|_2^2 + \|\Sigma_{k+1:p}^{1/2}(\hat{\beta}_{k+1:p} - \beta_{k+1:p}^*)\|_2^2.$$

The plan for the remainder of the section is to present the main theorem of [LS22] together with a sketch of its proof. In the latter, we display a proof technique to find solutions to regularized linear regression problems. We start by introducing the central theorems used by Lecué et al. to derive their main result.

2.5.3 Dvoretzky-Milman theorem and restricted isomorphy property

Dvoretzky-Milman theorem was initially a statement about the existence of Euclidean sections of convex bodies. More specifically, it was about finding the largest dimension

guaranteeing the existence of an Euclidean section in a fixed convex body K, say in $\mathbb{R}^p$. Roughly speaking, an Euclidean section of a convex body $K \subset \mathbb{R}^p$ is a subset of K that is the intersection between a subspace of $\mathbb{R}^p$ and K, and that approximately looks like a Euclidean ball. The largest dimension for this to hold is called the Dvoretzky dimension.

Definition 2.5.1. Let $\|\cdot\|$ be a norm on $\mathbb{R}^p$ and denote by B^* the dual ball with respect to the norm $\|\cdot\|$ (see B.1). The Dvoretzky dimension $d_*(B)$ is defined as follows.

$$d_*(B) \coloneqq \left(\frac{w(B^*)}{\text{diam}(B^*, l_2^p)}\right)^2,$$

where $\text{diam}(B^*, l_2^p) \coloneqq \sup(\|v\|_2 : v \in B^*)$ and $w(B^*)$ is the Gaussian width of B^* (see C.2).

Remark 2.5.2. Notice that this notion of Dvoretzky dimension is analogous to the notion of stable dimension discussed in $C.2$, and thus the intuition built for stable dimension carries over.

Example 2.5.1 (Dvoretzky dimension of an ellipsoid)**.** *Given $\Sigma \in \mathbb{R}^{p\times p}$ a positive semi-definite (PSD) matrix and the unit l_2-ball $B_2^p \subset \mathbb{R}^p$, consider the ellipsoid $\Sigma^{-1/2}B_2^p = \{v \in \mathbb{R}^p : \|\Sigma^{1/2}v\| \leq 1\}$. Then the Dvoretzky-Milman dimension of $\Sigma^{-1/2}B_2^p$ can be bounded as follows.*

$$\frac{\text{tr}(\Sigma)}{4\|\Sigma\|_{op}} \leq d_*(\Sigma^{-1/2}B_2^p) \leq \frac{\text{tr}(\Sigma)}{\|\Sigma\|_{op}}.$$

Proof. Fist of all,

$$\text{diam}((\Sigma^{-1/2}B_2^p)^*, l_2^p) = \text{diam}(\Sigma^{1/2}B_2^p, l_2^p) = \sqrt{\lambda_1} = \sqrt{\|\Sigma\|_{op}}.$$

Secondly,

$$\begin{aligned} w((\Sigma^{-1/2}B_2^p)^*) = w(\Sigma^{1/2}B_2^p) &= \mathbb{E}\sup_{v\in B_2^p}\langle \Sigma^{1/2}v, g\rangle = \mathbb{E}\sup_{v\in B_2^p}\langle v, \Sigma^{1/2}g\rangle \\ &= \mathbb{E}\|\Sigma^{1/2}g\|_2 \leq \sqrt{\mathbb{E}\|\Sigma^{1/2}g\|_2^2} = \sqrt{\text{tr}(\Sigma)}, \end{aligned}$$

where $g \sim \mathcal{N}(0, I)$ and we use Jensen's inequality.
It can also be shown that $\mathbb{E}\|\Sigma^{1/2}g\|_2 \geq \sqrt{\text{tr}(\Sigma)/2}$ but this requires tools that are outside the scope of our exposition.
Assembling everything together gives us the required bound. □

Remark 2.5.3. Notice that the quantity

$$\frac{\text{tr}(\Sigma)}{\|\Sigma\|_{op}}$$

coincides with the effective rank $r_0(\Sigma)$ introduced in the review of the paper [BLLT20]. Therefore, we can now give a new meaning to the condition $r_k(\Sigma) \gtrsim n$ corresponding

to the interpretation of Dvoretzky-Milman in terms of Euclidean sections. This condition is equivalent to asking that the Dvoretzky dimension is larger than n (up to a constant), and hence, that there exists a Euclidean section of that dimension of the ellipsoid $\Sigma_{k+1:p}^{-1/2} B_2^p$. This means that a certain projection of this ellipsoid in lower dimensional space looks like a Euclidean ball. Moreover, the statement of Dvoretzky-Milman we give here also shows that we can find such a projection with high probability, and thus that this phenomenon happens most of the time. This captures geometrically the behavior of the random matrices $\mathbb{X}\mathbb{X}^\top$ and $\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}$ that we care about in the linear regression problem, as demonstrated in Corollary 2.5.3 below.

The statement of Dvoretzky-Milman we present follows a more modern viewpoint, but is equivalent to the original statement. It is a probabilistic statement that emphasizes an important caveat of Dvoretzky-Milman theorem shared by several results of the same flavour, like Johnson-Lindenstrauss lemma (see for example [Ver18, Theorem 5.3.1]); even though the existence of such an Euclidean section is guaranteed, we don't know of any deterministic way to find it, and thus we resort to the use of probabilistic methods. These can often be stated in terms of random matrices that can be interpreted as random projections of vectors onto lower dimensional subspaces, and that is the case here.
According to [LS22, Theorem 3],

Theorem 2.5.2 (Dvoretzky-Milman)**.** *Let $\|\cdot\|$ be a norm on $\mathbb{R}^p$. Denote by $\mathbb{G} \coloneqq \mathbb{G}^{(n\times p)}$ the $n \times p$ matrix with i.i.d. $\mathcal{N}(0,1)$ Gaussian entries. There are absolute constants $\kappa_{DM} \leq 1$ and c_0 such that the following holds. Assume that $n \leq \kappa_{DM} d_*(B)$, then with probability at least $1 - \exp(-c_0 d_*(B))$, for every $v \in \mathbb{R}^n$,*

$$\frac{1}{\sqrt{2}}\|v\|_2 w(B^*) \leq \|\mathbb{G}^\top v\| \leq \sqrt{\frac{3}{2}}\|v\|_2 w(B^*).$$

The proof of Theorem 2.5.2 can be found in [Ver18] but must be adapted to hold with high probability.

Corollary 2.5.3. *Given $\mathbb{G}$ defined as in Theorem 2.5.2 and $\Gamma \in \mathbb{R}^{p\times p}$ a semi-definite matrix, let $\mathbb{X}_2 \coloneqq \mathbb{G}\Gamma^{1/2}$. Assume that $n \leq \kappa_{DM} d_*(\Gamma^{-1/2} B_2^p)$, then with probability at least $1 - \exp(-c_0 d_*(\Gamma^{-1/2} B_2^p))$,*

$$\left\|\mathbb{X}_2\mathbb{X}_2^\top - w(\Gamma^{1/2}B_2^p)^2 I\right\|_{op} \leq \frac{1}{2} w(\Gamma^{1/2}B_2^p)^2,$$

which implies that

$$\sqrt{s_1(\mathbb{X}_2\mathbb{X}_2^\top)} = s_1(\mathbb{X}_2) \leq \sqrt{3/2}\, w(\Sigma^{1/2}B_2^p) \leq \sqrt{3\,\mathrm{tr}(\Gamma)/2}$$
$$\sqrt{s_n(\mathbb{X}_2\mathbb{X}_2^\top)} = s_n(\mathbb{X}_2) \geq 1/\sqrt{2}\, w(\Sigma^{1/2}B_2^p) \geq \sqrt{\mathrm{tr}(\Gamma)}/2,$$

and

$$\frac{2}{\sqrt{\mathrm{tr}(\Gamma)}} \geq s_1(\mathbb{X}_2^\top(\mathbb{X}_2\mathbb{X}_2^\top)^{-1}) \geq s_n(\mathbb{X}_2^\top(\mathbb{X}_2\mathbb{X}_2^\top)^{-1}) \geq \sqrt{\frac{2}{3\,\mathrm{tr}(\Gamma)}},$$

where for any matrix M, $s_i(M)$ denotes its $i-$th singular value.

Proof. The proof is an immediate consequence of Theorem 2.5.2 together with the bound on the Dvoretzky dimension of an ellipsoid detailed above in Example 2.5.1. □

Remark 2.5.4. In particular, Corollary 2.5.4 allows us to express $\hat{\beta}_{1:k}$, the prediction component of the minimum l_2-norm interpolator, as a ridge estimator since we get that, for all $v \in \mathbb{R}^n$,

$$\|X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} v\|_2 \approx \left(\operatorname{tr}\left(\Sigma_{k+1:p}\right)\right)^{-1/2} \|v\|_2,$$

as previously discussed in Remark 2.5.1.

The other tool we need to understand and prove the main theorem of [LS22] is the following. According to [LS22, Corollary 1],

Theorem 2.5.4 (Restricted Isomorphy Property)**.** *There are absolute constants $0 < \kappa_{iso} < 1, c_0$, and c_1 such that the following holds. If Γ is a semi-definite $p \times p$ matrix of rank k such that $k \leq \kappa_{iso} n$, then for $\mathbb{X}_1 = \mathbb{G}^{(n\times p)}\Gamma^{1/2}$, with probability at least $1 - c_0 \exp(-c_1 n)$, for all $v \in$ range(Γ),*

$$\frac{1}{\sqrt{2}}\|\Gamma^{1/2} v\|_2 \leq \|\mathbb{X}_1 v\|_2 \leq \sqrt{\frac{3}{2}}\|\Gamma^{1/2} v\|_2.$$

Remark 2.5.5. The proof of Theorem 2.5.4 can be obtained by using an $\epsilon-$net argument.

2.5.4 Main theorem

We state the main theorem of Lecué et al., in [LS22] and discuss with varying degrees of details its proof as well as the tools that are required to understand its assumptions and proof. According to Lecué et al., [LS22, Theorem 4],

Theorem 2.5.5. *There are absolute constants c_0, c_1 and C_0 such that the following holds. We assume that $n \gtrsim \log p$ and that there exists $k \leq \kappa_{iso} n$ such that $n \leq \kappa_{DM} d_*(\Sigma_{k+1:p}^{-1/2} B_2^p)$. We also assume that $\lambda_1 n \geq \operatorname{tr}\left(\Sigma_{k+1:p}\right)$. Then the following holds for all such k. We define*

$$J_1 = \left\{ j \in [k] : \lambda_j \geq \frac{\operatorname{tr}(\Sigma_{k+1:p})}{n} \right\}, \quad J_2 = \left\{ j \in [k] : \lambda_j < \frac{\operatorname{tr}(\Sigma_{k+1:p})}{n} \right\}$$

and $\Sigma_{1,thres}^{-1/2} \coloneqq U \Lambda_{1,thres}^{-1/2} U^\top$ where U is the orthogonal matrix appearing in the SVD of Σ and

$$\Lambda_{1,thres}^{-1/2} \coloneqq \operatorname{diag}\left(\left(\lambda_1 \vee \frac{\operatorname{tr}\left(\Sigma_{k+1:p}\right)}{n}\right)^{-1/2}, \ldots, \left(\lambda_k \vee \frac{\operatorname{tr}\left(\Sigma_{k+1:p}\right)}{n}\right)^{-1/2}, 0, \ldots, 0 \right)$$

With probability at least $1 - c_0 \exp\left(-c_1 \left(|J_1| + n\left(\sum_{j\in J_2} \lambda_j\right) / \left(\operatorname{tr}\left(\Sigma_{k+1;p}\right)\right)\right)\right)$,

$$\left\|\Sigma^{1/2}\left(\hat{\beta} - \beta^*\right)\right\|_2 \lesssim \square + v_\xi \frac{\sqrt{n \operatorname{tr}\left(\Sigma_{k+1:p}^2\right)}}{\operatorname{tr}\left(\Sigma_{k+1:p}\right)} + \left\|\Sigma_{k+1:p}^{1/2} \beta_{k+1:p}^*\right\|_2$$

where

$$\square = C_0 \max\left\{ v_\xi \sqrt{\frac{|J_1|}{n}}, v_\xi \sqrt{\frac{\sum_{j\in J_2}\lambda_j}{\operatorname{tr}\left(\Sigma_{k+1:p}\right)}}, \left\|\Sigma_{k+1:p}^{1/2}\beta^*_{k+1:p}\right\|_2, \left\|\Sigma_{1,thres}^{-1/2}\beta^*_{1:k}\right\|_2 \frac{\operatorname{tr}\left(\Sigma_{k+1:p}\right)}{n} \right\}.$$

Remark 2.5.6. As discussed and proven in [LS22], the best feature space decomposition $\mathbb{R}^p = V_{J_1} \oplus V_{J_2}$ is given by $J_1 = \{1, \dots, k\}$, $J_2 = \{k+1, \dots, p\}$, that is, $\mathbb{R}^p = V_{1:k} \oplus V_{k+1:p}$. However, Lecué et al. show that similar results still hold for any decomposition into two subspaces of eigenvectors. Moreover, it is also shown that the best k is given by k^* defined in Theorem 2.3.1. This means that the best feature decomposition only depends on Σ and not on β^*. In that case, Theorem 2.5.5 resembles the earlier results Theorem 2.3.1 and Theorem 2.4.1.

Remark 2.5.7. The result also holds for the case when $\lambda_1 n < \operatorname{tr}\left(\Sigma_{k+1:p}\right)$, but in this case the quantity $\square$ is equal to something else.

$$\square = C_0 \max\left\{ v_\xi \sqrt{\frac{\operatorname{tr}\left(\Sigma_{1:k}\right)}{\operatorname{tr}\left(\Sigma_{k+1:p}\right)}}, \sqrt{\frac{n\lambda_1}{\operatorname{tr}\left(\Sigma_{k+1:p}\right)}} \left\|\Sigma_{k+1:p}^{1/2}\beta^*_{k+1:p}\right\|_2, \|\beta^*_{1:k}\|_2 \sqrt{\frac{\operatorname{tr}\left(\Sigma_{k+1:p}\right)}{n}} \right\}$$

However, this case must be thought of as pathological for us because it is the case where the ridge regularization coefficient $\operatorname{tr}\left(\Sigma_{k+1:p}\right)$ is so large that the regularization term dominates the least squares term $\|y - X_{1:k}\beta_{1:k}\|_2^2$. This leads the generalization error to come from a mixture of errors in the prediction component $\beta^*_{1:k}$ and the overfitting component $\beta^*_{k+1:p}$, to which the prediction component contributes significantly.This is not the regime we are interested in understanding.

Theorem 2.5.5 leads to a more refined definition of the benign overfitting regime (in the context of Theorem 2.5.5).

Definition 2.5.2. Overfitting is benign for the pair (Σ, β^*) if there exists $k^* = o(n)$ such that

$$\lambda_{k^*} n \le \operatorname{tr}(\Sigma_{k^*+1:p}), \quad \lambda_1 n \ge \operatorname{tr}(\Sigma_{k^*+1:p}) \quad \text{and} \quad n \operatorname{tr}(\Sigma^2_{k^*+1:p}) = o((\operatorname{tr}(\Sigma_{k^*+1:p}))^2),$$

$$\|\Sigma^{1/2}_{k^*+1:p}\beta^*_{k^*+1:p}\|_2 = o(1) \quad \text{and} \quad \|\Sigma^{-1/2}_{1:k^*}\beta^*_{1:k^*}\|_2 \frac{\operatorname{tr}(\Sigma_{k^*+1:p})}{n} = o(1).$$

The notable addition of this definition is that it depends not only on the covariance matrix Σ as was the case in the definitions provided in [BLLT20] and [TB20], but also on the true parameter β^*.

Let us now discuss the main differences of the approach in [LS22] with respect to what is done in [BLLT20] and [TB20]. The main distinction is the fact that Theorem 2.5.5 does not rely on the bias-variance decomposition that is made in both [BLLT20] and [TB20], but instead relies on a splitting into a prediction component and an overfitting component. Also, the fact that the phenomenon of benign overfitting depends on both Σ and β^* is an idea that is not captured in the other two previous papers. Furthermore, Theorem 2.5.5 shows that as soon as we are in the regime of benign overfitting,

we will see it happen with high probability. Lastly, although restricted to the case of anisotropic Gaussian design matrix, the results obtained by Lecué et al. may be generalizable to other regimes. Indeed, they mainly rely on Dvoretzky-Milman theorem, a result of geometric flavor that can be extended to random matrices whose entries come from other distributions. However, the results from Bartlett et al. and Tsigler et al. cannot be generalized in the same way because their proofs entirely rely on certain subgaussian assumptions.

2.5.5 Proof of the main theorem

We provide a sketch of the proof to get a big picture of how the main arguments fit together.
The goal is to control the square root of the excess risk which is given by the quantity

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2.$$

Following the excess risk decomposition in 2.5.2, the proof consists of two main parts; the first part is concerned with controlling $\|\Sigma_{1:k}^{1/2}(\hat{\beta}_{1:k} - \beta^*_{1:k})\|_2$ while the second consists in controlling $\|\Sigma_{k+1:p}^{1/2}(\hat{\beta}_{k+1:p} - \beta^*_{k+1:p})\|_2$. In both parts, the control will be achieved on a certain high probability event where Dvoretzky-Milman theorem and a restricted isomorphy property hold simultaneously.
To simplify notations, we define $\beta_1 \coloneqq \beta_{1:k}$ and $\beta_2 \coloneqq \beta_{k+1:p}$ for all $\beta \in \mathbb{R}^p$ such that $\beta = \beta_{1:k} + \beta_{k+1:p}$, where $\beta_{1:k} \in V_{1:k}$ and $\beta_{k+1:p} \in V_{k+1:p}$, and $X_1 \coloneqq X_{1:k}$, $X_2 \coloneqq X_{k+1:p}$. We also use the notation $\Sigma_1 \coloneqq \Sigma_{1:k}, \Sigma_2 \coloneqq \Sigma_{k+1:p}$.

Bound on the prediction component

This part is devoted to find a high probability upper bound on the prediction component. First of all, we place ourselves on a high probability event where we have strong control over the norms of $X_2^\top v$ for all $v \in \mathbb{R}^n$ and $X_1\beta_1$ thanks to Theorems 2.5.2 and 2.5.4 as follows.
Let Ω_0 be the event onto which the following hold.

1. for all $v \in \mathbb{R}^n$, $\frac{1}{2\sqrt{2}}\sqrt{\mathrm{tr}(\Sigma_2)}\|v\|_2 \leq \|X_2^\top v\|_2 \leq \frac{3}{2}\sqrt{\mathrm{tr}(\Sigma_2)}\|v\|_2$,
2. for all $\beta_1 \in V_{1:k}$, $\frac{1}{2}\|\Sigma_1^{1/2}\beta_1\|_2 \leq \frac{1}{\sqrt{n}}\|X_1\beta_1\| \leq \frac{3}{2}\|\Sigma_1^{1/2}\beta_1\|_2$.

By Theorems 2.5.2 and 2.5.4, if $n \leq \kappa_{DM} d_*(\Sigma_2^{-1/2}B_2^p)$ and $k \leq \kappa_{iso} n$, then $P[\Omega_0] \geq 1 - c_0\exp(-c_1 n)$.
Let $A \coloneqq X_2^\top(X_2X_2^\top)^{-1}$, then by Proposition 2.5.1,

$$\hat{\beta}_1 \in \underset{\beta_1 \in V_{1:k}}{\arg\min} \left(\|A(y - X_1\beta_1)\|_2^2 + \|\beta_1\|_2^2 \right), \tag{2.10}$$

because the minimizer $\hat{\beta}_1$ actually lives in $V_{1:k} \subset \mathbb{R}^p$.
Now let us give an overview of what the proof technique to bound $\|\Sigma_1^{1/2}(\hat{\beta}_1 - \beta_1^*)\|_2$ consists of. We define a certain convex function $\mathcal{L}$ of β_1, whose minimum is attained

at β_1^* and that admits a 'quadratic + multiplier + regularization' decomposition. Then the goal is to show that necessarily, $\hat{\beta}_1$ lives in a ball around β_1^* with respect to a certain norm that tells us something about $\|\Sigma_1^{1/2}(\hat{\beta}_1 - \beta_1^*)\|_2$ and $\|\hat{\beta}_1 - \beta_1^*\|_2$. In order to do so, we assume that β_1 lives outside this ball and show that this β_1 does not attain the minimum of $\mathcal{L}$. Therefore, by definition of $\hat{\beta}_1$, this β_1 cannot be $\hat{\beta}_1$, and that implies that $\hat{\beta}_1$ lives in the ball around β^*. Coincidentally, this yields a bound on $\|\Sigma_1^{1/2}(\hat{\beta}_1 - \beta_1^*)\|_2$ thanks to the careful choice of norm. The delicate part is showing that β_1 does not attain the minimum of $\mathcal{L}$ for all β_1 outside the ball centered around β_1^*; this part actually splits into two distinct cases that embody two different behaviors and that must be treated distinctly. Let us now define the function $\mathcal{L}$, its decomposition, as well as the norm whose ball we want to show $\hat{\beta}_1$ belongs to.

Definition 2.5.3. Let $\mathcal{L} : \mathbb{R}^p \to \mathbb{R}, \beta_1 \mapsto \mathcal{L}(\beta_1)$ with

$$\mathcal{L}(\beta_1) \coloneqq \|A(y - X_1\beta_1)\|_2^2 + \|\beta_1\|_2^2 - (\|A(y - X_1\beta_1^*)\|_2^2 + \|\beta_1^*\|_2^2).$$

It can be rearranged to obtain the decomposition

$$\mathcal{L}(\beta_1) = \|(X_2X_2^\top)^{-1/2}X_1(\beta_1 - \beta_1^*)\|_2^2 + 2\langle X_1^\top(X_2X_2^\top)^{-1}(X_2\beta_2^* + \xi) - \beta_1^*, \beta_1 - \beta_1^*\rangle + \|\beta_1 - \beta_1^*\|_2^2, \tag{2.11}$$

where $\langle \cdot, \cdot \rangle$ denote the usual inner product.

Remark 2.5.8. $\mathcal{L}(\hat{\beta}_1)$ is the empirical excess risk of the estimator $\hat{\beta}_1$ for the ridge regression problem (2.10). The decomposition (2.11) enables us to simultaneously control $\|\Sigma_1^{1/2}(\hat{\beta}_1 - \beta_1^*)\|_2$ and $\|\hat{\beta}_1 - \beta_1^*\|_2$, as is argued below. More precisely, we use it to prove that with high probability $\|\Sigma_1^{1/2}(\hat{\beta}_1 - \beta_1^*)\|_2 \leq \square$ and $\|\hat{\beta}_1 - \beta_1^*\|_2 \leq \triangle$ where $\square$ and $\triangle$ are real-valued parameters to be determined later.

Definition 2.5.4. For all $\beta_1 \in V_{1:k}$, let

$$\|\beta_1\|_m \coloneqq \max\left(\frac{\|\Sigma_1^{1/2}\beta_1\|_2}{\square}, \frac{\|\beta_1\|_2}{\triangle}\right).$$

Moreover, let B_m denote the intersection between $V_{1:k}$ and the unit ball with respect to $\|\cdot\|_m$, that is,

$$B_m = \{\beta_1 \in V_{1:k} : \|\beta_1\|_m \leq 1\}.$$

Therefore, with Definition 2.5.4, we wish to prove that $\hat{\beta}_1 \in \beta_1^* + B_m$, that is, $\hat{\beta}_1$ lives in the unit ball with respect to our 'max norm' centered at β_1^*. In order to do so, we show that if a vector $\beta_1 \in V_{1:k}$ is such that $\beta_1 \notin \beta_1^* + B_m$, then necessarily, $\mathcal{L}(\beta_1) > 0$. Note that by definition of $\mathcal{L}$, $\mathcal{L}(\hat{\beta}_1) \leq 0$ and thus proving that $\mathcal{L}(\beta_1) > 0$ shows that $\hat{\beta}_1 \in \beta_1^* + B_m$. By a homogeneity argument that we do not detail here, we can restrict our attention to vectors β_1 that live on the boundary $\beta^* + \partial B_m$. Hence, by the definition of the norm $\|\cdot\|_m$, the analysis must be split into two cases, namely,

1. $\|\Sigma_1^{1/2}(\beta_1 - \beta_1^*)\|_2 = \square$ and $\|\beta_1 - \beta_1^*\|_2 \leq \triangle$,
2. $\|\Sigma_1^{1/2}(\beta_1 - \beta_1^*)\|_2 \leq \square$ and $\|\beta_1 - \beta_1^*\|_2 = \triangle$.

Therefore, our goal is to show that in both cases, $\mathcal{L}(\beta_1) > 0$. Let us give names to the different quantities in the decomposition of $\mathcal{L}$. Let

$$\begin{aligned}
&\mathcal{Q}(\beta_1) \coloneqq \|(X_2X_2^\top)^{-1/2}X_1(\beta_1 - \beta_1^*)\|_2^2,\\
&\mathcal{M}(\beta_1) \coloneqq 2\langle X_1^\top(X_2X_2^\top)^{-1}(X_2\beta_2^* + \xi) - \beta_1^*, \beta_1 - \beta_1^*\rangle,\\
&\mathcal{C}(\beta_1) \coloneqq \|\beta_1 - \beta_1^*\|_2^2.
\end{aligned}$$

Notice that only $\mathcal{M}$ can take negative values, and thus to show that $\mathcal{L}(\beta_1) > 0$, we show that either $\mathcal{Q}(\beta_1) > |\mathcal{M}(\beta_1)|$ (which holds in case 1.), or $\mathcal{C}(\beta_1) > |\mathcal{M}(\beta_1)|$(which holds in case 2.).
For all $\beta_1 \in \beta_1^* + \partial B_m$, we have the following control over $\mathcal{M}$.

$$|\mathcal{M}(\beta_1)| \leq 2 \sup_{v \in B_m} \left|\langle X_1^\top(X_2X_2^\top)^{-1}(X_2\beta_2^* + \xi) - \beta_1^*, v\rangle\right|.$$

Then, by using some equivalence of norms, and the fact that for any symmetric matrix Γ the dual norm of $\|\Gamma\cdot\|$ is given by $\|\Gamma^{-1}\cdot\|$ (see Example B.1.3), we can get the following bound.

$$|\mathcal{M}(\beta_1)| \leq 2\sqrt{2}\left(\|\tilde{\Sigma}_1^{-1/2}X_1^\top(X_2X_2^\top)^{-1}X_2\beta_2^*\|_2 + \|\tilde{\Sigma}_1^{-1/2}X_1^\top(X_2X_2^\top)^{-1}\xi\|_2 + \|\tilde{\Sigma}_1^{-1/2}\beta_1^*\|_2\right), \tag{2.12}$$

where $\tilde{\Sigma}_1^{1/2} \coloneqq U\tilde{\Lambda}_1^{1/2}U^\top$ and

$$\tilde{\Lambda}_1^{1/2} \coloneqq \operatorname{diag}\left(\max\left(\frac{\sqrt{\lambda_1}}{\square}, \frac{1}{\triangle}\right), \ldots, \max\left(\frac{\sqrt{\lambda_k}}{\square}, \frac{1}{\triangle}, 0, \ldots, 0\right)\right).$$

Now we present two Lemmas which are used to control the first two terms of (2.12) by using classical concentration tools along with Dvoretzky-Milman theorem by placing ourselves on the high probability event Ω_0.

Lemma 2.5.6. *On the event Ω_0, the following holds with probability at least $1 - c_0\exp(-c_1 n)$.*

$$\|\tilde{\Sigma}_1^{-1/2}X_1^\top(X_2X_2^\top)^{-1}X_2\beta_2^*\|_2 \lesssim \frac{n\sigma(\square, \triangle)}{\operatorname{tr}(\Sigma_2)}\|\Sigma_2^{1/2}\beta_2^*\|,$$

where

$$\sigma(\square, \triangle) \coloneqq \begin{cases} \square & \text{if } \triangle\sqrt{\lambda_1} \geq \square, \\ \triangle\sqrt{\lambda_1} & \text{otherwise.} \end{cases}$$

Proof. We provide a proof sketch to show how we can use Theorem 2.5.2 and Theorem A.1.2 to control the quantity of interest.
Bernstein's inequality gives us that with probability at least $1 - c_0\exp(-c_1 n)$,

$$\|X_2\beta_2^*\|_2 \leq \frac{3\sqrt{n}}{2}\|\Sigma_2^{1/2}\beta_2^*\|_2. \tag{2.13}$$

To simplify notations, we show that $\|\mathbb{X}\beta\|_2 \leq \frac{3\sqrt{n}}{2}\|\Sigma^{1/2}\beta\|_2$ for any $\beta \in \mathbb{R}^p$ instead. Let $X_i \sim \mathcal{N}(0, \Sigma)$ denote the $i-$th row of $\mathbb{X}$. Also note that

$$\|\mathbb{X}\beta\|_2 \leq \frac{3\sqrt{n}}{2}\|\Sigma^{1/2}\beta\|_2 \Leftrightarrow \frac{1}{n}\|\mathbb{X}\beta\|_2^2 \leq \frac{9}{4}\|\Sigma^{1/2}\beta\|_2^2.$$

Since $\langle X_i, \beta\rangle \sim \mathcal{N}(0, \beta^\top \Sigma \beta) = \mathcal{N}(0, \|\Sigma^{1/2}\beta\|_2^2)$

$$\frac{1}{n}\|\mathbb{X}\beta\|_2^2 = \frac{1}{n}\sum_{i=1}^n \langle X_i, \beta\rangle^2 = \frac{1}{n}\sum_{i=1}^n \|\Sigma^{1/2}\beta\|_2^2 g_i^2,$$

where $g_i \sim \mathcal{N}(0,1)$ are i.i.d. standard Gaussian random variables. Therefore,

$$\begin{aligned}\frac{1}{n}\|\mathbb{X}\beta\|_2^2 - \mathbb{E}[\|\mathbb{X}\beta\|_2^2] &= \frac{1}{n}\sum_{i=1}^n (\langle X_i, \beta\rangle^2 - \mathbb{E}[\langle X_i, \beta\rangle^2]) = \frac{1}{n}\sum_{i=1}^n (\|\Sigma^{1/2}\beta\|_2^2 g_i^2 - \|\Sigma^{1/2}\beta\|_2^2) \\ &= \frac{1}{n}\|\Sigma^{1/2}\beta\|_2^2 \sum_{i=1}^n (g_i^2 - 1).\end{aligned}$$

Notice that $g_i^2 - 1$ is a subexponential random variable (see Lemma A.1.1). Now we want to show that with high probability,

$$\frac{1}{n}\|\mathbb{X}\beta\|_2^2 - \|\Sigma^{1/2}\beta\|_2^2 = \frac{1}{n}\|\mathbb{X}\beta\|_2^2 - \mathbb{E}[\|\mathbb{X}\beta\|_2^2] \leq \frac{5}{4}\mathbb{E}[\|\mathbb{X}\beta\|_2^2] = \frac{5}{4}\|\Sigma^{1/2}\beta\|_2^2.$$

This is equivalent to showing that with high probability,

$$\frac{1}{n}\sum_{i=1}^n (g_i^2 - 1) \leq \frac{5}{4},$$

which comes immediately from Bernstein's inequality thanks to the fact that $g_i^2 - 1$ are independent subexponential random variables.
Now that we understand how (2.13) is proven, we can use the fact that we are on the event Ω_0 to get the following.

$$\|\tilde{\Sigma}_1^{-1/2} X_1^\top\|_{op} = \|X_1 \tilde{\Sigma}_1^{-1/2}\|_{op} \leq \frac{3\sqrt{n}}{2}\|\Sigma_1^{1/2}\tilde{\Sigma}_1^{-1/2}\|_{op},$$

where the equality comes from the fact that the operator norm of a matrix is equal to the operator norm of its transpose and the inequality comes from the following.
By definition of the operator norm, there exists $\beta_1 \in V_{1:k}$ such that,

$$\|X_1\tilde{\Sigma}_1^{-1/2}\|_{op} = \|X_1\beta_1\|_2 \leq \frac{3\sqrt{n}}{2}\|\Sigma_1^{1/2}\beta_1\|_2 \leq \frac{3\sqrt{n}}{2}\|\Sigma_1^{1/2}\tilde{\Sigma}_1^{-1/2}\|_{op},$$

where the inequality comes from the isomorphic assumption on X_1.
Furthermore, since X_2 satisfies the Dvoretzky-Milman property, by Theorem 2.5.3,

$$\|(X_2 X_2^\top)^{-1}\|_{op} \leq 4(\mathrm{tr}(\Sigma_2))^{-1}.$$

The end of the proof consists in putting all these high probability bounds together to obtain the result.

$$\begin{aligned}\|\tilde{\Sigma}_1^{-1/2} X_1 (X_2 X_2^\top)^{-1} X_2 \beta_2^*\|_2 &\leq \|\tilde{\Sigma}_1^{-1/2} X_1^\top\|_{op}\|(X_2 X_2^\top)^{-1}\|_{op}\|X_2\beta_2^*\|_2 \\ &\lesssim \frac{n}{\mathrm{tr}(\Sigma_2)}\|\Sigma_1^{1/2}\tilde{\Sigma}_1^{-1/2}\|_{op}\|\Sigma_2^{1/2}\beta_2^*\|_2 = \frac{n}{\mathrm{tr}(\Sigma_2)}\|\Sigma_2^{1/2}\beta_2^*\|_2.\end{aligned}$$

□

Lemma 2.5.7. *On the event Ω_0, the following holds with high probability.*

$$\|\tilde{\Sigma}_1^{-1/2} X_1^\top (X_2 X_2^\top)^{-1} \xi\|_2 \lesssim \frac{\sqrt{n} v_\xi}{\mathrm{tr}(\Sigma_2)} \sqrt{|J_1| \square^2 + \triangle^2 \sum_{j \in J_2} \lambda_j},$$

where

$$J_1 \coloneqq \{j \in \{1, \dots k\} : \lambda_j \geq (\square/\triangle)^2\} \quad \textit{and} \quad J_2 \coloneqq \{1, \dots k\} \backslash J_2 = \{j \in \{1, \dots k\} : \lambda_j < (\square/\triangle)^2\}.$$

Proof. We also provide a proof sketch to show how the classical Borell-TIS theorem (Theorem C.1.1) can be used to provide the bounds we want. Let us use Borell-TIS theorem to show that for all $t > 0$, conditionally on $\mathbb{X}$, with probability (with respect only to the noise ξ) at least $1 - \exp(-t/2)$,

$$\|\tilde{\Sigma}_1^{-1/2} X_1^\top (X_2 X_2^\top)^{-1} \xi\|_2 \leq v_\xi \left(\sqrt{\mathrm{tr}(DD^\top)} + \sqrt{t} \|D\|_{op} \right), \tag{2.14}$$

where $D \coloneqq \tilde{\Sigma}_1^{-1/2} X_1^\top (X_2 X_2^\top)^{-1}$. Indeed,

$$\|\tilde{\Sigma}_1^{-1/2} X_1^\top (X_2 X_2^\top)^{-1} \xi\|_2 = \|D\xi\|_2 = \sup_{\lambda \in S^{k-1}} \langle D\xi, \lambda \rangle = \sup_{\lambda \in S^{k-1}} \langle \xi, D^\top \lambda \rangle,$$

where S^{k-1} denotes the $(k-1)-$dimensional sphere, that is, $S^{k-1} \coloneqq \{\lambda \in \mathbb{R}^{k-1} : \|\lambda\|_2 = 1\}$. Then, $G_\lambda \coloneqq \langle \xi, D^\top \lambda \rangle$ for all $\lambda \in S^{k-1}$ form a centered Gaussian process. Note that here, the index set is not countable as stipulated in Theorem C.1.1, however the theorem actually holds for a much broader class of index sets that contains S^{k-1}. Now,

$$\begin{aligned} \mathbb{E} \sup_{\lambda \in S^{k-1}} G_\lambda = \mathbb{E} \sup_{\lambda \in S^{k-1}} \langle D\xi, \lambda \rangle = \mathbb{E} \|D\xi\|_2 &\leq \left(\mathbb{E} \|D\xi\|_2^2 \right)^{1/2} \\ &= \sqrt{\mathbb{E}\, \mathrm{tr}(\xi^\top D^\top D \xi)} = \sqrt{\mathrm{tr}(D^\top D \mathbb{E}[\xi\xi^\top])} = v_\xi \sqrt{\mathrm{tr}(DD^\top)}, \end{aligned}$$

where we successively use Jensen's inequality and the linear and circular properties of the trace. Moreover,

$$\sigma^2 \coloneqq \sup_{\lambda \in S^{k-1}} \mathbb{E} \langle \xi, D^\top \lambda \rangle^2 = \sup_{\lambda \in S^{k-1}} v_\xi^2 \sum_{i=1}^n (D^\top \lambda)_i^2 = \sup_{\lambda \in S^{k-1}} v_\xi^2 \|D^\top \lambda\|_2^2 = v_\xi^2 \|D^T\|_{op}^2 = v_\xi^2 \|D\|_{op}^2.$$

Plugging everything in Theorem C.1.1, for $r = v_\xi \|D\|_{op} \sqrt{t}$ yields that

$$P\left(\sup_{\lambda \in S^{k-1}} G_\lambda - \mathbb{E} \sup_{\lambda \in S^{k-1}} G_\lambda \leq r \right) \geq 1 - \exp(-r^2/(2\sigma^2)) = 1 - \exp(-t/2),$$

which gives

$$P\left(\|D\xi\|_2 \leq v_\xi \left(\sqrt{t} \|D\|_{op} + \sqrt{\mathrm{tr}(DD^\top)} \right) \right) \geq 1 - \exp(-t/2).$$

Now that we have (2.14), the only thing that remains is to bound $\mathrm{tr}(DD^\top)$ and $\|D\|_{op}$. For that, we can use the fact that we are on the event Ω_0, to bound these quantities in a

way that is analogous to what is done in the proof of Lemma 2.5.6. Here we just write down the bounds we get.

$$\mathrm{tr}(DD^\top) \lesssim \frac{\sqrt{n}}{\mathrm{tr}(\Sigma_2)} \sqrt{|J_1|\,\square^2 + \triangle^2 \sum_{j\in J_2} \lambda_j} \quad \text{and} \quad \|D\|_{op} \lesssim \frac{\sqrt{n}\sigma(\square,\triangle)}{\mathrm{tr}(\Sigma_2)}.$$

Then to conclude the proof, there only remains to take the right parameter t. □

Now that we have both Lemma 2.5.6 and Lemma 2.5.7 available to control $|\mathcal{M}(\beta_1)|$, what remains to do is to choose $\triangle$ and $\square$ to ensure that, in case 1., $\mathcal{Q}(\beta_1) > |\mathcal{M}(\beta_1)|$, and in case 2., $\mathcal{C}(\beta_1) > |\mathcal{M}(\beta_1)|$. This choice of parameters $\triangle$ and $\square$ relies again on being on the event Ω_0 to control $\mathcal{Q}(\beta_1)$ and $\mathcal{C}(\beta_1)$ in a way that is similar to what is done to control the various quantities in the proof of Lemma 2.5.6 and we do not go into the details, since they do not contain any deep insight. One thing that is worth mentioning is that to constrain the choice of $\square$ and $\triangle$ further, we first fix $\square$, and then choose $\triangle$ so that $\square/\triangle = \sqrt{\mathrm{tr}(\Sigma_2)/n}$.

Bound on the overfitting component

The first thing is to write the crude upper bound

$$\|\Sigma_2^{1/2}(\hat\beta_2 - \beta_2^*)\|_2 \le \|\Sigma_2^{1/2}\hat\beta_2\|_2 + \|\Sigma_2^{1/2}\beta_2^*\|_2, \tag{2.15}$$

which indicates that we have no ambition to approximate β_2^* by $\hat\beta_2$, due to the fact that $\hat\beta_2$ is the minimum l_2-norm interpolator of $X_2\beta_2^* + \xi$, which we interpret as noise. The quantity $\|\Sigma_2^{1/2}(\hat\beta_2 - \beta_2^*)\|_2$ is the price we pay for overfitting the data.
The strategy is more straightforward than for the prediction component. Nevertheless, it similarly relies on bounding a decomposition of $\|\Sigma_2^{1/2}\hat\beta_2\|_2$ with high probability. This decomposition is the following. Using the closed form solution of $\hat\beta_2$ from Proposition 2.5.1, and denoting $X_2^\top(X_2X_2^\top)^{-1}$ by A,

$$\|\Sigma_2^{1/2}\hat\beta_2\|_2 = \|\Sigma_2^{1/2}A(y - X_1\hat\beta_1)\|_2 = \|\Sigma_2^{1/2}A(X_1\beta_1^* + X_2\beta_2^* + \xi - X_1\hat\beta_1)\|_2 \tag{2.16}$$

$$\le \|\Sigma_2^{1/2}AX_1(\beta_1^* - \hat\beta_1)\|_2 + \|\Sigma_2^{1/2}AX_2\beta_2^*\|_2 + \|\Sigma_2^{1/2}A\xi\|_2. \tag{2.17}$$

Our goal is now to control the three quantities appearing in (2.17) separately. First of all, we place define the high probability event Ω_1 as the event onto which for all $v \in \mathbb{R}^n$,

$$\|\Sigma_2^{1/2}X_2^\top v\|_2 \le 6\left(\sqrt{\mathrm{tr}(\Sigma_2^2)} + \sqrt{n}\|\Sigma_2\|_{op}\right)\|v\|_2.$$

It follows from a result that is similar to Dvoretzky-Milman theorem that Ω_1 holds with probability at least $1 - \exp(-n)$ (see [LS22, Proposition 2]). For the remainder of the proof sketch of the bound on the overfitting component, we place ourselves on the high probability event $\Omega_0 \cap \Omega_1$ where we have access to the same inequalities as in the proof of the bound of the prediction component, on top of the control over $\|\Sigma_2^{1/2}X_2^\top v\|_2$. Without going into the algebraic details, this yields

$$\|\Sigma_2^{1/2}AX_1(\beta_1^* - \hat\beta_1)\|_2 \lesssim \frac{\left(\sqrt{\mathrm{tr}(\Sigma_2^2)} + n\|\Sigma_2\|_{op}\right)}{\mathrm{tr}(\Sigma_2)}\|\Sigma_1^{1/2}(\beta_1^* - \hat\beta_1)\|_2. \tag{2.18}$$

Moreover, as shown in Lemma 2.5.6,

$$\|X_2\beta_2^*\|_2 \leq \frac{3\sqrt{n}}{2}\|\Sigma_2^{1/2}\beta_2^*\|_2,$$

holds with high probability by Bernstein's inequality, which yields that the following holds with high probability.

$$\|\Sigma_2^{1/2}AX_2\beta_2^*\|_2 \lesssim \frac{\left(\sqrt{\operatorname{tr}(\Sigma_2^2)} + \sqrt{n}\|\Sigma_2\|_{op}\right)}{\operatorname{tr}(\Sigma_2)}\|\Sigma_2^{1/2}\beta_2^*\|_2. \tag{2.19}$$

Finally, using Borell-TIS theorem as in Lemma 2.5.7, we have that for all $t > 0$,

$$\|D\xi\|_2 \leq v_\xi\left(\sqrt{\operatorname{tr}(DD^\top)} + \sqrt{t}\|D\|_{op}\right), \tag{2.20}$$

holds with probability at least $1-\exp(-t/2)$, where $D \coloneqq \Sigma_2^{1/2}A$. Then in a similar fashion as for Lemma 2.5.7, $\sqrt{\operatorname{tr}(DD^\top)}$ and $\|D\|_{op}$ can be bounded with high probability, thanks to Bernstein's inequality and Dvoretzky-Milman theorem, as follows.

$$\operatorname{tr}(DD^\top) \lesssim \frac{n\operatorname{tr}(\Sigma_2^2)}{(\operatorname{tr}(\Sigma_2))^2} \quad \text{and} \quad \|D\|_{op} \lesssim \frac{1}{\operatorname{tr}(\Sigma_2)}\left(\sqrt{\operatorname{tr}(\Sigma_2^2)} + \sqrt{n}\|\Sigma_2\|_{op}\right).$$

By assembling the three bounds (2.18), (2.19), and (2.20), we get the bound on $\|\Sigma_2^{1/2}(\hat{\beta}_2 - \beta_2^*)\|_2$.

End of proof of Theorem 2.5.5

The only thing that remains is to put the bounds on the prediction component and the overfitting components together, and to choose a suitable parameter k for the feature space decomposition, as well as to choose the right parameter t in (2.20).
We choose parameter k such that it satisfies

$$n \leq \frac{\operatorname{tr}(\Sigma_2)}{\|\Sigma_2\|_{op}},$$

which makes Theorem 2.5.2 hold by the bound on Dvoretzky dimension of the ellipsoid given in Example 2.5.1. In particular, $\sqrt{n\operatorname{tr}(\Sigma_2^2)} \leq \operatorname{tr}(\Sigma_2)$. Furthermore, by choosing $t = n\operatorname{tr}(\Sigma_1)/\operatorname{tr}(\Sigma_2)$, we get the result of Theorem 2.5.5, namely

$$\left\|\Sigma^{1/2}\left(\hat{\beta} - \beta^*\right)\right\|_2 \lesssim \square + v_\xi\frac{\sqrt{n\operatorname{tr}\left(\Sigma_{k+1:p}^2\right)}}{\operatorname{tr}\left(\Sigma_{k+1:p}\right)} + \left\|\Sigma_{k+1:p}^{1/2}\beta_{k+1:p}^*\right\|_2. \tag{2.21}$$

Remark 2.5.9. By choosing t differently in the pathological case mentioned in Remark 2.5.7, we recover the form $\square$ must take in that case.

Chapter 3

Towards a general treatment of Benign Overfitting

This chapter is devoted to the presentation of tentative results towards a generalization of the benign overfitting analysis in the context of minimum l_q-norm interpolators, for $q \in [1,2) \cup (2,+\infty]$. It mainly consists in an effort of generalization of the key results of [LS22] for the l_q-norm. The main difficulties of this approach are the absence of a closed-form solution for the minimum l_q-norm interpolator, as well as the identification of the best entanglement of (Σ, β) in an optimal basis. To circumvent the first difficulty, we make use of Dvoretzky-Milman theorem to find an approximate expression of the minimum l_q-norm interpolator. Concerning the second difficulty, we do not address it but we try to consider an unspecified basis as much as possible. We also specifically look at the canonical basis and the basis formed by the eigenvalues of the covariance matrix Σ.

In the case of the l_q-norm, for $q \in [1,\infty], q \neq 2$, the feature space decomposition cannot be done so easily because we do not have access to the Pythagorean theorem. What we can do however, is the following decomposition.

$$\|\beta\|_q^q = \sum_{i=1}^{p} |\beta_i|^q = \sum_{i=1}^{k} |\beta_i|^q + \sum_{i=k+1}^{p} |\beta_i|^q = \|P_{1:k}\beta\|_q^q + \|P_{k+1:p}\beta\|_q^q, \tag{3.1}$$

where $P_{1:k}$ is defined to be the orthogonal projection onto the first k vectors of the canonical basis, and $P_{k+1:p}$ is the orthogonal projection onto the last $p-k$ vectors of the canonical basis. However, one major caveat of this approach is that our split no longer occurs along the eigenvectors of the covariance matrix. Another idea would be to assume β is written in the eigenbasis already. In that case, we would obtain the same decomposition as in equation (3.1) but the split would be performed along the eigenvectors of the covariance matrix. In what follows, as long as we can afford to, we assume that $P_{1:k}$ and $P_{k+1:p}$ are orthogonal projection matrices that satisfy the relation (3.1), without specifying their form.

Notations

We denote $P_{1:k}\beta$ by $\beta_{1:k}$ and $P_{k+1:p}\beta$ by $\beta_{k+1:p}$, and in a similar way to what is done in the previous chapter, we write

$$\mathbb{X}\beta = \mathbb{X}(P_{1:k} + P_{k+1:p})\beta = \mathbb{X}P_{1:k}\beta + \mathbb{X}P_{k+1:p}\beta =: X_1\beta + X_2\beta.$$

3.1 Decomposition of the estimator for minimum l_q- interpolation

In this section, we attempt to achieve a decomposition of the estimator $\hat{\beta}$ in prediction and overfitting components, analogous to what is performed in Proposition 2.5.1, but for the minimum l_q-norm interpolator. The minimum l_q-norm interpolator is a solution to the minimization problem

$$\min_{\beta\in\mathbb{R}^p} \|\beta\|_q^q. \tag{3.2}$$

$$\text{s.t. } \mathbb{X}\beta = y \tag{3.3}$$

Following the split done in (3.1), we start by doing the following.

$$\begin{aligned}
\arg\min_{\beta\in\mathbb{R}^p}\{\|\beta\|_q^q : \mathbb{X}\beta = y\} &= \arg\min_{\beta\in\mathbb{R}^p}\{\|P_{1:k}\beta\|_q^q + \|P_{k+1:p}\beta\|_q^q : \mathbb{X}\beta = y\} \\
&= \arg\min_{\beta\in\mathbb{R}^p}\{\|\beta_{1:k}\|_q^q + \|\beta_{k+1:p}\|_q^q : X_1\beta + X_2\beta = y\} \\
&= \arg\min_{(\beta_1,\beta_2)\in\mathbb{R}^p\times\mathbb{R}^p}\{\|\beta_1\|_q^q + \|\beta_2\|_q^q : X_1\beta_1 + X_2\beta_2 = y\},
\end{aligned}$$

where in the last equality, one inequality is trivially true and for the other one, by definition of X_1 and X_2, the minimizers β_1 and β_2 must take values in range$(P_{1:k})$, respectively range$(P_{k+1:p})$ (which are orthogonal subspaces of $\mathbb{R}^p$), or else they would not minimize the l_q-norm. This argument is similar to what we do in the proof of Proposition 2.5.1. This allows us to decouple the minimization problem as in Proposition 2.5.1 and to optimize over β_2 first. Given $\beta_1 \in \mathbb{R}^p$ fixed,

$$\hat{\beta}_{k+1:p} \in \arg\min_{\beta_2\in\mathbb{R}^p}\{\|\beta_2\|_q^q : X_2\beta_2 = y - X_1\beta_1\}.$$

Although in the l_2-case, this problem is easy to solve since it admits a closed form solution, it is not the case here. To simplify the notations, let $\tilde{y} \coloneqq y - X_1\beta_1$. We use the dual formulation of the minimum norm interpolating problem, that we cover in detail in B.2 and obtain the following.

$$\min_{\beta_2\in\mathbb{R}^p}\{\|\beta_2\|_q : X_2\beta_2 = \tilde{y}\} = \max_{\gamma\in\mathbb{R}^n}\{\langle\gamma,\tilde{y}\rangle : \|X_2^\top\gamma\|_{q'} \leq 1\},$$

where q' is the Hölder conjugate of q, that is $\frac{1}{q} + \frac{1}{q'} = 1$, and where we use the fact that the dual l_q-norm is equal to the $l_{q'}-$norm (see Example B.1.2).

In essence, our idea is to use Theorem 2.5.2 to transform the condition $\|X_2^\top \gamma\|_{q'} \leq 1$ into something more amenable.

First attempt

The first thing we may try is to do something similar to the proof of Corollary 2.5.3. Consider the norm $\|\cdot\| \coloneqq \|\Sigma_2^{1/2}\cdot\|_{q'}$, then by definition of X_2,

$$\|X_2^\top \gamma\|_{q'} = \|\mathbb{G}^\top \gamma\|,$$

where $\mathbb{G}$ is a Gaussian matrix with i.i.d standard Gaussian entries. Hence, by Theorem 2.5.2, if $n \leq \kappa_{DM} d_*(\Sigma_2^{-1/2} B_{q'}^p)$, we have that with high probability,

$$\frac{1}{\sqrt{2}} w\left(\Sigma_2^{1/2} B_q^p\right) \|\gamma\|_2 \leq \|X_2^\top \gamma\|_{q'} \leq \sqrt{\frac{3}{2}} w\left(\Sigma_2^{1/2} B_q^p\right) \|\gamma\|_2.$$

Therefore,

$$\begin{aligned}\max_{\gamma \in \mathbb{R}^n} \left\{ \langle \gamma, \tilde{y} \rangle : \|\gamma\|_2 \leq \sqrt{\frac{2}{3}} \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)} \right\} &\leq \max_{\gamma \in \mathbb{R}^n} \left\{ \langle \gamma, \tilde{y} \rangle : \|X_2^\top \gamma\|_{q'} \leq 1 \right\} \\ &\leq \max_{\gamma \in \mathbb{R}^n} \left\{ \langle \gamma, \tilde{y} \rangle : \|\gamma\|_2 \leq \sqrt{2} \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)} \right\},\end{aligned}$$

which yields the following,

$$\sqrt{\frac{2}{3}} \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)} \|\tilde{y}\|_2 \leq \max_{\gamma \in \mathbb{R}^n} \left\{ \langle \gamma, \tilde{y} \rangle : \|X_2^\top \gamma\|_{q'} \leq 1 \right\} \leq \sqrt{2} \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)} \|\tilde{y}\|_2 \tag{3.4}$$

by noticing that the maximum on the left-hand side is achieved by the vector

$$\tilde{\gamma} = \sqrt{\frac{2}{3}} \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)} \frac{\tilde{y}}{\|\tilde{y}\|_2},$$

and analogously, the maximum on the right-hand side is achieved by

$$\tilde{\gamma} = \sqrt{2} \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)} \frac{\tilde{y}}{\|\tilde{y}\|_2}.$$

Therefore, (3.4) implies that with high probability,

$$\min_{\beta_2 \in \mathbb{R}^p} \{\|\beta_2\|_q : X_2 \beta_2 = \tilde{y}\} \asymp \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)} \|\tilde{y}\|_2 = \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)} \|y - X_1 \beta_1\|_2, \tag{3.5}$$

because the lower and upper bounds in Dvoretzky-Milman theorem can actually be made arbitrarily tight as p goes to ∞. Thus, we can express the norm of the estimator $\hat{\beta}_{k+1:p}$ as a function of β_1 as follows. With high probability,

$$\|\hat{\beta}_{k+1:p}\|_q \asymp \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)} \|y - X_1\beta_1\|_2, \tag{3.6}$$

Moreover, this leads the minimization problem (3.2) to take the form

$$\min_{\beta\in\mathbb{R}^p} \left\{\|\beta\|_q^q : \mathbb{X}\beta = y\right\} \asymp \min_{\beta_1\in\mathbb{R}^p} \left\{ \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)^q} \|y - X_1\beta_1\|_2^q + \|\beta_1\|_q^q \right\} \tag{3.7}$$

$$= \frac{1}{w\left(\Sigma_2^{1/2} B_q^p\right)^q} \min_{\beta_1\in\mathbb{R}^p} \left\{ \|y - X_1\beta_1\|_2^q + w\left(\Sigma_2^{1/2} B_q^p\right)^q \|\beta_1\|_q^q \right\}. \tag{3.8}$$

The minimization problem on the right-hand side of the equation above is similar to a regularized least squares minimization problem, apart from the fact that the least squares term is raised to the power q. Unfortunately, it is not possible to get rid of this power q since it is vital for the decomposition of the l_q-norm of β into two components. This prevents the use of results such as the one from Chinot in [Chi19] because the loss does not satisfy the typical regularity conditions we would expect, specifically in that case, because it is not Lipschitz. However, the function class we restrict to is not a general convex set as in [Chi19] but consists only of linear functions, which may facilitate the matter.
The approach taken in this first attempt seems interesting at first as it comes as a more or less natural generalization of what is done in [LS22]. Indeed, in the paper of Lecué et al., $q = q' = 2$, and that allows us to make the effective rank appear with the bounds on the Dvoretzky dimension (see Example 2.5.1 and remark below). Furthermore, it generalizes the observation in Remark 2.5.1 that $\hat{\beta}_{1:k}$ is approximately equal to

$$\arg\min_{\beta_1\in\mathbb{R}^p} \left(\|y - X_{1:k}\beta_1\|_2^2 + \mathrm{tr}(\Sigma_{k+1:p})\|\beta_1\|_2^2 \right),$$

as in that case, $\sqrt{\mathrm{tr}(\Sigma_{k+1:p})/2} \leq w(\Sigma_{k+1:p}^{1/2} B_2^p) \leq \sqrt{\mathrm{tr}(\Sigma_{k+1:p})}$.

However, this approach actually comes with caveats, and one of them is considerable. First of all, the Dvoretzky-Milman condition $n \leq \kappa_{DM} d_*(\Sigma_2^{-1/2} B_{q'}^p)$ is difficult to understand, because

$$d_*(\Sigma_2^{-1/2} B_{q'}^p) = \frac{w\left(\Sigma_2^{1/2} B_q^p\right)^2}{\mathrm{diam}\left(B_q^p \cap l_2^p\right)^2},$$

and neither the Gaussian width, nor the diameter of the ellipsoid $\Sigma_2^{1/2} B_q^p$, are easy to compute, as they involve the l_q-norm. Indeed, the Gaussian width is given by

$$w\left(\Sigma_2^{1/2} B_q^p\right) = \mathbb{E} \sup_{v\in B_q^p} \langle v, \Sigma_2^{1/2} g\rangle = \mathbb{E}\|\Sigma_2^{1/2} g\|_{q'},$$

where $g \sim \mathcal{N}(0, I)$. Moreover, the diameter is given by

$$\text{diam}\left(B_q^p, l_2^p\right) = \sup_{\|\Sigma_2^{-1/2} v\|_q \leq 1} \|v\|_2.$$

Secondly, since computing the Gaussian width of $\Sigma_2^{1/2} B_q^p$ is hard, it means that we do not have access to the regularization parameter $w(\Sigma_2^{1/2} B_q^p)$.
The major drawback of this approach however, is the simple fact that the Gaussian width of $\Sigma_2^{1/2} B_q^p$ does not give in general a useful Dvoretzky regime. Indeed, for $q = 1$ for example, in the case of l_1-regularization, the Dvoretzky condition we would have to satisfy is the too restrictive

$$n \lesssim log(p).$$

Second attempt

Another idea we explore, in order to avoid the impractical Dvoretzky condition from the first attempt, is to use some localization argument to obtain another asymptotically equivalent formulation of the minimization problem. This time we want to show with the help of Dvoretzky-Milman theorem that with high probability,

$$\max_{\gamma \in \mathbb{R}^n} \{\langle \gamma, \tilde{y} \rangle : \|X_2^\top \gamma\|_{q'} \leq 1\} \asymp \frac{1}{w(K_\mu)} \|\tilde{y}\|_2, \tag{3.9}$$

where $K_\mu \coloneqq B_q^p \cap \mu B_2^p$ is the intersection between the unitary l_q-ball and an l_2-ball for some μ that we define later. The use of the term 'localization' must be understood in the sense that we try to localize a certain vector with high probability in the set K_μ instead of the whole l_q-ball. As before, the right-hand side in (3.9) gives us an approximation of $\|\hat{\beta}_{k+1:p}\|_q$ in terms of β_1 which provides an easier minimization problem akin to a regularized least squares minimization problem in place of the minimum l_q-norm interpolation problem.
Let us explain how we arrive at the expression (3.9). We must take γ as large as possible, with the added constraint that $\|X_2^\top \gamma\|_{q'} \leq 1$. Therefore, we take some γ that is parallel to $\tilde{y}$ and we want to argue that the maximizer is of the form

$$\tilde{\gamma} \asymp \frac{1}{w(K_\mu)} \frac{\tilde{y}}{\|\tilde{y}\|_2}.$$

In order to do that, we must analyze the condition $\|X_2^\top \gamma\|_{q'} \leq 1$. This time, we go one level further than in the first attempt. Instead of using Dvoretzky-Milman theorem immediately to make the constraint nicer, we are using it later. By the dual formulation of the norm,

$$\|X_2^\top \gamma\|_{q'} = \max_{b \in B_q^p} \langle X_2^\top \gamma, b \rangle.$$

We would like to be able to localize the b that maximizes the quantity above in the following way. We would like to say that the typical b actually lives in a possibly much smaller space than B_q^p. Therefore, we observe that

$$\left\{\gamma : \|X_2^\top \gamma\|_{q'} \leq 1\right\} \subset \left\{\gamma : \max_{b \in B_q^p \cap \mu B_2^p} \langle X_2^\top \gamma, b \rangle \leq 1\right\},$$

for any $\mu > 0$. Hence, for any $\mu > 0$,

$$\max_{\gamma \in \mathbb{R}^n} \left\{ \langle \gamma, \tilde{y} \rangle : \|X_2^\top \gamma\|_{q'} \leq 1 \right\} \leq \max_{\gamma \in \mathbb{R}^n} \left\{ \langle \gamma, \tilde{y} \rangle : \max_{b \in B_q^p \cap \mu B_2^p} \langle X_2^\top \gamma, b \rangle \leq 1 \right\}.$$

Next, we use Dvoretzky-Milman theorem to control the quantity $\max_{b \in K_\mu} \langle X_2^\top \gamma, b \rangle$.

$$\max_{b \in K_\mu} \langle X_2^\top \gamma, b \rangle = \max_{b \in K_\mu} \langle \gamma, X_2 b \rangle = \max_{\eta \in X_2 K_\mu} \langle \gamma, \eta \rangle.$$

By Dvoretzky-Milman theorem (Theorem C.2.2), we have that with high probability,

$$X_2 K_\mu = \mathbb{G}^{(n \times p)} \Sigma_2^{1/2} K_\mu \approx w(\Sigma_2^{1/2} K_\mu) B_2^n, \tag{3.10}$$

and hence, when p tends to ∞, with high probability,

$$\max_{\eta \in X_2 K_\mu} \langle \gamma, \eta \rangle = \max_{\eta \in w(\Sigma_2^{1/2} K_\mu) B_2^n} \langle \gamma, \eta \rangle = w(\Sigma_2^{1/2} K_\mu) \|\gamma\|_2.$$

Therefore, still with high probability,

$$\max_{\gamma \in \mathbb{R}^n} \left\{ \langle \gamma, \tilde{y} \rangle : \|X_2^\top \gamma\|_{q'} \leq 1 \right\} \leq \max_{\gamma \in \mathbb{R}^n} \left\{ \langle \gamma, \tilde{y} \rangle : w(\Sigma_2^{1/2} K_\mu) \|\gamma\|_2 \leq 1 \right\} = \frac{1}{w(\Sigma_2^{1/2} K_\mu)} \|\tilde{y}\|_2,$$

which yields the high probability upper bound

$$\|\hat{\beta}_{k+1:p}\|_q \leq \frac{1}{w(\Sigma_2^{1/2} K_\mu)} \|y - X_1 \beta_1\|_2.$$

In order for (3.10) to be correct, the Dvoretzky condition must hold, that is, we must have

$$\sqrt{n} \lesssim \frac{w(\Sigma_2^{1/2} K_\mu)}{\operatorname{diam}(\Sigma_2^{1/2} K_\mu)}.$$

Observe that

$$\operatorname{diam}(\Sigma_2^{1/2} K_\mu) = \max_{v \in \Sigma_2^{1/2} K_\mu} \|v\|_2 = \max_{b \in K_\mu} \|\Sigma_2^{1/2} b\|_2 \leq \|\Sigma_2^{1/2}\|_{op} \max_{b \in K_\mu} \|b\|_2 = \sqrt{\lambda_{k+1}} \mu.$$

Therefore, to ensure Dvoretzky condition is satisfied, we can enforce the stronger condition

$$\sqrt{n} \lesssim \frac{w(\Sigma_2^{1/2} K_\mu)}{\|\Sigma_2^{1/2}\|_{op} \mu} = \frac{w(\Sigma_2^{1/2} K_\mu)}{\sqrt{\lambda_{k+1}} \mu},$$

which can be equivalently stated as

$$\mu \lesssim \frac{w(\Sigma_2^{1/2} K_\mu)}{\sqrt{n} \sqrt{\lambda_{k+1}}}.$$

This provides a sophisticated constraint on the choice of μ, since μ appears on both sides of the expression. The idea would be to pick μ that saturates Dvoretzky condition, that is, to take μ such that

$$\mu \approx \frac{w(\Sigma_2^{1/2} K_\mu)}{\sqrt{n} \sqrt{\lambda_{k+1}}} \approx \frac{w(K_\mu)}{\sqrt{n}}. \tag{3.11}$$

However, there remains to find a matching lower bound, and this inevitably puts further restrictions on the choice of localization parameter μ.

Remark 3.1.1. As indicated in [Ver18, Exercise 7.5.4], it can also be shown that

$$w\left(\Sigma_2^{1/2}K_\mu\right) \leq \|\Sigma_2^{1/2}\|_{op} w\left(K_\mu\right) = \sqrt{\lambda_{k+1}} w\left(K_\mu\right).$$

We need the matching lower bound

$$\|\hat{\beta}_{k+1:p}\|_q \geq \frac{1}{w(\Sigma_2^{1/2}K_\mu)}\|y - X_1\beta_1\|_2,$$

to be able to conclude that, with high probability,

$$\|\hat{\beta}_{k+1:p}\|_q = \frac{1}{w(\Sigma_2^{1/2}K_\mu)}\|y - X_1\beta_1\|_2.$$

In order to find such a lower bound, we must tackle an issue that we initially avoided when deriving the upper bound. Indeed, for a lower bound of this type to hold, we need to relate the maximum over the l_q-ball with the maximum under the localization constraint in a precise manner. That is, we must actually choose μ such that

$$\max_{b\in B_q^p}\langle X_2^\top\gamma, b\rangle = \max_{b\in B_q^p\cap\mu B_2^p}\langle X_2^\top\gamma, b\rangle. \tag{3.12}$$

For this equality to hold in general, we must pick $\mu \geq \|b^*(\gamma)\|_2$, where

$$b^*(\gamma) \coloneqq \arg\max_{b\in B_q^p}\langle X_2^\top\gamma, b\rangle.$$

However, this creates a dependence of μ on γ. But then, if μ depends on γ, it means that the localization we perform depends on γ as well. To fix this issue, we define for all suitable γ,

$$\mu(\gamma) \coloneqq \|b^*(\gamma)\|_2, \quad \text{and} \quad \tilde{\mu} \coloneqq \max_{\gamma\in\mathbb{R}^n}\{\mu(\gamma) : \|X_2^\top\gamma\|_{q'} \leq 1\}. \tag{3.13}$$

Note that we can express $b^*(\gamma)$ in terms of X_2 and γ, by arguing as in Example B.1.2. For all $1 \leq i \leq p$, define z_i as follows.

$$z_i = \text{sgn}((X_2^\top\gamma)_i)\left|(X_2^\top\gamma)_i\right|^{q-1},$$

where sgn denotes the sign function. Then

$$b^* = \frac{z}{\|z\|_{q'}}.$$

Unfortunately for us, the quantity $\hat{\mu}$ still depends on the condition $\|X_2^\top\gamma\|_{q'} \leq 1$, so it seems like we have just displaced the problem.
Another thing we could try is to not be aiming for the equality (3.12) to hold in general, but only with high probability. By Borell-TIS, we can obtain that the maximum of the Gaussian process

$$\langle X_2^\top\gamma, b\rangle$$

concentrates around its mean, that is, with high probability,

$$\|X_2^\top \gamma\|_{q'} = \max_{b \in B_q^p} \langle X_2^\top \gamma, b\rangle \asymp \mathbb{E} \max_{b \in B_q^p} \langle X_2^\top \gamma, b\rangle.$$

Then, we can focus our attention on making the following equality hold with high probability instead of (3.12).

$$\mathbb{E} \max_{b \in B_q^p} \langle X_2^\top \gamma, b\rangle = \mathbb{E} \max_{b \in B_q^p \cap \mu B_2^p} \langle X_2^\top \gamma, b\rangle,$$

for a suitable μ. This equality holds if

$$\mathbb{E}\|b^*(\gamma)\|_2 \leq \mu,$$

for all γ. Therefore we could adapt the definition of $\tilde{\mu}$ to become

$$\tilde{\mu} = \max_{\gamma \in \mathbb{R}^n} \{\mathbb{E}\|b^*(\gamma)\|_2 : \mathbb{E}\|X_2^\top \gamma\|_{q'} \leq 1\}. \tag{3.14}$$

Or, going even further, exploiting the form of b^* we could also assume that $\mu(\gamma)$ must only fulfill

$$\mathbb{E}[\|z\|_{q'}]\mu(\gamma) = \mathbb{E}\|z\|_2,$$

that is,

$$\mu(\gamma) = \frac{\mathbb{E}\|z\|_2}{\mathbb{E}\|z\|_{q'}}.$$

and hence define

$$\tilde{\mu} = \max_{\gamma \in \mathbb{R}^n} \{\mu(\gamma) : \mathbb{E}\|X_2^\top \gamma\|_{q'} \leq 1\}. \tag{3.15}$$

We are unable to find a way to guarantee that a $\tilde{\mu}$ defined as in (3.13), (3.14), or (3.15) and that simultaneously fulfills Dvoretzky condition (3.11) exists at this moment. But, assuming that such a $\tilde{\mu}$ exists, we would reach the goal of this section which is to get that the following holds with high probability.

$$\|\hat{\beta}_{k+1:p}\|_q = \frac{1}{w(\Sigma_2^{1/2} K_\mu)} \|y - X_1 \beta_1\|_2.$$

3.2 Control of the excess risk

In this section we assume that the l_q-minimum norm interpolator asymptotically behaves like a ridge estimator as follows, for some suitable μ (assumed to exist).

$$\min_{\beta \in \mathbb{R}^p} \left\{\|\beta\|_q^q : \mathbb{X}\beta = y\right\} \asymp \min_{\beta_1 \in \mathbb{R}^p} \left\{\frac{1}{w\left(K_\mu\right)^q} \|y - X_1\beta_1\|_2^q + \|\beta_1\|_q^q\right\} \tag{3.16}$$

$$= \frac{1}{w\left(K_\mu\right)^q} \min_{\beta_1 \in \mathbb{R}^p} \left\{\|y - X_1\beta_1\|_2^q + w\left(K_\mu\right)^q \|\beta_1\|_q^q\right\}. \tag{3.17}$$

We direct our focus on trying to show that the minimum l_q-norm interpolator is close enough to the true parameter β^*, and that its excess risk is small enough. In order

to control the excess risk, we try to proceed as in the proof of Theorem 2.5.5. We would like to confine the estimator $\hat{\beta}_1$ in a ball centered at the true parameter β_1^* whose associated norm provides control on the excess risk. In order to do that, we attempt to find a decomposition of a certain function $\mathcal{L}$ of β_1 similar to the decomposition (2.11). However, since we are working with the l_q-norm as well as expressions raised to the power q, the decomposition we are looking for necessarily looks nothing like the one in [LS22]. Let's try to do this decomposition with the function

$$\mathcal{L}(\beta_1) \coloneqq \|y - X_1\beta_1\|_2^q + w(K_\mu)^q\|\beta_1\|_q^q - \left(\|y - X_1\beta_1^*\|_2^q + w(K_\mu)^q\|\beta_1^*\|_q^q\right), \tag{3.18}$$

which would yield $\mathcal{L}(\hat{\beta}_1) \leq 0$ by definition. We immediately notice that in order for a term of the form $\|\beta_1 - \beta_1^*\|_q^q$ to appear in our decomposition, we would need to use very loose bounds such as

$$\|\beta_1 - \beta_1^*\|_q^q \geq (\|\beta_1\|_q - \|\beta_1^*\|_q)^q \geq \|\beta_1\|_q^q - \|\beta_1^*\|_q^q.$$

Not only is this bound loose, but more importantly, in order to imitate the proof technique of [LS22], we need to find a decomposition that is an equality. It may be possible to find ways to analyze the regularized least squares problem through the means of the methods presented in [Chi19] and [Men14]. The results of the paper of Mendelson [Men14] may be able to help us define a decomposition thanks to its minimal assumptions on the loss. Indeed, its results do not rely on the traditional contraction arguments that require a Lipschitz condition on the loss; it replaces it by asking for a small-ball property. However, the paper of Mendelson is only concerned about the setting with no regularization. The paper of Chinot deals with regularized empirical risk minimization for a wide range of regularization functions but assumes a Lipschitz condition on the loss, which is not satisfied in our setting. Regardless, the fact that the terms in (3.18) are raised to the power q simply does not allow us to express it in a way that enables the use of the classical statistical framework of regularized empirical risk minimization (RERM). Indeed,

$$\|y - X_1\beta_1\|_2^q = \left(\sum_{i=1}^n (y_i - \langle x_i, \beta_1\rangle)^2\right)^{q/2}$$

where x_i denotes the $i-$th row of X_1. However, in order for the problem to at least fit the framework of RERM, we would need to get rid of the power q to have the classical term

$$\|y - X_1\beta_1\|_2,$$

in which case we would be in the setting of regularized least squares. We unfortunately do not find a way to circumvent this major issue in our essay but there might be other functions $\mathcal{L}$ that would clear the problem.

Chapter 4

Conclusion

In this essay, we reviewed the relatively new concept of benign overfitting and uncovered the mechanisms that lie at the heart of this surprising phenomenon.
We first introduced the framework in which we examined benign overfitting and provided motivation for the study of this particular framework.
Then, we went through the main results of the papers in a chronological manner and presented how they get proven by highlighting the important steps in the proofs. We focused on the upper bounds on the excess risk of the minimum l_2-norm interpolator for the overparametrized linear regression problem, but the papers we studied also provide lower bounds demonstrating the sharpness of the control they obtain, and the paper [TB20] extends the results to the ridge regression setting. The papers [BLLT20] and [TB20] rely on a bias-variance decomposition of the excess risk, whereas the paper [LS22] takes an arguably more general approach, that is rooted in a geometric understanding of Gaussian random matrices. At the heart of the three papers resides the phenomenon of concentration of measures (see for example [Led01], [Ver18]) that is exploited to yield control over random quantities with high probability.
Finally, we attempted to build upon the approach of [LS22] to generalize the analysis of benign overfitting to the case of minimum l_q-norm interpolators. Our core idea to attack this problem was to use the fact that under Dvoretzky condition, the random projection of the unit l_q-ball by a Gaussian matrix looks like the unit l_2-ball thanks to Dvoretzky-Milman theorem. Hence, even for the l_q-norm, we expected the minimum l_q-norm interpolator to behave like an estimator for regularized least squares. There were two major obstacles that we would need to overcome to bring this idea to life. The first one was that, in order to leverage this property without imposing a too restrictive Dvoretzky condition on n, we had to use a localization trick. As it turned out, finding the right localization parameter μ is challenging. We found conditions that such a μ needed to fulfill but we did not show that this μ existed. The second challenge was that, due to the impractical decomposition of the minimum l_q-norm interpolator into prediction and overfitting components, the resulting regularized minimization problem did not have a satisfactory form, as it resembled the problem of regularized least squares but with the least squares term raised to the power q. As a result, it proved difficult to find some sort of 'quadratic+multiplier+regularization' decomposition, crucial for the idea behind the proof of [LS22] to be adapted. If we were to investigate the generalization of benign overfitting in the context of minimum

l_q-norm interpolators further, these two issues are the ones we would need to tackle first.
From the analysis made in [BLLT20], [TB20], and [LS22], we were able to convey why the notion of benign overfitting does not contradict the traditional viewpoint on overfitting leading to poor generalization; rather, it enriches the picture by providing explanations for linear regression models, and heuristics more generally, for the existence of benign overfitting in particular regimes. Furthermore, in the work of Lecué et al., it is demonstrated that when the right conditions are aligned, benign overfitting takes place with high probability and therefore we should expect to witness it consistently.

Appendix A

Random variables and random vectors

This appendix is considerably influenced by Chapter 2 and 3 from the remarkable book [Ver18].

A.1 Concentration of random variables

A.1.1 Subgaussian random variables

As their name suggest, subgaussian random variables are dominated my Gaussian random variables in some sense. Roughly speaking, the tail of a subgaussian random variable is smaller than the tail of a Gaussian random variable. Therefore, subgaussian random variables are at least as concentrated around their mean as Gaussian random variables are. Let us first define the notion of subgaussian norm before we formally introduce subgaussian random variables.

Definition A.1.1. The subgaussian norm $\|\cdot\|_{\psi_2}$ on the space of subgaussian random variables is defined as

$$\|X\|_{\psi_2} \coloneqq \inf\left\{t \geq 0 : \mathbb{E}\exp(X^2/t^2) \leq 2\right\}.$$

Definition A.1.2. A random variable X is subgaussian if it satisfies one of the following equivalent properties:

i) $P(|X| \geq t) \leq 2\exp(-ct^2/\|X\|_{\psi_2})$ for all $t \geq 0$,

ii) $\|X\|_{L^p} \leq C\|X\|_{\psi_2}\sqrt{p}$ for all $p \geq 1$,

iii) $\mathbb{E}\exp(X^2/\|X\|_{\psi_2}^2) \leq 2$,

iv) if $\mathbb{E}X = 0$ then $\mathbb{E}\exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_2}^2)$, for all $\lambda \in \mathbb{R}$.

In the above, C, c are some absolute constants. Furthermore, up to absolute constant factors, $\|X\|_{\psi_2}$ is the smallest possible number for which these inequalities are valid.

Example A.1.1. *A Gaussian random variable $X \sim \mathcal{N}(0, \sigma^2)$, for some $\sigma > 0$, is a subgaussian random variable with $\|X\|_{\psi_2} \leq C\sigma$.*

Example A.1.2. *Given a random vector $X = X_1, \dots, X_N$ with $\sigma-$subgaussian independent entries, and a vector $v \in \mathbb{R}^N$, $v^\top X$ is a $c_1\sigma-$subgaussian random variable, for some constant $c_1 > 0$.*

Proof.

$$\begin{aligned}\mathbb{E}\exp\left(\lambda v^\top X\right) = \mathbb{E}\exp\left(\lambda \sum_j^N v_j X_j\right) &= \prod_j^N \mathbb{E}\exp(\lambda v_j X_j) \leq \prod_j^N \mathbb{E}\exp(\lambda^2 v_j^2 c_1 \sigma^2) \\ &= \mathbb{E}\exp\left(\lambda^2 c_1 \sigma^2 \sum_j^N v_j^2\right) = \mathbb{E}\exp(\lambda^2 c_1 \sigma^2 \|v\|_2^2).\end{aligned}$$

□

Remark A.1.1. The smaller the subgaussian norm, the faster the tail of the distribution decays and the smaller the variance of the distribution becomes.

A.1.2 Subexponential random variables

We now introduce a notion capturing the behaviour of distributions that have heavier tail than the Gaussian distribution, but that still have thinner tails than the exponential distribution. We first introduce the subexponential norm, which is analogous to the subgaussian norm.

Definition A.1.3. The subexponential norm $\|\cdot\|_{\psi_1}$ on the space of subexponential random variables is defined as

$$\|X\|_{\psi_1} := \inf\{t > 0 : \mathbb{E}\exp(|X|/t) \leq 2\}. \tag{A.1}$$

Definition A.1.4. A random variable X is subexponential if it satisfies one of the following equivalent properties:

i) $P(|X| \geq t) \leq 2\exp(-ct/\|X\|_{\psi_1})$ for all $t \geq 0$,

ii) $\|X\|_{L^p} \leq C\|X\|_{\psi_1} p$ for all $p \geq 1$,

iii) $\mathbb{E}\exp(|X|/\|X\|_{\psi_1}) \leq 2$,

iv) if $\mathbb{E}X = 0$ then $\mathbb{E}\exp(\lambda X) \leq \exp(C\lambda^2\|X\|_{\psi_1}^2)$, for all λ such that $|\lambda| \leq \frac{1}{\|X\|_{\psi_1}}$.

In the above, C, c are some absolute constants. Furthermore, up to absolute constant factors, $\|X\|_{\psi_1}$ is the smallest possible number for which these inequalities are valid.

Remark A.1.2. A recurrent example of subexponential random variables that will be relevant for us is given by the sum of squares of subgaussian random variables. It often appears when we consider the squared norm of a random variable with subgaussian entries. For example the χ^2-distribution is a subexponential distribution arising as the squared norm of a Gaussian random vector.

Lemma A.1.1. *A random variable X is subgaussian if and only if X^2 is subexponential. Moreover,*

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$$

Proof. This can immediately be seen by comparing the definitions of subgaussian norm and subexponential norm. □

Remark A.1.3. It can also easily be shown that if a random variable X is subexponential, then so is $X - \mathbb{E}X$ and

$$\|X - \mathbb{E}X\|_{\psi_1} \leq C\|X\|_{\psi_1},$$

where C is an absolute constant.

A.1.3 Bernstein's concentration inequality

Bernstein's inequality illustrates the concentration phenomena in high dimension in the case of subexponential random variables. Its proof relies on bounding the moment generating function (MGF) of each random variable individually. In this subsection, we will present a few different forms of Bernstein's inequality that will be useful to follow the proof of multiple results presented in our paper.

According to [Ver18, Theorem 2.8.1],

Theorem A.1.2 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_N$ be independent mean-zero subexponential random variables. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left(|\sum_{i=1}^{N} X_i| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right)\right), \tag{A.2}$$

where $c > 0$ is an absolute constant.

Let us also write down another version of Bernstein's inequality, [Ver18, Theorem 2.8.2] that is more adequate for the proof of Theorem 2.3.1.

Theorem A.1.3 (Bernstein's inequality, alternative version)**.** *Let $X_1, \ldots, X_N$ be independent mean-zero subexponential random variables and let $a = (a_1 \ldots, a_N) \in \mathbb{R}^N$. Then for every $t \geq 0$, we have*

$$\mathbb{P}\left(|\sum_{i=1}^{N} a_i X_i| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\max_i\|X_i\|_{\psi_1}^2 \sum_i^N a_i^2}, \frac{t}{\max_i \|X_i\|_{\psi_1} \max_i a_i}\right)\right), \tag{A.3}$$

where $c > 0$ is an absolute constant.

Corollary A.1.4. *There exists an absolute constant $c > 0$ such that, for any $X_1, \ldots, X_N$ independent mean-zero $\sigma-$subexponential random variables, any $a = (a_1 \ldots, a_N) \in \mathbb{R}^N$ such that $a_1 \geq \ldots \geq a_N > 0$, and any $s > 0$, with probability at least $1 - 2e^{-s}$,*

$$\left|\sum_i a_i X_i\right| \leq c\sigma \max\left(sa_1, \sqrt{s\sum_i^N a_i^2}\right).$$

Proof. Take $s = c \min\left\{\frac{t^2}{\sigma^2 \sum_i^N a_i^2}, \frac{s}{\sigma a_1}\right\}$ in Theorem A.1.3 and observe that

$$t = \frac{1}{c}\sigma \max\left\{sa_1, \sqrt{s\sum_i^N a_i^2}\right\}.$$

□

A.2 Miscellaneous

Definition A.2.1. The covariance matrix of a random vector $X \in \mathbb{R}^p$ is defined as $\operatorname{cov}(X) \coloneqq \mathbb{E}(X-\mu)(X-\mu)^T = \mathbb{E}XX^T - \mu\mu^T$ where $\mu \coloneqq \mathbb{E}X$. The second moment matrix of a random vector X is defined as

$$\Sigma(X) \coloneqq \mathbb{E}XX^T. \tag{A.4}$$

Remark A.2.1. We will often assume that the random vector X is centered ($\mathbb{E}X = 0$) as it is easy to go back to the general case from the centered case. Therefore, we will often have that $\operatorname{cov}(X) = \Sigma(X)$ and for simplicity we will denote the covariance matrix of X as Σ, when it is unambiguous.

Remark A.2.2. Note that the covariance matrix and the second moment matrix are positive semi-definite symmetric. Hence all their eigenvalues are non-negative.

Appendix B

Dual formulation of minimization problem

B.1 Dual norm

Definition B.1.1. Given a norm $\|\cdot\|$ on a Banach space $\mathcal{X}$, we define the dual space $\mathcal{X}^*$ to be the space of continuous linear functionals $\mathcal{X} \to \mathbb{R}$. Moreover, we define the dual norm $\|\cdot\|_* : \mathcal{X}^* \to \mathbb{R}$ to be given by

$$\|f\|_* \coloneqq \sup_{x \in B} |f(x)|,$$

where B denotes the unit ball in $\mathcal{X}$ with respect to the norm $\|\cdot\|$.

Example B.1.1. *Given a norm $\|\cdot\|$ on $\mathbb{R}^p$, $(\mathbb{R}^p)^*$ is the space of linear functionals $\mathbb{R}^p \to \mathbb{R}$. In this case, one can observe that the linear functionals on $\mathbb{R}^p$ can be characterized by vectors in $\mathbb{R}^p$, that is, for any linear functional f on $\mathbb{R}^p$, there exists a unique vector y in $\mathbb{R}^p$ such that for all x in $\mathbb{R}^p$, $f(x) = y^\top x$ and thus $(\mathbb{R}^p)^*$ can be identified with $\mathbb{R}^p$. Moreover, with a slight abuse of notations, the dual norm $\|\cdot\|_* : (\mathbb{R}^p) \to \mathbb{R}$ then becomes*

$$\|y\|_* \coloneqq \sup_{x \in B} \left|y^\top x\right| = \sup_{x \in B} y^\top x,$$

where B denotes the unit ball in $\mathbb{R}^p$ with respect to the norm $\|\cdot\|$.

Example B.1.2. *More generally, given real numbers $p_1, p_2 > 1$ such that $\frac{1}{p_1} + \frac{1}{p_2} = 1$, the dual norm of the $l^{p_1}-$norm $\|\cdot\|_{p_1}$ on $\mathbb{R}^p$ is given by the $l^{p_2}-$norm $\|\cdot\|_{p_2}$ on $\mathbb{R}^p$. Moreover, the dual norm of the $l^\infty-$norm is the l^1-norm, and vice-versa. Furthermore, the vector $\tilde{x} \in B_2^{p_1}$ that achieves*

$$(\|y\|_{p_1})_* = \sup_{x \in B_{p_1}^p} y^\top x = \|y\|_{p_2}$$

is the following.
For all $1 \leq i \leq p$, define z_i as follows.

$$z_i = sgn(y_i)\, |y_i|^{p_1 - 1},$$

where sgn denotes the sign function. Then $\tilde{x} = \frac{z}{\|z\|_{p_2}}$.

Example B.1.3. *Given $\|\cdot\|_2$ the l^2-norm on $\mathbb{R}^p$ and a symmetric matrix $\Gamma \in \mathbb{R}^{p\times p}$, define the norm $\|\cdot\| \coloneqq \|\Gamma\cdot\|_2$. Then its dual norm is given by $\|\cdot\|_* = \|\Gamma^{-1}\cdot\|_2$.*

Proof. Given a vector y in $\mathbb{R}^p$,

$$\begin{aligned}\|y\|_* &= \sup_{\|x\|\leq 1} \langle y, x\rangle = \sup_{\|\Gamma x\|_2\leq 1} \langle y, x\rangle \\ &= \sup_{\|\Gamma x\|_2\leq 1} \langle y, \Gamma^{-1}\Gamma x\rangle = \sup_{\|\Gamma x\|_2\leq 1} \langle \Gamma^{-1}y, \Gamma x\rangle \\ &= \sup_{\|z\|_2\leq 1} \langle \Gamma^{-1}y, z\rangle = \|\Gamma^{-1}y\|_2,\end{aligned}$$

where the last equality holds by the dual characterization of the norm. □

In particular, we will use this result for positive definite diagonal matrices.

B.2 Derivation of the dual problem

Given a norm $\|\cdot\|$ on $\mathbb{R}^p$, a matrix $\mathbb{X} \in \mathbb{R}^{n\times p}$ with independent rows with $p \geq n$, and given vectors $\beta \in \mathbb{R}^p$ and $y \in \mathbb{R}^n$, the primal minimization problem we want to solve is given by

$$\min_{\beta\in\mathbb{R}^p} \|\beta\| \tag{B.1}$$

$$\text{s.t. } \mathbb{X}\beta = y \tag{B.2}$$

It can be stated as Lagrange dual maximization problem. We show what the dual problem is and that the duality gap is zero, that is, the primal and dual problems are equivalent.

In what follows, we more or less adapt the exposition in [CLvdG22, Proof of Lemma 2.2.].

The Lagrangian is defined as

$$\mathcal{L} : \mathbb{R}^p \longrightarrow \mathbb{R}, \quad (\beta, \gamma) \longmapsto \mathcal{L}(\beta, \gamma) \coloneqq \|\beta\| + \gamma^\top(\mathbb{X}\beta - y).$$

The primal problem can be rewritten as

$$\min_{\beta\in\mathbb{R}^P} \max_{\gamma\in\mathbb{R}^n} \mathcal{L}(\beta, \gamma), \tag{B.3}$$

and the dual problem is given by

$$\max_{\gamma\in\mathbb{R}^n} \min_{\beta\in\mathbb{R}^P} \mathcal{L}(\beta, \gamma) \tag{B.4}$$

Let us first rewrite the problem $\min_{\beta\in\mathbb{R}^P} \mathcal{L}(\beta, \gamma)$ in a more suitable form.

$$\min_{\beta\in\mathbb{R}^P} \mathcal{L}(\beta, \gamma) = \min_{\beta\in\mathbb{R}^P} \big(\|\beta\| + \gamma^\top(\mathbb{X}\beta - y)\big) = -\gamma^\top y - \max_{\beta\in\mathbb{R}^p} \big(\langle \beta, -\mathbb{X}^\top\gamma\rangle - \|\beta\|\big) \tag{B.5}$$

where in the second equality we use the fact that $\langle \gamma, \mathbb{X}\beta\rangle = \langle \beta, \mathbb{X}^\top\gamma\rangle$. Now let us introduce the notion of convex conjugate of a function.

Definition B.2.1. Given a function $f : \mathbb{R}^p \longrightarrow \mathbb{R}$, the convex conjugate f^* of f is defined as

$$f^* : \mathbb{R}^p \longrightarrow \mathbb{R}, \quad y \longmapsto f^*(y) \coloneqq \sup_{x \in \mathbb{R}^p} (\langle y, x \rangle - f(x)). \tag{B.6}$$

Let us consider $f(\beta) = \|\beta\|$, then $f^*(\delta) = \begin{cases} 0 & \text{if } \delta \in B^*, \\ +\infty & \text{otherwise,} \end{cases}$

where B^* is unit the ball with respect to the dual norm. Indeed, by definition of the dual norm and from the discussion in Example B.1.1,

$$\|\delta\|_* = \sup_{\|x\| \leq 1} \delta^\top x,$$

Thus if $\|\delta\|_* \leq 1, \langle \delta, x \rangle \leq \|x\|$ for all x in $\mathbb{R}^p$ and in particular, for $x = 0, \langle \delta, x \rangle = 0$, which implies that $f^*(\delta) = 0$. On the other hand, if $\|\delta\|_* > 1$, there exists x such that $\|x\| \leq 1$ and $\langle \delta, x \rangle > 1$. Therefore, for all $t > 0, f^*(\delta) \geq \langle \delta, tx \rangle - \|tx\| = t(\langle \delta, x \rangle) - \|x\|$, and the right-hand side tends to infinity as t goes to infinity.

Hence, (B.4) becomes

$$\max_{\gamma \in \mathbb{R}^n} (-\gamma^\top y - \max_{\beta \in \mathbb{R}^p} \left(\langle \beta, -\mathbb{X}^\top \gamma \rangle - \|\beta\| \right), \tag{B.7}$$

and from our analysis of the behaviour of the convex conjugate, we notice that the dual problem becomes

$$\begin{aligned} &\max_{\gamma \in \mathbb{R}^n} \gamma^\top y \\ &\text{s.t. } \|\mathbb{X}^\top \gamma\|_* \leq 1 \end{aligned}$$

where $\|\cdot\|_*$ denotes the dual norm of $\|\cdot\|$ as defined in B.1. Therefore, by weak duality, we have that the following holds.

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \|\beta\| \quad &\geq \quad \max_{\gamma \in \mathbb{R}^n} \gamma^\top y \\ \text{s.t. } \mathbb{X}\beta = y \quad &\text{s.t. } \|\mathbb{X}^\top \gamma\|_* \leq 1 \end{aligned}$$

Moreover, since by assumption, $p \geq n$ and the rows of $\mathbb{X}$ are independent, the Moore-Penrose pseudoinverse exists. This implies that Slater's condition (see for example subsection 5.2.3 in [BV04]) is fulfilled and that strong duality holds, that is, the primal problem and the dual problem coincide. Therefore we obtain the following.

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \|\beta\| \quad &= \quad \max_{\gamma \in \mathbb{R}^n} \gamma^\top y \\ \text{s.t. } \mathbb{X}\beta = y \quad &\text{s.t. } \|\mathbb{X}^\top \gamma\|_* \leq 1 \end{aligned}$$

Appendix C

Complexity measure and Dvoretzky-Milman theorem

Most of the content of this appendix is drawn from the exposition in the book [Ver18].

C.1 Random processes

Definition C.1.1 (Random process)**.** A random process is a family of random variables $(X_t)_{t\in T}$ on the same probability space, indexed by the elements of some set T.

Given a set T that indexes the random process $(X_t)_{t\in T}$, we can define the canonical metric on T to be given by

$$d(t,s) \coloneqq \|X_t - X_s\|_2 = (\mathbb{E}(X_t - X_s)^2)^{1/2}. \tag{C.1}$$

There are two main quantities characterizing random processes, the covariance function $\Sigma(s,t)$ and the increments $\|X_t - X_s\|_2$. These two quantities are related to one another by the following relations

$$\begin{aligned}
&\mathbb{E}(X_t - X_s)^2)^{1/2} = (\Sigma(t,t) - 2\Sigma(t,s) + \Sigma(s,s))^{1/2} \quad \text{and} \\
&\Sigma(t,s) = \frac{1}{4}\left(\mathbb{E}(X_t - (-X_s))^2) - \mathbb{E}(X_t - X_s)^2)\right) \quad \forall t,s \in T.
\end{aligned}$$

C.1.1 Gaussian processes

A Gaussian process is a random process $(X_t)_t$ indexed by a set T such that for every finite subset $T_0 \subset T$, $(X_t)_{t\in T_0}$ is a Gaussian random vector. The quintessential Gaussian process is the canonical Gaussian process.

Definition C.1.2. The canonical Gaussian process is a Gaussian process of the form

$$X_t = \langle g, t\rangle, \quad t \in T,$$

where T is a subset of $\mathbb{R}^N$ and g is a standard Gaussian random vector in $\mathbb{R}^N$.

All Gaussian processes can be written in the form of a canonical Gaussian process, because for any Gaussian random vector $Y = (Y_1, \dots, Y_N)$, there exists points $t_1, \dots, t_n \in \mathbb{R}^N$ such that Y and $(\langle g, t_i \rangle)_{i=1}^n$ have the same law, where g is a standard Gaussian random vector in $\mathbb{R}^N$.

There is a wide range of tools to bound Gaussian processes, one of them being the so-called Borell-TIS theorem. We give here a version corresponding to [Led01, Theorem 7.1].

Theorem C.1.1. *Let $(G_t)_{t\in T}$ be a centered Gaussian process indexed by a countable set T such that $\sup_{t\in T} G_t < +\infty$ almost surely. Then*

$$\mathbb{E} \sup_{t\in T} G_t < +\infty,$$

and for all $r \geq 0$,

$$\mathbb{P}\left(\sup_{t\in T} G_t \geq \mathbb{E} \sup_{t\in T} G_t + r \right) \leq \exp\left(-r^2/2\sigma^2\right),$$

where $\sigma^2 = \sup_{t\in T} \mathbb{E}(G_t^2) < +\infty$.

C.2 Gaussian width

The Gaussian width of a set T is a fundamental characteristic of the set, comparable in informativeness with the volume or the surface area. It naturally arises as an overall bound on the canonical Gaussian process on T.

Definition C.2.1 (Gaussian width)**.** The Gaussian width of a subset $T \subset \mathbb{R}^n$ is defined as

$$w(T) = \mathbb{E} \sup_{t\in T} \langle g, t \rangle, \tag{C.2}$$

where g is a standard Gaussian vector in $\mathbb{R}^n$.

It has strong properties of invariance and linearity directly coming from the invariance properties of the normal distribution.

Stable dimension

The standard notion of dimension of a subset T of $\mathbb{R}^n$ is very sensitive to small perturbations of the set T. Imagine if T lives in a one-dimensional subspace of $\mathbb{R}^n$, for example imagine that it's a collection of points on a line. Then, if T is a set of data, with some small measurement error (perturbation), maybe T would still be very close to being enclosed in the line, its diameter and its Gaussian width would not change much. However, its dimension would immediately jump from 1 to n with high probability, assuming that the noise is i.i.d. Gaussian for example. Therefore the usual notion of dimension captures poorly the complexity of T. That motivates the definition of stable dimension. This concept relies on a notion that is very closely related to Gaussian width.

Definition C.2.2.

$$h(T) := \left(\mathbb{E} \sup_{t \in T} \langle g, t \rangle^2 \right)^{1/2}, \tag{C.3}$$

where g is a standard Gaussian vector in $\mathbb{R}^n$.

We have the following relationship between Gaussian width and this new quantity:

Theorem C.2.1.

$$2w(T) \leq h(T - T) \leq 2Cw(T). \tag{C.4}$$

Definition C.2.3 (Stable dimension)**.** For a bounded subset T in $\mathbb{R}^n$, the stable dimension is defined as

$$d(T) := \frac{h(T-T)^2}{\operatorname{diam}(T)^2} \sim \frac{w(T)^2}{\operatorname{diam}(T)^2}. \tag{C.5}$$

Let us give a statement of Dvoretzky-Milman theorem, following the exposition of [Ver18, Theorem 11.3.3]

Theorem C.2.2. *Let $\mathbb{G}$ be an $n \times p$ Gaussian random matrix with i.i.d. $\mathcal{N}(0,1)$ entries, let $T \subset \mathbb{R}^p$ be a bounded set, and let $\epsilon \in (0,1)$. Suppose that*

$$n \leq c\epsilon \frac{w(T)^2}{diam(T)^2},$$

for some constant c. Then with high probability,

$$(1-\epsilon) w(T) B_2^p \subset conv(\mathbb{G}T) \subset (1+\epsilon) w(T) B_2^p,$$

where $conv(\mathbb{G}T)$ denotes the convex hull of the set $\mathbb{G}T$.

There are multiple proofs of Theorem C.2.2, and the approach chosen in [Ver18] is to prove it via a two-sided Chevet's inequality that is obtained through the tools provided by generic chaining and Talagrand's majorizing theorem.

Acknowledgments

I would like to thank my professor, Prof. Dr. Sara van de Geer, that recommended a great book about the mathematics of data science which subsequently encouraged me to do my Master's thesis under her supervision as one of her last Master students. She also found an exciting project for me to work on, which consolidated my appreciation for the subject of high dimensional probability and statistics. I would also like to thank Dr. Geoffrey Chinot, who supervised my progress attentively and guided me through the hurdles of working on my first research experience of this scale, while providing me with the insight of an experimented practitioner in the field. Lastly, I want to thank my family and friends for their support.

Bibliography

[BHMM19a] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, jul 2019.

[BHMM19b] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[BLLT20] Peter L. Bartlett, Philip M. Long, Gàbor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, apr 2020.

[BMR21] Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint, 2021.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Chi19] Geoffrey Chinot. Robust learning and complexity dependent bounds for regularized problems, 2019.

[CLvdG22] Geoffrey Chinot, Matthias Löffler, and Sara van de Geer. On the robustness of minimum norm interpolators and regularized empirical risk minimizers. *The Annals of Statistics*, 50(4):2306 – 2333, 2022.

[CRT05] Emmanuel Candes, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements, 2005.

[CW08] Emmanuel J. Candes and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[HTF09] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer, 2009.

[KL14] Vladimir Koltchinskii and Karim Lounici. Concentration inequalities and moment bounds for sample covariance operators, 2014.

[Led01] M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001.

[LS22] Guillaume Lecué and Zong Shang. A geometrical viewpoint on the benign overfitting property of the minimum l_2-norm interpolant estimator, 2022.

[Men14] Shahar Mendelson. Learning without concentration for general loss functions, 2014.

[Tao09] Terence Tao. Compressed sensing, 2009.

[TB20] A. Tsigler and P. L. Bartlett. Benign overfitting in ridge regression, 2020.

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[ZBH+16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016.

[ZBH+21] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, feb 2021.