A geometrical viewpoint on the benign overfitting property of the minimum ℓ_2-norm interpolant estimator.

Guillaume Lecué*and Zong Shang

email: *guillaume.lecue@ensae.fr*, *email:* *zong.shang@ensae.fr*
CREST, ENSAE, IPParis. 5, avenue Henry Le Chatelier, 91120 Palaiseau, France.

Abstract

Practitioners have observed that some deep learning models generalize well even with a perfect fit to noisy training data [5, 45, 44]. Since then many theoretical works have revealed some facets of this phenomenon [4, 2, 1, 8] known as *benign overfitting*. In particular, in the linear regression model, the minimum ℓ_2-norm interpolant estimator $\hat{\boldsymbol{\beta}}$ has received a lot of attention [1, 39] since it was proved to be consistent even though it perfectly fits noisy data under some condition on the covariance matrix Σ of the input vector. Motivated by this phenomenon, we study the generalization property of this estimator from a geometrical viewpoint. Our main results extend and improve the convergence rates as well as the deviation probability from [39]. Our proof differs from the classical bias/variance analysis and is based on the *self-induced regularization* property introduced in [2]: $\hat{\boldsymbol{\beta}}$ can be written as a sum of a ridge estimator $\hat{\boldsymbol{\beta}}_{1:k}$ and an overfitting component $\hat{\boldsymbol{\beta}}_{k+1:p}$ which follows a decomposition of the features space $\mathbb{R}^p = V_{1:k} \oplus^{\perp} V_{k+1:p}$ into the space $V_{1:k}$ spanned by the top k eigenvectors of Σ and the ones $V_{k+1:p}$ spanned by the $p-k$ last ones. We also prove a matching lower bound for the expected prediction risk. The two geometrical properties of random Gaussian matrices at the heart of our analysis are the Dvoretsky-Milman theorem and isomorphic and restricted isomorphic properties. In particular, the Dvoretsky dimension appearing naturally in our geometrical viewpoint coincides with the effective rank from [1, 39] and is the key tool to handle the behavior of the design matrix restricted to the sub-space $V_{k+1:p}$ where overfitting happens.

AMS subject classification: 62F35
Keywords: benign overfitting, deep learning models, self-induced regularization, Dvoretsky-Milman theorem, restricted isomorphy property,

1 Introduction

Chaldni patterns, Melde experiments and so on have been starting points of profound discoveries in Mathematics. Nowadays sources of observations and experiments include the behavior of algorithms if we think of computers as experimental devices. The numerous successes of deep learning methods [9] during the last decade can thus be seen as experiments that scientists of all kinds can try to understand and in particular mathematicians to explain. This may lead to a profound rethinking of statistical machine learning and potentially to new principles, quoting Joseph Fourier [16]: 'The principles of the theory are derived, as are those of rational mechanics, from a very small number of primary facts'.

One side of the 'deep learning experiment' that has received a lot of attention from the statistical community during the last five years is about an unexpected behavior of some deep learning models that perfectly fit (potentially noisy) data but can still generalize well [5, 45, 44]. This is a bit unexpected because it is quite rare to find a book about statistical learning written before the 'era of deep learning' which does not advise any engineer or student to avoid an algorithm that would tend to stick too much to the data. Overfitting on the training dataset was something that any statistician (the data scientists of that time) would try to avoid or to correct by (explicit) regularization when designing their learning models. This principle has been profoundly rethinked during the last five years.

This property is nowadays called the *benign overfitting* (BO) phenomenon [4, 2] and has been the subject of many recent works in the statistical community. Motivation is to identify situations where benign overfitting holds that is

*G.L. is supported by a grant of the French National Research Agency (ANR), "Investissements d'Avenir" (LabEx Ecodec/ANR-11-LABX-0047), by the Médiamétrie chair on "Statistical models and analysis of high-dimensional data" and by the French ANR PRC grant ADDS (ANR-19-CE48-0005).

when an estimator with a perfect fit on the training data can still generalize well. Several results have been obtained on BO in various statistical frameworks such as regression [7, 8, 34, 6, 17, 33, 43, 13, 46, 26, 29, 14, 47, 11, 34, 21] and classification [7, 37, 12, 27, 42]. All the point is to consider some estimators that perfectly fit the data – such estimators are called *interpolant estimators* – and to find conditions under which they can generalize well (that is to predict well on out-of-sample data); that is to prove that their excess risk or estimation risk tend to zero – in other words, when interpolant estimators are consistent.

In the linear regression model, it has been understood that BO happens principally for anisotropic design vectors having some special decay properties on their spectrum [1, 39, 46] except for the result from [41] where BO was achieved by the minimum ℓ_1-norm interpolent estimator in the isotropic case. Two important results on the BO in this model are Theorem 4 in [1] and Theorem 1 in [39] that we recall after introducing the model and the interpolant estimator.

The anistropic Gaussian design linear regression model with Gaussian noise. We consider $y = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$ where $\mathbb{X} : \mathbb{R}^p \to \mathbb{R}^N$ is a Gaussian matrix with i.i.d. $\mathcal{N}(0, \Sigma)$ row vectors, $\boldsymbol{\xi}$ is an independent Gaussian noise $\mathcal{N}(0, \sigma_\xi I_N)$ and $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the unknown parameter of the model. Dimension p is the number of features. We write

$$\mathbb{X} = \begin{pmatrix} (\Sigma^{1/2} G_1)^\top \\ \vdots \\ (\Sigma^{1/2} G_N)^\top \end{pmatrix} = \mathbb{G}^{(N\times p)}\Sigma^{1/2} \tag{1}$$

where $G_1, \ldots, G_N$ are N i.i.d. $\mathcal{N}(0, I_p)$ and $\mathbb{G}^{(N\times p)}$ is a $N \times p$ standard Gaussian matrix (with i.i.d. $\mathcal{N}(0,1)$ entries). We consider this model as a benchmark model because it is likely not reflecting real world data but it is the one that is expected to be universal in the sense that results obtained in other more realistic statistical model could be compared with or tend to the one obtained in this ideal benchmark Gaussian model. The relevance of the approximation of large neural networks by linear models via the neural tangent kernel [19, 18] feature map in some regimes has been discussed a lot in the machine learning community for instance in [4, 28, 1] and references therein.

The minimum ℓ_2^p-norm interpolant estimator. Interpolant estimators are estimators $\hat{\boldsymbol{\beta}}$ that perfectly fit the data, i.e. such that $\mathbb{X}\hat{\boldsymbol{\beta}} = y$. We denote by $\|\cdot\|_2$ the Euclidean norm on $\mathbb{R}^p$ and consider the interpolant estimator having the smallest ℓ_2^p-norm

$$\hat{\boldsymbol{\beta}} \in \underset{\boldsymbol{\beta}:\mathbb{X}\boldsymbol{\beta}=y}{\operatorname{argmin}} \|\boldsymbol{\beta}\|_2 \,. \tag{2}$$

The relevance of the minimum ℓ_2^p-norm interpolant estimator for neural networks models is that gradient descent algorithms converges to such estimators in some regimes (see Section 3.9 from [4] and Section 3 from [2] and references therein).

The goal is to identify situations (that is covariance matrices Σ and signal $\boldsymbol{\beta}^*$) for which $\hat{\boldsymbol{\beta}}$ generalizes well even though it perfectly fits the data. Generalization capacity of an estimator $\hat{\boldsymbol{\beta}}$ is measured via the excess risk:

$$\mathbb{E}[(Y - \langle X, \hat{\boldsymbol{\beta}}\rangle)^2 | (y, \mathbb{X})] - \mathbb{E}(Y - \langle X, \boldsymbol{\beta}^*\rangle)^2 = \left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2^2$$

where $Y = \langle X, \boldsymbol{\beta}^*\rangle + \xi$, $X \sim \mathcal{N}(0, \Sigma)$ and $\xi \sim \mathcal{N}(0, \sigma_\xi)$ is independent of X. In this model, the excess risk is an estimation risk with respect to the norm $\|\Sigma^{1/2}\cdot\|_2$. Our goal is to prove high probability upper bounds on the generalization error $\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2^2$ of the minimum ℓ_2-norm interpolant estimator $\hat{\boldsymbol{\beta}}$ from (2) and to identify cases (i.e. Σ and $\boldsymbol{\beta}^*$) where this bound (i.e. the rate of convergence) tends to zero as N and p go to infinity. These cases are all situations where the benign overfitting phenomenon takes place. The study of the BO requires to make both N and p tend to infinity; it means that the number p of features also increases with the sample size N. A way such asymptotic can be performed is by adding features in such a way that only the tail of the spectrum of Σ is modified by adding eigenvalues smaller to the previous one as p tends to infinity.

Two seminal results on benign overfitting in linear regression. We state the main upper bound results from [1] and [39] in the Gaussian linear model introduced above and for the minimum ℓ_2^p-norm interpolant estimator (2) even though they have been obtained under weaker assumptions such as sub-gaussian assumptions on the design and the noise or for a misspecified model or for the ridge estimator. Both results depend on two concepts of effective rank of Σ. To recall their definitions, we use the following notation. The SVD of Σ is $\Sigma = UDU^\top$ and $(f_1, \ldots, f_p)$ are the rows vectors of $U^\top$ and $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)$ is the $p \times p$ diagonal matrix with $\sigma_1 \geq \cdots \geq \sigma_p > 0$ such that $\Sigma f_j = \sigma_j f_j$ for all $j \in [p]$. For all $k \in [p]$ we denote $V_{1:k} = \operatorname{span}(f_1, \ldots, f_k)$ and $V_{k+1:p} = \operatorname{span}(f_{k+1}, \ldots, f_p)$. We denote by $P_{1:k} : \mathbb{R}^p \mapsto \mathbb{R}^p$ (resp. $P_{k+1:p}$) the orthogonal projection onto $V_{1:k}$ (resp. $V_{k+1:p}$) and for all $\boldsymbol{\beta} \in \mathbb{R}^p$, we denote $\boldsymbol{\beta}_{1:k} := P_{1:k}\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{k+1:p} := P_{k+1:p}\boldsymbol{\beta}$. We denote $\Sigma_{1:k} = UD_{1:k}U^\top$ and $\Sigma_{k+1:p} = UD_{k+1:p}U^\top$ where $D_{1:k} = \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0)$ and $D_{k+1:p} = \operatorname{diag}(0, \ldots, 0, \sigma_{k+1}, \ldots, \sigma_p)$. We also denote $X_{1:k} = \mathbb{G}^{(N\times p)}\Sigma_{1:k}^{1/2}$ and $X_{k+1:p} = \mathbb{G}^{(N\times p)}\Sigma_{k+1:p}^{1/2}$ so that $\mathbb{X} = X_{1:k} + X_{k+1:p}$. The two effective ranks used in [1, 39] are given by

$$r_k(\Sigma) = \frac{\operatorname{Tr}(\Sigma_{k+1:p})}{\|\Sigma_{k+1:p}\|_{op}} \text{ and } R_k(\Sigma) = \frac{(\operatorname{Tr}(\Sigma_{k+1:p}))^2}{\operatorname{Tr}(\Sigma_{k+1:p}^2)}. \tag{3}$$

Theorem 1. *[Theorem 4 in [1]] There are absolute constants $b, c, c_1 > 1$ for which the following holds. Define*

$$k_b^* = \min\left(k \geq 0 : r_k(\Sigma) \geq bN\right), \tag{4}$$

where the infimum of the empty set is defined as $+\infty$. Suppose $\delta < 1$ with $\log(1/\delta) < n/c$. If $k_b^* \geq n/c_1$ then $\mathbb{E}\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \geq \sigma_\xi/c$. Otherwise,

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \leq c\|\boldsymbol{\beta}^*\|_2\sqrt{\|\Sigma\|_{op}}\max\left(\left(\frac{r_0(\Sigma)}{N}\right)^{\frac{1}{4}}, \sqrt{\frac{r_0(\Sigma)}{N}}, \left(\frac{\log(1/\delta)}{N}\right)^{\frac{1}{4}}\right) + c\log(1/\delta)\sigma_\xi\left(\sqrt{\frac{k_b^*}{N}} + \sqrt{\frac{N}{R_{k_b^*}(\Sigma)}}\right) \tag{5}$$

with probability at least $1-\delta$, and $\mathbb{E}\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \geq (\sigma_\xi/c)\left(\sqrt{k_b^*/N} + \sqrt{N/R_{k_b^*}(\Sigma)}\right)$.

Theorem 1 is one of the first results proving that the BO phenomenon in the linear regression model can happen by identifying situations where the upper in (5) tends to zero: quoting [1]: '..$r_0(\Sigma)$ *should be small compared to the sample size N (from the first term) and $r_{k_b^*}(\Sigma)$ and $R_{k_b^*}(\Sigma)$ should be large compared to N. Together, these conditions imply that overparametrization is essential for BO in this setting: the number of non-zero eigenvalues should be large compared to N, they should have a small sum compared to N, and there should be many eigenvalues no larger than $\sigma_{k_b^*}$... For these reason, we say that a sequence of covariance operator Σ is benign if*

$$\lim_{N,p\to+\infty}\frac{r_0(\Sigma)}{N} = \lim_{N,p\to+\infty}\frac{k_b^*}{N} = \lim_{N,p\to+\infty}\frac{N}{R_{k_b^*}(\Sigma)} = 0. \tag{6}$$

The analysis of the BO from Theorem 1 is based uniquely on the behavior of the spectrum of Σ. In particular, it does not depend on the signal $\boldsymbol{\beta}^*$ (even though the norm $\|\boldsymbol{\beta}^*\|_2$ appears as a multiplying factor in (5) and this quantity may be large since $\boldsymbol{\beta}^*$ is a vector in $\mathbb{R}^p$ with p being large to allow for over-parametrization). This analysis was improved in the subsequent work from [39]. We recall the main result from [39] only for the ridge estimator with regularization parameter $\lambda = 0$ since in that case it coincides with the minimum ℓ_2-norm interpolant estimator $\hat{\boldsymbol{\beta}}$ from (2).

Theorem 2. *There are absolute constants such that the following holds. Let $\delta < 1 - 4\exp(-N/c_0)$ and $k \leq N/c_0$. We assume that with probability at least $1-\delta$, the condition number of $X_{k+1:p}X_{k+1:p}^\top$ is smaller than L. For all $t \in (1, N/c_0)$, with probability at least $1 - \delta - 20\exp(-t)$,*

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \lesssim L^2\left(\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_{k+1:p}^*\right\|_2 + \left\|\Sigma_{1:k}^{-1/2}\boldsymbol{\beta}_{1:k}^*\right\|_2\frac{\operatorname{Tr}(\Sigma_{k+1:p})}{N}\right) + \sigma_\xi\sqrt{t}L\left(\sqrt{\frac{k}{N}} + \frac{\sqrt{N\operatorname{Tr}(\Sigma_{k+1:p}^2)}}{\operatorname{Tr}(\Sigma_{k+1:p})}\right) \tag{7}$$

and

$$\frac{\operatorname{Tr}(\Sigma_{k+1:p})}{\|\Sigma_{k+1:p}\|_{op}} \geq \frac{cN}{L}. \tag{8}$$

Compare with Theorem 1, Theorem 2 shows that this is not the ℓ_2-norm of the entire vector $\boldsymbol{\beta}^*$ that has to be paid by $\hat{\boldsymbol{\beta}}$ in the bias term but the weighted ℓ_2 norm $\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_{k+1:p}^*\right\|_2$ of the signal $\boldsymbol{\beta}^*$ restricted to $V_{k+1:p}$, the space spanned by the $p-k$ smallest eigenvectors of Σ as well as $\left\|\Sigma_{1:k}^{-1/2}\boldsymbol{\beta}_{1:k}^*\right\|_2\operatorname{Tr}(\Sigma_{k+1:p})/N$. Theorem 2 also introduces a key idea that the feature space $\mathbb{R}^p$ should be decomposed as $V_{1:k}\oplus^\perp V_{k+1:p}$. This decomposition of $\mathbb{R}^p$ is actually associated to a decomposition of the estimator $\hat{\boldsymbol{\beta}}$ into a *prediction component* and an *overfitting component* following the *self-induced regularization* idea from [2] (see also the 'spiked-smooth' estimates from [42]).

The choice of k from [39, 2] is given by k_b^* from (4) and is called the *effective dimension* in [2]. For that choice of k, it is claimed in [2] that the self-induced regularization property of $\hat{\boldsymbol{\beta}}$ is to write it as a sum of an OLS over $V_{1:k}$ and an overfitting component over $V_{k+1:p}$. Moreover, for this choice of k, the upper bound on the variance term from the analysis of [39, 2] has been proved to be optimal and a lower bound on a Bayesian version of the bias term matching the upper bound on the bias term from [39] has been proved in [39]. In particular for benign overfitting, these results require $k_b^* = o(N)$.

Questions. Implicit in the work of [39] is the choice of decomposition $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ with J of the form $\{1, \dots, k\}$ and $k \lesssim N$ – where $V_J = \mathrm{span}(f_j : j \in J)$. One may however wonder if it is the best possible form of decomposition of $\mathbb{R}^p$ or if there are better choices of J (see [21])? Second, if it is the case that the best possible choice is of the form $J = \{1, \dots, k\}$ then what is the best possible choice of k? Given that the upper bounds from [1, 39] and ours below depend on the signal $\boldsymbol{\beta}^*$, one may expect the best choice of J and k (for $J = \{1, \dots, k\}$) would depend on $\boldsymbol{\beta}^*$. Several lower bound results on the variance term from [1, 39] may let us think that it is not the case. We will prove thanks to matching upper and lower bound on the prediction risk itself (for any $\boldsymbol{\beta}^*$) that the best decomposition is $\mathbb{R}^p = V_{1:k_b^*} \oplus^\perp V_{k_b^*+1:p}$ for k_b^* introduced in [1] (see (4)). Again, this fact is not obvious, given that the matching upper and lower bound depend on $\boldsymbol{\beta}^*$ and thus minimizing such a bound over all decomposition $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ for $J \subset [p]$ is expected to depend on $\boldsymbol{\beta}^*$ a priori. This is in fact not the case: whatever the signal is, $J = \{1, \dots, k_b^*\}$ is always the best choice made by $\hat{\boldsymbol{\beta}}$. This result shows that the two choices from [1, 39] about the best decomposition of $\mathbb{R}^p$ with the best choice of k are in fact optimal.

Our contribution. We improve the rates and probability deviations from Theorem 1 and Theorem 2. We extend the results to $k \gtrsim N$. This allows for a possibly better choice of the splitting parameter k that can lower the price of overfitting (overfitting happens on the space $V_{k+1:p}$). We also extend the features space decomposition (see Section 4.3) beyond the one of the form $\mathbb{R}^p = V_{1:k} \oplus^\perp V_{k+1:p}$ used in Theorem 2 to any features space decomposition of the form $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ – where $V_J = \mathrm{span}(f_j : j \in J)$ – with no restriction on the dimension $|J|$ of V_J (where estimation happens). We also improve the lower bound results from [1, 39] by removing unnecessary assumptions (see Section 4.4 for more detail) as well as obtaining a result for the risk itself and not a Bayesian one as in [39]. Indeed, we obtain a lower bound on the expected prediction risk which shows that the best choice of J is $\{1, \dots, k_b^*\}$. However, we emphasis through that rate that the benign overfitting property of $\hat{\boldsymbol{\beta}}$ depends on both Σ and $\boldsymbol{\beta}^*$ and in particular, on their interplay, i.e. the coordinates of $\boldsymbol{\beta}^*$ in a basis of eigenvectors of Σ (and not only on the spectrum of Σ). Our main contribution lies also in a technical point of view: our proofs are based upon the *self-regularization property* [2] of $\hat{\boldsymbol{\beta}}$ and not on a bias/variance decomposition as in [39, 1]. We also show that the Dvoretsky-Milman theorem is a key result to handle the behavior of the design matrix restricted to $V_{k+1:p}$. In particular, we recover the effective rank $r_k(\Sigma)$ from [1, 39] as a natural geometric parameter of the problem: the Dvortesky dimension of the ellipsoid $\Sigma_{k+1:p}^{-1/2} B_2^p$ (i.e. the one associated to the covariance matrix Σ restricted to the subspace of $\mathbb{R}^p$ where overfitting happens). The other geometric property that we introduce is a restricted isomorphic property of the design matrix $\mathbb{X}$ restricted to $V_{1:k}$ which is needed when $k \gtrsim N$. These two tools are at the basis of our 'geometrical viewpoint' on the benign overfitting phenomenon.

The paper is organized as follows. In the next section, we provide the two main geometrical tools we will use to prove the main results. In the third section, we provide a high-level description of the proofs of our main results. In Section 4, we provide the four main results of this paper, identify situations of benign overfitting and introduce a self-adaptive property of $\hat{\boldsymbol{\beta}}$. All proofs are given in the last section.

Notation. For all $q \geq 1$, $\|\cdot\|_q$ denotes the ℓ_q-norm in $\mathbb{R}^p$, B_q^p its unit ball and $\mathcal{S}_q^{p-1}$ its unit sphere. The operator norm of a matrix is denoted by $\|\cdot\|_{op}$. For $A \in \mathbb{R}^{m \times n}$, $s_1(A) \geq \cdots \geq s_{m \wedge n}(A)$ denote the singular values of A and its spectrum is the vector $\mathrm{spec}(A) = (s_1(A), \cdots, s_{m \wedge n}(A))$. The set of integers $\{1, \dots, p\}$ is denoted by $[p]$. Let Γ be a $p \times p$ semi-definite (not necessarily positive) matrix. We denote by Γ^{-1} the generalized inverse of this matrix.

2 Two geometrical properties of Gaussian matrices

We will use two properties of Gaussian matrices related to Euclidean sections of convex bodies. The first one is the Dvoretsky-Milman's theorem (which implies the existence of Euclidean sections of a body via a section of that body by the span of a Gaussian matrix) and the second one is a restricted isomorphy property of Gaussian matrices (which implies Euclidean sections via the kernel of a Gaussian matrix). We recall these two tools because they play a central role in our analysis of the minimum ℓ_2-norm interpolant estimator $\hat{\boldsymbol{\beta}}$.

The Dvoretsky-Milman's theorem for ellipsoids. The original aim of the Dvoretsky-Milman (DM) theorem is to show the existence of Euclidean sections of a convex body B and to identify the largest possible dimension (up to constant) of such Euclidean sections. Even though the terms of the problem are purely deterministic the only way known so far to obtain such sections are obtained via the range or the kernel of some random matrices such as standard Gaussian matrices and that is the reason why we will use this theorem for the properties of Gaussian random matrices they also use. We provide the general form of the DM theorem even though we will use it only for ellipsoids.

Definition 1. *Let $\|\|\cdot\|\|$ be some norm onto $\mathbb{R}^p$ denote by B its unit ball, by $\|\|\cdot\|\|_*$ its associated dual norm and by B^* its unit dual ball so that for all $x \in \mathbb{R}^p$, $\|\|x\|\| = \sup(\langle x, y\rangle, y \in B^*)$. The Gaussian mean width of B^* is $\ell^*(B^*) = \mathbb{E}\sup(\langle x, G\rangle : x \in B^*)$ where $G \sim \mathcal{N}(0, I_p)$ and the Dvoretsky dimension of B is*

$$d_*(B) = \left(\frac{\ell^*(B^*)}{\operatorname{diam}(B^*, \ell_2^p)}\right)^2$$

where $\operatorname{diam}(B^*, \ell_2^p) = \sup\left(\|v\|_2 : v \in B^*\right)$.

The Dvoretsky dimension $d_*(B)$ is up to absolute constants the largest dimension of an Euclidean section of B (see [35] for both upper and lower bounds). For instance, $d_*(B_1^p) = p$ means that there exists a subspace E of $\mathbb{R}^p$ of dimension $c_0 p$ where c_0 is some absolute constant such that $(c_1/\sqrt{p})B_2^p \cap E \subset B_1^p \cap E \subset (c_2/\sqrt{p})B_2^p \cap E$ – in other words the section $B_1^p \cap E$ of the convex body B_1^p by E looks like the Euclidean ball $B_2^p \cap E$ with radius $(1/\sqrt{p})$ up to absolute constants. Another example that we use below is the case of ellipsoids $\Sigma^{-1/2}B_2^p = \{v \in \mathbb{R}^p : \left\|\Sigma^{1/2}v\right\|_2 \leq 1\}$ where $\Sigma \in \mathbb{R}^{p\times p}$ is a PSD matrix. In that case, the Dvoretsky's dimension of $\Sigma^{-1/2}B_2^p$ is

$$\frac{\operatorname{Tr}(\Sigma)}{4\|\Sigma\|_{op}} \leq \frac{(\operatorname{Tr}(\Sigma) - 2\|\Sigma\|_{op}) \vee \|\Sigma\|_{op}}{\|\Sigma\|_{op}} \leq d_*(\Sigma^{-1/2}B_2^p) \leq \frac{\operatorname{Tr}(\Sigma)}{\|\Sigma\|_{op}} \tag{9}$$

because $\ell^*((\Sigma^{-1/2}B_2^p)^*) = \ell^*(\Sigma^{1/2}B_2^p) = \mathbb{E}\left\|\Sigma^{1/2}G\right\|_2 \leq \sqrt{\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2^2} = \sqrt{\operatorname{Tr}(\Sigma)}$ and $\mathbb{E}\left\|\Sigma^{1/2}G\right\|_2 \geq \sqrt{\operatorname{Tr}(\Sigma)/2}$. The quantity appearing in the right-hand side of (9) has been used several times in the litterature on interpolant estimators and was called there the effective rank or effective dimension [1, 39] (see $r_k(\Sigma)$ in (3)). As a consequence, assuming that $\|\Sigma\|_{op} N \lesssim \operatorname{Tr}(\Sigma)$ is equivalent to assume the existence of an Euclidean section of dimension of the order of N of the ellipsoid $\Sigma^{-1/2}B_2^p$.

Let us now state the general form of the Dvoretsky-Milman's Theorem that will be useful for our purpose that is the one that makes explicitly appearing a standard Gaussian matrix. The following form of Dvoretsky's theorem is due to V.Milman (up to the minor modification that [32] uses the surface measure on the Euclidean sphere $\mathcal{S}_2^{p-1}$ instead of the Gaussian measure on $\mathbb{R}^p$).

Theorem 3. *There are absolute constants $\kappa_{DM} \leq 1$ and c_0 such that the following holds. Denote by $\mathbb{G} := \mathbb{G}^{(N\times p)}$ the $N \times p$ standard Gaussian matrix with i.i.d. $\mathcal{N}(0, 1)$ Gaussian entries. Assume that $N \leq \kappa_{DM} d_*(B)$ then with probability at least $1 - \exp(-c_0 d_*(B))$, for every $\boldsymbol{\lambda} \in \mathbb{R}^N$,*

$$\frac{1}{\sqrt{2}}\|\boldsymbol{\lambda}\|_2 \ell^*(B^*) \leq \left\|\left\|\mathbb{G}^\top\boldsymbol{\lambda}\right\|\right\| \leq \sqrt{\frac{3}{2}}\|\boldsymbol{\lambda}\|_2 \ell^*(B^*). \tag{10}$$

Classical proofs of the Dvoretsky-Milman (DM) theorem hold only with constant probability as in [40] or [35] because DM's theorem is mainly used to prove an existence result: that $\operatorname{Im}(\mathbb{G}^\top)$ realizes an Euclidean section of the convex body B. However, we will need DM's theorem to hold with large probability and it is straightforward to get the right probability deviation as announced in Theorem 3 above for instance from the proofs in [40] or [35]. Theorem 3 has also implications on standard Gaussian matrices. We state such an outcome for anistropic Gaussian matrices.

Proposition 1. *Let $\mathbb{X}_2 = \mathbb{G}^{(N\times p)}\Gamma^{1/2}$ where $\mathbb{G}^{(N\times p)}$ is a $N \times p$ standard Gaussian matrice and Γ is a semi-definite matrix. Assume that $N \leq \kappa_{DM} d_*(\Gamma^{-1/2}B_2^d)$. With probability at least $1 - \exp(-c_0 d_*(\Gamma^{-1/2}B_2^d))$,*

$$\left\|\mathbb{X}_2\mathbb{X}_2^\top - \ell^*(\Gamma^{1/2}B_2^p)^2 I_N)\right\|_{op} \leq (1/2)\ell^*(\Gamma^{1/2}B_2^p)^2$$

which implies that

$$\sqrt{s_1(\mathbb{X}_2\mathbb{X}_2^\top)} = s_1(\mathbb{X}_2) \leq \sqrt{3/2}\ell^*(\Gamma^{1/2}B_2^p) \leq \sqrt{3\operatorname{Tr}(\Gamma)/2}$$
$$\sqrt{s_N(\mathbb{X}_2\mathbb{X}_2^\top)} = s_N(\mathbb{X}_2) \geq (1/\sqrt{2})\ell^*(\Gamma^{1/2}B_2^p) \leq \sqrt{\operatorname{Tr}(\Gamma)}/2$$

and

$$\frac{2}{\sqrt{\mathrm{Tr}(\Gamma)}} \geq s_1[\mathbb{X}_2^\top(\mathbb{X}_2\mathbb{X}_2^\top)^{-1}] \geq s_N[\mathbb{X}_2^\top(\mathbb{X}_2\mathbb{X}_2^\top)^{-1}] \geq \sqrt{\frac{2}{3\,\mathrm{Tr}(\Gamma)}}.$$

Proof. It follows from Theorem 3 for $\|\cdot\| = \left\|\Gamma^{1/2}\cdot\right\|_2$, (so that $B = \Gamma^{-1/2}B_2^p$ and $B^* = \Gamma^{1/2}B_2^p$) that with probability at least $1-\exp(-c_0 d_*(\Gamma^{-1/2}B_2^d))$,

$$\left\|\mathbb{X}_2\mathbb{X}_2^\top - \ell^*(\Gamma^{1/2}B_2^p)^2 I_N\right\|_{op} = \sup_{\|\boldsymbol{\lambda}\|_2=1}\left|\left\|\mathbb{X}_2^\top\boldsymbol{\lambda}\right\|_2^2 - \ell^*(\Gamma^{1/2}B_2^p)^2\left\|\boldsymbol{\lambda}\right\|_2^2\right| \leq \frac{\ell^*(\Gamma^{1/2}B_2^p)^2}{2}.$$

Moreover, we already saw that $\mathrm{Tr}(\Gamma)/2 \leq \ell^*(\Gamma^{1/2}B_2^p)^2 \leq \mathrm{Tr}(\Gamma)$ which proves the first statements. We finish the proof by using that for all $j=1,\dots,N$, $s_j[\mathbb{X}_2^\top(\mathbb{X}_2\mathbb{X}_2^\top)^{-1}] = \sqrt{s_j[(\mathbb{X}_2\mathbb{X}_2^\top)^{-1}]}$. ∎

A consequence of Proposition 1 is that in the Dvoretsky's regime (i.e. when $N \leq \kappa_{DM} d_*(\Gamma^{-1/2}B_2^d)$), $\mathbb{X}_2\mathbb{X}_2^\top = \mathbb{G}\Gamma\mathbb{G}^\top$ (where $\mathbb{G} = \mathbb{G}^{(N\times p)}$) behaves almost like the homothety $\ell^*(\Gamma^{1/2}B_2^p)^2 I_N$. DM's theorem holds for values of N small enough – which means from a statistical viewpoint that it is a high-dimensional property.

Remark 1 (Effective rank and the Dvoretsky dimension of an ellipsoid)**.** *The effective rank $r_k(\Sigma)$ (see (3)) and the Dvoretsky dimension of $\Sigma_{k+1:p}^{-1/2}B_2^p$ are up to absolute constants the same quantity. Note also that when the condition number L from Theorem 2 is a constant, the lower bound from (8) is a consequence of a general result on the upper bound on the dimension of an Euclidean section of a convex body which implies (8) in the case of Ellispoid (see Proposition 4.6 in [35]). Indeed, when L is like a constant, $X_{k+1:p}X_{k+1:p}^\top$ behaves like an isomorphy and so $\Sigma_{k+1:p}^{-1/2}B_2^p$ has an Euclidean section of dimension N given by the range of $X_{k+1:p}^\top$, hence, from Proposition 4.6 in [35], the Dvoretsky dimension of $\Sigma_{k+1:p}^{-1/2}B_2^p$ has to be larger than N up to some absolute constant and so (8) holds. The Dvoretsky-Milman theorem is behind several results previously used in literature about BO such as Lemma 9 and Lemma 10 in [1] and Lemma 4.3 in [2].*

There are cases where we want to avoid the condition '$N \leq \kappa_{DM} d_*(\Gamma^{-1/2}B_2^d)$' in DM theorem. Since this assumption is only used to obtain the left-hand side of (10) (which is sometimes called a lower isomorphic property), we can state a result on the right-hand side inequality of (10) without this assumption. The proof of the following statement is standard and follows the same scheme as the one of Theorem 3. We did not find a reference for it and so we sketch its proof for completeness.

Proposition 2. *Let $\|\cdot\|$ be some norm onto $\mathbb{R}^p$ and denote by B^* its unit dual ball. Denote by $\mathbb{G} := \mathbb{G}^{(N\times p)}$ the $N\times p$ standard Gaussian matrix with i.i.d. $\mathcal{N}(0,1)$ Gaussian entries. With probability at least $1-\exp(-N)$, for every $\boldsymbol{\lambda}\in\mathbb{R}^N$,*

$$\left\|\mathbb{G}^\top\boldsymbol{\lambda}\right\| \leq 2\left(\ell^*(B^*) + \sqrt{2N(1+\log(10))}\,\mathrm{diam}(B^*,\ell_2^p)\right)\|\boldsymbol{\lambda}\|_2.$$

Proof of Proposition 2. It is essentially the proof of the DM theorem which is followed here. Denote by $G_1,\dots,G_N$ the N i.i.d. $\mathcal{N}(0,I_p)$ row vectors of $\mathbb{G}$. Let $\boldsymbol{\lambda}\in\mathcal{S}_2^{N-1}$. It follows from Borell's inequality (see Theorem 7.1 in [24] or p.56-57 in [25]) that with probability at least $1-\exp(-t)$,

$$\left\|\mathbb{G}^\top\boldsymbol{\lambda}\right\| = \|G\| \leq \mathbb{E}\|G\| + \sqrt{2t}\sup_{x\in B^*}\sqrt{\mathbb{E}\langle G,x\rangle^2} = \ell^*(B^*) + \sqrt{2t}\,\mathrm{diam}(B^*,\ell_2^p)$$

where $G = \sum_{i=1}^N \lambda_i G_i \sim \mathcal{N}(0,I_p)$. Let $0<\epsilon<1/2$ and Λ_ϵ be an ϵ-net of $\mathcal{S}_2^{N-1}$ w.r.t. the ℓ_2^N-norm. Let $\boldsymbol{\lambda}^*\in\mathcal{S}_2^{N-1}$ be such that $\left\|\mathbb{G}^\top\boldsymbol{\lambda}^*\right\| = \sup\left(\left\|\mathbb{G}^\top\boldsymbol{\lambda}\right\| : \boldsymbol{\lambda}\in\mathcal{S}_2^{N-1}\right) := M$ and let $\boldsymbol{\lambda}_\epsilon^*\in\Lambda_\epsilon$ be such that $\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}_\epsilon^*\|_2\leq\epsilon$. We have

$$M = \left\|\mathbb{G}^\top\boldsymbol{\lambda}^*\right\| \leq \left\|\mathbb{G}^\top\boldsymbol{\lambda}_\epsilon^*\right\| + \left\|\mathbb{G}^\top(\boldsymbol{\lambda}^*-\boldsymbol{\lambda}_\epsilon^*)\right\| \leq \sup_{\boldsymbol{\lambda}_\epsilon\in\Lambda_\epsilon}\left\|\mathbb{G}^\top\boldsymbol{\lambda}_\epsilon\right\| + M\epsilon$$

and so $M \leq (1-\epsilon)^{-1}\sup_{\boldsymbol{\lambda}_\epsilon\in\Lambda_\epsilon}\left\|\mathbb{G}^\top\boldsymbol{\lambda}_\epsilon\right\|$. It follows from the bound above derived from Borell's inequality and an union bound that with probability at least $1-|\Lambda_\epsilon|\exp(-t)$, $M \leq (1-\epsilon)^{-1}\left(\ell^*(B^*)+\sqrt{2t}\,\mathrm{diam}(B^*,\ell_2^p)\right)$. It follows from a volume argument [35] that $|\Lambda_\epsilon|\leq(5/\epsilon)^N$ hence, for $\epsilon=1/2$ and $t=N(1+\log(10))$ with probability at least $1-\exp(-N)$,

$$\sup\left(\left\|\mathbb{G}^\top\boldsymbol{\lambda}\right\| : \boldsymbol{\lambda}\in\mathcal{S}_2^{N-1}\right) \leq 2\left(\ell^*(B^*) + \sqrt{2N(1+\log(10))}\,\mathrm{diam}(B^*,\ell_2^p)\right).$$

∎

Isomorphy and Restricted isomorphy properties. We use our budget of data, i.e. N, to insure some isomorphic or restricted isomorphic property of a matrix $\mathbb{X}_1 = \mathbb{G}^{(N\times p)}\Gamma^{1/2}$ for some semi-definite matrix Γ. Let B be the unit ball of some norm $\|\cdot\|$. Let κ_{RIP} be some absolute constant. We define the following family of fixed points: for all $\rho > 0$,

$$r(\rho) = \inf\left(r > 0 : \ell^*\left((\rho\Gamma^{1/2}B)\cap(rB_2^p)\right) \le \kappa_{RIP} r\sqrt{N}\right) \tag{11}$$

where $\ell^*\left((\rho\Gamma^{1/2}B)\cap(rB_2^p)\right)$ is the Gaussian mean width of $(\rho\Gamma^{1/2}B)\cap(rB_2^p)$. We see that for all $\rho > 0$, $r(\rho) = \rho R_N(\Gamma^{1/2}B)$ where

$$R_N(\Gamma^{1/2}B) = \inf\left(R > 0 : \ell^*\left((\Gamma^{1/2}B)\cap(RB_2^p)\right) \le \kappa_{RIP} R\sqrt{N}\right). \tag{12}$$

The relevance of the fixed point $R_N(\Gamma^{1/2}B)$ may be seen in the following result. This result follows from a line of research in empirical process theory on the quadratic process [20, 31, 30, 15, 3].

Theorem 4. *There are absolute constants $0 < \kappa_{RIP} < 1$ and c_0 such that for the fixed point $R_N(\Gamma^{1/2}B)$ defined in (12) the following holds. With probability at least $1 - \exp(-c_0 N)$,*

$$\frac{1}{2}\left\|\Gamma^{1/2}v\right\|_2^2 \le \frac{1}{N}\|\mathbb{X}_1 v\|_2^2 \le \frac{3}{2}\left\|\Gamma^{1/2}v\right\|_2^2, \tag{13}$$

for all $v \in \mathbb{R}^p$ such that $R_N(\Gamma^{1/2}B)\|v\| \le \left\|\Gamma^{1/2}v\right\|_2$.

Proof of Theorem 4. For all $v \in \mathbb{R}^p$, we set $f_v : x \in \mathbb{R}^p \mapsto \langle x, v\rangle$ and denote by $F := \{f_v : v \in \mathbb{R}^p, \|v\| \le R_N(\Gamma^{1/2}B)^{-1}$ and $\left\|\Gamma^{1/2}v\right\|_2 = 1\}$ when $R_N(\Gamma^{1/2}B) \neq 0$ and $F := \{f_v : v \in \mathbb{R}^p, \left\|\Gamma^{1/2}v\right\|_2 = 1\}$ when $R_N(\Gamma^{1/2}B) = 0$. We denote by $\mu = \mathcal{N}(0,\Gamma)$ the probability measure of the i.i.d. p-dimensional rows vectors $Y_1, \cdots, Y_N$ of $\mathbb{X}_1$. According to Theorem 5.5 in [15] there is an absolute constant $C \ge 1$ such that for all $t \ge 1$, with probability at least $1 - \exp(-t)$,

$$\sup_{f\in F}\left|\frac{1}{N}\sum_{i=1}^N f^2(Y_i) - \mathbb{E}f^2(Y_1)\right| \le C\left(\frac{\operatorname{diam}(F, L_2(\mu))\ell^*(F)}{\sqrt{N}} + \frac{\ell^*(F)^2}{N} + \operatorname{diam}(F, L_2(\mu))^2\left(\sqrt{\frac{t}{N}} + \frac{t}{N}\right)\right) \tag{14}$$

where $\operatorname{diam}(F, L_2(\mu)) := \sup(\|f\|_{L_2(\mu)} : f \in F)$. Note that we used the majorizing measure theorem [38] to get the equivalence between Talagrand's γ_2-functional and the Gaussian mean width and that for the Gaussian measure μ the Orlicz space $L_{\psi_2}(\mu)$ is equivalent to the Hilbert space $L_2(\mu)$. One can check that $\operatorname{diam}(F, L_2(\mu)) = 1$ and the result follows for $t = N/(64C^2)$ and $\kappa_{RIP} = 1/[8\sqrt{C}]$. ■

Let us give some insight about Theorem 4 in the case that is interesting to us that is for Γ of rank k and $B = B_2^p$. When $\kappa_{RIP}^2 N \ge k$ we have $\ell^*\left((\Gamma^{1/2}B_2^p)\cap(RB_2^p)\right) \le \ell^*\left(RB_2^p \cap \operatorname{range}(\Gamma)\right) \le R\sqrt{k} \le \kappa_{RIP}R\sqrt{N}$ for all R and so one has $R_N(\Gamma^{1/2}B_2^p) = 0$ which means that $\mathbb{X}_1$ satisfies an isomorphic property (see (13)) on the entire space $\operatorname{range}(\Gamma)$. We will use this property to study $\hat{\boldsymbol{\beta}}$ in the case where $k \le \kappa_{RIP}N$; we therefore state it in the following result.

Corollary 1. *There are absolute constants 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 $k \le \kappa_{RIP}N$ 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 $v \in \operatorname{range}(\Gamma)$, $(1/\sqrt{2})\left\|\Gamma^{1/2}v\right\|_2 \le \|\mathbb{X}_1 v\|_2/\sqrt{N} \le \sqrt{3/2}\left\|\Gamma^{1/2}v\right\|_2$.*

Corollary 1 can be proved using a straightforward epsilon net argument and there is no need for Theorem 4 in that case. However the other case '$k > N$' does not follow in general from such an argument. When $p > 4\kappa_{RIP}^2 N$, we necessarily have $R_N(\Gamma^{1/2}B) > 0$ because there exists a $R^* > 0$ such that $R^* B_2^p \cap \operatorname{range}(\Gamma) \subset \Gamma^{1/2}B_2^p \cap \operatorname{range}(\Gamma)$ (take R^* for instance to be the smallest non-zero singular value of Γ) and so for all $0 < R \le R^*$, we have $\ell^*\left((\Gamma^{1/2}B_2^p \cap \operatorname{range}(\Gamma))\cap(RB_2^p)\right) = \ell^*(RB_2^p \cap \operatorname{range}(\Gamma)) \ge R\sqrt{k}/2 > c_0 R\sqrt{N}$ and so one necessarily has $R_N(\Gamma^{1/2}B) \ge R^* > 0$. This is something expected because when $k > N$, $\mathbb{X}_1$ has a none trivial kernel and therefore $\mathbb{X}_1$ cannot be an isomorphy on the entire space $\operatorname{range}(\Gamma)$. However, Theorem 4 shows that $\mathbb{X}_1$ can still act as an isomorphy on a *restricted* subset of $\operatorname{range}(\Gamma)$ and given the homogeneity property of an isomorphic relation, this set is a cone; in the setup of Theorem 4 this cone is the one endowed by $B_2^p \cap R_N(\Gamma^{1/2}B)\Gamma^{-1/2}S_2^{p-1}$ since one can check that

$$\mathcal{C} := \operatorname{cone}\left(B_2^p \cap R_N(\Gamma^{1/2}B)\Gamma^{-1/2}S_2^{p-1}\right) = \left\{v \in \operatorname{range}(\Gamma) : R_N(\Gamma^{1/2}B_2^p)\|v\| \le \left\|\Gamma^{1/2}v\right\|_2\right\} \tag{15}$$

(where we denote $\operatorname{cone}(T) = \{\lambda x : x \in T, \lambda \ge 0\}$ for all $T \subset \mathbb{R}^p$). We therefore speak about a 'restricted isomorphy property' (RIP) in reminiscence to the classical RIP [10] used in Compressed sensing.

As we said, Theorem 4 is used below only for k-dimensional ellipsoids – that is when Γ is of rank k and $B = B_2^p$. For this type of convex body, there exists a sharp (up to absolute constants) computation of their Gaussian mean width intersected with an unit Euclidean ball: denote by $\sigma_1 \geq \cdots \geq \sigma_k$ the k non zero singular values of Γ associated with k eigenvectors $f_1, \ldots, f_k$, it follows from Proposition 2.5.1 from [38] that there exists some absolute constant $C_0 \geq 1$ such that

$$\ell^*\left(\Gamma^{1/2} B_2^p \cap R B_2^p\right) \leq C_0 \sqrt{\sum_{j=1}^{k} \min(\sigma_j, R^2)} \tag{16}$$

(this estimate is sharp up to absolute constants) and so $R_N(\Gamma^{1/2} B_2^p) \leq \inf\left(R : \sum_{j=1}^k \min(\sigma_j, R^2) \leq c_0 R^2 N\right)$ where $c_0 = (\kappa_{RIP}/C_0)^2$. We may now identify three regimes for the geometric parameter $R_N(\Gamma^{1/2} B_2^p)$:

A) when $k \leq c_0 N$ (which is up to absolute constants the case studied in Corollary 1);

B) when there exists $k_0 \in \{1, \cdots, \lfloor c_0 N \rfloor\}$ such that $\sum_{j=k_0}^k \sigma_j \leq (c_0 N - k_0 + 1)\sigma_{k_0}$ and in that case we define

$$k^{**} = \max\left(k_0 \in \{1, \cdots, \lfloor c_0 N \rfloor\} : \sum_{j=k_0}^{k} \sigma_j \leq (c_0 N - k_0 + 1)\sigma_{k_0}\right); \tag{17}$$

C) when for all $k_0 \in \{1, \cdots, \lfloor c_0 N \rfloor\}$ we have $\sum_{j=k_0}^k \sigma_j > (c_0 N - k_0 + 1)\sigma_{k_0}$.

Depending on these three cases, we find the following upper bound on the complexity fixed point

$$R_N(\Gamma^{1/2} B_2^p) \leq \inf\left(R > 0 : \sum_{j=1}^{k} \min(\sigma_j, R^2) \leq c_0^2 R^2 N\right) = \begin{cases} 0 & \text{in case A)} \\ \sqrt{\sigma_{k^{**}}} & \text{in case B)} \\ \sqrt{\frac{\mathrm{Tr}(\Gamma)}{c_0 N}} & \text{in case C).} \end{cases} \tag{18}$$

We note that k^{**} (when it exists) has a geometric interpretation: if $\sum_{j=k^{**}}^k \sigma_j \sim (c_0 N - k^{**} + 1)\sigma_k^{**}$ and $k^{**} \leq c_0 N/2$ then $k - k^{**}$ is the smallest dimension of an ellipsoid of the form $\Gamma_{k^{**}:k}^{-1/2} B_2^p$ having an Euclidean section of dimension of the order of N since we see that the Dvoretsky dimension of $\Gamma_{k^{**}:k}^{-1/2} B_2^p$ is $d_*(\Gamma_{k^{**}:k}^{-1/2} B_2^p) \sim \mathrm{Tr}(\Gamma_{k^{**}:k}) / \|\Gamma_{k^{**}:k}\|_{op} = \sum_{j=k^{**}}^k \sigma_j / \sigma_{k^{**}}$ and so $N \sim d_*(\Gamma_{k^{**}:k}^{-1/2} B_2^p)$ which is the regime of existence of Euclidean sections of dimension of the order of N for $\Gamma_{k^{**}:k}^{-1/2} B_2^p$.

We also note that in *case B)*, $\mathrm{span}(f_1, \ldots, f_{k^{**}}) \subset \mathcal{C}$ where $\mathcal{C}$ is the cone defined in (15) and in *case C)*, $\mathrm{span}(f_1, \ldots, f_r) \subset \mathcal{C}$ where r is the largest integer such that $c_0 N \sigma_r \geq \mathrm{Tr}(\Gamma)$. Hence, a way to understand the cone $\mathcal{C}$ is to look at it as a cone surrounding a space endowed by the top eigenvectors of Γ up to some dimension k^{**} or r depending on case B) or C). On that cone, $\mathbb{X}_1 = \mathbb{G}^{(N \times p)} \Gamma^{1/2}$ behaves like an isomorphy.

3 The approach: $\hat{\boldsymbol{\beta}}$ is a sum of a ridge estimator and an overfitting component.

Our proof strategy relies on a decomposition of $\hat{\boldsymbol{\beta}}$ that has been observed in several works. It is called the *self-induced regularization phenomenon* in [2]: the minimum ℓ_2-norm interpolant estimator $\hat{\boldsymbol{\beta}}$ can be written as $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{1:k} + \hat{\boldsymbol{\beta}}_{k+1:p}$ where $\hat{\boldsymbol{\beta}}_{1:k}$ is the projection of $\hat{\boldsymbol{\beta}}$ onto the space $V_{1:k}$ spanned by the top k eigenvectors of Σ and $\hat{\boldsymbol{\beta}}_{k+1:p}$ is the projection onto the space spanned by the last $p-k$ eigenvectors of Σ. Both $\hat{\boldsymbol{\beta}}_{1:k}$ and $\hat{\boldsymbol{\beta}}_{k+1:p}$ play very different role: $\hat{\boldsymbol{\beta}}_{1:k}$ is used for estimation whereas $\hat{\boldsymbol{\beta}}_{k+1:p}$ is used for overfitting. In the next result, a key feature of the ℓ_2-norm is used to write $\hat{\boldsymbol{\beta}}_{1:k}$ in a way which makes explicit its role as a ridge estimator of $\boldsymbol{\beta}^*_{1:k}$ and of $\hat{\boldsymbol{\beta}}_{k+1:p}$ as a minimum ℓ_2-norm estimator of $X_{k+1:p}\boldsymbol{\beta}^*_{k+1:p} + \boldsymbol{\xi}$. It can be seen as a formal statement of Eq.(39) in [2].

Proposition 3. *We have* $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{1:k} + \hat{\boldsymbol{\beta}}_{k+1:p}$ *where*

$$\hat{\boldsymbol{\beta}}_{1:k} \in \underset{\boldsymbol{\beta}_1 \in \mathbb{R}^p}{\mathrm{argmin}} \left(\left\| X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} (y - X_{1:k}\boldsymbol{\beta}_1) \right\|_2^2 + \|\boldsymbol{\beta}_1\|_2^2 \right) \tag{19}$$

and

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

Proof. Since $\hat{\boldsymbol{\beta}}$ minimizes the ℓ_2-norm over the set of vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $\mathbb{X}\boldsymbol{\beta} = y$ and since for all $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_2^2 = \|\boldsymbol{\beta}_{1:k}\|_2^2 + \|\boldsymbol{\beta}_{k+1:p}\|_2^2$ it is clear that $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{1:k} + \hat{\boldsymbol{\beta}}_{k+1:p}$ and

$$
\begin{aligned}
(\hat{\boldsymbol{\beta}}_{1:k}, \hat{\boldsymbol{\beta}}_{k+1:p}) &\in \underset{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)\in V_{1:k}\times V_{k+1:p}}{\operatorname{argmin}} \left(\|\boldsymbol{\beta}_1\|_2^2 + \|\boldsymbol{\beta}_2\|_2^2 : X_{1:k}\boldsymbol{\beta}_1 + X_{k+1:p}\boldsymbol{\beta}_2 = y \right) \\
&\in \underset{(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)\in \mathbb{R}^p\times\mathbb{R}^p}{\operatorname{argmin}} \left(\|\boldsymbol{\beta}_1\|_2^2 + \|\boldsymbol{\beta}_2\|_2^2 : X_{1:k}\boldsymbol{\beta}_1 + X_{k+1:p}\boldsymbol{\beta}_2 = y \right)
\end{aligned}
$$

The result follows by optimizing separately first in $\boldsymbol{\beta}_2$ and then in $\boldsymbol{\beta}_1$. ■

Proposition 3 is our starting point to the analysis of the prediction properties of $\hat{\boldsymbol{\beta}}$. It is also a key result to understand the roles played by $\hat{\boldsymbol{\beta}}_{1:k}$ and $\hat{\boldsymbol{\beta}}_{k+1:p}$.

Risk decomposition. We prove our main theorems below by using the risk decomposition that follows from the estimator decomposition $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{1:k} + \hat{\boldsymbol{\beta}}_{k+1:p}$:

$$
\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2^2 = \left\|\Sigma_{1:k}^{1/2}(\hat{\boldsymbol{\beta}}_{1:k} - \boldsymbol{\beta}_{1:k}^*)\right\|_2^2 + \left\|\Sigma_{k+1:p}^{1/2}(\hat{\boldsymbol{\beta}}_{k+1:p} - \boldsymbol{\beta}_{k+1:p}^*)\right\|_2^2. \tag{20}
$$

Estimation results will follow from high probability upper bounds on the two terms in the right-hand side of (20). As in [2], the *prediction component* $\hat{\boldsymbol{\beta}}_{1:k}$ is expected to be a good estimator of $\boldsymbol{\beta}_{1:k}^*$, that is of the k components of $\boldsymbol{\beta}^*$ in the basis of the top k eigenvectors of Σ. These k components are the most important ones to estimate because they are associated with the largest weights in the prediction norm $\|\Sigma^{1/2}\cdot\|_2$. We will see that $\hat{\boldsymbol{\beta}}_{1:k}$ estimates $\boldsymbol{\beta}_{1:k}^*$ as a ridge estimator. On the contrary, the *overfitting component* $\hat{\boldsymbol{\beta}}_{k+1:p}$ is not expected to be a good estimator of anything (and not of $\boldsymbol{\beta}_{k+1:p}^*$ in particular), it is here to (over)fit the data (and in particular the noise) and to make $\hat{\boldsymbol{\beta}}$ an interpolant estimator.

$\hat{\boldsymbol{\beta}}_{1:k}$: a ridge estimator of $\boldsymbol{\beta}_{1:k}^*$. One of the two components of our analysis based on the *self-induced regularization* property of $\hat{\boldsymbol{\beta}}$ as exhibited in Proposition 3 is to prove that $\hat{\boldsymbol{\beta}}_{1:k}$ is a good estimator of $\boldsymbol{\beta}_{1:k}^*$ and to that end, we rely on the following observation: if k is chosen so that $N \leq \kappa_{DM} d_*(\Sigma_{k+1:p}^{-1/2} B_2^p)$ then, it follows from DM's theorem (see Theorem 3), that with large probability $\left\|X_{k+1:p}^\top (X_{k+1:p}X_{k+1:p}^\top)^{-1}\cdot\right\|_2$ is isomorphic (i.e. equivalent up to absolute constants) to $(\mathrm{Tr}(\Sigma_{k+1:p}))^{-1/2}\|\cdot\|_2$ and so, according to Proposition 3, $\hat{\boldsymbol{\beta}}_{1:k}$ is expected to behave like

$$
\underset{\boldsymbol{\beta}_1\in V_{1:k}}{\operatorname{argmin}} \left(\|y - X_{1:k}\boldsymbol{\beta}_1\|_2^2 + \mathrm{Tr}(\Sigma_{k+1:p})\|\boldsymbol{\beta}_1\|_2^2 \right), \tag{21}
$$

which is a **ridge estimator** with regularization parameter $\mathrm{Tr}(\Sigma_{k+1:p})$ and with a random design matrix $X_{1:k}$. When $k \lesssim N$, with high probability, the spectrum of $X_{1:k}$ is up to absolute constants given by $\{\sqrt{N\sigma_1}, \dots, \sqrt{N\sigma_k}, 0, \cdots, 0\}$ (because, w.h.p. for all $\boldsymbol{\beta}_1 \in V_{1:k}, \|X_{1:k}\boldsymbol{\beta}_1\|_2 \sim \sqrt{N}\left\|\Sigma_{1:k}^{1/2}\boldsymbol{\beta}_1\right\|_2$ and for all $\boldsymbol{\beta}_2 \in V_{k+1:p}, X_{1:k}\boldsymbol{\beta}_2 = 0$). As a consequence the ridge estimator with regularization parameter $\mathrm{Tr}(\Sigma_{k+1:p})$ from (21) will be like an OLS only when $\sqrt{N\sigma_k} \gtrsim \mathrm{Tr}(\Sigma_{k+1:p})$. But since, we chose k such that $\sqrt{N\sigma_{k+1}} \lesssim \mathrm{Tr}(\Sigma_{k+1:p})$ (to apply the DM's theorem), $\hat{\boldsymbol{\beta}}_{1:k}$ will be an OLS estimator of $\boldsymbol{\beta}_{1:k}^*$ only when $N \sim \mathrm{Tr}(\Sigma_{k+1:p})/\sigma_{k+1} \sim d_*(\Sigma_{k+1:p}^{-1/2}B_2^p)$ unless there is a big gap between σ_k and σ_{k+1}. Otherwise, in general we only have $N \leq \kappa_{DM} d_*(\Sigma_{k+1:p}^{-1/2}B_2^p)$ and so (in general) $\hat{\boldsymbol{\beta}}_{1:k}$ behaves like a ridge estimator and the regularization term $\mathrm{Tr}(\Sigma_{k+1:p})\|\boldsymbol{\beta}_{1:k}\|_2^2$ has an impact on the estimation properties of $\boldsymbol{\beta}_{1:k}^*$ by $\hat{\boldsymbol{\beta}}_{1:k}$ (this differs from the comment from [2] because in [2] $k = k_b^*$ so the ridge regularization term in (21) has almost no effect because the regularization parameter $\mathrm{Tr}(\Sigma_{k+1:p})$ is smaller than the square of the smallest singular value of $X_{1:k}$ restricted to $V_{1:k}$). Our proof strategy is therefore to analyze $\hat{\boldsymbol{\beta}}_{1:k}$ as a ridge estimator with a random anisotropic design with the extra two difficulties: the operator $X_{k+1:p}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}$ appearing in (19) and the output y equals $\mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}$ and not $X_{1:k}\boldsymbol{\beta}_{1:k}^* + \boldsymbol{\xi}$. We will handle the first difficulty thanks to the DM's theorem which implies that $X_{k+1:p}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}$ is an 'isomorphic' operator and the second difficulty will be handled by looking at y as $y = X_{1:k}\boldsymbol{\beta}_{1:k}^* + (X_{k+1:p}\boldsymbol{\beta}_{k+1:p}^* + \boldsymbol{\xi})$; which means that $X_{k+1:p}\boldsymbol{\beta}_{k+1:p}^* + \boldsymbol{\xi}$ is considered as a noise. In fact, in all our analysis $X_{k+1:p}\boldsymbol{\beta}_{k+1:p}^* + \boldsymbol{\xi}$ is considered as a noise even from the viewpoint of $\hat{\boldsymbol{\beta}}_{k+1:p}$. In particular, $\hat{\boldsymbol{\beta}}$ does not aim at estimating $\boldsymbol{\beta}_{k+1:p}^*$ well even via $\hat{\boldsymbol{\beta}}_{k+1:p}$ as we are now explaining thanks to Proposition 3.

$\hat{\boldsymbol{\beta}}_{k+1:p}$**: a minimum ℓ_2-norm interpolant estimator of $X_{k+1:p}\boldsymbol{\beta}^*_{k+1:p}+\boldsymbol{\xi}$.** The *overfitting component* $\hat{\boldsymbol{\beta}}_{k+1:p}$ (see [2]), even though it appears as an estimator of $\boldsymbol{\beta}^*_{k+1:p}$ in the decomposition (20) is not expected to be a good estimator of $\boldsymbol{\beta}^*_{k+1:p}$. In fact, the remaining space $V_{k+1:p}$ is (automatically) used by $\hat{\boldsymbol{\beta}}$ to interpolate $X_{k+1:p}\boldsymbol{\beta}^*_{k+1:p}+\boldsymbol{\xi}$: we see from Proposition 3 that if $\hat{\boldsymbol{\beta}}_{1:k}$ was an exact estimator of $\boldsymbol{\beta}^*_1$ then $\hat{\boldsymbol{\beta}}_{k+1:p}$ would be equal to $X^\top_{k+1:p}(X_{k+1:p}X^\top_{k+1:p})^{-1}(X_{k+1:p}\boldsymbol{\beta}_{k+1:p}+\boldsymbol{\xi})$, which is the minimum ℓ_2-norm interpolant of $X_{k+1:p}\boldsymbol{\beta}^*_{k+1:p}+\boldsymbol{\xi}$ – and not just of $X_{k+1:p}\boldsymbol{\beta}^*_{k+1:p}$. It is therefore the place where overfitting takes place. This overfitting property of $\hat{\boldsymbol{\beta}}_{k+1:p}$ and so of $\hat{\boldsymbol{\beta}}$ has a price in terms of generalization which can be measured by the price to pay for 'bad estimation' of $\boldsymbol{\beta}^*_{k+1:p}$ by $\hat{\boldsymbol{\beta}}_{k+1:p}$ in the second term $\left\|\Sigma^{1/2}_{k+1:p}(\hat{\boldsymbol{\beta}}_{k+1:p}-\boldsymbol{\beta}^*_{k+1:p})\right\|_2$ of (20). However, this price is expected to be small because this estimation error is associated with the smallest weights in the prediction norm $\left\|\Sigma^{1/2}\cdot\right\|_2$. This will be indeed the case under an extra assumption on the spectrum of Σ, that $N\,\mathrm{Tr}(\Sigma^2_{k+1:p})=o(\mathrm{Tr}(\Sigma_{k+1:p})^2)$ which essentially says that the spectrum of $\Sigma_{k+1:p}$ needs to be well-spread, i.e. that it cannot be well approximated by a N-sparse vector. We will therefore use the estimation properties of $\hat{\boldsymbol{\beta}}_{1:k}$ and the bound

$$\left\|\Sigma^{1/2}_{k+1:p}(\hat{\boldsymbol{\beta}}_{k+1:p}-\boldsymbol{\beta}^*_{k+1:p})\right\|_2\leq\left\|\Sigma^{1/2}_{k+1:p}\hat{\boldsymbol{\beta}}_{k+1:p}\right\|_2+\left\|\Sigma^{1/2}_{k+1:p}\boldsymbol{\beta}^*_{k+1:p}\right\|_2 \tag{22}$$

to handle this second term. In particular, it is clear from (22) that we will not try to estimate $\boldsymbol{\beta}^*_{k+1:p}$ with $\hat{\boldsymbol{\beta}}_{k+1:p}$.

The key role of parameter k. Decomposition of the feature space $\mathbb{R}^p=V_{1:k}\oplus^\perp V_{k+1:p}$ is natural when $\boldsymbol{\beta}^*$ is estimated with respect to the prediction/weighted norm $\left\|\Sigma^{1/2}\cdot\right\|_2$ and when $\boldsymbol{\beta}^*$ has most of its energy in $V_{1:k}$. This features space decomposition is at the heart of the decomposition in (20) and is the one used in [39]. In the approach described above, parameter k is the parameter of a trade-off between estimation of $\boldsymbol{\beta}^*_{1:k}$ (by $\hat{\boldsymbol{\beta}}_{1:k}$) and the lack of estimation of $\boldsymbol{\beta}^*_{k+1:p}$ (by $\hat{\boldsymbol{\beta}}_{k+1:p}$) that permits overfitting. Both properties happen simultaneously inside $\hat{\boldsymbol{\beta}}$ and so k needs to be chosen so that the price for the estimation of $\boldsymbol{\beta}^*_1$ and the price of overfitting have equal magnitude. From this viewpoint, there is a priori no reason to take k smaller than N; in particular, $\left\|\Sigma^{1/2}_{k+1:p}\boldsymbol{\beta}^*_{k+1:p}\right\|_2$ will be part of the price of overfitting (since $\boldsymbol{\beta}^*_{k+1:p}$ is not estimated by $\hat{\boldsymbol{\beta}}_{k+1:p}$) and so having k large will be beneficial for this term – in particular, when the signal is strong in $V_{k+1:p}$. Hence, we will explore the properties of $\hat{\boldsymbol{\beta}}$ beyond the case $k\lesssim N$. This is different from the previous works on the benign overfitting phenomenon of $\hat{\boldsymbol{\beta}}$ [1, 39] which do not study this case. That is the reason why there are two subsections in the next section covering the two cases '$k\lesssim N$' and '$k\gtrsim N$'. From a stochastic viewpoint, the two regimes are different because in the first case $X_{1:k}$ behaves like an isomorphism on the entire space $V_{1:k}$ whereas, in the second case, $X_{1:k}$ cannot be an isomorphism on the entire space $V_{1:k}$ anymore but it is an isomorphism restricted to a cone as proved in Theorem 4; a geometric property we will use to prove Theorem 6, that is for the case $k\gtrsim N$.

4 Main results

In this section, we provide two estimation results of $\boldsymbol{\beta}^*$ by $\hat{\boldsymbol{\beta}}$ with respect to $\left\|\Sigma^{1/2}\cdot\right\|$ depending on the value of the parameter k driving the space decomposition $\mathbb{R}^p=V_{1:k}\oplus^\perp V_{k+1:p}$ with respect to the sample size N. We start with the case $k\lesssim N$ and then we state the result in the case $k\gtrsim N$. We show that these two results also hold for any features space decomposition of the form $\mathbb{R}^p=V_J\oplus^\perp V_{J^c}$ in Section 4.3. Given that the minimum ℓ_2-norm interpolant estimator does not depend on any parameter nor on any features space decomposition this shows that it can find the 'best features space decomposition' by itself. In a final section, we obtain a lower bound on the expected prediction risk which matches the upper bound for the choice $J=\{1,\dots,k^*_b\}$ where k^*_b has been introduced in [1] (see (4)) for some constant b. This shows that choices of features space decomposition as well as the introduction of k^*_b from [1, 39] are the right choices to make regarding the minimum ℓ_2-norm interpolant estimator in the Gaussian linear model.

4.1 The small dimensional case $k\lesssim N$

Theorem 5. *[**the** $k\lesssim N$ **case.**] There are absolute constants c_0, c_1 and $C_0=9216$ such that the following holds. We assume that $N\geq 5\log p$ and that there exists $k\leq\kappa_{iso}N$ such that $N\leq\kappa_{DM}d_*(\Sigma^{-1/2}_{k+1:p}B^p_2)$, then the following holds for all such k's. We define*

$$J_1:=\left\{j\in[k]:\sigma_j\geq\frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right\},\quad J_2:=\left\{j\in[k]:\sigma_j<\frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right\}$$

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

$$D_{1,thres}^{-1/2} := \operatorname{diag}\left(\left(\sigma_1 \vee \frac{\operatorname{Tr}(\Sigma_{k+1:p})}{N}\right)^{-1/2}, \dots, \left(\sigma_k \vee \frac{\operatorname{Tr}(\Sigma_{k+1:p})}{N}\right)^{-1/2}, 0, \dots, 0\right).$$

With probability at least $1 - c_0 \exp\left(-c_1\left(|J_1| + N\left(\sum_{j\in J_2}\sigma_j\right)/\left(\operatorname{Tr}(\Sigma_{k+1:p})\right)\right)\right)$,

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \leq \square + \sigma_\xi \frac{17\sqrt{N\operatorname{Tr}(\Sigma_{k+1:p}^2)}}{\operatorname{Tr}(\Sigma_{k+1:p})} + \left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_{k+1:p}^*\right\|_2$$

where

i) if $\sigma_1 N < \operatorname{Tr}(\Sigma_{k+1:p})$ *then*

$$\square = C_0^2 \max\left\{\sigma_\xi\sqrt{\frac{\operatorname{Tr}(\Sigma_{1:k})}{\operatorname{Tr}(\Sigma_{k+1:p})}}, \sqrt{\frac{N\sigma_1}{\operatorname{Tr}(\Sigma_{k+1:p})}}\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_{k+1:p}^*\right\|_2, \|\boldsymbol{\beta}_{1:k}^*\|_2\sqrt{\frac{\operatorname{Tr}(\Sigma_{k+1:p})}{N}}\right\} \tag{23}$$

ii) if $\sigma_1 N \geq \operatorname{Tr}(\Sigma_{k+1:p})$ *then*

$$\square = C_0^2 \max\left\{\sigma_\xi\sqrt{\frac{|J_1|}{N}}, \sigma_\xi\sqrt{\frac{\sum_{j\in J_2}\sigma_j}{\operatorname{Tr}(\Sigma_{k+1:p})}}, \left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_{k+1:p}^*\right\|_2, \left\|\Sigma_{1,thres}^{-1/2}\boldsymbol{\beta}_{1:k}^*\right\|_2\frac{\operatorname{Tr}(\Sigma_{k+1:p})}{N}\right\}. \tag{24}$$

Several comments on Theorem 5 are in order and we list some of them below.

Effective dimension and the Dvoretsky dimension of an ellipsoid. We recall that $\operatorname{Tr}(\Sigma_{k+1:p})/(4\sigma_{k+1}) \leq d_*(\Sigma_{k+1:p}^{-1/2}B_2^p) \leq \operatorname{Tr}(\Sigma_{k+1:p})/\sigma_{k+1}$ so that the choice of k in Theorem 5 is to take it so that N is smaller than the effective rank $r_k(\Sigma)$ as in [1, 39] (see (3)). Theorem 5 holds for all such k's so that the self-induced regularization phenomenon holds true for every space decomposition $\mathbb{R}^p = V_{1:k} \oplus^\perp V_{k+1:p}$ simultaneously such that this condition holds. One can therefore optimize either the rate or the deviation parameter over this parameter k. This is an interesting adaptive property of the minimum ℓ_2-norm interpolant estimator that will be commented later.

Price of overfitting. The price paid by $\hat{\boldsymbol{\beta}}$ for overfitting on the training set in terms of generalization capacity is measured by the 'estimation' error $\left\|\Sigma_{k+1:p}^{1/2}(\hat{\boldsymbol{\beta}}_{k+1:p} - \boldsymbol{\beta}_{k+1:p}^*)\right\|_2$. Because, unlike none interpolant and reasonable estimators that would try either to estimate $\boldsymbol{\beta}_{k+1:p}^*$ or to take values 0 on $V_{k+1:p}$ (such as a hard thresholding estimator which avoids an unecessary variance term on $V_{k+1:p}$), $\hat{\boldsymbol{\beta}}_{k+1:p}$ is in fact interpolating the data (if we had $\hat{\boldsymbol{\beta}}_{1:k} = \boldsymbol{\beta}_{1:k}^*$ then we would still have $X_{k+1:p}\hat{\boldsymbol{\beta}}_{k+1:p} = X_{k+1:p}\boldsymbol{\beta}_{k+1:p}^* + \boldsymbol{\xi}$). One may find an upper bound on this rate in Proposition 5 below. In the final estimation rate obtained in Theorem 5 only remains the sum

$$\sigma_\xi \frac{17\sqrt{N\operatorname{Tr}(\Sigma_{k+1:p}^2)}}{\operatorname{Tr}(\Sigma_{k+1:p})} + \left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2 \tag{25}$$

which is what we call the price for overfitting from the convergence rate in Theorem 5.

Persisting regularization in the ridge estimator $\hat{\boldsymbol{\beta}}_{1:k}$. Another way to look at the self-induced regularization property of $\hat{\boldsymbol{\beta}}$ from [2] is to look at it as a persisting regularization property of $\hat{\boldsymbol{\beta}}$ because $\hat{\boldsymbol{\beta}}_{1:k}$ is a ridge estimator. Indeed, $\hat{\boldsymbol{\beta}}$ is the limit of ridge estimators when the regularization parameter tends to zero, hence, since $\hat{\boldsymbol{\beta}}_{1:k}$ is also a ridge estimator, $\hat{\boldsymbol{\beta}}$ still keeps a part of the space $\mathbb{R}^p$ onto which it performs a ridge regularization. In particular, there are two regimes in Theorem 5 ('$\sigma_1 N \leq \operatorname{Tr}(\Sigma_{k+1:p})$' or '$\sigma_1 N \geq \operatorname{Tr}(\Sigma_{k+1:p})$') because there are two regimes for a ridge estimator: either the regularization parameter $\operatorname{Tr}(\Sigma_{k+1:p})$ in the ridge estimator (21) is larger than the square of the largest singular value of $X_{1:k}$, i.e. $\sigma_1 N$ (with high probability and up to absolute constants) or not. In the first case '$\sigma_1 N \leq \operatorname{Tr}(\Sigma_{k+1:p})$', the regularization parameter in (21) is so large that it is mainly the regularization term '$\operatorname{Tr}(\Sigma_{k+1:p})\|\boldsymbol{\beta}_{1:k}\|_2^2$' which is minimized. So for $\hat{\boldsymbol{\beta}}_{1:k}$ to be a good estimator of $\boldsymbol{\beta}_{1:k}^*$, we require $\boldsymbol{\beta}_{1:k}^*$ to be close to zero, that is why we pay a price proportional to $\|\boldsymbol{\beta}_{1:k}^*\|_2$ in that case in Theorem 5. In particular, BO will require in that case that $\|\boldsymbol{\beta}^*_{1:k}\|_2^2 = o\left(\mathrm{Tr}(\Sigma_{k+1:p})/N\right)$. However, this case should be looked at a pathological one since it is a case where the regularization parameter of a ridge estimator is too large (larger than the square of the largest singular value of the design matrix) so that the data fitting term $\|y - X_{1:k}\boldsymbol{\beta}_{1:k}\|_2^2$ does not play an important role in the definition (21) of the ridge estimator compared with the regularization term. In particular, sources of generalization errors are due to a bad estimation of $\boldsymbol{\beta}^*_{1:k}$ (when $\boldsymbol{\beta}^*_{1:k}$ is not close to 0) as well as overfitting. Since our aim is to identify when overfitting is benign, this case adds some extra difficulties which are not at the heart of the purpose and so we look at it as pathological even though it is possible to obtain a convergence rates also in that case from Theorem 5.

In the general case where ridge regularization parameter is not too large, i.e. the second case '$\sigma_1 N \geq \mathrm{Tr}(\Sigma_{k+1:p})$', then the regularization term appears in the rate through the two sets J_1 and J_2 as well as in the thresholded matrix $\Sigma_{1,thres}^{-1/2}$. This is the interesting case, because it shows that benign overfitting happens when $\sigma_1 N \geq \mathrm{Tr}(\Sigma_{k+1:p})$, $N\,\mathrm{Tr}(\Sigma_{k+1:p}^2) = o\left(\mathrm{Tr}^2(\Sigma_{k+1:p})\right)$,

$$|J_1| = o(N), \sum_{j\in J_2}\sigma_j = o\left(\mathrm{Tr}(\Sigma_{k+1:p})\right), \left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}^*_{k+1:p}\right\|_2 = o(1) \text{ and } \left\|\Sigma_{1,thres}^{-1/2}\boldsymbol{\beta}^*_{1:k}\right\|_2 \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N} = o(1). \tag{26}$$

In particular, situation where BO happens depends on both the behavior of Σ as well as $\boldsymbol{\beta}^*$. Compare with Theorem 1 from [39] (recalled in Theorem 2), we observe that the convergence rate from Theorem 5 is better because it fully exploits the thresholding effect of the spectrum of $X_{1:k}$ by the ridge regularization; indeed, we have

$$\sqrt{\frac{|J_1|}{N}} + \sqrt{\frac{\sum_{j\in J_2}\sigma_j}{\mathrm{Tr}(\Sigma_{k+1:p})}} \leq \sqrt{\frac{k}{N}} \text{ and } \left\|\Sigma_{1,thres}^{-1/2}\boldsymbol{\beta}^*_{1:k}\right\|_2 \leq \left\|\Sigma_{1:k}^{-1/2}\boldsymbol{\beta}^*_{1:k}\right\|_2$$

and the other terms are the same in Theorem 1 and Theorem 5 so that Theorem 5 indeed improves the upper bound result from [39] for any $k \lesssim N$ (it also improves the deviation rate, see below). We may also check that the two results coincide when $J_2 = \emptyset$; which is what happens for the choice of $k = k_b^*$ for $b = 1$. However, even in that case our proof is different from the one in [39] since it is based on the self-induced regularization property of $\hat{\boldsymbol{\beta}}$ and not a bias/variance trade-off. It also improves the deviation parameter from constant to exponentially small in k_b^*: $1 - c_0\exp(-c_1 k_b^*)$ is the deviation estimate expected for an OLS in the Gaussian linear model over $\mathbb{R}^{k_b^*}$.

The benign overfitting phenomenom happens with large probability. The generalization bounds obtained in Theorem 5 hold with exponentially large probability. This shows that when Σ and $\boldsymbol{\beta}^*$ are so that benign overfitting holds then it is very likely to see it happening on data that is to see the good generalization property of $\hat{\boldsymbol{\beta}}$ on a test set (even though it interpolates the training data). In Theorem 1 or Theorem 2 the rates are multiplied by the deviation parameter t so that they essentially hold with constant probability – unless one is willing to increase the rates – and does not explain why benign overfitting happens very often in practice.

4.2 The large dimensional case $k \gtrsim N$.

Theorem 6. *[**the** $k \gtrsim N$ **case.**] There are absolute constants c_0, c_1, c_2, c_3 and C_0 such that the following holds. We assume that there exists $k \in [p]$ such that $N \leq \kappa_{DM} d_*(\Sigma_{k+1:p}^{-1/2} B_2^p)$ and $R_N(\Sigma_{1:k}^{1/2} B_2^p) \leq \sqrt{\mathrm{Tr}(\Sigma_{k+1:p})/N}$. Then the following result holds for all such k's: with probability at least $1 - c_0\exp\left(-c_1\left(|J_1| + N\left(\sum_{j\in J_2}\sigma_j\right)/\left(\mathrm{Tr}(\Sigma_{k+1:p})\right)\right)\right)$,*

$$\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \leq \square + \sigma_\xi \frac{c_2\sqrt{N\,\mathrm{Tr}(\Sigma_{k+1:p}^2)}}{\mathrm{Tr}(\Sigma_{k+1:p})} + c_3\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}^*_{k+1:p}\right\|_2$$

where J_1 and J_2 are defined in Theorem 5 and

i) if $\sigma_1 N < \mathrm{Tr}(\Sigma_{k+1:p})$ and $\mathrm{Tr}(\Sigma_{1:k}) \leq N\sigma_1$ then $\square$ is defined in (23)

ii) if $\sigma_1 N \geq \mathrm{Tr}(\Sigma_{k+1:p})$ and $\sum_{j\in J_2}\sigma_j \leq \mathrm{Tr}(\Sigma_{k+1:p})(1 - |J_1|/N)$ then $\square$ is defined in (24).

As mentioned previously, Theorem 6 is the main result if one wants to lower the price of overfitting by considering features space decomposition $\mathbb{R}^p = V_{1:k} \oplus^\perp V_{k+1:p}$ beyond the case $k \lesssim N$, which was the only case studied in the literature so far to our knowledge. In particular, we can now identify situations where benign overfitting happens thanks to Theorem 6 but before that let us comment on the assumptions in Theorem 6.

Assumptions in Theorem 6. Both assumptions '$N \leq \kappa_{DM} d_*(\Sigma_{k+1:p}^{-1/2} B_2^p)$' and '$R_N(\Sigma_{1:k}^{1/2} B_2^p) \leq \sqrt{\operatorname{Tr}(\Sigma_{k+1:p})/N}$' are of geometric nature: the first one involved the Dvoretsky dimension of the ellipsoid $\Sigma_{k+1:p}^{-1/2} B_2^p$ and the second one is used to define the cone (15) onto which $X_{1:k}$ is an isomorphy. Following (18), The latter condition is implied by the slightly stronger condition:

$$\begin{cases} \operatorname{Tr}(\Sigma_{k+1:p}) \geq N\sigma_{k^{**}} & \text{when } k^{**} = \max\left(k_0 \in \{1, \cdots, \lfloor c_0 N \rfloor\} : \sum_{j=k_0}^{k} \sigma_j \leq (c_0 N - k_0 + 1)\sigma_{k_0}\right) \text{ exists} \\ c_0 \operatorname{Tr}(\Sigma_{k+1:p}) \geq \operatorname{Tr}(\Sigma_{1:k}) & \text{when for all } k_0 \in \{1, \cdots, \lfloor c_0 N \rfloor\}, \sum_{j=k_0}^{k} \sigma_j > (c_0 N - k_0 + 1)\sigma_{k_0} \end{cases}$$

In case *ii)* of Theorem 6 which is the most interesting case to us, the extra condition '$\sum_{j \in J_2} \sigma_j \leq \operatorname{Tr}(\Sigma_{k+1:p})(1 - |J_1|/N)$' (compared with the case $k \lesssim N$) is a very weak one given that the term $\sqrt{\sum_{j \in J_2} \sigma_j / \operatorname{Tr}(\Sigma_{k+1:p})}$ appears in $\square$ and is therefore asked to tend to 0 to see the BO happening.

Comparison with the previous results. The reader may be interested in the benign overfitting phenomenon when $k \gtrsim N$, which seems to be contradict with the lower bound of Theorem 4 in [1] (recalled in Theorem 1), that is for k_b^* recalled in (4) and some absolute constants $b, c > 1$:

$$\mathbb{E}\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2 \geqslant \frac{\sigma_\xi}{c}\left(\sqrt{\frac{k_b^*}{N}} + \frac{\sqrt{N \operatorname{Tr}(\Sigma_{k_b^*+1:p}^2)}}{\operatorname{Tr}(\Sigma_{k_b^*+1:p})}\right).$$

However, this is not the case. This is because our choice of k is different from theirs. We choose k such that $\operatorname{Tr}(\Sigma_{k+1:p}) \gtrsim N \|\Sigma_{k+1:p}\|_{op}$ and $R_N(\Sigma_{1:k}^{1/2} B_2^p) \lesssim \sqrt{\operatorname{Tr}(\Sigma_{k+1:p})/N}$. It is therefore (not yet) optimized and not taken equal to k_b^* or $k^{**} + 1$ – it is a free parameter only asked to satisfy the conditions of Theorem 6. In fact, $k^{**} + 1$ (where k^{**} is defined in Equation (18)) plays a role similar to k_b^*. Indeed, by definition, $k^{**} \in \{1, \cdots, c_0 N\}$ satisfies

$$\frac{\operatorname{Tr}(\Sigma_{k^{**}+1:p})}{\|\Sigma_{k^{**}+1:p}\|_{op}} = \frac{\operatorname{Tr}(\Sigma_{k^{**}+1:k})}{\|\Sigma_{k^{**}+1:p}\|_{op}} + \frac{\operatorname{Tr}(\Sigma_{k+1:p})}{\|\Sigma_{k^{**}+1:p}\|_{op}} > c_0 N - k^{**} + 1 + \frac{\operatorname{Tr}(\Sigma_{k+1:p})}{\|\Sigma_{k^{**}+1:k}\|_{op}} > N$$

and so $k_b^* \leq k^{**} + 1$ for $b = 1$. Therefore, the k_b^* defined in [1] plays a similar role to the geometric parameter $k^{**} + 1$, which is also assumed to be smaller than $c_0 N + 1$ in one case considered in Theorem 6.

As remarked previously in [1], a necessary condition for overfitting is the existence of a $k_b^* \lesssim N$ such that $\operatorname{Tr}(\Sigma_{k_b^*+1:p}) \gtrsim N\sigma_{k_b^*+1}$. However, the existence of such a k_b^* does not necessarily forces us to take it equal to k in the decomposition $\mathbb{R}^p = V_{1:k} \oplus^{\perp} V_{k+1:p}$. A priori, one could find a better features space decomposition $\mathbb{R}^p = V_{1:k} \oplus^{\perp} V_{k+1:p}$ that in particular lower the price for overfitting. Since $\left\|\Sigma^{1/2}\boldsymbol{\beta}_{1:k}^*\right\|_2$ is part of this price (25) and that it decreases when k increases, one may look for larger k's, and Theorem 6 shows that it is possible to take it larger than an order of N.

Benign overfitting in the regime $k \gtrsim N$. It follows from Theorem 6 that BO happens for choices of k larger than N. Such a situation holds when the spectrum of Σ is as follows: let $k_0 < N < k$, $a < b < c$, $0 < \alpha < 1$ and

$$\sigma_j = a, \forall j = 1, \ldots, k_0; \quad \sigma_j = \frac{b}{j^\alpha}, \forall j = k_0 + 1, \ldots, k \text{ and } \sigma_j = c, \forall j \geq k + 1. \tag{27}$$

It follows from Theorem 6 that $\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2$ tends to zero when $N, p \to +\infty$ when $c \sim bN/[k_0^\alpha p]$, $(k/k_0)^\alpha < p/N$, $a > b/k_0^\alpha$, $k_0 = o(N)$, $k^{1-\alpha}k_0^\alpha = o(1)$ and $N = o(p - k)$ (which is the well-known 'over-parametrized' regime needed for BO) and

$$\left\|\boldsymbol{\beta}_{1:k_0}^*\right\|_2 = o\left(\frac{\sqrt{a}k_0^\alpha}{b}\right), \quad \left\|\boldsymbol{\beta}_{k_0+1:k}^*\right\|_2 = o\left(\sqrt{\frac{k_0^\alpha}{b}}\right) \text{ and } \left\|\boldsymbol{\beta}_{k+1:p}^*\right\|_2 = o\left(\sqrt{\frac{k_0^\alpha p}{bN}}\right) \tag{28}$$

where we used that $\left\|\Sigma_{1,thres}^{-1/2}\boldsymbol{\beta}_{1:k}^*\right\|_2^2 \leqslant \left\|\boldsymbol{\beta}_{J_1}^*\right\|_2^2/a + (N/\operatorname{Tr}(\Sigma_{k+1:p}))\left\|\boldsymbol{\beta}_{J_2}^*\right\|_2^2$ when $a > \operatorname{Tr}(\Sigma_{k+1:p})/N$ and $b < \operatorname{Tr}(\Sigma_{k+1:p})/N$ so that $J_1 = \{1, \ldots, k_0\}$ and $J_2 = \{k_0 + 1, \ldots, k\}$. Note that under these assumptions, k^{**} (defined in (17)) exists and is of the order of N so that one can apply Theorem 6 in that case. In general, we do not want to make any assumption on $\boldsymbol{\beta}_{1:k_0}^*$, which is where we expect most of the information on the signal $\boldsymbol{\beta}^*$ to lie, hence, for the k_0-dimensional vector $\boldsymbol{\beta}_{1:k_0}^*$, we expect $\left\|\boldsymbol{\beta}_{1:k_0}^*\right\|_2$ to be of the order of $\sqrt{k_0}$. This will be the case when $\alpha > 1/2$ or when $a >> b^2 k_0^{1/2-\alpha}$. The remarkable point here is that BO happens even for values of k larger than N.

Other situations of BO in the case $k \gtrsim N$ may be found with various speed of decay on the three regimes introduced in (27).

A priori, given the bound obtained in Theorem 1, 2, 5 or 6, the benign overfitting phenomenon depends on both Σ and $\boldsymbol{\beta}^*$. Ideal situations for BO are when for some $k \in [p]$ (not necessarily smaller than N), we have

(PO1) $\boldsymbol{\beta}^*$ is mostly supported on $V_{1:k}$ so that $\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_{k+1:p}^*\right\|_2 = o(1)$,

(PO2) the spectrum of $\Sigma_{k+1:p}$, denoted by $\mathrm{spec}(\Sigma_{k+1:p})$, has to be such that its ℓ_2/ℓ_1-ratio is negligible in front of $1/\sqrt{N}$. This type of condition means that $\mathrm{spec}(\Sigma_{k+1:p})$ is not compressible: it cannot be well approximated by a N-sparse vector. In other word, $\mathrm{spec}(\Sigma_{k+1:p})$ is asked to be a well-spread vector.

(PE1) the cardinality of J_1 (the set of eigenvalues of Σ larger than $\mathrm{Tr}(\Sigma_{k+1:p})/N$) is negligible in front of N and the remaining of the spectrum of $\Sigma_{1:k}$ is such that $\sum_{j\in J_2}\sigma_j = o(\mathrm{Tr}(\Sigma_{k+1:p}))$,

(PE2) top eigenvalues of Σ are large so that $\sigma_1 N \geq \mathrm{Tr}(\Sigma_{k+1:p})$ and $\left\|\Sigma_{1,thres}^{-1/2}\boldsymbol{\beta}_{1:k}^*\right\|_2$ is negligible in front of $N/\mathrm{Tr}(\Sigma_{k+1:p})$.

The price for not estimating $\boldsymbol{\beta}_{k+1:p}^*$ is part of the price for overfitting as well as the bias term in the estimation of $\boldsymbol{\beta}_{1:k}^*$ by the 'ridge' estimator $\hat{\boldsymbol{\beta}}_{1:k}$; the term $\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_{k+1:p}^*\right\|_2$ appearing in both components of the risk decomposition (20). One way to lower it is to ask for a condition like *(PO1)*. Together with *(PO2)*, they are the two conditions for BO that come from the price for overfitting. The other *(PE)* conditions come from the estimation of $\boldsymbol{\beta}_{1:k}^*$ by $\hat{\boldsymbol{\beta}}_{1:k}$: *(PE1)* is a condition on its variance term and *(PE2)* as well as *(PO1)* are the conditions for BO coming from the control of its bias term.

4.3 On the choice of k and the self-adaptive property

The choice of k in [39] is limited by the constraint $k \lesssim N$ (see Theorem 2) and so the recommendation from [39] is to take k for which $\rho_k := \mathrm{Tr}(\Sigma_{k+1:p})/[N\left\|\Sigma_{k+1:p}\right\|_{op}]$ is of order of a constant and if such k doesn't exist, one should take the smallest k for which ρ_k is larger than a constant. However, this recommendation does not take into account all the quantities depending on the signal even though $\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_{k+1:p}^*\right\|_2$ appears explicitly in the upper bound from [39] (see Theorem 2). A consequence of Theorem 5 and Theorem 6 is that there is no constraint to choose $k \lesssim N$ and so all features space splitting $\mathbb{R}^p = V_{1:k} \oplus^{\perp} V_{k+1:p}$ are allowed even for $k \gtrsim N$ as long as the two geometrical conditions of Theorem 6 hold. In particular, one can chose any k satisfying the geometric assumptions of these theorems and optimize the upper bound, including signal dependent terms such as $\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_{k+1:p}^*\right\|_2$. The best choice of k is a priori making a trade-off between three terms coming from the estimation of $\boldsymbol{\beta}_{J_1}^*$ by the OLS part of $\hat{\boldsymbol{\beta}}_{1:k}$, from the estimation of $\boldsymbol{\beta}_{J_2}^*$ by the 'over-regularized' part of $\hat{\boldsymbol{\beta}}_{1:k}$ (where J_1 and J_2 have been introduced in Theorem 5) and the none-estimation of $\boldsymbol{\beta}_{k+1:p}^*$ by the overfitting component $\hat{\boldsymbol{\beta}}_{k+1:p}$. On top of that this trade-off is particularly subtle since the ridge regularization parameter of $\hat{\boldsymbol{\beta}}_{1:k}$ is of the order of $\mathrm{Tr}(\Sigma_{k+1:p})$ and therefore depends on k (as well as the spectrum of Σ). Fortunately $\hat{\boldsymbol{\beta}}$ does this trade-off by itself.

It follows from the analysis of BO from the last subsection that cases where BO happens depend on the coordinates of $\boldsymbol{\beta}^*$ in the basis of eigenvectors of Σ and that the best cases are obtained when $\boldsymbol{\beta}^*$ is sparse in this basis with all its energy supported on the first top k eigenvectors. However, such a configuration may not be a typical situation for real-world data. Fortunately, the decomposition of the features space $\mathbb{R}^p$ as $V_{1:k} \oplus^{\perp} V_{k+1:p}$ is arbitrary and can in fact be more adapted to $\boldsymbol{\beta}^*$. It is indeed possible to obtain all the results (Proposition 3 and Theorems 5 and 6) for any decomposition $\mathbb{R}^p = V_J \oplus^{\perp} V_{J^c}$ where $J \subset [p]$, $J^c = [p]\backslash J$ and $V_J = \mathrm{span}(f_j, j \in J)$ as in [21]. The key observation here is that Proposition 3 still holds for this decomposition: one can still write $\hat{\boldsymbol{\beta}}$ as a sum $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_J + \hat{\boldsymbol{\beta}}_{J^c}$ where

$$\hat{\boldsymbol{\beta}}_J \in \underset{\boldsymbol{\beta}_1\in\mathbb{R}^p}{\mathrm{argmin}}\left(\left\|X_{J^c}^\top(X_{J^c}X_{J^c}^\top)^{-1}(y - X_J\boldsymbol{\beta}_1)\right\|_2^2 + \left\|\boldsymbol{\beta}_1\right\|_2^2\right) \text{ and } \hat{\boldsymbol{\beta}}_{J^c} = X_{J^c}^\top(X_{J^c}X_{J^c}^\top)^{-1}(y - X_J\hat{\boldsymbol{\beta}}_J). \tag{29}$$

where $X_J = \mathbb{G}^{(N\times p)}\Sigma_J^{1/2}$ and $X_{J^c} = \mathbb{G}^{(N\times p)}\Sigma_{J^c}^{1/2}$ (so that $X_J + X_{J^c} = \mathbb{X}$) and $\Sigma_J = UD_JU^\top$ and $\Sigma_{J^c} = UD_{J^c}U^\top$ where $D_J = \mathrm{diag}(\sigma_1 I(j\in J), \dots, \sigma_p I(j \in J))$ and $D_{J^c} = \mathrm{diag}(\sigma_1 I(j \in J^c), \dots, \sigma_p I(j\in J^c))$.

As a consequence, Theorem 5 and Theorem 6 still holds if one replaces the subsets $\{1, \cdots, k\}$ and $\{k+1, \cdots, p\}$ respectively by J and J^c.

Theorem 7. *[**features space decomposition** $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$.] There are absolute constants c_0, c_1, c_2, c_3 and C_0 such that the following holds. Let $J \sqcup J^c$ be a partition of $[p]$. We assume that $N \leq \kappa_{DM} d_*(\Sigma_{J^c}^{-1/2} B_2^p)$ and $R_N(\Sigma_J^{1/2} B_2^p) \leq \sqrt{\mathrm{Tr}(\Sigma_{J^c})/N}$ (note that $R_N(\Sigma_J^{1/2} B_2^p) = 0$ when $|J| \leq \kappa_{RIP} N$ so that this conditions holds trivially in that case). We define*

$$J_1 := \left\{ j \in J : \sigma_j \geq \frac{\mathrm{Tr}(\Sigma_{J^c})}{N} \right\}, \quad J_2 := \left\{ j \in J : \sigma_j < \frac{\mathrm{Tr}(\Sigma_{J^c})}{N} \right\}$$

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

$$D_{J,thres}^{-1/2} := \mathrm{diag}\left(\left(\sigma_1 \vee \frac{\mathrm{Tr}(\Sigma_{J^c})}{N} \right)^{-1/2} I(1 \in J), \ldots, \left(\sigma_p \vee \frac{\mathrm{Tr}(\Sigma_{J^c})}{N} \right)^{-1/2} I(p \in J) \right).$$

Then the following result holds for all such space decomposition $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$: with probability at least $1 - c_0 \exp\left(-c_1 \left(|J_1| + N \left(\sum_{j \in J_2} \sigma_j \right) / \left(\mathrm{Tr}(\Sigma_{J^c}) \right) \right) \right)$,

$$\left\| \Sigma^{1/2} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \right\|_2 \leq \square(J) + \sigma_\xi \frac{c_2 \sqrt{N \, \mathrm{Tr}(\Sigma_{J^c}^2)}}{\mathrm{Tr}(\Sigma_{J^c})} + c_3 \left\| \Sigma_{J^c}^{1/2} \boldsymbol{\beta}_{J^c}^* \right\|_2$$

where

- *i) $\square(J)$ is defined by $\square$ in (23) where $\{1 : k\}$ (resp. $\{k+1 : p\}$) is replaced by J (resp. J^c) when $|J| \lesssim N$ and $\|\Sigma_J\|_{op} N < \mathrm{Tr}(\Sigma_{J^c})$ or when $|J| \gtrsim N$, $\|\Sigma_J\|_{op} N < \mathrm{Tr}(\Sigma_{J^c})$ and $\mathrm{Tr}(\Sigma_J) \leq N \|\Sigma_J\|_{op}$;*
- *ii) $\square(J)$ is defined by $\square$ in (24) where $\{1 : k\}$ (resp. $\{k+1 : p\}$) is replaced by J (resp. J^c) when $|J| \lesssim N$ and $\|\Sigma_J\|_{op} N \geq \mathrm{Tr}(\Sigma_{J^c})$ or $|J| \gtrsim N$, $\|\Sigma_J\|_{op} N \geq \mathrm{Tr}(\Sigma_{J^c})$ and $\sum_{j \in J_2} \sigma_j \leq \mathrm{Tr}(\Sigma_{J^c}) (1 - |J_1|/N)$.*

It follows from Theorem 7 that the range where benign overfitting happens can be extended to cases where: a) estimation of $\boldsymbol{\beta}_J^*$ by $\hat{\boldsymbol{\beta}}_J$ is good enough (this happens under the same conditions as for instance in (28) except that $\{1, \cdots, k\}$ should be replaced by J) and b) the price for overfitting is low that is when $\|\Sigma_{J^c} \boldsymbol{\beta}_{J^c}^*\|_2$ is small and when the ℓ_2/ℓ_1 ratio of the spectrum of Σ_{J^c} is smaller than $o(1/\sqrt{N})$.

Since $\hat{\boldsymbol{\beta}}$ is a parameter free estimator, we observe that the best features space decomposition $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ is performed automatically in $\hat{\boldsymbol{\beta}}$. This is a remarkable property of $\hat{\boldsymbol{\beta}}$ since, given the upper from Theorem 7, this best features space decomposition should a priori depends on both Σ and $\boldsymbol{\beta}^*$. However, in the next sub-section, we will see that in fact the best choice of J is for $\{1, \ldots, k_b^*\}$ for some constant b so that it only depends on Σ and not on $\boldsymbol{\beta}^*$.

This subsection and the previous ones lead to the following problem: what is the optimal way to decompose the features space? Does it necessarily have to be a direct sum of (two or more) eigenspaces of Σ? It may be the case that there is a better way to entangle Σ and $\boldsymbol{\beta}^*$ in some $(\Sigma, \boldsymbol{\beta}^*)$-adapted basis $\mathcal{B}$ such that on the top k eigenvectors of $\mathcal{B}$, $\boldsymbol{\beta}^*$ and Σ have simultaneously most of their energy and that the restriction of $\hat{\boldsymbol{\beta}}$ to this k-dimensional space is a good (OLS or ridge) estimator of the restriction of $\boldsymbol{\beta}^*$ to it and the rest of the space is used for overfitting (as long as it has the properties for benign overfitting, i.e. the restricted spectrum is well-spread and the restriction of $\boldsymbol{\beta}^*$ has low energy). We answer this question in the next subsection.

4.4 Lower bound on the prediction risk of $\hat{\boldsymbol{\beta}}$ and the best features space decomposition

In this section, we obtain a lower bound on the expectation of the prediction risk of $\hat{\boldsymbol{\beta}}$. This lower bound improves the lower bounds obtained previously in [1] and [39]. It removes some (a posteriori) unnecessary assumptions on the condition number of $X_{k+1:p}$ as well as the smallest singular values of A_{-j} (see Lemma 3 in [39]) and, more importantly, it is a lower bound on the prediction risk of $\hat{\boldsymbol{\beta}}$, not a lower bound on a Bayesian prediction risk as in [1, 39]. This lower bound shows that the best space decomposition is of the form $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ where $J = \{1, \ldots, k\}$ and for the optimal choice of $k = k_b^*$ as previously announced in [1]. This answers the question asked above and proves that the two intuitions from [1] and [39] that the best split of $\mathbb{R}^p$ is $V_J \oplus^\perp V_{J^c}$ for $J = \{1, \ldots, k\}$ and it is best for $k = k_b^*$ and some well chosen constant b. In particular, this optimal choice depends only on Σ and not at all on $\boldsymbol{\beta}^*$ even though the convergence rate depends as well on $\boldsymbol{\beta}^*$.

Theorem 8. *There exists absolute constants $c_0, c_1 > 0$ such that the following holds. If $N \geq c_0$ and Σ is such that $k_b^* < N/4$ for some $b \geq \max(4/\kappa_{DM}, 24)$ then*

$$\mathbb{E}\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2^2 \geq \frac{c_0}{b^2}\max\left\{\frac{\sigma_\xi^2 k_b^*}{N}, \frac{\sigma_\xi^2 N \operatorname{Tr}(\Sigma_{k_b^*+1:p}^2)}{\operatorname{Tr}^2(\Sigma_{k_b^*+1:p})}, \left\|\Sigma_{k_b^*+1:p}^{1/2}\boldsymbol{\beta}_{k_b^*+1:p}^*\right\|_2^2, \left\|\Sigma_{1:k_b^*}^{-1/2}\boldsymbol{\beta}_{1:k_b^*}^*\right\|_2^2\left(\frac{\operatorname{Tr}(\Sigma_{k_b^*+1:p})}{N}\right)^2\right\}.$$

This result holds for the generalization excess risk of $\hat{\boldsymbol{\beta}}$ for any $\boldsymbol{\beta}^*$. In particular, Theorem 8 differs from the previous lower bound results from [1, 39] on some Bayesian risks where models on $\boldsymbol{\beta}^*$ are assumed on the signal. The upper bound on the generalization risk from Theorem 5 with $\square$ from (24) for $k = k_b^*$ matches (up to an absolute constant) the lower bound from Theorem 8. This indeed shows that optimal features space decomposition is $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ for $J = \{1, \dots, k_b^*\}$ and that the convergence to zero of the rate obtained in Theorem 5 for $k = k_b^*$ and Theorem 8 is almost a necessary and sufficient condition for benign overfitting of $\hat{\boldsymbol{\beta}}$ when consistency is defined as the convergence to zero in probability of $\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\right\|_2$. We will therefore say that **overfitting is benign for** $(\Sigma, \boldsymbol{\beta}^*)$ when there exists $k_b^* = o(N)$ such that $\sigma_{k_b^*+1}N \leq \operatorname{Tr}(\Sigma_{k_b^*+1:p})$, $\sigma_1 N \geq \operatorname{Tr}(\Sigma_{k_b^*+1:p})$, $N\operatorname{Tr}(\Sigma_{k_b^*+1:p}^2) = o\left(\operatorname{Tr}^2(\Sigma_{k_b^*+1:p})\right)$,

$$\left\|\Sigma_{k_b^*+1:p}^{1/2}\boldsymbol{\beta}_{k_b^*+1:p}^*\right\|_2 = o(1) \text{ and } \left\|\Sigma_1^{-1/2}\boldsymbol{\beta}_{1:k_b^*}^*\right\|_2 \frac{\operatorname{Tr}(\Sigma_{k_b^*+1:p})}{N} = o(1). \tag{30}$$

The main point in this definition is that it depends on both Σ and $\boldsymbol{\beta}^*$ unlike the previous one given in [1] or [21] which depend only on Σ. Finally, we also emphasis that, once again, our proof of Theorem 8 highlights the role played by the DM theorem on the overfitting part of the features space.

5 Conclusion

Our main results and their proofs brings the following ideas on the benign overfitting phenomenon:

- It was known [39, 2] that the minimum ℓ_2-norm interpolant estimator $\hat{\boldsymbol{\beta}}$ is the sum of an estimator $\hat{\boldsymbol{\beta}}_{1:k}$ and an overfitting component $\hat{\boldsymbol{\beta}}_{k+1:p}$. It is indeed the case that for the optimal features space decomposition $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ for $J = \{1, \dots, k_b^*\}$, $\hat{\boldsymbol{\beta}}_{1:k_b^*}$ is a ridge estimator with a regularization parameter $\operatorname{Tr}(\Sigma_{k_b^*+1:p})$ negligible in front of the square of the smallest singular value of the design matrix so it is essentially an OLS (as announced in [2]) and $V_{k+1:p}$ is used for interpolation. For other choices of J it is in general a ridge estimator over V_J (we showed that all the result can be extended to a more general features space decomposition $\mathbb{R}^p = V_J \oplus^\perp V_{J^c}$ even when $|J| \gtrsim N$). Our findings support the previous ideas of 'over-parametrization' (i.e. $p >> N$) which is necessary for the existence of a space $V_{k+1:p}$ used for interpolation as well as the one that the features space contains a space $V_{1:k_b^*}$ with a small complexity (because $k_b^* \lesssim N$) where most of the estimation of $\boldsymbol{\beta}^*$ happens.
- the proofs of the main results (Theorem 5 and Theorem 6) follow from the risk decomposition (20) which follows from the idea that $\hat{\boldsymbol{\beta}}$ can be decomposed as a sum of an estimator and an overfitting component [2]. In particular, it does not follow from a bias/variance analysis as in [1, 39] because we wanted to put forward that for interpolant estimators, there is part of the feature space $\mathbb{R}^p$ which is not use for estimation but for interpolation (see Section 6). The two geometrical tools we used (i.e. Dvoretsky-Milman theorem and isomophic and restricted isomorphic properties) are classical results from the local geometry of Banach spaces which are both using Gaussian mean widths as a complexity measure. The study of the benign overfitting phenomenon may not require new complexity measure tools but just a better understanding of the feature space decomposition (in contrast with the discussion from [44]) together with a localization argument (which is used here to remove all non plausible candidates in a model). Whereas isomophic and restricted isomorphic properties have been often used in statistical learning theory (and we used it on $V_{1:k}$, the part of the features space where 'estimation happens'), we don't know of any example where the Dvoretsky-Milman theorem has been used in statistical learning theory to obtain convergence rates. It may be because it is used to describe the behavior of the minimum ℓ_2-norm interpolant estimator regarding its overfitting component $\hat{\boldsymbol{\beta}}_{k+1:p}$ on $V_{k+1:p}$ which is not an estimator and therefore cannot be analyzed with classical estimation tools (see (22)). The Dvoretsky-Milman theorem may be a missing tool in the 'classical theory' that may help understand interpolant estimators and the benign overfitting phenomenon (see [44, 4, 2] and references therein for discussions regarding the 'failure' of classical statistical learning theory constructed during the 1990's and 2000's). Both tools are however dealing only with the property of the design matrix $\mathbb{X}$ and not of the output or the signal. We believe that it is because the linear model that we are considering should be seen as a construction coming for instance

after a linear approximation of more complex models (such as the NTK approximation of neural network in some cases) and that it may be the case that for this construction of a features space (that we could look at a feature learning step coming before our linear model) one may require a complexity measure depending on $\mathbb{X}$ and y (or on Σ and $\boldsymbol{\beta}^*$). Somehow, the linear model considered here comes after this features learning step and does not require any other tools than the one considered in this work.

- Even though the optimal features space decomposition is totally independent of $\boldsymbol{\beta}^*$, the benign overfitting phenomenon depends on both Σ (spectrum and eigenvectors) and $\boldsymbol{\beta}^*$ as well as their interplay; indeed benign overfitting (see the definition in (30)) requires that $k_b^* = o(N)$, the spectrum of $\Sigma_{k_b^*+1:p}$ is well spread, the singular values of $\Sigma_{1:k_b^*}$ should be much larger than the one of $\Sigma_{k_b^*+1:p}$ and that $\boldsymbol{\beta}^*$ and Σ need to be well 'aligned': ideally $\boldsymbol{\beta}^*$ should be supported on $V_{1:k_b^*}$.

- the benign overfitting phenomenon happens with large probability,

- The minimum ℓ_2-norm interpolant estimator $\hat{\boldsymbol{\beta}}$ automatically adapts to the best decomposition of the features space: it 'finds' the best split $\mathbb{R}^p = V_{1:k_b^*} \oplus^\perp V_{k_b^*+1:p}$ by itself. However, since this optimal split is independent of the signal, it shows that $\hat{\boldsymbol{\beta}}$ does not learn the best features in $\mathbb{R}^p$ that could predict the output in the best possible way. In other words, $\hat{\boldsymbol{\beta}}$ does not do any features learning by itself and so it needs an upstream procedure to do it for it; that is a construction of a space $\mathbb{R}^p$, design matrix $\mathbb{X}$ with covariance Σ and signal $\boldsymbol{\beta}^*$ that can predict well the output Y as well as the couple $(\Sigma, \boldsymbol{\beta}^*)$ allows for benign overfitting.

Acknowledgement. We would like to thank Matthieu Lerasle, Jaouad Mourtada and Alexandre Tsybakov for useful discussions on interpolant estimators and the benign overfitting phenomenon.

6 Proof of Theorem 5.

In this section, we provide a proof of Theorem 5 which relies on the prediction risk decomposition from (20). To make this scheme analysis described in Section 3 works we need $X_{1:k}$ to behave like an isomorphy onto $V_{1:k}$: for all $\boldsymbol{\beta}_1 \in V_{1:k}, \|X_{1:k}\boldsymbol{\beta}_1\|_2 \sim \sqrt{N}\left\|\Sigma_{1:k}^{1/2}\boldsymbol{\beta}_1\right\|_2$; and we need $\left\|X_{k+1:p}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}\cdot\right\|_2$ to be isomorphic to $(\mathrm{Tr}(\Sigma_{k+1:p}))^{-1/2}\|\cdot\|_2$. These two properties hold on a event that we are now introducing.

Stochastic event behind Theorem 5. We denote by Ω_0 the event onto which we have:

- for all $\boldsymbol{\lambda} \in \mathbb{R}^N$, $(1/(2\sqrt{2}))\sqrt{\mathrm{Tr}(\Sigma_{k+1:p})}\|\boldsymbol{\lambda}\|_2 \le \left\|X_{k+1:p}^\top\boldsymbol{\lambda}\right\|_2 \le (3/2)\sqrt{\mathrm{Tr}(\Sigma_{k+1:p})}\|\boldsymbol{\lambda}\|_2$

- for all $\boldsymbol{\beta}_1 \in V_{1:k}$, $(1/2)\left\|\Sigma_{1:k}^{1/2}\boldsymbol{\beta}_1\right\|_2 \le (1/\sqrt{N})\|X_{1:k}\boldsymbol{\beta}_1\|_2 \le (3/2)\left\|\Sigma_{1:k}^{1/2}\boldsymbol{\beta}_1\right\|_2$.

It follows from Theorem 3 and Corollary 1 that if $N \le \kappa_{DM} d_*(\Sigma_{k+1:p}^{-1/2}B_2^p)$ and $k \le \kappa_{iso}N$ then $\mathbb{P}[\Omega_0] \ge 1 - c_0\exp(-c_1 N)$. We place ourselves on the event Ω_0 up to the end of the proof.

To make the presentation simpler we denote $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_{1:k}^*$, $\boldsymbol{\beta}_2^* = \boldsymbol{\beta}_{k+1:p}^*$, $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{1:k}$ and $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_{k+1:p}$.

6.1 Estimation properties of the 'ridge estimator' $\hat{\boldsymbol{\beta}}_{1:k}$

Our starting point is (19):

$$\hat{\boldsymbol{\beta}}_{1:k} \in \underset{\boldsymbol{\beta}_1 \in V_{1:k}}{\mathrm{argmin}}\left(\|A(y - X_{1:k}\boldsymbol{\beta}_1)\|_2^2 + \|\boldsymbol{\beta}_1\|_2^2\right) \tag{31}$$

where we set $A = X_{k+1:p}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}$ and where we used that the minimum over $\mathbb{R}^p$ in (19) is actually achieved in $V_{1:k}$. Next, we use a 'quadratic + multiplier + regularization decomposition' of the excess regularized risk associated with the RERM (31) similar to the one from [23] but with the difference that the regularization term in (31) is the square of a norm and not directly a norm. This makes a big difference (otherwise we will need a lower bound on the quantity $\mathrm{Tr}(\Sigma_{k+1:p})$ which will play the role of the regularization term in (31), a condition we want to avoid here). We therefore write for all $\boldsymbol{\beta}_1 \in V_{1:k}$,

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\beta}_1} &= \|A(y - X_{1:k}\boldsymbol{\beta}_1)\|_2^2 + \|\boldsymbol{\beta}_1\|_2^2 - \left(\|A(y - X_{1:k}\boldsymbol{\beta}_1^*)\|_2^2 + \|\boldsymbol{\beta}_1^*\|_2^2\right) \\ &= \|AX_{1:k}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\|_2^2 + 2\big\langle A(y - X_{1:k}\boldsymbol{\beta}_1^*), AX_{1:k}(\boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1)\big\rangle + \|\boldsymbol{\beta}_1\|_2^2 - \|\boldsymbol{\beta}_1^*\|_2^2 \qquad (32) \\ &= \left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1/2}X_{1:k}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right\|_2^2 + 2\big\langle X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}(X_{k+1:p}\boldsymbol{\beta}_2^* + \boldsymbol{\xi}) - \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\big\rangle + \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2 \\ &\qquad (33)\end{aligned}$$

where we used that for all $\boldsymbol{\lambda} \in \mathbb{R}^N, \|A\boldsymbol{\lambda}\|_2 = \left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1/2}\boldsymbol{\lambda}\right\|_2$, $A^\top A = (X_{k+1:p}X_{k+1:p}^\top)^{-1}$ and $\|\boldsymbol{\beta}_1\|_2^2 - \|\boldsymbol{\beta}_1^*\|_2^2 = \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2 - 2\langle\boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1^* - \boldsymbol{\beta}_1\rangle$. The last equality is a modification on the regularization term '$\|\boldsymbol{\beta}_1\|_2^2 - \|\boldsymbol{\beta}_1^*\|_2^2$' as well as the multiplier term (i.e. the second term in (32)) in the classical 'quadratic + multiplier + regularization decomposition' of the excess regularized risk written in (32). As mentionned previously, this modification is key to our analysis (we may look at it as a 'quadratic + multiplier decomposition' of the excess regularization term).

We will use the excess risk decomposition from (33) to prove that with high probability

$$\left\|\Sigma_{1:k}^{1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)\right\|_2 \leq \Box \text{ and } \left\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right\|_2 \leq \triangle \qquad (34)$$

where $\Box$ and $\triangle$ are two quantities that we will choose later. In other words, we want to show that $\hat{\boldsymbol{\beta}}_1 \in \boldsymbol{\beta}_1^* + B$ where $B = \{\boldsymbol{\beta} \in V_{1:k} : |||\boldsymbol{\beta}||| \leq 1\}$ and for all $\boldsymbol{\beta} \in V_{1:k}$,

$$|||\boldsymbol{\beta}||| := \max\left(\frac{\left\|\Sigma_{1:k}^{1/2}\boldsymbol{\beta}\right\|_2}{\Box}, \frac{\|\boldsymbol{\beta}\|_2}{\triangle}\right). \qquad (35)$$

To do that we show that if $\boldsymbol{\beta}_1$ is a vector in $V_{1:k}$ such that $\boldsymbol{\beta}_1 \notin \boldsymbol{\beta}_1^* + B$ then necessarily $\mathcal{L}_{\boldsymbol{\beta}_1} > 0$. This is what we are doing now and we start with an homogeneity argument similar to the one in [22].

Denote by ∂B the border of B in $V_{1:k}$. Let $\boldsymbol{\beta}_1 \in V_{1:k}$ be such that $\boldsymbol{\beta}_1 \notin \boldsymbol{\beta}_1^* + B$. There exists $\boldsymbol{\beta}_0 \in \partial B$ and $\theta > 1$ such that $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^* = \theta(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_1^*)$. Using (33), it is clear that $\mathcal{L}_{\boldsymbol{\beta}_1} \geq \theta\mathcal{L}_{\boldsymbol{\beta}_0}$. As a consequence, if we prove that $\mathcal{L}_{\boldsymbol{\beta}_1} > 0$ for all $\boldsymbol{\beta}_1 \in \boldsymbol{\beta}_1^* + \partial B$ this will imply that $\mathcal{L}_{\boldsymbol{\beta}_1} > 0$ for all $\boldsymbol{\beta}_1 \notin \boldsymbol{\beta}_1^* + B$. Hence, we only need to show the positivity of the excess regularized risk $\mathcal{L}_{\boldsymbol{\beta}_1}$ on the border $\boldsymbol{\beta}_1^* + \partial B$ and since an element $\boldsymbol{\beta}_1$ in the border $\boldsymbol{\beta}_1^* + \partial B$ may have two different behaviors, we introduce two cases:

a) $\left\|\Sigma_{1:k}^{1/2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right\|_2 = \Box$ and $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq \triangle$;

b) $\left\|\Sigma_{1:k}^{1/2}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right\|_2 \leq \Box$ and $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 = \triangle$.

All the point is to show that in the two cases $\mathcal{L}_{\boldsymbol{\beta}_1} > 0$. If we look at (33) among the three terms in this equation only the 'multiplier term' $\mathcal{M}_{\boldsymbol{\beta}_1} := 2\big\langle X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}(X_{k+1:p}\boldsymbol{\beta}_2^* + \boldsymbol{\xi}) - \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\big\rangle$ can possibly be negative whereas the two others quadratic term $\mathcal{Q}_{\boldsymbol{\beta}_1} := \left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1/2}X_{1:k}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*)\right\|_2^2$ and 'regularization term' $\mathcal{R}_{\boldsymbol{\beta}_1} := \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2^2$ are positive. As a consequence, we will show that $\mathcal{L}_{\boldsymbol{\beta}_1} > 0$ because either $\mathcal{Q}_{\boldsymbol{\beta}_1} > |\mathcal{M}_{\boldsymbol{\beta}_1}|$ (this will hold in case *a)*) or $\mathcal{R}_{\boldsymbol{\beta}_1} > |\mathcal{M}_{\boldsymbol{\beta}_1}|$ (this will hold in case *b)*).

Let us first control the multiplier term $\mathcal{M}_{\boldsymbol{\beta}_1}$ for $\boldsymbol{\beta}_1 \in \boldsymbol{\beta}_1^* + \partial B$. We have

$$\begin{aligned}|\mathcal{M}_{\boldsymbol{\beta}_1}|/2 &= |\big\langle X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}(X_{k+1:p}\boldsymbol{\beta}_2^* + \boldsymbol{\xi}) - \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\big\rangle| \\ &\leq \sup_{v \in B} |\big\langle X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}(X_{k+1:p}\boldsymbol{\beta}_2^* + \boldsymbol{\xi}) - \boldsymbol{\beta}_1^*, v\big\rangle|\end{aligned}$$

where we recall that B is the unit ball of $|||\cdot|||$ intersected with $V_{1:k}$. It is straightforward to check that for all $\boldsymbol{\beta} \in V_{1:k}$, $|||\boldsymbol{\beta}||| \leq \left\|\tilde{\Sigma}_1^{1/2}\boldsymbol{\beta}\right\|_2 \leq \sqrt{2}|||\boldsymbol{\beta}|||$ where $\tilde{\Sigma}_1^{1/2} = U\tilde{D}_1^{1/2}U^\top$ and

$$\tilde{D}_1^{1/2} = \operatorname{diag}\left(\max\left(\frac{\sqrt{\sigma_1}}{\Box}, \frac{1}{\triangle}\right), \dots, \max\left(\frac{\sqrt{\sigma_k}}{\Box}, \frac{1}{\triangle}\right), 0, \dots, 0\right). \qquad (36)$$

Therefore, $|||\cdot|||$'s dual norm $|||\cdot|||_*$ is also equivalent to $\left\|\tilde\Sigma_1^{1/2}\cdot\right\|_2$'s dual norm which is given by $\left\|\tilde\Sigma_1^{-1/2}\cdot\right\|_2$: for all $\boldsymbol\beta\in V_{1:k}$, $(1/\sqrt2)|||\boldsymbol\beta|||_*\le\left\|\tilde\Sigma_1^{-1/2}\boldsymbol\beta\right\|_2\le|||\boldsymbol\beta|||_*$. Hence, we have for all $\boldsymbol\beta_1\in\boldsymbol\beta_1^*+\partial B$,

$$\begin{aligned}|\mathcal M_{\boldsymbol\beta_1}|&\le 2\sqrt2\left\|\tilde\Sigma_1^{-1/2}\left(X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}(X_{k+1:p}\boldsymbol\beta_2^*+\boldsymbol\xi)-\boldsymbol\beta_1^*\right)\right\|_2\\&\le 2\sqrt2\left(\left\|\tilde\Sigma_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}X_{k+1:p}\boldsymbol\beta_2^*\right\|_2+\left\|\tilde\Sigma_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}\boldsymbol\xi\right\|_2+\left\|\tilde\Sigma_1^{-1/2}\boldsymbol\beta_1^*\right\|_2\right).\end{aligned}\tag{37}$$

Next, we handle the first two terms in (37) in the next two lemmas.

Lemma 1. *Assume that $N\le\kappa_{DM}d_*(\Sigma_{k+1:p}^{-1/2}B_2^p)$ and $k\le\kappa_{iso}N$. With probability at least $1-c_0\exp(-c_1N)$,*

$$\left\|\tilde\Sigma_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}X_{k+1:p}\boldsymbol\beta_2^*\right\|_2\le\frac{18N\sigma(\square,\triangle)}{\mathrm{Tr}(\Sigma_{k+1:p})}\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol\beta_2^*\right\|_2$$

where

$$\sigma(\square,\triangle):=\begin{cases}\square & \text{if } \triangle\sqrt{\sigma_1}\ge\square\\ \triangle\sqrt{\sigma_1} & \text{otherwise.}\end{cases}\tag{38}$$

Proof. It follows from Bernstein's inequality that with probability at least $1-c_0\exp(-c_1N)$,

$$\|X_{k+1:p}\boldsymbol\beta_2^*\|_2\le\frac{3\sqrt N}{2}\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol\beta_2^*\right\|_2.$$

On the event Ω_0, we have $\left\|\tilde\Sigma_{1:k}^{-1/2}X_{1:k}^\top\right\|_{op}=\left\|X_{1:k}\tilde\Sigma_{1:k}^{-1/2}\right\|_{op}\le(3/2)\sqrt N\left\|\Sigma_{1:k}^{1/2}\tilde\Sigma_{1:k}^{-1/2}\right\|_{op}$ because of the isomorphic property of $X_{1:k}$ and $\left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1}\right\|_{op}\le(3/2)(\mathrm{Tr}(\Sigma_{k+1:p}))^{-1}$ because of the Dvoretsky-Milman's property satisfied by $X_{k+1:p}$ (see Proposition 1). As a consequence, with probability at least $1-c_0\exp(-c_1N)$,

$$\begin{aligned}&\left\|\tilde\Sigma_{1:k}^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}X_{k+1:p}\boldsymbol\beta_2^*\right\|_2\le\left\|\tilde\Sigma_{1:k}^{-1/2}X_{1:k}^\top\right\|_{op}\left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1}\right\|_{op}\|X_{k+1:p}\boldsymbol\beta_2^*\|_2\\&\le\frac{18N}{\mathrm{Tr}(\Sigma_{k+1:p})}\left\|\Sigma_{1:k}^{1/2}\tilde\Sigma_{1:k}^{-1/2}\right\|_{op}\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol\beta_2^*\right\|_2=\frac{18N\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol\beta_2^*\right\|_2}{\mathrm{Tr}(\Sigma_{k+1:p})}\end{aligned}$$

because $\left\|\Sigma_{1:k}^{1/2}\tilde\Sigma_{1:k}^{-1/2}\right\|_{op}$ is smaller than $\max\left(\max(\square\mathbb 1\left(\triangle\sqrt{\sigma_j}\ge\square\right),\triangle\sqrt{\sigma_j}\mathbb 1\left(\triangle\sqrt{\sigma_j}\le\square\right)):j\in[k]\right)$ which is equal to $\sigma(\square,\triangle)$ as defined in (38). ■

To control the second term in (37), we recall the definition of the following two subsets

$$J_1:=\left\{j\in[k]:\sigma_j\geqslant(\square/\triangle)^2\right\}\text{ and }J_2:=[k]\backslash J_1=\left\{j\in[k]:\sigma_j<(\square/\triangle)^2\right\}.$$

Lemma 2. *Assume that $N\le\kappa_{DM}d_*(\Sigma_{k+1:p}^{-1/2}B_2^p)$ and $k\le\kappa_{iso}N$. With probability at least $1-\exp(-t(\square,\triangle)/2)-c_0\exp(-c_1N)$,*

$$\left\|\tilde\Sigma_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}\boldsymbol\xi\right\|_2\le\frac{32\sqrt N\sigma_\xi}{\mathrm{Tr}(\Sigma_{k+1:p})}\sqrt{|J_1|\,\square^2+\triangle^2\sum_{j\in J_2}\sigma_j}$$

where $t(\square,\triangle):=\left(|J_1|\,\square^2+\triangle^2\sum_{j\in J_2}\sigma_j\right)/\sigma^2(\square,\triangle)$ and $\sigma(\square,\triangle)$ has been defined in Lemma 1.

Proof. It follows from the Borell-TIS's inequality (see Theorem 7.1 in [24] or p.56-57 in [25]) that for all $t>0$, conditionnaly on $\mathbb X$, with probability (w.r.t. the randomness of the noise $\boldsymbol\xi$) at least $1-\exp(-t/2)$,

$$\left\|\tilde\Sigma_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}\boldsymbol\xi\right\|_2\le\sigma_\xi\left(\sqrt{\mathrm{Tr}(DD^\top)}+\sqrt t\,\|D\|_{op}\right)\tag{39}$$

where $D = \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1}$. For the basis $(f_j)_{j=1}^p$ of eigenvectors of Σ, on the event Ω_0, we have

$$\operatorname{Tr}(DD^\top) = \sum_{j=1}^p \left\| D^\top f_j \right\|_2^2 = \sum_{j=1}^k \left(\frac{\sqrt{\sigma_j}}{\square} \vee \frac{1}{\triangle} \right)^{-2} \left\| \left(X_{k+1:p} X_{k+1:p}^\top \right)^{-1} X_{1:k} f_j \right\|_2^2$$
$$\leq \sum_{j=1}^k \left(\frac{\sqrt{\sigma_j}}{\square} \vee \frac{1}{\triangle} \right)^{-2} \left\| \left(X_{k+1:p} X_{k+1:p}^\top \right)^{-1} \right\|_{op}^2 \| X_{1:k} f_j \|_2^2 \leq \sum_{j=1}^k \left(\frac{\sqrt{\sigma_j}}{\square} \vee \frac{1}{\triangle} \right)^{-2} \frac{256 N \sigma_j}{(\operatorname{Tr} \Sigma_{k+1:p})^2}$$

and therefore

$$\sqrt{\operatorname{Tr}(DD^\top)} \leq \frac{16\sqrt{N}}{\operatorname{Tr}(\Sigma_{k+1:p})} \sqrt{|J_1| \, \square^2 + \triangle^2 \sum_{j \in J_2} \sigma_j}.$$

Using similar arguments and that $\left\| \Sigma_{1:k}^{1/2} \tilde{\Sigma}_{1:k}^{-1/2} \right\|_{op}$ is smaller than $\sigma(\square, \triangle)$, we also have (still on the event Ω_0)

$$\|D\|_{op} = \left\| D^\top \right\|_{op} \leq \left\| (X_{k+1:p} X_{k+1:p}^\top)^{-1} \right\|_{op} \left\| X_{1:k} \tilde{\Sigma}_1^{-1/2} \right\|_{op} \leq \frac{16\sqrt{N} \left\| \Sigma_{1:k}^{1/2} \tilde{\Sigma}_{1:k}^{-1/2} \right\|_{op}}{\operatorname{Tr}(\Sigma_{k+1:p})} \leq \frac{16\sqrt{N} \sigma(\square, \triangle)}{\operatorname{Tr}(\Sigma_{k+1:p})}.$$

We use the later upper bounds on $\sqrt{\operatorname{Tr}(DD^\top)}$ and $\|D\|_{op}$ and take $t = \left(|J_1| \, \square^2 + \triangle^2 \sum_{j \in J_2} \sigma_j \right) / \sigma^2(\square, \triangle)$ in (39) to conclude. ■

Let us now show that $\mathcal{L}_{\boldsymbol{\beta}_1} > 0$ for $\boldsymbol{\beta}_1 \in \boldsymbol{\beta}_1^* + \partial B$ using the upper bounds on the multiplier part $\mathcal{M}_{\boldsymbol{\beta}_1}$ which follows from the last two lemmas. Let us first analyze *case a)*: let $\boldsymbol{\beta}_1 \in V_{1:k}$ be such that $\left\| \Sigma^{1/2} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \right\|_2 = \square$ and $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 \leq \triangle$. In that case, we now show that $\mathcal{Q}_{\boldsymbol{\beta}_1} > \mathcal{M}_{\boldsymbol{\beta}_1}$. On the event Ω_0, we have

$$\mathcal{Q}_{\boldsymbol{\beta}_1} = \left\| (X_{k+1:p} X_{k+1:p}^\top)^{-1/2} X_{1:k} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \right\|_2^2 \geq \frac{N}{32 \operatorname{Tr}(\Sigma_{k+1:p})} \left\| \Sigma_{1:k}^{1/2} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \right\|_2^2 = \frac{N \square^2}{32 \operatorname{Tr}(\Sigma_{k+1:p})}. \tag{40}$$

It follows from (37) that to get $\mathcal{Q}_{\boldsymbol{\beta}_1} > \mathcal{M}_{\boldsymbol{\beta}_1}$ (and so $\mathcal{L}_{\boldsymbol{\beta}_1} > 0$), it is enough to have

$$\frac{N \square^2}{64\sqrt{2} \operatorname{Tr}(\Sigma_{k+1:p})} > \left\| \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} X_{k+1:p} \boldsymbol{\beta}_2^* \right\|_2 + \left\| \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} \boldsymbol{\xi} \right\|_2 + \left\| \tilde{\Sigma}_1^{-1/2} \boldsymbol{\beta}_1^* \right\|_2. \tag{41}$$

Let us now assume that we choose $\square$ and $\triangle$ so that $\sigma(\square, \triangle) = \square$, which means that $\triangle \sqrt{\sigma_1} \geq \square$ (we will explore the other case later). Using Lemma 1 and 2, inequality (41) holds if we take $\square$ such that

$$\square \geq C_0 \max \left\{ \sigma_\xi \sqrt{\frac{|J_1|}{N}}, \left(\triangle \sigma_\xi \sqrt{\frac{1}{N} \sum_{j \in J_2} \sigma_j} \right)^{1/2}, \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2, \sqrt{\left\| \tilde{\Sigma}_1^{-1/2} \boldsymbol{\beta}_1^* \right\|_2 \frac{\operatorname{Tr}(\Sigma_{k+1:p})}{N}} \right\} \tag{42}$$

where $C_0 = 4608\sqrt{2}$.

Next, let us analyse *case b)*: let $\boldsymbol{\beta}_1 \in V_{1:k}$ be such that $\left\| \Sigma^{1/2} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*) \right\|_2 \leq \square$ and $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^*\|_2 = \triangle$. In that case, we show that $\mathcal{L}_{\boldsymbol{\beta}_1} > 0$ by proving that $\mathcal{R}_{\boldsymbol{\beta}_1} > \mathcal{M}_{\boldsymbol{\beta}_1}$. Since $\mathcal{R}_{\boldsymbol{\beta}_1} = \triangle^2$, it follows from (37) that $\mathcal{R}_{\boldsymbol{\beta}_1} > \mathcal{M}_{\boldsymbol{\beta}_1}$ holds when

$$\triangle^2 > 2\sqrt{2} \left(\left\| \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} X_{k+1:p} \boldsymbol{\beta}_2^* \right\|_2 + \left\| \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} \boldsymbol{\xi} \right\|_2 + \left\| \tilde{\Sigma}_1^{-1/2} \boldsymbol{\beta}_1^* \right\|_2 \right).$$

Using Lemma 1 and 2, the latter inequality holds when

$$\triangle^2 \geq C_1 \max \left\{ \frac{\sigma_\xi \square \sqrt{|J_1| N}}{\operatorname{Tr}(\Sigma_{k+1:p})}, \frac{\sigma_\xi^2 N}{\operatorname{Tr}(\Sigma_{k+1:p})^2} \sum_{j \in J_2} \sigma_j, \frac{N \square}{\operatorname{Tr}(\Sigma_{k+1:p})} \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2, \left\| \tilde{\Sigma}_1^{-1/2} \boldsymbol{\beta}_1^* \right\|_2 \right\} \tag{43}$$

for $C_1 = 4608$.

We set $\triangle$ so that $\square / \triangle = \sqrt{\operatorname{Tr}(\Sigma_{k+1:p})/N}$ and take

$$\square = C_0 \max \left\{ \sigma_\xi \sqrt{\frac{|J_1|}{N}}, \sigma_\xi \sqrt{\frac{\sum_{j \in J_2} \sigma_j}{\operatorname{Tr}(\Sigma_{k+1:p})}}, \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2, \sqrt{\left\| \tilde{\Sigma}_1^{-1/2} \boldsymbol{\beta}_1^* \right\|_2 \frac{\operatorname{Tr}(\Sigma_{k+1:p})}{N}} \right\}. \tag{44}$$

In particular, we check that for this choice $\triangle$ satisfies (43) and $\triangle\sqrt{\sigma_1} \geq \square$ holds iff $\sigma_1 \geq \mathrm{Tr}(\Sigma_{k+1:p})/N$. As a consequence, in the case where $\sigma_1 \geq \mathrm{Tr}(\Sigma_{k+1:p})/N$, $\square$ is a valid upper bound on the rate of convergence of $\hat{\boldsymbol{\beta}}_1$ toward $\boldsymbol{\beta}_1^*$ with respect to the $\left\|\Sigma_{1:k}^{1/2}\cdot\right\|_2$-norm. This upper bound holds with probabilty at least $1 - c_0\exp(-c_1 N) - \exp(-t(\square,\triangle)/2)$ where in that case

$$t(\square,\triangle) = |J_1| + N\frac{\sum_{j\in J_2}\sigma_j}{\mathrm{Tr}(\Sigma_{k+1:p})}.$$

We also note that because $\square/\triangle = \sqrt{\mathrm{Tr}(\Sigma_{k+1:p})/N}$, one has $\tilde{\Sigma}_1^{-1/2} = U\tilde{D}_1^{-1/2}U^\top$ with

$$\tilde{D}_1^{-1/2} = \square\mathrm{diag}\left(\left(\sigma_1 \vee \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right)^{-1/2}, \cdots, \left(\sigma_k \vee \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right)^{-1/2}, 0, \dots, 0\right) := \square D_{1,thres}^{-1/2}$$

and so one can simplify the choice of $\square$ from (44) and take

$$\square = C_0^2 \max\left\{\sigma_\xi\sqrt{\frac{|J_1|}{N}}, \sigma_\xi\sqrt{\frac{\sum_{j\in J_2}\sigma_j}{\mathrm{Tr}(\Sigma_{k+1:p})}}, \left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2, \left\|\Sigma_{1,thres}^{-1/2}\boldsymbol{\beta}_1^*\right\|_2 \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right\} \tag{45}$$

where $\Sigma_{1,thres}^{-1/2} = UD_{1,thres}^{-1/2}U^\top$.

Let us now consider the other case that is when $\square$ and $\triangle$ are chosen so that $\sigma(\square,\triangle) = \triangle\sqrt{\sigma_1}$, which means that $\triangle\sqrt{\sigma_1} \leq \square$. In that case, J_1 is empty, $J_2 = [k]$, $t(\square,\triangle) = \mathrm{Tr}(\Sigma_1)/\sigma_1$, $\tilde{\Sigma}_1^{1/2} = U\tilde{D}_1^{1/2}U^\top$ with $\tilde{D}_1^{1/2} = (1/\triangle)\mathrm{diag}(1,\cdots,1,0,\cdots,0)$ (with k ones and $p-k$ zeros). In *case a)*, it follows from Lemma 1, Lemma 2 and inequality (41) that $\mathcal{Q}_{\boldsymbol{\beta}_1} > \mathcal{M}_{\boldsymbol{\beta}_1}$ holds when

$$\square^2 \geq C_0^2 \max\left\{\triangle\sigma_\xi\sqrt{\frac{\mathrm{Tr}(\Sigma_{1:k})}{N}}, \triangle\sqrt{\sigma_1}\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2, \triangle\left\|\boldsymbol{\beta}_1^*\right\|_2\frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right\} \tag{46}$$

where $C_0 = 4608\sqrt{2}$. In *case b)*, we will have $\mathcal{R}_{\boldsymbol{\beta}_1} > \mathcal{M}_{\boldsymbol{\beta}_1}$ when

$$\triangle^2 \geq C_1\max\left\{\frac{\sigma_\xi^2 N\,\mathrm{Tr}(\Sigma_{1:k})}{\mathrm{Tr}(\Sigma_{k+1:p})^2}, \sigma_1\left(\frac{N}{\mathrm{Tr}(\Sigma_{k+1:p})}\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2\right)^2, \left\|\boldsymbol{\beta}_1^*\right\|_2^2\right\} \tag{47}$$

for $C_1 = 4608$. We choose $\triangle = \square\sqrt{N/\mathrm{Tr}(\Sigma_{k+1:p})}$ and take

$$\square = C_0^2\max\left\{\sigma_\xi\sqrt{\frac{\mathrm{Tr}(\Sigma_{1:k})}{\mathrm{Tr}(\Sigma_{k+1:p})}}, \sqrt{\frac{N\sigma_1}{\mathrm{Tr}(\Sigma_{k+1:p})}}\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2, \left\|\boldsymbol{\beta}_1^*\right\|_2\sqrt{\frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}}\right\} \tag{48}$$

In particular, we check that for this choice, $\triangle$ satisfies (47). Therefore, when $\triangle\sqrt{\sigma_1} \leq \square$ – which is equivalent to $\sigma_1 N \leq \mathrm{Tr}(\Sigma_{k+1:p})$ – the choice of $\square$ from (48) provides an upper bound on $\left\|\Sigma_1^{1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)\right\|_2$ with probability at least $1 - c_0\exp(-c_1 N)$ – note that $t(\square,\triangle) = \mathrm{Tr}(\Sigma_1)/\sigma_1 \geq N \geq k$ so that $\exp(-t(\square,\triangle)/2) \leq \exp(-k/2)$.

We are now gathering our finding on the estimation properties of $\hat{\boldsymbol{\beta}}_1$ in the next result. We state the result for both norms $\left\|\Sigma_1^{1/2}\cdot\right\|_2$ and $\|\cdot\|_2$.

Proposition 4. *There are absolute constants c_0, c_1 and $C_0 = 4608\sqrt{2}$ such that the following holds. We assume that there exists $k \leq \kappa_{iso}N$ such that $N \leq \kappa_{DM} d_*(\Sigma_{k+1:p}^{-1/2}B_2^p)$, then the following holds for all such k's. We define*

$$J_1 := \left\{j \in [k] : \sigma_j \geq \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right\}, \quad J_2 := \left\{j\in[k] : \sigma_j < \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right\}$$

and $\Sigma_{1,thres}^{-1/2} := UD_{1,thres}^{-1/2}U^\top$ where

$$D_{1,thres}^{-1/2} := \mathrm{diag}\left(\left(\sigma_1 \vee \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right)^{-1/2}, \dots, \left(\sigma_k \vee \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}\right)^{-1/2}, 0, \dots, 0\right).$$

With probability at least $1 - p^*$, $\left\|\Sigma_1^{1/2}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*)\right\|_2 \leq \square$ and $\left\|\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*\right\|_2 \leq \square\sqrt{N/\mathrm{Tr}(\Sigma_{k+1:p})}$ where, *i) if* $\sigma_1 N \leq \mathrm{Tr}(\Sigma_{k+1:p})$, $p^* = c_0 \exp(-c_1 N)$ *and*

$$\square = C_0^2 \max\left\{ \sigma_\xi \sqrt{\frac{\mathrm{Tr}(\Sigma_{1:k})}{\mathrm{Tr}(\Sigma_{k+1:p})}}, \sqrt{\frac{N\sigma_1}{\mathrm{Tr}(\Sigma_{k+1:p})}} \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2, \|\boldsymbol{\beta}_1^*\|_2 \sqrt{\frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}} \right\}$$

ii) if $\sigma_1 N \geq \mathrm{Tr}(\Sigma_{k+1:p})$, $p^* = c_0 \exp\left(-c_1 \left(|J_1| + N\left(\sum_{j\in J_2} \sigma_j\right) / (\mathrm{Tr}(\Sigma_{k+1:p}))\right)\right)$ *and*

$$\square = C_0^2 \max\left\{ \sigma_\xi \sqrt{\frac{|J_1|}{N}}, \sigma_\xi \sqrt{\frac{\sum_{j\in J_2}\sigma_j}{\mathrm{Tr}(\Sigma_{k+1:p})}}, \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2, \left\| \Sigma_{1,thres}^{-1/2} \boldsymbol{\beta}_1^* \right\|_2 \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N} \right\}.$$

The two cases *i)* and *ii)* appear naturally in the study of the ridge estimator (21): in *case i)*, the regularization parameter $\mathrm{Tr}(\Sigma_{k+1:p})$ is larger than the square of the top singular value of $X_{1:k}$ and so the ridge estimator is mainly minimizing the regularization norm. Whereas in the other case, the regularization terms is doing a shrinkage on the spectrum of $X_{1:k}$ and one can see this effect through the threshold operator $\Sigma_{1,thres}$ appearing in *case ii)* from Proposition 4.

6.2 Upper bound on the price for overfitting.

Following the risk decomposition (20), the last term we need to handle is $\left\| \Sigma_{k+1:p}^{1/2} (\hat{\boldsymbol{\beta}}_{k+1:p} - \boldsymbol{\beta}_{k+1:p}^*) \right\|_2$. As we said above, we do not expect $\hat{\boldsymbol{\beta}}_{k+1:p}$ to be a good estimator of $\boldsymbol{\beta}_2^* := \boldsymbol{\beta}_{k+1:p}^*$ because the minimum ℓ_2-norm estimator $\hat{\boldsymbol{\beta}}$ is using the 'reamaining part' of $\mathbb{R}^p$ endowed by the $p-k$ smallest eigenvectors of Σ (we denoted this space by $V_{k+1:p}$) to interpolate the noise $\boldsymbol{\xi}$ and not to estimate $\boldsymbol{\beta}_2^*$ that is why we call the error term $\left\| \Sigma_{k+1:p}^{1/2} (\hat{\boldsymbol{\beta}}_{k+1:p} - \boldsymbol{\beta}_2^*) \right\|_2$ a price for noise interpolation instead of an estimation error. A consequence is that we just upper bound this term by

$$\left\| \Sigma_{k+1:p}^{1/2} (\hat{\boldsymbol{\beta}}_{k+1:p} - \boldsymbol{\beta}_2^*) \right\|_2 \leq \left\| \Sigma_{k+1:p}^{1/2} \hat{\boldsymbol{\beta}}_{k+1:p} \right\|_2 + \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2.$$

Then we just need to find a high probability upper bound on $\left\| \Sigma_{k+1:p}^{1/2} \hat{\boldsymbol{\beta}}_{k+1:p} \right\|_2$. It follows from Proposition 3 that $\hat{\boldsymbol{\beta}}_{k+1:p} = X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} (y - X_{1:k}\hat{\boldsymbol{\beta}}_{1:k})$, hence, we have for $A := X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1}$,

$$\begin{aligned} \left\| \Sigma_{k+1:p}^{1/2} \hat{\boldsymbol{\beta}}_{k+1:p} \right\|_2 &= \left\| \Sigma_{k+1:p}^{1/2} A (y - X_{1:k}\hat{\boldsymbol{\beta}}_{1:k}) \right\|_2 \\ &\leq \left\| \Sigma_{k+1:p}^{1/2} A X_{1:k} (\boldsymbol{\beta}_1^* - \hat{\boldsymbol{\beta}}_{1:k}) \right\|_2 + \left\| \Sigma_{k+1:p}^{1/2} A X_{k+1:p} \boldsymbol{\beta}_2^* \right\|_2 + \left\| \Sigma_{k+1:p}^{1/2} A \boldsymbol{\xi} \right\|_2 \end{aligned} \tag{49}$$

and now we obtain high probability upper bounds on the three terms in (49).

We denote by Ω_1 the event onto which for all $\boldsymbol{\lambda} \in \mathbb{R}^N$,

$$\left\| \Sigma_{k+1:p}^{1/2} X_{k+1:p}^\top \boldsymbol{\lambda} \right\|_2 \leq 6 \left(\sqrt{\mathrm{Tr}(\Sigma_{k+1:p}^2)} + \sqrt{N} \|\Sigma_{k+1:p}\|_{op} \right) \|\boldsymbol{\lambda}\|_2. \tag{50}$$

It follows from Proposition 2 that $\mathbb{P}[\Omega_1] \geq 1 - \exp(-N)$ (and this result holds without any extra assumption on N).

On $\Omega_0 \cap \Omega_1$, we have

$$\begin{aligned} \left\| \Sigma_{k+1:p}^{1/2} A X_{1:k} (\boldsymbol{\beta}_1^* - \hat{\boldsymbol{\beta}}_{1:k}) \right\|_2 &= \left\| \Sigma_{k+1:p}^{1/2} X_{k+1:p}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} X_{1:k} (\boldsymbol{\beta}_1^* - \hat{\boldsymbol{\beta}}_{1:k}) \right\|_2 \\ &\leq \left\| \Sigma_{k+1:p}^{1/2} X_{k+1:p}^\top \right\|_{op} \left\| (X_{k+1:p} X_{k+1:p}^\top)^{-1} \right\|_{op} \left\| X_{1:k} (\boldsymbol{\beta}_1^* - \hat{\boldsymbol{\beta}}_{1:k}) \right\|_2 \\ &\leq 30\sqrt{2} \frac{\left(\sqrt{N \,\mathrm{Tr}(\Sigma_{k+1:p}^2)} + N \|\Sigma_{k+1:p}\|_{op} \right)}{\mathrm{Tr}(\Sigma_{k+1:p})} \left\| \Sigma_{1:k}^{1/2} (\boldsymbol{\beta}_1^* - \hat{\boldsymbol{\beta}}_{1:k}) \right\|_2 \end{aligned} \tag{51}$$

It follows from Bernstein's inequality that $\|X_{k+1:p}\boldsymbol{\beta}_2^*\|_2 \leq (3/2)\sqrt{N} \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2$ holds with probability at least $1 - c_0 \exp(-c_1 N)$. Hence, with probability at least $1 - c_0 \exp(-c_1 N) - \mathbb{P}[(\Omega_0 \cap \Omega_1)^c]$,

$$\begin{aligned} \left\| \Sigma_{k+1:p}^{1/2} A X_{k+1:p} \boldsymbol{\beta}_2^* \right\|_2 &\leq \left\| \Sigma_{k+1:p}^{1/2} X_{k+1:p}^\top \right\|_{op} \left\| (X_{k+1:p} X_{k+1:p}^\top)^{-1} \right\|_{op} \|X_{k+1:p}\boldsymbol{\beta}_2^*\|_2 \\ &\leq 30\sqrt{2} \frac{\sqrt{N \,\mathrm{Tr}(\Sigma_{k+1:p}^2)} + N \|\Sigma_{k+1:p}\|_{op}}{\mathrm{Tr}(\Sigma_{k+1:p})} \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2. \end{aligned}$$

Finally, it follows from Borell's inequality that, conditionnaly on $\mathbb{X}$, for all $t>0$ with probability at least $1-\exp(-t/2)$, $\|D\boldsymbol{\xi}\|_2 \leq \sigma_\xi\left(\sqrt{\operatorname{Tr}(DD^\top)}+\sqrt{t}\,\|D\|_{op}\right)$ where $D=\Sigma_{k+1:p}^{1/2}A$. Moreover, it follows from Bernstein's inequality that if $N\geq 5\log p$, with probability at least $1-c_0\exp(-c_1N)$, for all $j\in[p]$, $\|X_{k+1:p}f_j\|_2 \leq (3/2)\sqrt{N}\left\|\Sigma_{k+1:p}^{1/2}f_j\right\|_2 = 3\sqrt{N\sigma_j}/2$, so

$$\begin{aligned}\operatorname{Tr}(DD^\top) &= \sum_{j=1}^p \|D^\top f_j\|_2^2 = \sum_{j=1}^p \left\|A^\top \Sigma_{k+1:p}^{1/2} f_j\right\|_2^2 = \sum_{j=k+1}^p \|(X_{k+1:p}X_{k+1:p}^\top)^{-1}X_{k+1:p}f_j\|_2^2\sigma_j\\ &\leq \sum_{j=k+1}^p \|(X_{k+1:p}X_{k+1:p}^\top)^{-1}\|_{op}^2\|X_{k+1:p}f_j\|_2^2\sigma_j \leq \frac{20N}{\operatorname{Tr}^2(\Sigma_{k+1:p})}\sum_{j=k+1}^p\sigma_j^2 = \frac{20N\operatorname{Tr}(\Sigma_{k+1:p}^2)}{\operatorname{Tr}^2(\Sigma_{k+1:p})}\end{aligned}$$

and

$$\begin{aligned}\|D\|_{op} &= \left\|\Sigma_{k+1:p}^{1/2}X_{k+1:p}^\top (X_{k+1:p}X_{k+1:p}^\top)^{-1}\right\|_{op} \leq \left\|\Sigma_{k+1:p}^{1/2}X_{k+1:p}^\top\right\|_{op}\left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1}\right\|_{op}\\ &\leq \frac{12}{\operatorname{Tr}(\Sigma_{k+1:p})}\left(\sqrt{\operatorname{Tr}(\Sigma_{k+1:p}^2)}+\sqrt{N}\,\|\Sigma_{k+1:p}\|_{op}\right).\end{aligned}$$

As a consequence, if $N\geq 5\log p$ then for all $t>0$ with probability at least $1-\exp(-t/2)$,

$$\left\|\Sigma_{k+1:p}^{1/2}A\boldsymbol{\xi}\right\|_2 \leq \sigma_\xi\left(\frac{\sqrt{20N\operatorname{Tr}(\Sigma_{k+1:p}^2)}}{\operatorname{Tr}(\Sigma_{k+1:p})}+\frac{12\sqrt{t}}{\operatorname{Tr}(\Sigma_{k+1:p})}\left(\sqrt{\operatorname{Tr}(\Sigma_{k+1:p}^2)}+\sqrt{N}\,\|\Sigma_{k+1:p}\|_{op}\right)\right).$$

Gathering the last three upper bounds in (49), we obtain the following result on the cost for noise interpolation.

Proposition 5. *There are absolute constants c_0 and c_1 such that the following holds. We assume that $N\geq 5\log p$ and that there exists $k\leq \kappa_{iso}N$ such that $N\leq \kappa_{DM}d_*(\Sigma_{k+1:p}^{-1/2}B_2^p)$, then the following holds for all such k's. For all $t>0$, with probability at least $1-c_0\exp(-c_1N)-\exp(-t/2)$,*

$$\begin{aligned}\left\|\Sigma_{k+1:p}^{1/2}(\hat{\boldsymbol{\beta}}_{k+1:p}-\boldsymbol{\beta}_2^*)\right\|_2 &\leq 40\sqrt{2}\frac{\sqrt{N\operatorname{Tr}(\Sigma_{k+1:p}^2)}+N\,\|\Sigma_{k+1:p}\|_{op}}{\operatorname{Tr}(\Sigma_{k+1:p})}\left(\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2+\left\|\Sigma_{1:k}^{1/2}(\boldsymbol{\beta}_1^*-\hat{\boldsymbol{\beta}}_{1:k})\right\|_2\right)+\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2\\ &+\sigma_\xi\left(\frac{\sqrt{20N\operatorname{Tr}(\Sigma_{k+1:p}^2)}}{\operatorname{Tr}(\Sigma_{k+1:p})}+\frac{12\sqrt{t}}{\operatorname{Tr}(\Sigma_{k+1:p})}\left(\sqrt{\operatorname{Tr}(\Sigma_{k+1:p}^2)}+\sqrt{N}\,\|\Sigma_{k+1:p}\|_{op}\right)\right).\end{aligned}$$

6.3 End of the proof of Theorem 5.

Parameter k is chosen so that $N\|\Sigma_{k+1:p}\|_{op}\leq \operatorname{Tr}(\Sigma_{k+1:p})$ (because $\kappa_{DM}\leq 1$), in particular, $\sqrt{N\operatorname{Tr}(\Sigma_{k+1:p}^2)}\leq \operatorname{Tr}(\Sigma_{k+1:p})$. As a consequence, under the assumption of Proposition 5, we have, with probability at least $1-c_0\exp(-c_1N)-\exp(-t/2)$,

$$\begin{aligned}\left\|\Sigma_{k+1:p}^{1/2}(\hat{\boldsymbol{\beta}}_{k+1:p}-\boldsymbol{\beta}_2^*)\right\|_2 \leq& 80\sqrt{2}\left\|\Sigma_{1:k}^{1/2}(\boldsymbol{\beta}_1^*-\hat{\boldsymbol{\beta}}_{1:k})\right\|_2+(80\sqrt{2}+1)\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2+\sigma_\xi\frac{\sqrt{20N\operatorname{Tr}(\Sigma_{k+1:p}^2)}}{\operatorname{Tr}(\Sigma_{k+1:p})}\\ &+\sigma_\xi\left(\frac{12\sqrt{t}}{\operatorname{Tr}(\Sigma_{k+1:p})}\left(\sqrt{\operatorname{Tr}(\Sigma_{k+1:p}^2)}+\sqrt{N}\,\|\Sigma_{k+1:p}\|_{op}\right)\right).\end{aligned} \tag{52}$$

Theorem 5 follows from the last result by choosing $t=N\operatorname{Tr}(\Sigma_{1:k})/\operatorname{Tr}(\Sigma_{k+1:p})$ when $\sigma_1N\leq\operatorname{Tr}(\Sigma_{k+1:p})$ and $t=|J_1|+N\left(\sum_{j\in J_2}\sigma_j\right)/(\operatorname{Tr}(\Sigma_{k+1:p}))$ when $\sigma_1N\geq\operatorname{Tr}(\Sigma_{k+1:p})$. Because, in the first case, when $t=N\operatorname{Tr}(\Sigma_{1:k})/\operatorname{Tr}(\Sigma_{k+1:p})$ we have $t\leq N$ and

$$\frac{\sqrt{tN}\,\|\Sigma_{k+1:p}\|_{op}}{\operatorname{Tr}(\Sigma_{k+1:p})}\leq\frac{\sqrt{N\operatorname{Tr}(\Sigma_{k+1:p}^2)}}{\operatorname{Tr}(\Sigma_{k+1:p})}$$

and when $t=|J_1|+N\left(\sum_{j\in J_2}\sigma_j\right)/(\operatorname{Tr}(\Sigma_{k+1:p}))$ we have $t\leq N$ and

$$\frac{\sqrt{tN}\,\|\Sigma_{k+1:p}\|_{op}}{\operatorname{Tr}(\Sigma_{k+1:p})}\leq\sqrt{\frac{|J_1|}{N}}+\sqrt{\frac{\sum_{j\in J_2}\sigma_j}{\operatorname{Tr}(\Sigma_{k+1:p})}}.$$

Therefore, in both cases, the term in (52) is negligible in front of $12\sigma_\xi\sqrt{N\,\mathrm{Tr}(\Sigma_{k+1:p}^2)}/\,\mathrm{Tr}(\Sigma_{k+1:p})$ and $\square$ defined in Proposition 4.

7 Proof of Theorem 6

The proof of Theorem 6 relies on the decomposition given in Proposition 3 of the estimator $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\beta}}_2$ like in the proof of Theorem 5. It therefore follows the same path as the one of the proof in the previous section (we will therefore use the same notation and detail only the main differences); in particular, it uses the excess risk decomposition (20). However, because $k > N$, $X_{1:k}$ cannot act anymore as an isomorphy over the entire space $V_{1:k}$ since $V_{1:k}$ is of dimension k and $X_{1:k}$ has only $N < k$ rows and is therefore of rank at most N; in particular, $\ker(X_{1:k}) \cap V_{1:k}$ is none trivial.

However, following Theorem 4 there is a cone in $V_{1:k}$ onto which $X_{1:k}$ behaves like an isomorphy; it is given by $\mathcal{C} := \left\{ v \in V_{1:k} : R_N(\Sigma_{1:k}^{1/2} B_2^p)\left\| v \right\|_2 \leq \left\| \Sigma_{1:k}^{1/2} v \right\|_2 \right\}$. We will use this *restricted isomorphy property* to lower bound the quadratic process $\mathcal{Q}_{\boldsymbol{\beta}_1}$ in *case a)* (it is the case defined in the previous section where the quadratic process is used to dominate the multiplier process); we will therefore need $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_1^* \in \mathcal{C}$ in *case a)*. The other difference with the proof from the previous section (that is for the case $k < N$) deals with the upper bound on the multiplier process: we will use an upper bound on $\left\| \tilde{\Sigma}_{1:k}^{-1/2} X_{1:k}^\top \right\|_{op}$ that does not follow from an isomorphic property of $X_{1:k}$ but from the upper side of DM's theorem like the one given in Proposition 2. The stochastic properties we use on $X_{k+1:p}$ in the case $k > N$ are the same as the one used in the previous proof since having $k < N$ or $k > N$ does not play any role on the behavior of $X_{k+1:p}$.

We gather all the stochastic properties we need on $X_{1:k}$ and $X_{k+1:p}$ to obtain our estimation results for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ in the case $k > N$ in the following event.

Stochastic event behind Theorem 6. We denote by Ω_0' the event onto which the following three geometric properties hold:

- for all $\boldsymbol{\lambda} \in \mathbb{R}^N$, $(1/(2\sqrt{2}))\sqrt{\mathrm{Tr}(\Sigma_{k+1:p})}\left\| \boldsymbol{\lambda} \right\|_2 \leq \left\| X_{k+1:p}^\top \boldsymbol{\lambda} \right\|_2 \leq (3/2)\sqrt{\mathrm{Tr}(\Sigma_{k+1:p})}\left\| \boldsymbol{\lambda} \right\|_2$
- for all $\boldsymbol{\beta}_1$ in the cone $\mathcal{C} := \left\{ v \in V_{1:k} : R_N(\Sigma_{1:k}^{1/2} B_2^p)\left\| v \right\|_2 \leq \left\| \Sigma_{1:k}^{1/2} v \right\|_2 \right\}$,
$$\frac{1}{2}\left\| \Sigma_{1:k}^{1/2} \boldsymbol{\beta}_1 \right\|_2 \leq \frac{1}{\sqrt{N}}\left\| X_{1:k}\boldsymbol{\beta}_1 \right\|_2 \leq \frac{3}{2}\left\| \Sigma_{1:k}^{1/2} \boldsymbol{\beta}_1 \right\|_2,$$
- for all $\boldsymbol{\lambda} \in \mathbb{R}^N$,
$$\left\| \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top \boldsymbol{\lambda} \right\|_2 \leq 6\left(\sqrt{\square^2 |J_1| + \triangle^2 \sum_{j\in J_2} \sigma_j} + \sqrt{N}\sigma(\square, \triangle) \right)\left\| \boldsymbol{\lambda} \right\|_2 \tag{53}$$
where $\tilde{\Sigma}_1^{1/2}$ has been introduced in (36) and $\sigma(\square, \triangle)$ in Lemma 1. However, in the following we will always assume that
$$\sqrt{\square^2 |J_1| + \triangle^2 \sum_{j\in J_2} \sigma_j} \leqslant \sqrt{N}\sigma(\square, \triangle) \tag{54}$$
where $\sigma(\square, \triangle)$ is defined in Lemma 1. We will explain the reason in Section 7.1.

If $N \leq \kappa_{DM} d_*(\Sigma_{k+1:p}^{-1/2} B_2^p)$, we know from Theorem 3 and Theorem 4 that the first two points defining the event Ω_0' hold simultaneously with probability at least $1 - c_0\exp(-c_1 N)$. It only remains to handle the last point. To that end we use Proposition 2: with probability at least $1 - c_0\exp(-c_1 N)$, for all $\boldsymbol{\lambda} \in \mathbb{R}^N$,

$$\left\| \tilde{\Sigma}_{1:k}^{-1/2} X_{1:k}^\top \boldsymbol{\lambda} \right\|_2 = \left\| \tilde{\Sigma}_{1:k}^{-1/2}\Sigma_{1:k}^{1/2}\mathbb{G}^\top \boldsymbol{\lambda} \right\|_2 \leq 2\left(\sqrt{\mathrm{Tr}\left((\tilde{\Sigma}_{1:k}^{-1/2}\Sigma_{1:k}^{1/2})(\tilde{\Sigma}_{1:k}^{-1/2}\Sigma_{1:k}^{1/2})^\top \right)} + \sqrt{c_2 N}\left\| \tilde{\Sigma}_{1:k}^{-1/2}\Sigma_{1:k}^{1/2} \right\|_{op} \right)\left\| \boldsymbol{\lambda} \right\|_2.$$

where $\mathbb{G}$ is a $N \times p$ standard Gaussian matrix. The spectrum of $\tilde{\Sigma}_{1:k}^{-1/2}\Sigma_{1:k}^{1/2}$ is the same as the one of $\tilde{D}_1^{-1/2} D_{1:k}^{1/2}$ given by $p - k$ zeros and $\left(\square\sqrt{\sigma_j}/\max(\sqrt{\sigma_j}, (\square/\triangle)) : j = 1, \ldots, k\right)$. Therefore, the third point of event Ω_0' holds with probability at least $1 - \exp(-N)$. We conclude that $\mathbb{P}[\Omega_0'] \geq 1 - 2c_0\exp(-c_1 N)$.

Note that for the choice of $\Box$ and $\triangle$ such that $(\Box/\triangle)^2 = \mathrm{Tr}(\Sigma_{k+1:p})/N$ that we will make later, (54) is equivalent to

$$\begin{cases} \sum_{j\in J_2}\sigma_j \leq \mathrm{Tr}(\Sigma_{k+1:p})\left(1-\frac{|J_1|}{N}\right) & \text{when } N\sigma_1 \geq \mathrm{Tr}(\Sigma_{k+1:p}) \\ \mathrm{Tr}(\Sigma_{1:k}) \leq N\sigma_1 & \text{otherwise.} \end{cases} \tag{55}$$

We place ourselves on the event Ω_0' up to the end of the proof. As in the proof of Theorem 5, we split our analysis into three subsections: one for the study of $\hat{\boldsymbol{\beta}}_1$, one for $\hat{\boldsymbol{\beta}}_2$ and the last one where the two previous sections are merged.

7.1 Estimation properties of the 'ridge estimator' $\hat{\boldsymbol{\beta}}_{1:k}$; case $k \gtrsim N$

We are now providing some details on the arguments we need to prove Theorem 6 which are different from the one of Theorem 5. Let us first handle the lower bound we need on the quadratic process in *case a)* that is when $\boldsymbol{\beta}_1$ is such that $\left\|\Sigma^{1/2}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right\|_2 = \Box$ and $\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2 \leq \triangle$. We can only use the restricted isomorphy property satisfied by $X_{1:k}$ over the cone $\mathcal{C}$ so we have to insure that $\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*$ lies in that cone. That is the case when $\Box/\triangle \geq R_N(\Sigma_{1:k}^{1/2}B_2^p)$ since, in that case, we have

$$R_N(\Sigma_{1:k}^{1/2}B_2^p)\,\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*\|_2 \leq R_N(\Sigma_{1:k}^{1/2}B_2^p)\triangle = R_N(\Sigma_{1:k}^{1/2}B_2^p)\frac{\triangle}{\Box}\left\|\Sigma^{1/2}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right\|_2 \leq \left\|\Sigma^{1/2}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right\|_2.$$

That is the reason why, in the case $k>N$ we do have the extra condition $\Box/\triangle \geq R_N(\Sigma_{1:k}^{1/2}B_2^p)$ (note that when $k \leq \kappa_{RIP}N$ then $R_N(\Sigma_{1:k}^{1/2}B_2^p) = 0$ and so there is no need for this condition). As a consequence, when $\Box/\triangle \geq R_N(\Sigma_{1:k}^{1/2}B_2^p)$ we have $\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^* \in \mathcal{C}$ in *case a)* and so (on the event Ω_0'),

$$\mathcal{Q}_{\boldsymbol{\beta}_1} = \left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1/2}X_{1:k}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\right\|_2^2 \geq \frac{1}{8\,\mathrm{Tr}(\Sigma_{k+1:p})}\|X_{1:k}(\boldsymbol{\beta}_1-\boldsymbol{\beta}_1^*)\|_2^2 \geq \frac{N\Box^2}{32\,\mathrm{Tr}(\Sigma_{k+1:p})}.$$

We therefore recover the same lower bound as in the proof of Theorem 5 (see (40)).

Let us now handle the multiplier process. It follows from (37) that we need to upper bound the two quantities $\left\|\tilde{\Sigma}_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}X_{k+1:p}\boldsymbol{\beta}_2^*\right\|_2$ and $\left\|\tilde{\Sigma}_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}\boldsymbol{\xi}\right\|_2$ as we did in Lemma 1 and Lemma 2 but without the isomorphic property of $X_{1:k}$ on $V_{1:k}$. We know from the third point of the event Ω_0' that

$$\left\|\tilde{\Sigma}_{1:k}^{-1/2}X_{1:k}^\top\right\|_{op} \leq 6\sqrt{N}\sigma(\Box,\triangle). \tag{56}$$

If we use either $\sqrt{\Box^2|J_1| + \triangle^2\sum_{j\in J_2}\sigma_j}+\sqrt{N}\sigma(\Box,\triangle)$ or $\sqrt{\Box^2|J_1| + \triangle^2\sum_{j\in J_2}\sigma_j}$ as the upper bound of $\left\|\tilde{\Sigma}_1^{-1/2}X_{1:k}^\top\right\|_{op}$, this will lead to the symmetric of the true noise σ_ξ and the "formal" noise $\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2$, which further leads to the fact that: all the terms containing σ_ξ or $\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2$ in Proposition 4 has to be replaced by $\sigma_\xi \vee \left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2$. Therefore, if the two noises are symmetric, there will never be benign overfitting phenomenon. The aim of using $\sqrt{N}\sigma(\Box,\triangle)$ instead of $\sqrt{\Box^2|J_1| + \triangle^2\sum_{j\in J_2}\sigma_j}$ is to keep them asymmetric as that in section 6.1.

As a consequence, using similar argument as in the proof of Lemma 1, we have with probability at least $1-c_0\exp(-c_1N)$,

$$\begin{aligned}&\left\|\tilde{\Sigma}_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}X_{k+1:p}\boldsymbol{\beta}_2^*\right\|_2 \leq \left\|\tilde{\Sigma}_{1:k}^{-1/2}X_{1:k}^\top\right\|_{op}\left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1}\right\|_{op}\|X_{k+1:p}\boldsymbol{\beta}_2^*\|_2 \\ &\leq \sigma(\Box,\triangle)\frac{54N}{\mathrm{Tr}(\Sigma_{k+1:p})}\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2\end{aligned} \tag{57}$$

which is up to absolute constants the same result as in Lemma 1.

Next, we prove a high probability upper bound on $\left\|\tilde{\Sigma}_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}\boldsymbol{\xi}\right\|_2$. It follows from Borell's inequality that for all $t>0$, with probability at least $1-\exp(-t/2)$,

$$\left\|\tilde{\Sigma}_1^{-1/2}X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}\boldsymbol{\xi}\right\|_2 \leq \sigma_\xi\left(\sqrt{\mathrm{Tr}(DD^\top)}+\sqrt{t}\,\|D\|_{op}\right)$$

where $D = \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1}$. We know from Bernstein's inequality that with probability at least $1 - \exp(-c_1 N)$, for all $j = 1, \cdots, k$ $\|X_{1:k} f_j\|_2 \le 2\sqrt{N}$ as long as $N \ge c_2 \log(k)$. Hence, using this latter argument in place of the 'isomorphic' argument used in the proof of Lemma 2 we have with probability at least $1 - c_0 \exp(-c_1 N)$,

$$\sqrt{\mathrm{Tr}(DD^\top)} \le \frac{16\sqrt{N}}{\mathrm{Tr}(\Sigma_{k+1:p})} \sqrt{|J_1| \square^2 + \triangle^2 \sum_{j \in J_2} \sigma_j}.$$

We also have

$$\|D\|_{op} \le \left\| \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top \right\|_{op} \left\| (X_{k+1:p} X_{k+1:p}^\top)^{-1} \right\|_{op} \le \frac{24\sqrt{N}\sigma(\square, \triangle)}{\mathrm{Tr}(\Sigma_{k+1:p})}.$$

Finally, using the same other arguments as in the proof of Lemma 2 we obtain the very same bound as in Lemma 2: with probability at least $1 - c_0 \exp(-c_1 N) - \exp(-t(\square, \triangle)/2)$,

$$\left\| \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top (X_{k+1:p} X_{k+1:p}^\top)^{-1} \boldsymbol{\xi} \right\|_2 \le \frac{32\sqrt{N}\sigma_\xi}{\mathrm{Tr}(\Sigma_{k+1:p})} \sqrt{|J_1| \square^2 + \triangle^2 \sum_{j \in J_2} \sigma_j} \tag{58}$$

which is up to absolute constants the same result as in Lemma 2.

As a consequence, all the machinery used in Section 6.1 also applies for the same choice of $\square$ and $\triangle$(up to absolute constants) under the three extra conditions that $\square/\triangle \geqslant R_N(\Sigma_{1:k}^{1/2} B_2^p)$, that (55) and $N \ge c_0 \log k$. We therefore end up with almost the same result as in Proposition 4:

Proposition 6. *There are absolute constants c_0, c_1 and C_0 such that the following holds. We assume that there exists $k \in [p]$ such that $N \le \kappa_{DM} d_*(\Sigma_{k+1:p}^{-1/2} B_2^p)$, (55) holds and $\mathrm{Tr}(\Sigma_{k+1:p}) \ge R_N(\Sigma_{1:k}^{1/2} B_2^p)^2 N$, then the following holds for all such k's. We define*

$$J_1 := \left\{ j \in [k] : \sigma_j \ge \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N} \right\}, \quad J_2 := \left\{ j \in [k] : \sigma_j < \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N} \right\}$$

and $\Sigma_{1,thres}^{-1/2} := U D_{1,thres}^{-1/2} U^\top$ where

$$D_{1,thres}^{-1/2} := \mathrm{diag}\left(\left(\sigma_1 \vee \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N} \right)^{-1/2}, \dots, \left(\sigma_k \vee \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N} \right)^{-1/2}, 0, \dots, 0 \right).$$

With probability at least $1 - p^*$, $\left\| \Sigma_1^{1/2} (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^*) \right\|_2 \le \square$ and $\left\| \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1^* \right\|_2 \le \square \sqrt{N / \mathrm{Tr}(\Sigma_{k+1:p})}$ where,

i) if $\sigma_1 N \le \mathrm{Tr}(\Sigma_{k+1:p})$, $p^* = c_0 \exp(-c_1 N)$ and

$$\square = C_0^2 \max \left\{ \sigma_\xi \sqrt{\frac{\mathrm{Tr}(\Sigma_{1:k})}{\mathrm{Tr}(\Sigma_{k+1:p})}}, \sqrt{\frac{N\sigma_1}{\mathrm{Tr}(\Sigma_{k+1:p})}} \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2, \|\boldsymbol{\beta}_1^*\|_2 \sqrt{\frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N}} \right\}$$

ii) if $\sigma_1 N \ge \mathrm{Tr}(\Sigma_{k+1:p})$, $p^* = c_0 \exp\left(-c_1 \left(|J_1| + N \left(\sum_{j \in J_2} \sigma_j \right) / (\mathrm{Tr}(\Sigma_{k+1:p})) \right) \right)$ and

$$\square = C_0^2 \max \left\{ \sigma_\xi \sqrt{\frac{|J_1|}{N}}, \sigma_\xi \sqrt{\frac{\sum_{j \in J_2} \sigma_j}{\mathrm{Tr}(\Sigma_{k+1:p})}}, \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2, \left\| \Sigma_{1,thres}^{-1/2} \boldsymbol{\beta}_1^* \right\|_2 \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N} \right\}.$$

Remark 2. *If we use*

$$\left\| \tilde{\Sigma}_1^{-1/2} X_{1:k}^\top \right\|_{op} \le 6 \left(\sqrt{\square^2 |J_1| + \triangle^2 \sum_{j \in J_2} \sigma_j} \right),$$

which holds when $N \le \kappa_{DM} t(\square, \triangle)$ instead of (53), we will end up with

$$\square = C_0^2 \max \left\{ \left(\sigma_\xi \vee \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2 \right) \sqrt{\frac{|J_1|}{N}}, \left(\sigma_\xi \vee \left\| \Sigma_{k+1:p}^{1/2} \boldsymbol{\beta}_2^* \right\|_2 \right) \sqrt{\frac{\sum_{j \in J_2} \sigma_j}{\mathrm{Tr}(\Sigma_{k+1:p})}}, \left\| \Sigma_{1,thres}^{-1/2} \boldsymbol{\beta}_1^* \right\|_2 \frac{\mathrm{Tr}(\Sigma_{k+1:p})}{N} \right\}.$$

However, the latter rate does not showcase the benign overfitting phenomenom. Indeed, if one has $N \le \kappa_{DM} t(\square, \triangle)$ then

- *when* $\sigma(\Box,\triangle)=\Box$*: in this case,* $t(\Box,\triangle)=|J_1|+N\sum_{j\in J_2}\sigma_j/\operatorname{Tr}(\Sigma_{k+1:p})\gtrsim N$*, which leads to* $\sum_{j\in J_2}\sigma_j/\operatorname{Tr}(\Sigma_{k+1:p})\geqslant 1-|J_1|/N$*. As* $|J_1|/N$ *tends to* 0*, we have* $\sum_{j\in J_2}\sigma_j/\operatorname{Tr}(\Sigma_{k+1:p})$ *tends to* 1*, on which we will never have BO.*
- *when* $\sigma(\Box,\triangle)=\triangle\sqrt{\sigma_1}$*, which is equivalent to* $\sigma_1 N\leqslant \operatorname{Tr}(\Sigma_{k+1:p})$*. In this case,* $\operatorname{Tr}(\Sigma_{1:k})/\sigma_1=t(\Box,\triangle)\gtrsim N$*. Therefore,* $\operatorname{Tr}(\Sigma_{1:k})/\operatorname{Tr}(\Sigma_{k+1:p})\geqslant\sigma_1 N/\operatorname{Tr}(\Sigma_{k+1:p})$*. However, as* $\operatorname{Tr}(\Sigma_{k+1:p})/N$ *tends to* 0*, this will not converge.*

7.2 Upper bound on the price for noise interpolation; the case $k\gtrsim N$

The aim of this section is to obtain a high probability bound on $\left\|\Sigma_{k+1:p}^{1/2}(\hat{\boldsymbol\beta}_{k+1:p}-\boldsymbol\beta^*_{k+1:p})\right\|_2$. We use the same decomposition as in (22) together with the closed-form of $\hat{\boldsymbol\beta}_{k+1:p}$ given in Proposition 3:

$$\begin{aligned}\left\|\Sigma_{k+1:p}^{1/2}(\hat{\boldsymbol\beta}_{k+1:p}-\boldsymbol\beta^*_{k+1:p})\right\|_2&\le\left\|\Sigma_{k+1:p}^{1/2}A(y-X_{1:k}\hat{\boldsymbol\beta}_{1:k})\right\|_2+\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol\beta^*_{k+1:p}\right\|_2\\&\le\left\|\Sigma_{k+1:p}^{1/2}AX_{1:k}(\boldsymbol\beta^*_1-\hat{\boldsymbol\beta}_{1:k})\right\|_2+\left\|\Sigma_{k+1:p}^{1/2}AX_{k+1:p}\boldsymbol\beta^*_2\right\|_2+\left\|\Sigma_{k+1:p}^{1/2}A\boldsymbol\xi\right\|_2+\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol\beta^*_{k+1:p}\right\|_2\end{aligned}\tag{59}$$

where $A=X_{k+1:p}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}$. We will handle the last three terms of (59) as in the Section 6.2 because they do not depend on $X_{1:k}$. However, for the first term in (59), we used in the proof of Theorem 5 the isomorphic property of $X_{1:k}$ in (51): $\left\|X_{1:k}(\boldsymbol\beta^*_1-\hat{\boldsymbol\beta}_{1:k})\right\|_2\le 3\sqrt{N}\left\|\Sigma_{1:k}^{1/2}(\boldsymbol\beta^*_1-\hat{\boldsymbol\beta}_{1:k})\right\|_2/2$. We cannot use this isomorphic property in the case $k\gtrsim N$ since it does not hold on the entire space $V_{1:k}$. However, this last inequality still holds if one can show that $\boldsymbol\beta^*_1-\hat{\boldsymbol\beta}_{1:k}$ lies in the cone $\mathcal{C}$; that is to show that $R_N(\Sigma_{1:k}^{1/2}B_2^p)\left\|\boldsymbol\beta^*_1-\hat{\boldsymbol\beta}_{1:k}\right\|_2\le\left\|\Sigma_{1:k}^{1/2}(\boldsymbol\beta^*_1-\hat{\boldsymbol\beta}_{1:k})\right\|_2$. This type of condition usually holds when we regularize by a norm (this is for instance the case of the LASSO) but it is not clear if this holds when the regularization function is the *square of a norm* as it is the case for the 'ridge estimator' $\hat{\boldsymbol\beta}_1$. Therefore, we cannot use this argument here. The way we will handle this issue is by showing an upper bound on $\left\|X_{1:k}(\boldsymbol\beta^*_1-\hat{\boldsymbol\beta}_{1:k})\right\|_2$ directly without going through a norm equivalence with $\left\|\Sigma_1^{1/2}\cdot\right\|_2$.

We know from Proposition 3 that $\mathcal{L}_{\hat{\boldsymbol\beta}_1}\le 0$ where $\mathcal{L}_{\hat{\boldsymbol\beta}_1}=\mathcal{Q}_{\hat{\boldsymbol\beta}_1}+\mathcal{M}_{\hat{\boldsymbol\beta}_1}+\mathcal{R}_{\hat{\boldsymbol\beta}_1}$ and for all $\boldsymbol\beta_1\in V_{1:k}$,

$$\mathcal{Q}_{\boldsymbol\beta_1}=\left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1/2}X_{1:k}(\boldsymbol\beta_1-\boldsymbol\beta^*_1)\right\|_2^2,\mathcal{M}_{\boldsymbol\beta_1}=2\left\langle X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}(X_{k+1:p}\boldsymbol\beta^*_2+\boldsymbol\xi)-\boldsymbol\beta^*_1,\boldsymbol\beta_1-\boldsymbol\beta^*_1\right\rangle$$

and $\mathcal{R}_{\boldsymbol\beta_1}=\|\boldsymbol\beta_1-\boldsymbol\beta^*_1\|_2^2$. We therefore have $\mathcal{Q}_{\hat{\boldsymbol\beta}_1}+\mathcal{R}_{\hat{\boldsymbol\beta}_1}\le|\mathcal{M}_{\hat{\boldsymbol\beta}_1}|$. We are now proving a lower bound on $\mathcal{Q}_{\hat{\boldsymbol\beta}_1}$ and an upper bound on $|\mathcal{M}_{\hat{\boldsymbol\beta}_1}|$. For the lower bound on $\mathcal{Q}_{\hat{\boldsymbol\beta}_1}$, we use that on Ω_0',

$$\mathcal{Q}_{\hat{\boldsymbol\beta}_1}=\left\|(X_{k+1:p}X_{k+1:p}^\top)^{-1/2}X_{1:k}(\hat{\boldsymbol\beta}_1-\boldsymbol\beta^*_1)\right\|_2^2\ge\frac{1}{32\operatorname{Tr}(\Sigma_{k+1:p})}\left\|X_{1:k}(\hat{\boldsymbol\beta}_1-\boldsymbol\beta^*_1)\right\|_2^2.\tag{60}$$

For the upper bound on $|\mathcal{M}_{\hat{\boldsymbol\beta}_1}|$, we consider the norm (restricted to $V_{1:k}$) defined for all $\boldsymbol\beta\in V_{1:k}$

$$|||\boldsymbol\beta|||:=\max\left(\left\|\Sigma_{1:k}^{1/2}\boldsymbol\beta\right\|_2,\sqrt{\frac{\operatorname{Tr}(\Sigma_{k+1:p})}{N}}\|\boldsymbol\beta\|_2\right)\tag{61}$$

and we set $|||\boldsymbol\beta|||=0$ for all $\boldsymbol\beta\in V_{k+1:p}$. On the event Ω_0', we have $\left|\left|\left|\hat{\boldsymbol\beta}_1-\boldsymbol\beta^*_1\right|\right|\right|\le\Box$ where $\Box$ is defined in Proposition 6. Therefore, if we define $\theta:=\sup\left(|\mathcal{M}_{\boldsymbol\beta_1}|:|||\boldsymbol\beta_1-\boldsymbol\beta^*_1|||\le 1\right)$ we have

$$\left\|\hat{\boldsymbol\beta}_1-\boldsymbol\beta^*_1\right\|_2^2+\frac{1}{32\operatorname{Tr}(\Sigma_{k+1:p})}\|X_{1:k}(\boldsymbol\beta_1-\boldsymbol\beta^*_1)\|_2^2\le\theta\Box.$$

and so we have $\left\|X_{1:k}(\hat{\boldsymbol\beta}_1-\boldsymbol\beta^*_1)\right\|_2\leqslant 4\sqrt{2}\sqrt{\theta\Box\operatorname{Tr}(\Sigma_{k+1:p})}$ (we also have $\left\|\hat{\boldsymbol\beta}_1-\boldsymbol\beta^*_1\right\|_2\le\sqrt{\theta\Box}$ however we will not use it here). The last result we need to prove is a high probability upper bound on θ. It appears that we already did it in the previous Section 7.1 since for the choice of $\Box$ and $\triangle$ such that $(\Box/\triangle)^2=\operatorname{Tr}(\Sigma_{k+1:p})/N$ we have $\Box|||\boldsymbol\beta|||=|||\boldsymbol\beta|||$ for all $\boldsymbol\beta\in\mathbb{R}^p$. Hence, if we denote by $|||\cdot|||_*$ the dual norm of $|||\cdot|||$, we have $|||\cdot|||_*=\Box|||\cdot|||_*$ and so $|||\cdot|||_*$ is equivalent to $\Box^{-1}\left\|\tilde\Sigma_1^{-1/2}\cdot\right\|_2$ where $\tilde\Sigma_1^{-1/2}$ is defined in (36). Therefore, using (57) and (58) in the decomposition

$$\theta/2\le|||\boldsymbol\beta^*_1|||^*+\left|\left|\left|X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}X_{k+1:p}\boldsymbol\beta^*_2\right|\right|\right|^*+\left|\left|\left|X_{1:k}^\top(X_{k+1:p}X_{k+1:p}^\top)^{-1}\boldsymbol\xi\right|\right|\right|^*$$

we obtain the following result:

Proposition 7. *Under the same assumptions as in Proposition 6, with probability at least* $1-p^*$, $\left\|X_{1:k}(\hat{\boldsymbol{\beta}}_{1:k}-\boldsymbol{\beta}_1^*)\right\|_2 \leq 4\sqrt{N}\square$ *where* p^* *and* $\square$ *are defined in Proposition 6.*

The following result follows from the decomposition (59) and Section 6.2 and Proposition 7, for the control of the four terms in this decomposition. Namely, if $N \geqslant 5\log p$, then

$$\left\|\Sigma_{k+1:p}^{1/2}AX_{k+1:p}\boldsymbol{\beta}_2^*\right\|_2 \leq 30\sqrt{2}\frac{\sqrt{N\,\mathrm{Tr}(\Sigma_{k+1:p}^2)}+N\left\|\Sigma_{k+1:p}\right\|_{op}}{\mathrm{Tr}(\Sigma_{k+1:p})}\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2,$$

and

$$\left\|\Sigma_{k+1:p}^{1/2}A\boldsymbol{\xi}\right\|_2 \leq \sigma_\xi\left(\frac{\sqrt{20N\,\mathrm{Tr}(\Sigma_{k+1:p}^2)}}{\mathrm{Tr}(\Sigma_{k+1:p})}+\frac{12\sqrt{t}}{\mathrm{Tr}(\Sigma_{k+1:p})}\left(\sqrt{\mathrm{Tr}(\Sigma_{k+1:p}^2)}+\sqrt{N}\left\|\Sigma_{k+1:p}\right\|_{op}\right)\right)$$

holds with probability at least $1-\exp(-t/2)-c_0\exp(-c_1N)-\mathbb{P}[(\Omega_0'\cap\Omega_1)^c]$, where Ω_1 is defined in (50).

Proposition 8. *There are absolute constants* c_0 *and* c_1 *such that the following holds. We assume that* $N\geq 5\log p$ *and that there exists* $k>N$ *such that* $N\leq \kappa_{DM}d_*(\Sigma_{k+1:p}^{-1/2}B_2^p))$*, then the following holds for all such* k*'s. For all* $t>0$*, with probability at least* $1-2c_0\exp(-c_1N)-\exp(-t(\square,\triangle)/2)-\exp(-t/2)$*,*

$$\begin{aligned}\left\|\Sigma_{k+1:p}^{1/2}(\hat{\boldsymbol{\beta}}_{k+1:p}-\boldsymbol{\beta}_2^*)\right\|_2 &\leq 40\sqrt{2}\frac{\sqrt{N\,\mathrm{Tr}(\Sigma_{k+1:p}^2)}+N\left\|\Sigma_{k+1:p}\right\|_{op}}{\mathrm{Tr}(\Sigma_{k+1:p})}\left(\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2+\left\|\Sigma_{1:k}^{1/2}(\boldsymbol{\beta}_1^*-\hat{\boldsymbol{\beta}}_{1:k})\right\|_2\right)+\left\|\Sigma_{k+1:p}^{1/2}\boldsymbol{\beta}_2^*\right\|_2\\ &+\sigma_\xi\left(\frac{\sqrt{20N\,\mathrm{Tr}(\Sigma_{k+1:p}^2)}}{\mathrm{Tr}(\Sigma_{k+1:p})}+\frac{12\sqrt{t}}{\mathrm{Tr}(\Sigma_{k+1:p})}\left(\sqrt{\mathrm{Tr}(\Sigma_{k+1:p}^2)}+\sqrt{N}\left\|\Sigma_{k+1:p}\right\|_{op}\right)\right).\end{aligned}$$

7.3 End of the proof of Theorem 6

We use the decomposition of the excess from (20), the result on the estimation property of $\hat{\boldsymbol{\beta}}_1$ from Proposition 6 and the one on $\hat{\boldsymbol{\beta}}_2$ from Proposition 8 to conclude. Since the upper bounds on $\left\|\Sigma_{k+1:p}^{1/2}(\hat{\boldsymbol{\beta}}_{k+1:p}-\boldsymbol{\beta}_2^*)\right\|_2$ and $\left\|\Sigma_{1:k}^{1/2}(\hat{\boldsymbol{\beta}}_{1:k}-\boldsymbol{\beta}_1^*)\right\|_2$ are the same up to absolute constants as that in Section 6.1 and Section 6.2, all the choices of t from Section 6.3 also applies for the case $k\gtrsim N$ under the two extra assumptions: $\mathrm{Tr}(\Sigma_{k+1:p})\geqslant R_N(\Sigma_{1:k}^{1/2}B_2^p)^2N$ and (55).

8 Proof of Theorem 8

The proof of Theorem 8 is based on a slightly adapted randomization argument from Proposition 1 in [36] that allows to replace the Bayesian risk from [1, 39] by the true risk as well as the argument from [1, 39] on $V_{1:k_b^*}$ and the DM theorem on $V_{k_b^*+1:p}$. We denote by Ω^* the event onto which for all $\boldsymbol{\lambda}\in\mathbb{R}^N$,

$$(1/(2\sqrt{2}))\sqrt{\mathrm{Tr}(\Sigma_{k_b^*+1:p})}\left\|\boldsymbol{\lambda}\right\|_2\leq\left\|X_{k_b^*+1:p}^\top\boldsymbol{\lambda}\right\|_2\leq(3/2)\sqrt{\mathrm{Tr}(\Sigma_{k_b^*+1:p})}\left\|\boldsymbol{\lambda}\right\|_2.$$

It follows from Theorem 3 and (9) that when $b\geq 4/\kappa_{DM}$, $\mathbb{P}[\Omega^*]\geq 1-\exp(-c_0N)$.

We start with a bias/variance decomposition of the risk as in [1, 39]. We have

$$\mathbb{E}\left\|\Sigma^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*)\right\|_2^2=\mathbb{E}\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X}-I_p)\boldsymbol{\beta}^*\right\|_2^2+\mathbb{E}\left\|\Sigma^{1/2}\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\boldsymbol{\xi}\right\|_2^2 \quad (62)$$

where we used that $\hat{\boldsymbol{\beta}}=\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}y=\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}(\mathbb{X}\boldsymbol{\beta}^*+\boldsymbol{\xi})$.

For the variance term (second term in the right-hand side of (62)) we use Lemma 2 and Theorem 5 from [39] (see also Lemma 16 in [1] for a similar result).

Proposition 9. *[Lemma 2 and Theorem 5 in [39]]There exists some absolute constant* $c_0>0$ *such that the following holds. Let* $b\geq 1/N$ *and assume that* $k_b^*<N/4$*. We have*

$$\mathbb{E}\left\|\Sigma^{1/2}\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\boldsymbol{\xi}\right\|_2^2\geq\frac{c_0\sigma_\xi^2}{\max((2+b)^2,(1+2b^{-1})^2)}\left(\frac{k_b^*}{N}+\frac{N\,\mathrm{Tr}(\Sigma_{k_b^*+1:p}^2)}{\mathrm{Tr}^2(\Sigma_{k_b^*+1:p})}\right).$$

For the bias term (first term in the right-hand side of (62)), we cannot use the results from [1, 39] because they hold either for one specific $\boldsymbol{\beta}^*$ (see Theorem 4 in [1]) or for some Bayesian risk (see Lemma 3 in [39]). Moreover Lemma 3 in [39] requires some extra assumptions on the smallest singular values of some matrix (see the condition *'...for any $j > k$, w.p. at least $1-\delta$, $\mu_n(A_{-j}) \geq L^{-1}(\sum_{j>k}\sigma_j)$.'* in there). The following lower bound holds for the prediction risk of $\hat{\boldsymbol{\beta}}$ for the estimation of $\boldsymbol{\beta}^*$ itself. It holds for any given $\boldsymbol{\beta}^*$ and not a random one. It also holds under only the assumption that $k_b^* \lesssim N$ (which is a necessary condition for BO according to Theorem 1 in [1]).

Proposition 10. *There is an absolute constant $c_0 > 0$ such that the following holds. If $N \geq c_0$ then for any $b \geq \max(4/\kappa_{DM}, 24)$, we have*

$$\mathbb{E}\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)\boldsymbol{\beta}^*\right\|_2^2 \geq \frac{1}{18b^2}\left(\left\|\Sigma_{k_b^*+1:p}^{1/2}\boldsymbol{\beta}^*_{k_b^*+1:p}\right\|_2^2, \left\|\Sigma_{1:k_b^*}^{-1/2}\boldsymbol{\beta}^*_{1:k_b^*}\right\|_2^2\left(\frac{\mathrm{Tr}(\Sigma_{k_b^*+1:p})}{N}\right)^2\right).$$

Proof of Proposition 10. We show that the bias term can be written as a 'Bayesian bias term' using a similar argument as in Proposition 1 from [36]. Let U be an $p \times p$ orthogonal matrix which commutes with Σ. We have

$$\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)U\boldsymbol{\beta}^*\right\|_2 = \left\|\Sigma^{1/2}((\mathbb{X}U)^\top((\mathbb{X}U)(\mathbb{X}U)^\top)^{-1}(\mathbb{X}U) - I_p)\boldsymbol{\beta}^*\right\|_2^2.$$

Moreover, $\mathbb{X}U$ has the same distribution as $\mathbb{X}$ and so $\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)U\boldsymbol{\beta}^*\right\|_2$ has the same distribution as $\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)\boldsymbol{\beta}^*\right\|_2$. In particular, they have the same expectation with respect to $\mathbb{X}$. Now, let us consider the random matrix $U = \sum_{j=1}^p \varepsilon_j f_j f_j^\top$ where $(\epsilon_j)_{j=1}^p$ is a family of p i.i.d. Rademacher variables (and $(f_j)_j$ is a basis of eigenvectors of Σ). Since $Uf_j = \varepsilon_j f_j$ for all j, U commutes with Σ and it is an orthogonal matrix. Therefore, we get

$$\mathbb{E}\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)\boldsymbol{\beta}^*\right\|_2^2 = \mathbb{E}_U\mathbb{E}_{\mathbb{X}}\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)U\boldsymbol{\beta}^*\right\|_2^2 \tag{63}$$

where $\mathbb{E}_U$ (resp. $\mathbb{E}_{\mathbb{X}}$) denotes the expectation w.r.t. U (resp. $\mathbb{X}$).

We could now use Lemma 3 from [39] to handle the Bayesian bias term from (63) (that is the right hand side term). However, we want to avoid some (unnecessary) conditions required for that result to hold. To do so we use the DM theorem on $V_{k_b^*+1:p}$.

We denote for all $j \in [p], \boldsymbol{\beta}_j^* = \langle \boldsymbol{\beta}^*, f_j\rangle, z_j = \sigma_j^{-1/2}\mathbb{X}f_j$, that is $z_j = \mathbb{G}f_j$ where we recall that $\mathbb{X} = \mathbb{G}\Sigma^{1/2}$. We have

$$\mathbb{E}_U\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)U\boldsymbol{\beta}^*\right\|_2^2 = \sum_{j=1}^p(\boldsymbol{\beta}_j^*)^2\left\|\Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)f_j\right\|_2^2$$

$$\geq \sum_{j=1}^p(\boldsymbol{\beta}_j^*)^2\langle f_j, \Sigma^{1/2}(\mathbb{X}^\top(\mathbb{X}\mathbb{X}^\top)^{-1}\mathbb{X} - I_p)f_j\rangle^2 = \sum_{j=1}^p(\boldsymbol{\beta}_j^*)^2\sigma_j\left(1 - \sigma_j\left\|(\mathbb{X}\mathbb{X}^\top)^{-1/2}z_j\right\|_2^2\right)^2.$$

We lower bound the terms from the overfitting part (i.e. $j \geq k_b^*+1$) using the DM theorem and the one from the estimating part (i.e. $1 \leq j \leq k_b^*$) using the strategy from [39]. Let us start with the overfitting part.

Let $j \geq k_b^*+1$. On the event Ω^*, it follows from Proposition 1 that

$$\left\|(\mathbb{X}\mathbb{X}^\top)^{-1/2}z_j\right\|_2 \leq s_1[(\mathbb{X}\mathbb{X}^\top)^{-1/2}]\left\|z_j\right\|_2 = \frac{\left\|z_j\right\|_2}{\sqrt{s_N[\mathbb{X}\mathbb{X}^\top]}} \leq \frac{2\left\|z_j\right\|_2}{\sqrt{\mathrm{Tr}(\Sigma_{k_b^*+1:p})}}$$

where we used that $\mathbb{X}\mathbb{X}^\top = \mathbb{X}_{1:k_b^*}\mathbb{X}_{1:k_b^*}^\top + \mathbb{X}_{k_b^*+1:p}\mathbb{X}_{k_b^*+1:p}^\top \succeq \mathbb{X}_{k_b^*+1:p}\mathbb{X}_{k_b^*+1:p}^\top \succeq (\mathrm{Tr}(\Sigma_{k_b^*+1:p})/4)I_N$ (thanks to Proposition 1). It follows from Borell's inequality that with probability at least $1-\exp(-N)$, $\left\|z_j\right\|_2 = \left\|\mathbb{G}f_j\right\|_2 \leq 3\sqrt{N}$. Hence, we obtain that with probability at least $1-\exp(-c_0N)$, $\left\|(\mathbb{X}\mathbb{X}^\top)^{-1/2}z_j\right\|_2 \leq 6\sqrt{N/\mathrm{Tr}(\Sigma_{k_b^*+1:p})}$ and so

$$\left(1 - \sigma_j\left\|(\mathbb{X}\mathbb{X}^\top)^{-1/2}z_j\right\|_2^2\right)^2 \geq 1 - 2\sigma_j\left\|(\mathbb{X}\mathbb{X}^\top)^{-1/2}z_j\right\|_2^2 \geq 1 - \frac{12\sigma_{k_b^*+1}N}{\mathrm{Tr}(\Sigma_{k_b^*+1:p})} \geq 1 - \frac{12}{b} \geq \frac{1}{2}$$

when $b \geq 24$. Therefore, when N is larger than some absolute constant so that $1-\exp(-c_0N) \geq 1/2$, we get that

$$\mathbb{E}_{\mathbb{X}}\left[\sum_{j=k_b^*+1}^p(\boldsymbol{\beta}_j^*)^2\sigma_j\left(1 - \sigma_j\left\|(\mathbb{X}\mathbb{X}^\top)^{-1/2}z_j\right\|_2^2\right)^2\right] \geq \frac{1}{4}\sum_{j=k_b^*+1}^p(\boldsymbol{\beta}_j^*)^2\sigma_j = \frac{\left\|\Sigma_{k_b^*+1:p}^{1/2}\boldsymbol{\beta}^*_{k_b^*+1:p}\right\|_2^2}{8}$$

which is the expected lower bound on the overfitting part of the bias term. Let us now turn to the estimation part.

Let $1 \leq j \leq k_b^*$. To obtain the desired lower bound we use the same strategy as in Lemma 3 from [39] together with Borell's inequality and the DM theorem. It follows from the Sherman-Morrison formulae (see the proof of Lemma 15 in [39]) that

$$\sigma_j \left(1 - \sigma_j \left\|(\mathbb{X}\mathbb{X}^\top)^{-1/2} z_j\right\|_2^2\right)^2 = \frac{\sigma_j}{\left(1 + \sigma_j z_j^\top A_{-j}^{-1} z_j\right)^2}$$

where $A_{-j} = \mathbb{X}\mathbb{X}^\top - \sigma_j z_j z_j^\top$. We have $z_j^\top A_{-j}^{-1} z_j \leq \|z_j\|_2 \left\|A_{-j}^{-1} z_j\right\|_2 \leq \|z_j\|_2^2 s_1[A_{-j}^{-1}] = \|z_j\|_2^2 / s_N[A_{-j}]$. Since $j \leq k_b^*$, we see that $A_{-j} \succeq X_{k_b^*+1:p} X_{k_b^*+1:p}^\top$ and so $s_N[A_{-j}] \geq s_N[X_{k_b^*+1:p} X_{k_b^*+1:p}^\top]$. On the event Ω^*, $s_N[X_{k_b^*+1:p} X_{k_b^*+1:p}^\top] \geq \mathrm{Tr}(\Sigma_{k_b^*+1:p})/4$ (see Proposition 1). The last result together with Borell's inequality yields with probability at least $1 - \exp(-c_0 N)$, $z_j^\top A_{-j}^{-1} z_j \leq 36N/\mathrm{Tr}(\Sigma_{k_b^*+1:p})$. Furthermore, by definition of k_b^* and since $j \leq k_b^*$, we have $bN\sigma_j > \mathrm{Tr}(\Sigma_{j:p}) \geqslant \mathrm{Tr}(\Sigma_{k_b^*:p})$. Gathering the two pieces together we get that with probability at least $1 - \exp(-c_0 N)$,

$$\frac{\sigma_j}{\left(1 + \sigma_j z_j^\top A_{-j}^{-1} z_j\right)^2} \geq \frac{\sigma_j}{\left(\frac{(36+b)\sigma_j N}{\mathrm{Tr}(\Sigma_{k_b^*+1:p})}\right)^2} = \frac{\mathrm{Tr}^2(\Sigma_{k_b^*+1:p})}{(36+b)^2 \sigma_j N^2}.$$

Therefore, when N is larger than some absolute constant so that $1 - \exp(-c_0 N) \geq 1/2$, we get that

$$\mathbb{E}_{\mathbb{X}}\left[\sum_{j=1}^{k_b^*} (\boldsymbol{\beta}_j^*)^2 \sigma_j \left(1 - \sigma_j \left\|(\mathbb{X}\mathbb{X}^\top)^{-1/2} z_j\right\|_2^2\right)^2\right] \geq \frac{1}{2(36+b)^2} \sum_{j=1}^{k_b^*} \frac{(\boldsymbol{\beta}_j^*)^2}{\sigma_j} \left(\frac{\mathrm{Tr}(\Sigma_{k_b^*+1:p})}{N}\right)^2$$

which is the expected lower bound on the estimation of $\boldsymbol{\beta}^* 1 : k_b^*$ part of the bias term.

References

[1] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proc. Natl. Acad. Sci. USA*, 117(48):30063–30070, 2020.

[2] Peter L. Bartlett, Andreas Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. Technical report, ArXiv 2103.09177, 2021.

[3] Wiltord Bednorz. Concentration via chaining method and its applications. 2014.

[4] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *CoRR*, abs/2105.14368, 2021.

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proc. Natl. Acad. Sci. USA*, 116(32):15849–15854, 2019.

[6] Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM J. Math. Data Sci.*, 2(4):1167–1180, 2020.

[7] Mikhail Belkin, Daniel J. Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2306–2317, 2018.

[8] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. Does data interpolation contradict statistical optimality? In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 1611–1619. PMLR, 2019.

[9] Yoshua Bengio, Yann LeCun, and Geoffrey E. Hinton. Deep learning for AI. *Commun. ACM*, 64(7):58–65, 2021.

[10] Emmanuel J. Candes and Terence Tao. Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[11] Emmanuel Caron and Stéphane Chrétien. A finite sample analysis of the benign overfitting phenomenon for ridge function estimation. Technical report, 2020.

[12] Geoffrey Chinot, Felix Kuchelmeister, Matthias Löffler, and Sara A. van de Geer. Adaboost and robust one-bit compressed sensing. *CoRR*, abs/2105.02083, 2021.

[13] Geoffrey Chinot, Matthias Löffler, and Sara A. van de Geer. On the robustness of minimum-norm interpolators. *CoRR*, abs/2012.00807, 2020.

[14] Geoffrey Chinot and Lerasle Matthieu. On the robustness of the minimum l2 interpolator. Technical report, ENSAE, CREST, IPParis, 2020.

[15] Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20, 2015.

[16] Joseph Fourier. *Analytical theory of heat.* Dover Publications, Inc., New York, 1955. Translated, with notes, by Alexander Freeman.

[17] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *CoRR*, abs/1903.08560, 2019.

[18] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks (invited paper). In Samir Khuller and Virginia Vassilevska Williams, editors, *STOC '21: 53rd Annual ACM SIGACT Symposium on Theory of Computing, Virtual Event, Italy, June 21-25, 2021*, page 6. ACM, 2021.

[19] Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8580–8589, 2018.

[20] B. Klartag and S. Mendelson. Empirical processes and random projections. *J. Funct. Anal.*, 225(1):229–245, 2005.

[21] Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds, and benign overfitting. *CoRR*, abs/2106.09276, 2021.

[22] Guillaume Lecué and Shahar Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

[23] Guillaume Lecué and Shahar Mendelson. Regularization and the small-ball method I: Sparse recovery. *Ann. Statist.*, 46(2):611–641, 2018.

[24] Michel Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.

[25] Michel Ledoux and Michel Talagrand. *Probability in Banach spaces*. Classics in Mathematics. Springer-Verlag, Berlin, 2011. Isoperimetry and processes, Reprint of the 1991 edition.

[26] Tengyuan Liang and Alexander Rakhlin. Just interpolate: kernel "ridgeless" regression can generalize. *Ann. Statist.*, 48(3):1329–1347, 2020.

[27] Tengyuan Liang and Benjamin Recht. Interpolating classifiers make few mistakes. *CoRR*, abs/2101.11815, 2021.

[28] Chaoyue Liu, Libin Zhu, and Mikhail Belkin. On the linearity of large non-linear models: when and why the tangent kernel is constant. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[29] Song Mei and Andreas Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. Technical report, Stanford University, 2020.

[30] Shahar Mendelson. Upper bounds on product and multiplier empirical processes. *Stochastic Process. Appl.*, 126(12):3652–3680, 2016.

[31] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.*, 17(4):1248–1282, 2007.

[32] V. D. Milman. A new proof of A. Dvoretzky's theorem on cross-sections of convex bodies. *Funkcional. Anal. i Priložen.*, 5(4):28–37, 1971.

[33] Partha P. Mitra. Understanding overfitting peaks in generalization error: Analytical risk curves for l_2 and l_1 penalized interpolation. *CoRR*, abs/1906.03667, 2019.

[34] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE J. Sel. Areas Inf. Theory*, 1(1):67–83, 2020.

[35] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1989.

[36] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge(less) regression under general source condition. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 3889–3897. PMLR, 2021.

[37] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. *Ann. Statist.*, 26(5):1651–1686, 1998.

[38] Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer, Heidelberg, 2014. Modern methods and classical problems.

[39] Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. 2021.

[40] Roman Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. An introduction with applications in data science, With a foreword by Sara van de Geer.

[41] Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum l1-norm interpolation of noisy data. *CoRR*, abs/2111.05987, 2021.

[42] Abraham J. Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *J. Mach. Learn. Res.*, 18:48:1–48:33, 2017.

[43] Ji Xu and Daniel J. Hsu. On the number of variables to use in principal component regression. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5095–5104, 2019.

[44] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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

[46] Lijia Zhou, Frederic Koehler, Danica J. Sutherland, and Nathan Srebro. Optimistic rates: A unifying theory for interpolation learning and regularization in linear regression. *CoRR*, abs/2112.04470, 2021.

[47] Difan Zou, Jingfeng Wu, Vladimir Braverman, Quanquan Gu, and Sham M. Kakade. Benign overfitting of constant-stepsize SGD for linear regression. In Mikhail Belkin and Samory Kpotufe, editors, *Conference on Learning Theory, COLT 2021, 15-19 August 2021, Boulder, Colorado, USA*, volume 134 of *Proceedings of Machine Learning Research*, pages 4633–4635. PMLR, 2021.