On the robustness of the minimim l2 interpolator

Geoffrey Chinot, Matthieu Lerasle

On the robustness of the minimim ℓ_2 interpolator

GEOFFREY CHINOT[1] and MATTHIEU LERASLE[2]

[1] *ETHZ, Rämistrasse 101, 8092 Zürich.*
E-mail: geoffrey.chinot@stat.math.ethz.ch

[2] *CNRS, ENSAE, CREST, 5 avenue Henri Chatelier 91120 Palaiseau, France.*
E-mail: matthieu.lerasle@ensae.fr

We analyse the interpolator with minimal ℓ_2-norm $\hat{\beta}$ in a general high dimensional linear regression framework where $\mathbb{Y} = \mathbb{X}\beta^* + \xi$ with $\mathbb{X}$ a random $n \times p$ matrix with independent $\mathcal{N}(0, \Sigma)$ rows. We prove that, with high probability, without assumption on the noise vector $\xi \in \mathbb{R}^n$, the ellipsoid risk $\|\hat{\beta} - \beta^*\|_\Sigma^2 = (\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*)$ is bounded from above by $(\|\beta^*\|_2^2 r_{cn}(\Sigma) \vee \|\xi\|^2)/n$, where c is an absolute constant and, for any $k \geqslant 1$, $r_k(\Sigma) = \sum_{i \geq k} \lambda_i(\Sigma)$ is the tail sum of the eigenvalues of Σ. These bounds show a transition in the rates. For high signal to noise ratios, the rates $\|\beta^*\|_2^2 r_{cn}(\Sigma)/n$ broadly improve the existing ones. For low signal to noise ratio, we also provide lower bound holding with large probability. General lower bounds are proved under minor restrictions on the noise ξ (see Theorem 1). Under assumptions on the sprectrum of Σ, this lower bound is of order $\|\xi\|_2^2/n$, matching the upper bound. Consequently, in the large noise regime, we are able to precisely track the ellipsoid risk with large probability. These results give new insight when the interpolation can be harmless in high dimensions.

Keywords: Interpolation problems, statistical learning, robustness.

1. Introduction

In this paper, we consider the problem of estimating a high dimensional vector $\beta^* \in \mathbb{R}^p$ from a few possibly noisy observations of random projections of it. Let $\mathbb{X} \in \mathbb{R}^{n \times p}$ denote a random matrix with rows X_i^T. The observations can therefore equivalently be written

$$y_i = \langle X_i, \beta^* \rangle + \xi_i, \qquad i \in \{1, \dots, n\} \ ,$$

or, in the matrix form

$$\mathbb{Y} = \mathbb{X}\beta^* + \xi \ .$$

The vector $\xi = (\xi_1, \dots, \xi_n)^T$ is called the noise. This problem is classical in signal processing and in statistics where it is known as the linear regression problem. In particular, the Gaussian linear regression problem is the problem of recovering β^* when ξ is independent from $\mathbb{X}$ and Gaussian. The arguably most famous estimator is the least-squares estimator defined as

$$\hat{\beta} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|\mathbb{Y} - \mathbb{X}\beta\|_2^2 \ ,$$

where $\|\cdot\|_2$ denotes the usual Euclidean norm in $\mathbb{R}^d$, whatever $d \geqslant 2$. The quality of $\hat{\beta}$ can be assessed through upper and lower bounds on the estimation error $\|\hat{\beta} - \beta^*\|_2$. When the rows X_i^T of $\mathbb{X}$ are i.i.d. with second moment matrix $\Sigma = \mathbb{E}[X_1 X_1^T]$, another popular quality measure is the ellispoid risk

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

In this formulation, X denote an independent copy of X_1, independent from $\mathbb{X}, \mathbb{Y}$. The ellipsoid risk therefore measures how far are the predictions at a typical point X by $\hat{\beta}$: $\langle X, \hat{\beta} \rangle$ and by the actual signal β^*: $\langle X, \beta^* \rangle$. Both risks are random variables and upper and lower bounds for these risks, in expectation and with high probability are now well known in the small dimensional Gaussian linear problem where $p < n$.

These bounds deteriorate as the dimension grows and the least-squares estimator behaves poorly when $p \asymp n$. This can be understood as follows: In high dimension where $p \geqslant n$, the set of least-squares estimators is typically infinite. Actually, when the matrix $\mathbb{X}$ has full rank n, its null space is non trivial and any solution in the set $\{\mathbb{X}^g \mathbb{Y}\}$, where $\mathbb{X}^g$ describes all pseudo-inverses of $\mathbb{X}$ satisfies $\mathbb{X}\mathbb{X}^g\mathbb{Y} = \mathbb{Y}$. In other words, in large dimension, least-squares estimators interpolate data. This kind of behavior is typically undesirable in statistics, as the estimators clearly overfit the observed dataset, and have usually poor generalization abilities. The least-squares estimators are not the only estimators suffering this kind of limitation, actually, this feature is shared with any estimator without further assumptions on the model, a phenomenon known as the curse of dimensionality in statistics.

The classical trick in high dimensional statistics to bypass this issue is to assume structural assumptions on β^*, such as sparsity or regularity assumptions. This has given rise to an impressive literature these last decades. We cannot review here this massive literature. The interested reader can find comprehensive introductions to these topics in the textbooks [11, 21, 37, 38, 19] and the references therein. Let us just mention that this approach was proved efficient, for example in the high dimensional Gaussian regression problem. Among popular such algorithms, one can mention basis pursuit [16], ridge regression [22, 14], the LASSO [34, 35, 9] and the elastic net [44, 17].

We do not pursue this path in this paper: We want to tackle the problem in high dimension, that is when $p \geqslant n$, without any assumption on β^*. As we said, bounding the estimation error $\|\hat{\beta} - \beta^*\|_2$ in this context remains impossible in general [36]. However, and perhaps counter-intuitively, [4] discovered that, when the dimension p is large in front of n, the prediction risk $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2$ can be small for the least-squares estimator $\hat{\beta} = \mathbb{X}^\dagger \mathbb{Y}$, where $\mathbb{X}^\dagger$ is the Moore-Penrose pseudo-inverse of $\mathbb{X}$. They proved that this holds in the Gaussian regression problem, when the lines X_i^T of $\mathbb{X}$ are i.i.d. with Gaussian distribution $\mathcal{N}(0, \Sigma)$, under some conditions on the spectrum of Σ. An important take home message of [4] is therefore the following

In high dimension, even without structural assumptions on β^*, it is still possible to predict well.

This interesting phenomenon has given rise to a rapidly growing literature these last

months, see [5, 6, 7, 8, 12, 20, 27, 28]. This success is not surprising as many algorithms in machine learning require to fit a huge number of parameters with a smaller number of data. The most famous examples are neural networks for which it has been repeatedly observed empirically that enlarging the network, hence, the number of parameters, may help to improve prediction performance [1, 5, 42]. Of course, the linear prediction problem here is much simpler than understanding the predictions of neural networks, but it is interesting to understand when and how high dimension helps prediction. Moreover, several recent works have shown that the analysis of linear models can be relevant for over-parametrized neural networks, see for example [15]. A reason is that, when neural networks are trained by gradient descent properly initialized, they are well approximated by a linear model in a Hilbert space. This method is known as *neural tangent kernel* approach [23, 10, 3, 26]. Understanding the generalization of over-parametrized linear models could therefore be seen as a first step in the direction of understanding deep learning.

In this paper, as in [4], we analyse the least-squares estimator $\mathbb{X}^\dagger\mathbb{Y}$, where $\mathbb{X}^\dagger$ denotes the Penrose Moore inverse of $\mathbb{X}$, which is the least-squares estimator with minimal Euclidean norm. We also assume that $\mathbb{X}$ has i.i.d. Gaussian $\mathcal{N}(0,\Sigma)$ lines. We will assume all along the paper that we are in the high dimensional regime where $p \geq n$ and that Σ has rank larger than n, which implies in particular that $\mathbb{X}$ has a.s. full rank n. In this setting, the estimator can be equivalently defined as the minimum norm interpolator

$$\hat{\beta} = \underset{\beta\in\mathbb{R}^p}{\operatorname{argmin}} \|\beta\|_2 \quad \text{subject to} \quad \mathbb{X}\beta = \mathbb{Y} \ . \tag{1}$$

The main difference with [4] is that all upper bounds are proved without assumption on the noise ξ. This shows the robustness of $\hat{\beta}$ to various contaminations of the response data $\mathbb{Y}$, as the noise, for example, can be deterministic (and hence null), random with any distributions (hence allowing heavy tailed perturbations), or even adversarial. Similar assumptions have been considered in the compressed sensing community [39], in a different problem and with the Euclidean norm replaced by the ℓ_1-norm.

This setting can be used to compare the predictions of any linear predictor, based on how well the predictions agree on the training set: For any vector β^*, it holds that $\mathbb{Y} = \mathbb{X}\beta^* + \xi$ with ξ simply defined as the associated vector of residuals $\mathbb{Y} - \mathbb{X}\beta^*$. Although we usually think of β^* as playing a key role in the data-generating process, here it works to be simply arbitrary. Nevertheless, the most interesting choice of β^* in our bounds would of course be the one from the true data-generating process.

Our main results give upper and lower bounds on the ellipsoid risk of $\hat{\beta}$, $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2$. The bounds are typically dependent on the noise $\|\xi\|_2$ and are therefore random in general. To deduce deterministic bounds, an independent analysis of this term should be performed in each situation. As in [4], the bounds are interesting under assumptions on the spectrum of the covariance matrix Σ. These assumptions involve the tail sum of singular values of the matrix Σ defined for any $k \geqslant 1$ by $r_k(\Sigma) = \sum_{i=k}^p \lambda_i(\Sigma)$. Our bounds exhibit a phase transition when the signal to noise ratio $\mathrm{SNR} = \|\beta^*\|_2^2/\|\xi\|_2^2$ becomes larger than a threshold $t = 1/r_{cn}(\Sigma)$, where c is an absolute constant.

- For high signal to noise ratios: SNR $> t$, the prediction risk of the estimator satisfies, $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \|\beta^*\|^2 r_{cn}(\Sigma)/n$, with large probability. This result improves the one presented in [4] in two ways: First they only reached in this regime the more pessimistic upper bound $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \|\beta^*\|^2\sqrt{\mathrm{Tr}(\Sigma)/n}$. Notice that, in a different framework the improvement on the rate from $\sqrt{1/n}$ to $1/n$ for interpolators already appeared in [12]. Second, we prove that, for interpolators, these fast rates of convergence hold with probability $1 - ce^{-n/c}$, for some absolute constant c, while in [4] the results where established with probability $1 - ce^{-c\mathrm{Tr}(\Sigma)}$.
- When the SNR is low, SNR $\leq t$, we show that with probability $1 - ce^{-n/c}$,

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \leq c\frac{\|\xi\|_2^2}{n} \ .$$

  In Gaussian linear regression with $\xi \sim \mathcal{N}(0, \sigma^2\mathbb{I}) \in \mathbb{R}^n$, this rates becomes σ^2 with probability $1 - ce^{-n/c}$, which matches the minimax rate of convergence for this confidence level [25, Theorem A']. Besides, we prove that this bound cannot be improved in general. Indeed, when the noise is independent from $\mathbb{X}$ (whatever its distribution), for a well chosen $\bar{k} \leqslant p$ (see Section 2 for a precise definition),

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \geqslant c\frac{\|\xi\|_2^2}{n \wedge \bar{k}} \ .$$

  Therefore, in the particular case where $p = cn$ for example, this yields $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \asymp \|\xi\|_2^2/n$. More generally, for any spectrum of Σ, such that $\bar{k} \leq cn$ (we provide an example at the end of Section 2 where this holds while $n = o(p)$), upper and lower bounds match $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \asymp \|\xi\|_2^2/n$. Interestingly, the lower bounds are also obtained with probability $1 - ce^{-n/c}$. In this regime, the comparison with [4] is less clear. On one hand, our results are more general as they allow any kind of noise and, when the noise is Gaussian, improve the bounds of [4] at confidence levels $1 - ce^{-n/c}$. On the other hand, when the noise is Gaussian $\xi \sim \mathcal{N}(0, \sigma^2\mathbb{I})$, our bounds do not shrink to 0 as $n \to \infty$ while those in [4] might at smaller confidence levels.
  Our lower bound depends on a new parameter $\bar{k}$ that was not present in the previous work of [4]. It gives new insights when the overfitting can be harmless.

Our extension to general noise is based on the following observation: As $\hat{\beta}$ interpolates, we have

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\langle X_i, \hat{\beta} - \beta^*\rangle^2 - \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2}_{\text{deviation}} = \frac{\|\xi\|_2^2}{n} \ . \tag{2}$$

The main technical contribution of the paper is then an upper bound on the deviation term that holds independently to the form of the noise ξ. The control of the deviation term is possible using a preliminary result showing that dimension may help to localize

this estimator with respect to the estimation norm $\|\hat{\beta} - \beta\|_2$. Heuristically, since the dimension of the set of interpolators increases with the dimension p, it is expected that $\|\hat{\beta}\|_2$ also decreases with the dimension. We show that $\|\hat{\beta} - \beta^*\|_2 \leq \|\beta^*\|_2 + \Delta(\Sigma)$, where $\Delta(\Sigma)$ is a remainder term controlling the improvement with the dimension p.

The paper is divided in two parts: Section 2 presents the assumptions and main results, and discuss the general case in particular situations of interest. The main proofs are gathered in Section 3.

Notations For any symmetric matrix $A \in \mathbb{R}^{n\times n}$, we denote by $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$ its eigenvalues in the non-increasing order and, for any $k \geqslant 1$, by $r_k(A) = \sum_{i=k}^n \lambda_i(A)$. More generally, for any matrix $B \in \mathbb{R}^{n\times p}$, we denote by $\sigma_1(B) \geq \cdots \geq \sigma_{min}(B) > 0$, its positive singular values in the non-increasing order. The operator norm of B is denoted by $\|B\| = \sigma_1(B)$. For any symmetric positive semi-definite matrix A, let $\|\beta\|_A = \sqrt{\beta^T A \beta}$. Let $S(r)$ (resp. $S_A(r)$) denote the sphere in $\mathbb{R}^p$ with radius r with respect to the Euclidean norm $\|\cdot\|_2$ (resp. with respect to the semi-norm $\|\cdot\|_A$). Define similarly $B(r)$ and $B_A(r)$ to be the balls with radius r. All along the paper, $c, c_1, c_2 \cdots$ denote absolute positive constants whose values may change from one instance to another.

2. Main results

This section provides our main contributions. Before stating our main result, let us introduce quantities that will drive the prediction risk of $\hat{\beta}$. Let c_0 denote an absolute constant that should be large enough and let

$$\rho = \|\beta^*\|_2 + \frac{4\|\xi\|_2}{\sqrt{r_{k^*}(\Sigma)}}, \qquad \text{where} \qquad k^* = \inf\left\{k \in \{1, \cdots, p\} : \frac{r_k(\Sigma)}{\lambda_k(\Sigma)} \geq c_0 n\right\} , \quad (3)$$

with the convention that $\inf \emptyset = +\infty$. Also, for constants $\eta, \gamma > 0$, let us define the following two complexity parameters:

$$r^*(\eta) = \inf\left\{r > 0 : \sum_{i=1}^p \lambda_i(\Sigma) \wedge r^2 \leq \eta n r^2\right\} \quad (4)$$

$$\bar{r}(\gamma) = \sup\left\{r > 0 : \sum_{i=1}^p \lambda_i(\Sigma)\rho^2 \wedge r^2 \leq \gamma\|\xi\|_2^2\right\} . \quad (5)$$

We are now in position to state our main theorem.

Theorem 1. *Assume $k^* \leq cn$, where $c > 0$ is a small enough absolute constant. For η small enough, there exist absolute constants c_1, c_2, c_3 such that with probability larger than $1 - c_1 e^{-c_2 n}$, the estimator $\hat{\beta}$ defined in Equation (1) satisfies*

$$\|\hat{\beta} - \beta^*\|_2 \leq \rho \quad \text{and} \quad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \leq \rho r^*(\eta) \vee c_3 \frac{\|\xi\|_2}{\sqrt{n}} .$$

Moreover, let us assume that $X_i|\xi$ *are i.i.d.* $\mathcal{N}(0,\Sigma)$ *(recall that* X_i^T *is the i-th row of* $\mathbb{X}$*). There exist absolute constants* $\gamma > 0$*,* c_1, c_2, c_3 *such that with probability larger than* $1 - c_1 e^{-c_2 n}$*, the estimator* $\hat{\beta}$ *defined in Equation (1) satisfies*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \geq \bar{r}(\gamma) \wedge c_3 \frac{\|\xi\|_2}{\sqrt{n}} \quad .$$

Theorem 1 is proved in Section 3.2. This proof is split into two parts. First, we use the structure of $\hat{\beta}$ (solution with minimal norm interpolating the data) and a spectral analysis of $\mathbb{X}$ to control the estimation error $\|\hat{\beta} - \beta^*\|_2$. Then, using uniform concentration arguments, we bound from above its ellipsoid risk $\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2$ with high probability. In this sense, our results naturally fit in the recent literature that shows generalization bounds for overparametrized models using uniform convergence arguments [43, 32, 41].

The estimation bound ρ does not converge to 0. This is not surprising in high dimension without sparsity assumption. However, it is interesting to see that it may decrease, up to a certain threshold, with the dimension p. In particular, when the signal to noise ratio $\|\beta^*\|^2/\|\xi\|_2^2$ is larger than the threshold $1/r_{k^*}(\Sigma)$, $\|\hat{\beta} - \beta^*\|_2$ is at most of order $\|\beta^*\|_2$.

As stressed before, the upper bound in Theorem 1 holds without assumption on the ξ_i's. The noise can be deterministic or depend on $\mathbb{X}$. This is a major difference with previous results in the literature such as [4, 12] where this noise was always sub-Gaussian and independent from $\mathbb{X}$. Here, the results can be applied to the following example, where $\mathbb{Y}$ is itself the output of a prediction algorithm, that is, when $\mathbb{Y} = f(\mathbb{X}) = (f_1(\mathbb{X}_1), \ldots, f_n(\mathbb{X}_n))^T$. In this case, the upper bounds becomes

$$\frac{\|\xi\|_2^2}{n} = \frac{\|f(\mathbb{X}) - \mathbb{X}\beta^*\|_2^2}{n} = \frac{1}{n}\sum_{i=1}^{n} \left(\langle \mathbb{X}_i, \beta^* \rangle - f_i(\mathbb{X}_i)\right)^2 \quad .$$

This error measures how far the initial prediction is from the linear model.

While results without assumptions on the noise have not been proved for interpolators before, they are on the other hand classical in the compressed sensing literature [13, 40]. In compressed sensing, the goal is to recover a low dimensional signal from noisy observations, so the natural risk is the estimation risk $\|\hat{\beta} - \beta^*\|_2$ and the signal β^* satisfies some sparsity assumption. For example, in [40], the author shows that a s-sparse vector β^* (with s small enough) can be recovered from n noisy observations $y_i = \langle X_i, \beta^* \rangle + \xi_i$, $X_i \sim \mathcal{N}(0, I_p)$, without assumption on the noise $\xi \in \mathbb{R}^n$ and with optimal rate $\|\xi\|_2/\sqrt{n}$ (see [13] for optimality). As the covariance Σ in this example is I_p, it turns out that estimation and ellipsoid risks are the same in their framework. In the same spirit, our result holds without assumption on the error vector $\xi \in \mathbb{R}^n$. Our results complement the compressed sensing literature in this setting as we replace the sparsity assumption by assumptions on the covariance matrix Σ. In this new setting, the estimation and ellipsoid risk do not match any more. The vector β^* itself cannot be recovered ($\|\hat{\beta} - \beta^*\|_2$ does not converged to 0), but an estimator with converging ellipsoid risk can be built. Actually, this estimator can even be obtained without regularization and yet reach the optimal

rate $\|\xi\|_2/\sqrt{n}$ (in ellipsoid risk). However, in this new setting, an extra bias term appears in the risk bound in addition to the variance term $\|\xi\|_2/\sqrt{n}$. Theorem 1 shows that this bias can be bounded from above by $\rho r^*(\eta)$.

To discuss further this upper bound on the ellipsoid risk, it is useful to give the following corollary. It shows a phase transition in the ellipsoid risk when the signal to noise ratio $\mathrm{SNR} = \|\beta^*\|^2/\|\xi\|^2$ becomes larger than the threshold $t = 1/r_{cn}(\Sigma)$.

Corollary 1. *Grant the assumptions and notations of Theorem 1. The estimator $\hat{\beta}$ defined in Equation (1) satisfies, with probability larger than $1 - c_1 e^{-c_2 n}$,*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \lesssim \|\beta^*\|_2^2 \frac{r_{cn}(\Sigma)}{n} \vee \frac{\|\xi\|_2^2}{n} \ .$$

Corollary 1 is proved in Section 3.3. It can be used to compare our results with [4].

1. If the signal to noise ratio is large enough, $\|\beta^*\|_2^2/\|\xi\|_2^2 \geq 1/r_{cn}(\Sigma)$ the bounds in [4] are always larger than ours. Indeed, for large SNR, our bounds are of order $\|\beta^*\|_2^2 r_{cn}(\Sigma)/n$ while theirs are of order $\|\beta^*\|_2^2\sqrt{\mathrm{Tr}(\Sigma)/n}$. The improvement can even be exponential as shown in the example below.
2. For small signal to noise ratios, $\mathrm{SNR} < t$, our prediction rates are of order $\frac{\|\xi\|_2}{\sqrt{n}}$. This bound holds without any assumption on the noise ξ. Contrary to [4], this rate does not converge to 0 as $n \to \infty$ when $\xi \sim \mathcal{N}(0, \sigma^2 \mathbb{I})$ is independent from $\mathbb{X}$. However, in this case, $\|\xi\|_2^2/n \asymp \sigma^2$ matches the optimal rate holding with probability larger than $1 - c_1 \exp(-c_2 n)$ (see [25, Theorem A']).

Moreover, when the noise is null ($\|\xi\|_2 = 0$), the condition on k^* is no longer required (because $\hat{\beta} = \mathbb{X}^\dagger \mathbb{X} \beta^*$ and Lemma 2 is not longer needed). In this case, Corollary 1 gives the upper bound $\|\beta^*\|_2^2 r_{cn}/n$. For isotropic design ($\Sigma = I_p$), our upper bound becomes $\|\beta^*\|_2^2$ and matches the lower bound obtain in Theorem 1 in [6]. As a consequence, without further assumption, our bias term cannot be improved.

To illustrate their upper bounds, [4] provide several examples of "benign matrices" where the different quantities of interest in Theorem 1 can easily be computed. We compute the quantities appearing in one of these examples now.
Assume that there exist $\epsilon, \tau > 0$ such that $p = \bar{C} n$, $\tau \log(1/\epsilon) \leq \bar{c} n$ for $\bar{c} > 0$ (resp. $\bar{C} > 0$) small enough (resp. large enough) and

$$\forall k \geq 1, \qquad \lambda_k(\Sigma) = e^{-k/\tau} + \epsilon \ .$$

Let $c_1 > 0$ be an absolute constant. Its value may change from one line to another. In this case, for any k,

$$\frac{r_k}{\lambda_k} \geq c_1 \frac{(p - k + 1)\epsilon + \tau \exp(-k/\tau)}{e^{-k/\tau} + \epsilon} \ ,$$

Therefore, for $k = \tau \log(1/\epsilon)$,

$$\frac{r_k}{\lambda_k} \geq c_1 \frac{p\epsilon + \tau\epsilon}{\epsilon} \geq c_1 \bar{C} n, \qquad \text{so} \qquad k^* \leq \tau \log(1/\epsilon) \leq \bar{c} n \ .$$

Moreover, $r_{cn}(\Sigma) \leq c_1 \left(p\epsilon + \tau \exp(-cn/\tau)\right)$ and Corollary 1 shows in this example that with probability larger than $1 - c_1 e^{-n}$,

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \leq c_1 \left(\|\beta^*\|_2^2 \frac{p\epsilon + \tau \exp(-cn/\tau)}{n} \vee \frac{\|\xi\|_2^2}{n} \right) \ .$$

Our rates of convergence in this example can therefore be as fast as $\epsilon \vee \tau \exp(-cn/\tau)/n$, while [4, Theorem 6] gives in this setting a rate $\sqrt{\epsilon \vee \tau \exp(-1/\tau)/n}$ leading to a potential exponential improvement (a multiplication by e^{-cn} of the rates) for small ϵ.

Let us turn to the study of lower bound in the large noise regime.

Corollary 2. *Grant the assumptions and notations of Theorem 1, with γ chosen such that the conclusion of the Theorem holds. Assume that $X_i|\xi \sim \mathcal{N}(0, \Sigma)$ and are independent conditionally on ξ. If the signal to noise ratio is small, $\|\beta^*\|_2^2/\|\xi\|_2^2 \leq 1/r_{k^*}(\Sigma)$, then, the estimator $\hat{\beta}$ defined in Equation (1) satisfies, with probability larger than $1 - c_1 e^{-c_2 n}$,*

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \geq c_3 \frac{\|\xi\|_2^2}{n \vee \bar{k}} \ ,$$

where

$$\bar{k} = \inf \left\{ k \geq k^* : \sum_{i=k}^{p} \lambda_i(\Sigma) \leq \frac{\gamma}{2} \sum_{i=k^*}^{p} \lambda_i(\Sigma) \right\} \ , \tag{6}$$

with the convention that $\bar{k} = p + 1$ if the set is empty.

Corollary 2 is proved in Section 3.3. Note that when $p = cn$, we have $\bar{k} \leq cn$. In this case, in the large noise regime,

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \gtrsim \frac{\|\xi\|_2^2}{n} \ ,$$

with probability larger than $1 - c_1 \exp(-c_2 n)$, which matches the upper bound.

Let us give an example of spectrum where p might be much larger than n while $\bar{k}$ remains smaller than n. Let k_1, c, $k_2 = k_1 + cn + 1$, $\varepsilon_1, \varepsilon_2$ real numbers and assume that

$$\lambda_i = \begin{cases} 1 & \text{if } i \leqslant k_1 - 1 \ , \\ \varepsilon_1 & \text{if } i \in \{k_1, \cdots, k_2 - 1\} \ , \\ \varepsilon_2 & \text{if } i \geqslant k_2 \ . \end{cases}$$

In this case,

$$\frac{r_{k_1}}{\lambda_{k_1}} \geqslant cn + (p - k_1 - cn)\frac{\varepsilon_2}{\varepsilon_1} \geq cn \ ,$$

It follows that $k^* \leqslant k_1$. Finally

$$\frac{\sum_{i=k_2}^p \lambda_i}{\sum_{i=k_1}^p \lambda_i} \leqslant \frac{(p-k_2+1)\varepsilon_2}{cn\varepsilon_1 + (p-k_2+1)\varepsilon_2} \ .$$

This ratio is smaller than $\gamma/2$ if

$$\varepsilon_2 \leq \frac{\gamma}{2-\gamma}\frac{cn}{p-k_2+1}\varepsilon_1 \ .$$

This proves that $\bar{k} \leqslant k_1 + cn$. If $k_1 \leq n$, we have therefore, in this example, if $\|\beta^*\|_2^2/\|\xi\|_2^2 \leq r_{cn}(\Sigma)$

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

3. Proofs of the main results

The remaining of the paper is devoted to the proofs of the main results. Section 3.1 (resp. 3.2) shows the estimation bound (resp. the prediction bounds) in Theorem 1.

3.1. Proof of the estimation bound of Theorem 1

The following theorem establishes the bound on the estimation error in Theorem 1. In the following section, this preliminary estimate will be used to "localize" the interpolator which will lead to a more precise analysis of the ellipsoid risk of $\hat{\beta}$. This approach is now classical in statistical learning, it has been applied successfully, for example, in [24, 29, 30, 31]. This first step of our analysis is derived using the exact formula defining the interpolator and a spectral analysis of the matrix $\mathbb{X}$, similar bounds cannot be derived for any interpolator.

Theorem 2. *There exists $c > 0$ such that, if $c_0 \geq c$ in the definition of k^*, the estimator $\hat{\beta}$ defined in Equation (1) satisfies*

$$\mathbb{P}\left(\|\hat{\beta} - \beta^*\|_2 \leq \|\beta^*\|_2 + \frac{4\|\xi\|_2}{\sqrt{r_{k*}(\Sigma)}}\right) \geq 1 - 2\exp\big(-n\big) \ . \tag{7}$$

Proof of Theorem 2. The proof starts with the following lemma.

Lemma 1. *The estimator $\hat{\beta}$ verifies*

$$\|\hat{\beta} - \beta^*\|_2 \leq \|\beta^*\|_2 + \frac{\|\xi\|_2}{\sigma_n(\mathbb{X})} \ . \tag{8}$$

Proof of Lemma 1. Recall that

$$\hat{\beta} = \mathbb{X}^\dagger \mathbb{Y} = \mathbb{X}^\dagger \mathbb{X}\beta^* + \mathbb{X}^\dagger \xi \ ,$$

where $\mathbb{X}^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathbb{X}$. Therefore,

$$\|\hat{\beta} - \beta^*\|_2 = \|(\mathbb{X}^\dagger \mathbb{X} - I_p)\beta^* - \mathbb{X}^\dagger \xi\|_2 \leq \|\beta^*\|_2 + \|\mathbb{X}^\dagger \xi\|_2 \ , \tag{9}$$

where the last inequality follows from the triangular inequality and the fact that $\mathbb{X}^\dagger\mathbb{X} - I_p$ is the projection matrix onto the null-space of $\mathbb{X}$. Since $\|\mathbb{X}^\dagger \xi\|_2 \leq \|\mathbb{X}^\dagger\| \|\xi\|_2$ it follows that

$$\|\hat{\beta} - \beta^*\|_2 \leq \|\beta^*\|_2 + \|\xi\|_2 \|\mathbb{X}^\dagger\| = \|\beta^*\|_2 + \frac{\|\xi\|_2}{\sigma_n(\mathbb{X})} \ ,$$

where the last identity holds because $\mathrm{rank}(\mathbb{X}) = n$ a.s. and thus $\|\mathbb{X}^\dagger\| = \sigma_n^{-1}(\mathbb{X})$. □

Lemma 1 provides a random bound on the estimation error of $\hat{\beta}$. To prove Theorem 2, it remains to bound from below, with high probability, the n-th singular value $\sigma_n(\mathbb{X})$ of $\mathbb{X}$. This control is obtained in the following lemma.

Lemma 2. *With probability larger than $1 - 2\exp(-n)$, if c_0 in the definition (3) of k^* is large enough, we have*

$$\sigma_n(\mathbb{X}) \geq \frac{\sqrt{r_{k^*}(\Sigma)}}{4} \ .$$

Proof. The matrix $\mathbb{X}^T$ is distributed as $\Sigma^{1/2}G$, where $G \in \mathbb{R}^{p\times n}$ is a random matrix with i.i.d standard Gaussian variables, hence $\sigma_n(\mathbb{X}) = \sigma_n(\mathbb{X}^T)$ is distributed as $\sigma_n(\Sigma^{1/2}G)$. Let S^{n-1} denote the unit sphere in $\mathbb{R}^n$. From the Courant-Fischer-Weyl min-max principle, we have

$$\sigma_n(\Sigma^{1/2}G) = \min_{x\in S^{n-1}} \|\Sigma^{1/2}Gx\|_2 \ .$$

Let $\Lambda = \mathrm{diag}(\lambda_1(\Sigma), \cdots, \lambda_p(\Sigma))$. By the spectral theorem, there exists an orthogonal matrix P such that $\Sigma^{1/2} = P\Lambda^{1/2}P^T$, so, for any $x \in S^{n-1}$, $\|\Sigma^{1/2}Gx\|_2^2 = \|P\Lambda^{1/2}P^TGx\|_2^2 = \|\Lambda^{1/2}P^TGx\|_2^2$. Hence, by rotation invariance of Gaussian random vectors, $\|\Sigma^{1/2}Gx\|_2^2 = \sum_{i=1}^p \lambda_i(\Sigma)g_i^2$, where $g_1, \cdots, g_p$ are i.i.d standard Gaussian random variables. It follows that

$$\|\Sigma^{1/2}Gx\|_2^2 = \sum_{i=1}^p \lambda_i(\Sigma)g_i^2 \geq \sum_{i=k^*}^p \lambda_i(\Sigma)g_i^2 = \|\Lambda_{k^*}^{1/2}Gx\|_2^2 \ , \tag{10}$$

where k^* is defined in Theorem 1 and $\Lambda_{k^*} = \mathrm{diag}(0, \cdots, 0, \lambda_{k^*}(\Sigma), \cdots, \lambda_p(\Sigma))$.

Let $\varepsilon \in (0,1)$ and $\mathcal{N}_\varepsilon$ be an ε-net of S^{n-1}. A classical volume argument (see for example [37, Corollary 4.2.13]) shows that we can choose $\mathcal{N}_\varepsilon$ with $|\mathcal{N}_\varepsilon| \le (3/\varepsilon)^n$. For any $x \in S^{n-1}$, there exists $y \in N_\varepsilon$ such that $\|x-y\|_2 \le \varepsilon$, so

$$\|\Sigma^{1/2}Gx\|_2 \ge \|\Lambda_{k^*}^{1/2}Gx\|_2 \ge \|\Lambda_{k^*}^{1/2}Gy\|_2 - \varepsilon\|\Lambda_{k^*}^{1/2}G\| \ .$$

Hence,

$$\sigma_n(\Sigma^{1/2}G) \ge \sigma_n(\Lambda_{k^*}^{1/2}G) \geqslant \min_{y\in N_\varepsilon} \|\Lambda_{k^*}^{1/2}Gy\|_2 - \varepsilon\|\Lambda_{k^*}^{1/2}G\| \ .$$

Besides (see for example [37, Lemma 4.4.1])

$$\|\Lambda_{k^*}^{1/2}G\| \le \frac{1}{1-\varepsilon}\max_{y\in N_\varepsilon}\|\Lambda_{k^*}^{1/2}Gy\|_2 \ ,$$

so

$$\sigma_n(\Sigma^{1/2}G) \ge \min_{y\in N_\varepsilon} \|\Lambda_{k^*}^{1/2}Gy\|_2 - \frac{\varepsilon}{1-\varepsilon}\max_{y\in N_\varepsilon}\|\Lambda_{k^*}^{1/2}Gy\|_2 \ . \tag{11}$$

Elementary computations show that, for any i, $\lambda_i(\Sigma)g_i^2$ is sub-exponential (see Definition 1) with parameters $(2\lambda_i(\Sigma), 4\lambda_i(\Sigma))$. As these variables are independent, by Proposition 1, $\sum_{i=k^*}^p \lambda_i(\Sigma)g_i^2$ is sub-exponential with parameters $\big(2\big(\sum_{i=k^*}^p \lambda_i^2(\Sigma)\big)^{1/2}, 4\lambda_{k^*}(\Sigma)\big)$. Therefore, by Proposition 2, with probability $1-2\exp(-t)$,

$$\begin{aligned}
\Big|\sum_{i=k^*}^p \lambda_i(\Sigma)g_i^2 - r_{k^*}(\Sigma)\Big| &\le \max\left(\sqrt{8t\sum_{i=k^*}^p \lambda_i^2(\Sigma)}, 8t\lambda_{k^*}(\Sigma)\right) \\
&\le \max\left(\sqrt{8\lambda_{k^*}(\Sigma)\sum_{i=k^*}^p \lambda_i(\Sigma)}, 8t\lambda_{k^*}(\Sigma)\right) \\
&\le \frac{r_k^*(\Sigma)}{2} + 12t\lambda_{k^*}(\Sigma) \ ,
\end{aligned} \tag{12}$$

where we used the inequality $\sqrt{ab} \le a/2 + b/2$, for any $a, b > 0$ to get the last expression. A union bound shows therefore that, with probability $1-2\exp(-t+n\log(3/\epsilon))$,

$$\frac{r_k^*(\Sigma)}{2} - 12t\lambda_{k^*}(\Sigma) \le \min_{y\in N_\varepsilon}\|\Lambda_{k^*}^{1/2}Gy\|_2 \le \max_{y\in N_\varepsilon}\|\Lambda_{k^*}^{1/2}Gy\|_2 \le \frac{3r_k^*(\Sigma)}{2} - 12t\lambda_{k^*}(\Sigma) \ .$$

Plugging this bound into (11) yields

$$\sigma_n(\Sigma^{1/2}Gx) \ge \left(1-\frac{\sqrt{3}\varepsilon}{1-\varepsilon}\right)\sqrt{\frac{r_{k^*}(\Sigma)}{2}} - 2\left(1+\frac{\sqrt{3}\varepsilon}{1-\varepsilon}\right)\sqrt{3t\lambda_{k^*}(\Sigma)} \ .$$

For $\varepsilon = 1/4$, $t = n(1+\log(3/\varepsilon))$, this yields the result if c_0 in the definition (3) of k^* is large enough. □

Theorem 2 then follows directly from Lemmas 1 and 2.

□

3.2. Proof of the upper bounds on the ellipsoid risk in Theorem 1

We start this section with two Lemmas. Lemma 3 enables to control the deviation of a quadratic process uniformly over a subset of $B_\Sigma(r)$, $r>0$. The main quantity driving this deviation is the Gaussian mean width that we introduce now. For any set $\mathcal{B}\subset\mathbb{R}^p$ we define the Gaussian mean width of $\mathcal{B}$ as

$$w(\mathcal{B}) = \mathbb{E}\sup_{\beta\in\mathcal{B}}\langle\beta, g\rangle, \quad g\sim\mathcal{N}(0,I_p)\ . \tag{13}$$

The Gaussian mean width serves as a measure of effective dimension of the set $\mathcal{B}$ (see [2] for equivalent formulations).

Lemma 3. *Let $r,\rho>0$ and $\delta\geq\exp(-n)$. Define $H_{r,\rho}=B(r)\cap B_{\Sigma^{-1/2}}(\rho)$, where we recall that $B_{\Sigma^{-1/2}}(\rho)=\{\beta\in\mathbb{R}^p:\sqrt{\beta^T\Sigma^{-1}\beta}\leq\rho\}$ and $B(r)=\{\beta\in\mathbb{R}^p:\sqrt{\beta^T\beta}\leq r\}$. There exists an absolute constant $c>0$ such that with probability larger than $1-\delta$*

$$\sup_{\substack{\|\beta-\beta^*\|_2\leq\rho\\ \|\Sigma^{1/2}(\beta-\beta^*)\|_2\leq r}}\left|\frac{1}{n}\sum_{i=1}^n\langle X_i,\beta-\beta^*\rangle^2-\mathbb{E}\langle X_i,\beta-\beta^*\rangle^2\right| \tag{14}$$

$$\leq c\left(\frac{w^2\big(H_{r,\rho}\big)}{n}+r\sqrt{\frac{w^2\big(H_{r,\rho}\big)}{n}}+r^2\sqrt{\frac{\log(1/\delta)}{n}}\right)\ , \tag{15}$$

Proof. Let $\delta\geq\exp(-n)$, $\mathcal{B}\subset\mathbb{R}^p$ and $r>0$. Let $(g_i)_{i=1}^n$ be n i.i.d centered standard Gaussian vectors in $\mathbb{R}^p$. From [18, Theorem 5.5], there exists an absolute constant $c>0$ such that with probability larger than $1-\delta$

$$\sup_{\beta\in B(r)\cap\mathcal{B}}\left|\frac{1}{n}\sum_{i=1}^n\langle g_i,\beta\rangle^2-\mathbb{E}\langle g_i,\beta\rangle^2\right|$$

$$\leq c\left(\frac{w^2(B(r)\cap\mathcal{B})}{n}+r\sqrt{\frac{w^2(B(r)\cap\mathcal{B})}{n}}+r^2\sqrt{\frac{\log(1/\delta)}{n}}\right)\ . \tag{16}$$

The rest of the proof simply consists in rewritting the empirical process (14). Since X_i is distributed as $\Sigma^{1/2}g_i$, where $g_i\sim\mathcal{N}(0,I_p)$ it follows that

$$\sup_{\substack{\|\beta-\beta^*\|_2\leq\rho\\ \|\Sigma^{1/2}(\beta-\beta^*)\|_2=r}}\left|\frac{1}{n}\sum_{i=1}^n\langle X_i,\beta-\beta^*\rangle^2-\mathbb{E}\langle X_i,\beta-\beta^*\rangle^2\right|$$

$$=\sup_{\substack{\|\Sigma^{-1/2}\beta\|_2\leq\rho\\ \|\beta\|_2=r}}\left|\frac{1}{n}\sum_{i=1}^n\langle g_i,\beta\rangle^2-\mathbb{E}\langle g_i,\beta\rangle^2\right|\ .$$

Applying (16) to the right-hand term concludes the proof. □

Lemma 4. *Let $r, \rho > 0$. The following holds*

$$w\big(B(r) \cap B_{\Sigma^{-1/2}}(\rho)\big) \leq \sqrt{2}\Big(\sum_{i=1}^{p} \lambda_i(\Sigma)\rho^2 \wedge r^2\Big)^{1/2} \ . \tag{17}$$

Proof. From Equation (13), we have, for $g \sim \mathcal{N}(0, I_p)$

$$w\big(B(r) \cap B_{\Sigma^{-1/2}}(\rho)\big) = \mathbb{E} \sup_{\beta \in B(r) \cap B_{\Sigma^{-1/2}}(\rho)} \langle g, t \rangle \ .$$

Moreover

$$\begin{aligned} B(r) \cap B_{\Sigma^{-1/2}}(\rho) &= \{\beta \in \mathbb{R}^p : \|\beta\|_2 \leq r, \|\Sigma^{-1/2}\beta\|_2 \leq \rho\} \\ &= \left\{\beta \in \mathbb{R}^p : \sum_{i=1}^{p} \frac{\beta_i^2}{\lambda_i(\Sigma)\rho^2} \leq 1, \ \sum_{i=1}^{p} \frac{\beta_i^2}{r^2} \leq 1\right\} \\ &\subset \left\{\beta \in \mathbb{R}^p : \sum_{i=1}^{p} \frac{\beta_i^2}{\lambda_i(\Sigma)\rho^2 \wedge r^2} \leq 2\right\} \ . \end{aligned}$$

The Gaussian mean-width of an ellipsoid is given by [33, Proposition 2.5.1] and it follows that

$$w\big(B(r) \cap B_{\Sigma^{-1/2}}(\rho)\big) \leq \sqrt{2}\Big(\sum_{i=1}^{p} \lambda_i(\Sigma)\rho^2 \wedge r^2\Big)^{1/2} \ . \tag{18}$$

□

Theorem 3. *For $\gamma > 0$, let us define*

$$\bar{r}(\gamma) = \sup\left\{r > 0 : \sum_{i=1}^{p} \lambda_i(\Sigma)\rho^2 \wedge r^2 \leq \gamma\|\xi\|_2^2\right\} \ .$$

Assume that $X_i|\xi \sim \mathcal{N}(0, \Sigma)$. There exist $c_1, c_2, c_3 > 0$ such that if γ is small enough, then with probability larger than $1 - c_1\exp(-cn_2)$

$$\bar{r}^2(\gamma) \wedge c_3\frac{\|\xi\|_2^2}{n} \leq \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2 \ .$$

Proof. Let $r = \bar{r}^2(\gamma) \wedge c_3\frac{\|\xi\|_2^2}{n}$ and define

$$\Omega_1 = \{\|\hat{\beta} - \beta^*\|_2 \leq \rho\}, \quad \rho = \|\beta^*\|_2 + \frac{4\|\xi\|_2}{\sqrt{r_{k^*}(\Sigma)}} \ . \tag{19}$$

From Theorem 2, the event Ω_1 holds with probability larger than $1-2\exp(-n)$. Until the end of the proof, we place ourselves on the event Ω_1. Since $\hat{\beta}$ is an interpolator, we have $\mathbb{X}(\hat{\beta}-\beta^*)=\xi$ and it follows that

$$\frac{1}{n}\sum_{i=1}^{n}\langle X_i,\hat{\beta}-\beta^*\rangle^2=\frac{\|\xi\|_2^2}{n}\ . \tag{20}$$

Assume that $\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2\le r$. On Ω_1, conditionally on ξ, we have

$$\begin{aligned}\frac{\|\xi\|_2^2}{n}&\le \sup_{\substack{\|\beta-\beta^*\|_2\le\rho\\ \|\Sigma^{1/2}(\beta-\beta^*)\|_2\le r}}\frac{1}{n}\sum_{i=1}^{n}\langle X_i,\beta-\beta^*\rangle^2\\ &\le r^2+\sup_{\substack{\|\beta-\beta^*\|_2\le\rho\\ \|\Sigma^{1/2}(\beta-\beta^*)\|_2\le r}}\left|\frac{1}{n}\sum_{i=1}^{n}\langle X_i,\beta-\beta^*\rangle^2-\mathbb{E}\langle X_i,\beta-\beta^*\rangle^2\right|\\ &\le r^2+c\left(\frac{w^2\big(H_{r,\rho}\big)}{n}+r\sqrt{\frac{w^2\big(H_{r,\rho}\big)}{n}}+r^2\sqrt{\frac{\log(1/\delta)}{n}}\right)\ ,\end{aligned}$$

where the last inequality holds with probabiliy larger than $1-\delta$, for $\delta\ge\exp(-n)$, according to Lemma 3. Now, from Lemma 4 and using the inequality $\sqrt{ab}\le a/2+b/2$ for $a,b>0$ we obtain

$$\begin{aligned}\frac{\|\xi\|_2^2}{n}&\le c\left[r^2\left(1+\frac{\log(1/\delta)}{n}\right)\right)+\frac{w^2\big(H_{r,\rho}\big)}{n}\Bigg]\\ &\le c\left[r^2\left(1+\frac{\log(1/\delta)}{n}\right)+2\frac{\sum_{i=1}^{p}\lambda_i(\Sigma)\rho^2\wedge r^2}{n}\right]\\ &\le cc_3\frac{\|\xi\|_2^2}{n}\left(1+\frac{\log(1/\delta)}{n}\right)+2c\gamma\frac{\|\xi\|_2^2}{n}\ ,\end{aligned}$$

where the last inequality holds because of the definition of $\bar{r}(\gamma)$. Taking $\delta=\exp(-n)$ and $\gamma=1/(4c)$ leads to a contradiction for c_3 small enough. □

Theorem 4. *For $\eta>0$, let us define*

$$r^*(\eta)=\inf\left\{r>0:\sum_{i=1}^{p}\lambda_i(\Sigma)\wedge r^2\le\eta n r^2\right\}\ .$$

There exists $c_1,c_2,c_3>0$ such that if η is small enough, then with probability larger than $1-c_1\exp(-cn_2)$,

$$\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2\le\big(\rho r^*(\eta)\big)^2\vee c_3\frac{\|\xi\|_2^2}{n}\ .$$

Proof. Until the end of the proof, we place ourselves on the event Ω_1, defined in (19). The proof is splitted in two parts.

1) Consider first the case where $\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2 \leq r^*(\eta)\|\hat{\beta}-\beta^*\|_2$.
Then, on Ω_1, it follows that $\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2 \leq r^*(\eta)\rho$ so the conclusion of Theorem 4 holds.

2) Now, consider the case where $\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2 \geq r^*(\eta)\|\hat{\beta}-\beta^*\|_2$.
From (20), we have

$$\frac{\|\xi\|_2^2}{n} = \frac{\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2}{nr^*(\eta)^2}\sum_{i=1}^n \left\langle X_i, r^*(\eta)\frac{\hat{\beta}-\beta^*}{\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2}\right\rangle^2 .$$

Let us define $\tilde{\beta}-\beta^* = r^*(\eta)(\hat{\beta}-\beta^*)/\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2$. Since, $\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2 \geq r^*(\eta)\|\hat{\beta}-\beta^*\|_2$, we have $\|\tilde{\beta}-\beta^*\|_2 \leq 1$ and $\|\Sigma^{1/2}(\tilde{\beta}-\beta^*)\|_2 = r^*(\eta)$ and it follows that

$$\begin{aligned}\frac{\|\xi\|_2^2}{n} &= \frac{\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2}{nr^*(\eta)^2}\sum_{i=1}^n \langle X_i, \tilde{\beta}-\beta^*\rangle^2 \\ &\geq \frac{\|\Sigma^{1/2}(\hat{\beta}-\beta^*)\|_2^2}{r^*(\eta)^2} \inf_{\substack{\|\beta-\beta^*\|_2\leq 1 \\ \|\Sigma^{1/2}(\beta-\beta^*)\|_2 = r^*(\eta)}} \frac{1}{n}\sum_{i=1}^n \langle X_i, \beta-\beta^*\rangle^2 . \end{aligned} \tag{21}$$

Morover, we have

$$\begin{aligned}&\inf_{\substack{\|\beta-\beta^*\|_2\leq 1 \\ \|\Sigma^{1/2}(\beta-\beta^*)\|_2 = r^*(\eta)}} \frac{1}{n}\sum_{i=1}^n \langle X_i, \beta-\beta^*\rangle^2 \\ &\qquad \geq r^*(\eta)^2 - \underbrace{\sup_{\substack{\|\beta-\beta^*\|_2\leq 1 \\ \|\Sigma^{1/2}(\beta-\beta^*)\|_2 = r^*(\eta)}} \left|\frac{1}{n}\sum_{i=1}^n \langle X_i, \beta-\beta^*\rangle^2 - \mathbb{E}\langle X_i, \beta-\beta^*\rangle^2\right|}_{\star} . \end{aligned} \tag{22}$$

Finally, from Lemmas 3 and 4 and the definition of $r^*(\eta)$, we have

$$\begin{aligned}\star &\leq c\left(\frac{\sum_{i=1}^p \lambda_i(\Sigma)\wedge r^*(\eta)^2}{n} + r^*(\eta)\sqrt{\frac{\sum_{i=1}^p \lambda_i(\Sigma)\wedge r^*(\eta)^2}{n}} + r^*(\eta)^2\sqrt{\frac{\log(1/\delta)}{n}}\right) \\ &\leq c\eta r^*(\eta)^2 + c\sqrt{\eta}r^*(\eta)^2 + cr^*(\eta)^2\sqrt{\frac{\log(1/\delta)}{n}} \leq \frac{r^*(\eta)^2}{2} .\end{aligned}$$

The last inequality holds if η is small enough and $\delta = \exp(-c_1 n)$ with $c_1 > 0$ small enough. Putting this bound into (22) yields

$$\inf_{\substack{\|\beta-\beta^*\|_2\leq 1 \\ \|\Sigma^{1/2}(\beta-\beta^*)\|_2 = r^*(\eta)}} \frac{1}{n}\sum_{i=1}^n \langle X_i, \beta-\beta^*\rangle^2 \geqslant \frac{r^*(\eta)^2}{2} .$$

Together with (21), this finally leads to

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

□

3.3. Proof of Corollaries 1 and 2

Proof of Corollary 1 For any $r>0$ and $k=\lfloor \eta n/2\rfloor := cn$,

$$\sum_{i=1}^{p}\lambda_i(\Sigma)\wedge r^2 \le r_k(\Sigma)+\big(\lfloor \eta n/2\rfloor\big)r^2 \le r_k(\Sigma)+\big(\eta n/2\big)r^2 \ ,$$

and it follows that

$$\sum_{i=1}^{p}\lambda_i(\Sigma)\wedge r^2 \le \eta n r^2, \qquad \text{if} \qquad r^2 \ge \frac{2}{\eta}\frac{r_k(\Sigma)}{n} \ .$$

Hence,

$$r^*(\eta)^2 \lesssim \frac{r_{cn}(\Sigma)}{n} \ .$$

Thus,

$$\rho^2 r^*(\eta)^2 \le \left(\frac{\|\beta^*\|_2^2 r_{cn}(\Sigma)}{n}\right)\vee\left(\frac{\|\xi\|_2^2 r_{cn}(\Sigma)}{n r_{k*}(\Sigma)}\right) \le \left(\frac{\|\beta^*\|_2^2 r_{cn}(\Sigma)}{n}\right)\vee\left(\frac{\|\xi\|_2^2}{n}\right) \ .$$

Therefore the result follows from Theorem 1

Prood Corollary 2 Assume that $\|\beta^*\|_2^2/\|\xi\|_2^2 \le 1/r_{k^*}(\Sigma)$. Thus, we have $\rho \lesssim \|\xi\|_2/\sqrt{r_{k^*}(\Sigma)}$. For any $r>0$, from the definition of $\bar{k}$ given in (6)

$$\sum_{i=1}^{p}\lambda_i(\Sigma)\rho^2\wedge r^2 \lesssim \|\xi\|_2^2\frac{r_{\bar{k}}(\Sigma)}{r_{k^*}(\Sigma)}+\bar{k}r^2 \le \frac{\gamma}{2}\|\xi\|_2^2+\bar{k}r^2 \ ,$$

and it follows that $\bar{r}^2(\gamma)\gtrsim \|\xi\|_2^2/(\gamma\bar{k})$ and therefore the result follows from Theorem 1.

Appendix A: Supplementary material

A.1. Sub-exponential random variables: definitions and properties

The following definition and propositions can be found in [38].

Definition 1. *A random variable X with mean $\mathbb{E}[X] = \mu$ is called sub-exponential with non-negative parameters (ν, b) if*

$$\mathbb{E}\big[e^{\lambda(X-\mu)}\big] \leq e^{\nu^2\lambda^2/2} \quad \text{for all} \quad |\lambda| \leq 1/b \ . \tag{23}$$

Proposition 1. *Let $X_1, \cdots, X_n$ be independent random variables such that X_i is sub-exponential with parameters (ν_i, b_i). Then $Y = \sum_{i=1}^n X_i$ is sub-exponential with parameters $\big((\sum_{i=1}^n \nu_i^2)^{1/2}, \max_{i=1,\cdots,n} b_i\big)$.*

Proposition 2 (Sub-exponential tail bound)**.** *Suppose that X is sub-exponential with parameters (ν, b). Then*

$$\mathbb{P}\big(|X-\mu| \geq t\big) \leq \begin{cases} 2e^{-t^2/(2\nu^2)} & \text{if } 0 < t \leq \nu^2/b \ , \\ 2e^{-t/(2b)} & \text{if } t \geq \nu^2/b \ . \end{cases} \tag{24}$$

References

[1] Madhu S Advani and Andrew M Saxe, *High-dimensional dynamics of generalization error in neural networks*, arXiv preprint arXiv:1710.03667 (2017).

[2] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp, *Living on the edge: Phase transitions in convex programs with random data*, Information and Inference: A Journal of the IMA **3** (2014), no. 3, 224–294.

[3] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang, *On exact computation with an infinitely wide neural net*, Advances in Neural Information Processing Systems, 2019, pp. 8139–8148.

[4] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler, *Benign overfitting in linear regression*, arXiv preprint arXiv:1906.11300 (2019).

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, *Reconciling modern machine-learning practice and the classical bias–variance trade-off*, Proceedings of the National Academy of Sciences **116** (2019), no. 32, 15849–15854.

[6] Mikhail Belkin, Daniel Hsu, and Ji Xu, *Two models of double descent for weak features*, arXiv preprint arXiv:1903.07571 (2019).

[7] Mikhail Belkin, Daniel J Hsu, and Partha Mitra, *Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate*, Advances in neural information processing systems, 2018, pp. 2300–2311.

[8] Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov, *Does data interpolation contradict statistical optimality?*, arXiv preprint arXiv:1806.09471 (2018).

[9] Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al., *Simultaneous analysis of lasso and dantzig selector*, The Annals of Statistics **37** (2009), no. 4, 1705–1732.

[10] Alberto Bietti and Julien Mairal, *On the inductive bias of neural tangent kernels*, Advances in Neural Information Processing Systems, 2019, pp. 12873–12884.

[11] Peter Bühlmann and Sara Van De Geer, *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media, 2011.

[12] Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp, *Interpolation under latent factor regression models*, arXiv preprint arXiv:2002.02525 (2020).
[13] Emmanuel J Candes, Justin K Romberg, and Terence Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences **59** (2006), no. 8, 1207–1223.
[14] George Casella, *Minimax ridge regression estimation*, The Annals of Statistics (1980), 1036–1056.
[15] Niladri S Chatterji, Philip M Long, and Peter L Bartlett, *When does gradient descent with logistic loss find interpolating two-layer networks?*, arXiv preprint arXiv:2012.02409 (2020).
[16] S.S. Chen, D.L. Donoho, and M.A. Saunders, *Atomic decomposition by basis pursuit*, SIAM J. Sci. Comput. **20** (1998), 33–61.
[17] Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco, *Elastic-net regularization in learning theory*, Journal of Complexity **25** (2009), no. 2, 201–230.
[18] Sjoerd Dirksen et al., *Tail bounds via generic chaining*, Electronic Journal of Probability **20** (2015).
[19] Jianqing Fan, Runze Li, Cun-Hui Zhang, and Hui Zou, *Statistical foundations of data science*, CRC press, 2020.
[20] Vitaly Feldman, *Does learning require memorization? a short tale about a long tail*, arXiv preprint arXiv:1906.05271 (2019).
[21] Christophe Giraud, *Introduction to high-dimensional statistics*, vol. 138, CRC Press, 2014.
[22] Arthur E Hoerl and Robert W Kennard, *Ridge regression: Biased estimation for nonorthogonal problems*, Technometrics **12** (1970), no. 1, 55–67.
[23] Arthur Jacot, Franck Gabriel, and Clément Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, Advances in neural information processing systems, 2018, pp. 8571–8580.
[24] Vladimir Koltchinskii and Shahar Mendelson, *Bounding the smallest singular value of a random matrix without concentration*, Int. Math. Res. Not. IMRN (2015), no. 23, 12991–13008. MR3431642
[25] Guillaume Lecué and Shahar Mendelson, *Learning subgaussian classes: Upper and minimax bounds*, arXiv preprint arXiv:1305.4825 (2013).
[26] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington, *Wide neural networks of any depth evolve as linear models under gradient descent*, Advances in neural information processing systems, 2019, pp. 8570–8581.
[27] Tengyuan Liang and Alexander Rakhlin, *Just interpolate: Kernel" ridgeless" regression can generalize*, arXiv preprint arXiv:1808.00387 (2018).
[28] Song Mei and Andrea Montanari, *The generalization error of random features regression: Precise asymptotics and double descent curve*, arXiv preprint arXiv:1908.05355 (2019).
[29] Shahar Mendelson, *Learning without concentration*, Conference on Learning Theory, 2014, pp. 25–39.

[30] ———, *Upper bounds on product and multiplier empirical processes*, Stochastic Processes and their Applications **126** (2016), no. 12, 3652–3680.
[31] ———, *On multiplier processes under weak moment assumptions*, Geometric Aspects of Functional Analysis, Springer, 2017, pp. 301–318.
[32] Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel Roy, *In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors*, International Conference on Machine Learning, PMLR, 2020, pp. 7263–7272.
[33] Michel Talagrand, *Upper and lower bounds for stochastic processes: modern methods and classical problems*, vol. 60, Springer Science & Business Media, 2014.
[34] Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B (Methodological) **58** (1996), no. 1, 267–288.
[35] Sara A Van de Geer et al., *High-dimensional generalized linear models and the lasso*, The Annals of Statistics **36** (2008), no. 2, 614–645.
[36] Roman Vershynin, *Estimation in high dimensions: a geometric perspective*, Sampling theory, a renaissance, Springer, 2015, pp. 3–66.
[37] ———, *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge university press, 2018.
[38] Martin J Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge University Press, 2019.
[39] P. Wojtaszczyk, *Stability and Instance Optimality for Gaussian Measurements in Compressed Sensing*, Found. Comput. Math. **10** (2010), 1–13.
[40] P Wojtaszczyk, *Stability and instance optimality for gaussian measurements in compressed sensing*, Foundations of Computational Mathematics **10** (2010), no. 1, 1–13.
[41] Zitong Yang, Yu Bai, and Song Mei, *Exact gap between generalization error and uniform convergence in random feature models*, arXiv preprint arXiv:2103.04554 (2021).
[42] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, *Understanding deep learning requires rethinking generalization*, arXiv preprint arXiv:1611.03530 (2016).
[43] Lijia Zhou, Danica J Sutherland, and Nathan Srebro, *On uniform convergence and low-norm interpolation learning*, arXiv preprint arXiv:2006.05942 (2020).
[44] Hui Zou and Trevor Hastie, *Regularization and variable selection via the elastic net*, Journal of the royal statistical society: series B (statistical methodology) **67** (2005), no. 2, 301–320.