Benign overfitting
in the large deviation regime

Geoffrey Chinot and Matthieu Lerascle

Abstract: We investigate the benign overfitting phenomenon in the large deviation regime where the bounds on the prediction risk hold with probability $1 - e^{-\zeta n}$, for some absolute constant $\zeta$. We prove that these bounds can converge to 0 for the quadratic loss. We obtain this result by a new analysis of the interpolating estimator with minimal Euclidean norm, relying on a preliminary localization of this estimator with respect to the Euclidean norm. This new analysis complements and strengthens particular cases obtained in [4] for the square loss and is extended to other loss functions. To illustrate this, we also provide excess risk bounds for the Huber and absolute losses, two widely spread losses in robust statistics.

Keywords and phrases: Interpolation problems, statistical learning.

1. Introduction

In this paper, we consider Gaussian regression problems where one observes a dataset $D_n$ of i.i.d. random vectors $(x_i, y_i), i \in \{1, \ldots, n\}$ such that $y_i = \langle x_i, \beta^* \rangle + \xi_i$, where $\beta^* \in \mathbb{R}^p$ is an unknown vector, $x \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^p$ and $\xi \sim \mathcal{N}(0, \sigma^2) \in \mathbb{R}$ are independent random variables. Defining the matrix $X$ with lines $x_T$ and the vector $Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$, the set of least-squares estimators is defined by

$$\hat{\beta} \in \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - \langle x_i, \beta \rangle)^2 = \text{argmin}_{\beta \in \mathbb{R}^p} \|X\beta^* - Y\|_2^2.$$  

The solutions of this problem are $\hat{\beta} = X^gY$, where $X^g$ is any pseudo-inverse of $X$. When the dimension $p$ of $\beta$ is smaller than $n$, the least-squares estimator is typically unique and has a risk of order $O(\sigma^2 p/n)$, which deteriorates with the dimension $p$. This deterioration is unavoidable in general, a phenomenon known as the “curse of dimensionality” in statistical textbooks.

To bypass this issue, statisticians have focused on situations where $\beta^*$ satisfies some sparsity conditions, meaning that it belongs, or is close, to a known set $S$ of small dimensional subspaces $S \subset \mathbb{R}^p$. In many of these situations, least-squares estimators can be improved, by considering minimizers of regularized least-squares criteria of the form $\|X\beta - Y\|_2^2 + \Omega(\beta)$. Several examples of such procedures have been studied in the literature. Among the most popular ones, one can mention ridge regression [18, 12], the LASSO [29, 30, 9] and the elastic...
net \([34, 15]\). Regularization ensures that both the prediction risk
\[
E[(x, \hat{\beta} - \beta^*)^2 | D_n] = (\hat{\beta} - \beta^*)^T \Sigma (\hat{\beta} - \beta^*) = \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2^2
\]
and the estimation risk \(\|\hat{\beta} - \beta^*\|_2^2\) are controlled. These results hold even if \(p \geq n\) provided that \(\beta^*\) is close to a linear subspace \(S \subset \mathbb{R}^p\) with dimension \(s < n\).

When the dimension \(p \geq n\), the set of least-squares estimators is typically infinite. Actually, the matrix \(X\) in this case has typically full rank (and a non trivial kernel) and any solution in the set \(\{X^gY\}\), where \(X^g\) describes all pseudo-inverses of \(X\) satisfy \(XX^gY = Y\). 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. However, and perhaps counter-intuitively, it turns out that, when the dimension \(p\) is large in front of \(n\), the risk of prediction can become smaller for some of these solutions. This interesting phenomenon has given rise to a rapidly growing literature these last months, see \([5, 6, 7, 8, 11, 17, 23, 24]\). 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 \([1, 5, 33]\).

Of course, linear regression is much simpler than neural networks and the results proved here are not sufficient to explain the amazing prediction properties of these algorithms, but it is interesting to understand when and how high dimension helps prediction, at least in this simpler example. Moreover, several recent works have shown that the analysis of linear models can be relevant for over-parametrized neural networks. 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 \([19, 10, 3, 22]\). 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, we consider more precisely the problem of \([4]\) where the least-squares solution with minimal Euclidean norm is analysed. It is well known that this solution is \(\hat{\beta} = X^+ Y\), where \(X^+\) is the Moore-Penrose pseudo inverse of \(X\). Our main results complement those in \([4]\) in the following sense. First, our results are derived in the large deviation regime, meaning that they hold with probability \(1 - e^{-\zeta n}\), for some absolute constant \(\zeta\). This regime is considered in \([4]\) but the bounds there don’t converge to 0 as \(n \to \infty\). On the contrary, our bounds can converge to 0 under proper assumptions on the spectrum of the covariance matrix \(\Sigma = \mathbb{E}[xx^T]\). These assumptions involve the rest of the series of singular values of the matrix \(\Sigma\), \(r_{k^*}(\Sigma) = \sum_{k=k^*}^p \lambda_i(\Sigma)\) for a well chosen index \(k^*\) as in \([4]\). The index \(k^*\) in our result is typically slightly larger than the one in \([4]\) by a logarithmic factor, see \([4]\) for a definition of \(k^*\) and the discussion at the end of Section 3.1 for a precise comparison between the \(k^*\) in
a particular example. Besides considering the large deviation regime, our new bounds improve those of [4] in typical examples where benign overfitting occurs, see the discussion following Corollary 1. These improvements are made possible by a new analysis of the estimator \( \hat{\beta} \), that relies on preliminary results showing that dimension may help to localize this estimator with respect to the estimation norm \( \| \hat{\beta} - \beta \|_2 \), see Theorem 3. This localization allows, for example, to prove rates of convergence that can be as fast as \( 1/n \) for this estimator, while the bounds in [4] only allow to reach \( 1/\sqrt{n} \). Our bounds exhibit a phase transition of the rates of convergence when the signal to noise ratio \( \text{SNR} = \| \beta^* \|_2^2 / \sigma^2 \) becomes larger than a threshold \( t = n/r_{k^*}(\Sigma) \) (this threshold typically grows to infinity in the examples). When \( \text{SNR} > t \), the prediction risk of the estimator satisfies, in the large deviation regime, \( \| \Sigma^{1/2}(\hat{\beta} - \beta^*) \|_2^2 \lesssim \| \beta^* \|_2^2 \text{Tr}(\Sigma)/n \). This rate can be exponentially better than the one in [4] for some spectrum of the covariance matrix \( \Sigma \), even if it holds with probability \( 1 - e^{-\zeta n} \) in our result and with constant probability in [4] (see the example following Corollary 1). On the other hand, when the SNR is too low, \( \text{SNR} \leq t \), these rates deteriorate into \( \| \Sigma^{1/2}(\hat{\beta} - \beta^*) \|_2^2 \lesssim \sigma^2 + k^*/n \). In this case, our rates improve those of [4] which are always larger than \( \sigma^2 k^* \) in the large deviation regime and actually met the optimal rate \( \sigma^2 \) as proved in [21, Theorem A'].

Besides the least-squares loss, our new strategy can be easily applied to analyse the excess risk of interpolating estimators with respect to other loss functions. This extension was mentioned as a relevant conjecture in [4]. We illustrate this by providing a short analysis of the excess risk of \( \hat{\beta} \) with respect to the Huber loss and the absolute loss, two widely spread methods in robust statistics. The bounds obtained on the excess risk of \( \hat{\beta} \) with respect to these losses involve the same quantities as for the quadratic loss. They are gathered in Theorem 2.

The remainder of the paper is decomposed as follows. Section 2 sets the main notations and recall the construction of the estimator \( \hat{\beta} \). Section 3 gathers the main results of the paper, the upper bounds on the excess risk of the estimator \( \beta \) with respect to the quadratic, absolute and Huber losses. The proofs of these results are gathered in Section 4.

2. Setting

Let \( (x, y), (x_i, y_i)_{i=1,...,n} \) denote i.i.d random vectors generated according to the following Gaussian linear model,

\[
y = x^T \beta^* + \xi ,
\]

where \( \beta^* \in \mathbb{R}^p \) is the signal of interest, the design \( x \) is a Gaussian vector \( x \sim \mathcal{N}(0, \Sigma) \in \mathbb{R}^p \) and the noise \( \xi \) is a Gaussian random variable \( \xi \sim \mathcal{N}(0, \sigma^2) \), independent of \( x \). Let \( X \in \mathbb{R}^{n \times p} \) denote the matrix with lines \( x_1^T, \ldots, x_n^T \). Let \( Y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n \) and \( \xi = (\xi_1, \ldots, \xi_n)^T \). Using these notations, the dataset \( D_n = \{ (x_1, y_1), \ldots, (x_n, y_n) \} \) can be represented in the matrix form as

\[
Y = X \beta^* + \xi .
\]
The set of interpolating vectors $H_n \subset \mathbb{R}^p$ is defined as $H_n = \{ \beta \in \mathbb{R}^p : X \beta = Y \}$. We analyse the estimator defined as the interpolating vector with minimal Euclidean norm, that is

$$\hat{\beta} = \arg\min_{\beta \in H_n} \| \beta \|_2,$$

where $\| \cdot \|_2$ denotes the Euclidean norm in $\mathbb{R}^p$. This estimator is defined only when the set $H_n$ is non-empty. In general, this occurs only when $X$ has full rank $n$, which holds almost surely when the dimension $p$ is larger than the number of observations $n$, provided that $\Sigma$ has rank at least $n$. In the following, we assume therefore that $p \geq 4n$ and that $\Sigma$ has rank at least $n$. The constant 4 has no particular meaning here, it could be replaced by any constant strictly larger than 1 without affecting the results.

Our main results give upper bounds on the prediction loss of $\hat{\beta}$. Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ denotes a loss function such that $\ell(y, y) = 0$ for all $y \in \mathbb{R}$ and $\ell(y, y') > 0$ if $y \neq y'$. It is also assumed that the function $y \mapsto \ell(y, y')$ is convex for any $y \in \mathbb{R}$. In the first part of the paper, $\ell$ will be the square loss $\ell(y, y') = (y - y')^2$. Other losses will be considered in Section 3.2. For any $\beta \in \mathbb{R}^p$ and any $(u, v) \in \mathbb{R}^p \times \mathbb{R}$, let $\ell_\beta(u, v) = \ell((u, \beta)^T, v)$ and let $\mathcal{L}_\beta(u, v) = \ell_\beta(u, v) - \ell_{\beta^*}(u, v)$. For any function $f : \mathbb{R}^p \times \mathbb{R} \to \mathbb{R}$, let $P f = \mathbb{E}[f(x, y)]$. The excess risk is then defined as:

$$\mathbb{E}\left[ \ell\left(\langle x, \hat{\beta} \rangle, y \right) - \ell\left(\langle x, \beta^* \rangle, y \right) \bigg| D_n \right] = P(\ell_\beta - \ell_{\beta^*}) = P\mathcal{L}_\beta.$$

As usual, the expectation is taken over the random variables $(x, y)$ only, so the excess risk is a random variable. In this paper, we provide risk bounds for the estimator $\hat{\beta}$ that hold in the large deviation regime. This means that we build deterministic upper bounds $r_n$ on $P\mathcal{L}_\beta$ such that $P(P\mathcal{L}_\beta > r_n) \leq \exp(-\zeta n)$, for some absolute constant $\zeta$.

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 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 $r$. Let also, for any subset $B$ of $\mathbb{R}^p$, denote by $\beta + B = \{ u \in \mathbb{R}^p : \exists v \in B \text{ such that } u = \beta + v \}$. All along the paper, $c$ and $\zeta$ denote absolute positive constants. Typically, $\zeta$ denotes a small constant while $c$ denotes a large one.

3. Main results

This section provides our main contributions. Prediction bounds for the square loss are provided in Section 3.1 and for other loss functions in Section 3.2.
3.1. Prediction with least-squares loss

The following theorem is the main result of this paper.

**Theorem 1.** Let

\[ k^* = \inf \left\{ k \in \{1, \ldots, p\} : \frac{r_k(\Sigma)}{\lambda_k(\Sigma)} \geq \frac{32n \log \left( 1 + \frac{44}{3} \sqrt{p} \|\Sigma\| \right)}{r_k(\Sigma)} \right\}. \]  

(4)

Let \( \zeta > 0 \) be an absolute constant. Define the parameter \( v \), the estimation rate \( \rho \) and the prediction rate \( r^* \) by

\[ v = \frac{r_{k^*}(\Sigma)}{32n \lambda_{k^*}(\Sigma)}, \quad \rho = \|\beta^*\|_2 + \sigma \sqrt{\frac{32n}{r_{k^*}(\Sigma)}}, \]  

(5)

\[ r^* = \inf \left\{ r > 0 : \sum_{i=1}^p r^2 \wedge \lambda_i(\Sigma) \rho^2 \leq \zeta nr^2 \right\}. \]  

(6)

If \( k^* \leq cn \), for \( c > 0 \) an absolute constant, then, with probability larger than \( 1 - 7e^{-\left(v \wedge \zeta\right)n} \), the estimator \( \hat{\beta} \) defined in Equation (2) satisfies

\[ \|\hat{\beta} - \beta^*\|_2 \leq \rho \quad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \leq r^*. \]

Theorem 1 is proved in Section 4.2. The estimation bound \( \rho \) does not converge to 0, which is not surprising in our high dimensional setting, in absence of 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 \( \|\beta^*\|^2/\sigma^2 \) is larger that the threshold \( n/r_{k^*}(\Sigma) \), \( \|\hat{\beta} - \beta^*\|_2 \) is at most of order \( \|\beta^*\|_2 \) when the dimension is large enough.

To discuss the prediction bounds, it is useful to give the following corollary, whose proof is a direct consequence of Theorem 1 left as an exercise. The corollary shows a phase transition in the rates of convergence when the signal to noise ratio \( \text{SNR} = \|\beta^*\|^2/\sigma^2 \) becomes larger than the threshold \( t = n/r_{k^*}(\Sigma) \).

**Corollary 1.** Grant the assumptions and notations of Theorem 1.

- If the signal to noise ratio is large enough, \( \|\beta^*\|^2/\sigma^2 \geq n/r_{k^*}(\Sigma) \), the estimator \( \hat{\beta} \) defined in Equation (2) satisfies, with probability larger than \( 1 - 7e^{-(v \wedge \zeta)n} \),

\[ \|\hat{\beta} - \beta^*\|_2 \lesssim \|\beta^*\|_2, \quad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \lesssim \|\beta^*\|_2 \frac{\text{Tr}(\Sigma)}{n}. \]

- On the other hand, if the signal to noise ratio is too small, \( \|\beta^*\|^2/\sigma^2 \leq n/r_{k^*}(\Sigma) \), then, the estimator \( \hat{\beta} \) defined in Equation (2) satisfies, with probability larger than \( 1 - 7e^{-(v \wedge \zeta)n} \),

\[ \|\hat{\beta} - \beta^*\|_2 \lesssim \sigma \sqrt{\frac{n}{r_{k^*}(\Sigma)}}, \quad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \lesssim \left( \sigma^2 + \frac{k^*}{n} \right). \]
Corollary 1 can be used to compare our results with those in [4].

1. The upper bounds in Corollary 1 hold with probability larger than $1 - \exp(-\zeta n)$ and may converge to 0 while those in [4] are always larger than a constant at these confidence levels.

2. For high signal to noise ratios, $\text{SNR} = \frac{\|\beta^*\|^2}{\|\Sigma\|} > t = \frac{n}{r_k^*} / \sqrt{\text{Tr}(\Sigma) / n}$, Corollary 1 improves the results provided in [4], since the main term in this case here is $\|\beta^*\|^2 \frac{\text{Tr}(\Sigma)}{n}$ while it is $\|\beta^*\|^2 \sqrt{\text{Tr}(\Sigma) / n}$ in this paper.

3. For small signal to noise ratios, $\text{SNR} < t$, our rates are of order $\frac{\sigma^2 + k^*/n}{n}$, which improve the result of [4] at confidence levels $e^{-\zeta n}$. An interesting feature of the results in [4] is that it provides upper bounds that can converge to 0 at smaller confidence levels. On the other hand, [21, Theorem A'] shows that $\sigma^2$ is the optimal rate that can hold with probability larger than $1 - \exp(-\zeta n)$.

4. The parameter $k^*$ in Theorem 1 is slightly larger in general than the one in [4], since they only require that $r_k^* / \lambda_k^* > \lambda_k^* / (\Sigma)$, while we have an extra logarithmic factor in the definition (4).

To illustrate the 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 $\epsilon = o(1)$ and $\tau = \Omega(1)$ such that, for any $k$,

$$\lambda_k^*(\Sigma) = e^{-k/\tau} + \epsilon,$$

with $\tau \log(1/\epsilon) < n$, $p = cn \log(1/\epsilon)$.

In this case, for any $k$ and $\gamma = \tau/(1 - e^{-\tau})$,

$$\frac{r_k}{\lambda_k} = \frac{(p - k)\epsilon + \gamma(e^{-k/\tau} - e^{-p/\tau})}{e^{-k/\tau} + \epsilon},$$

$$\frac{p\|\Sigma\|}{r_k^*(\Sigma)} = \frac{p}{(p - k)\epsilon + \gamma(e^{-k/\tau} - e^{-p/\tau})}.$$

Therefore, for $k = \tau \log(1/\epsilon) < p/2$ and $c$ large enough,

$$\frac{r_k}{\lambda_k} \geq \frac{pe/2 + \gamma \epsilon}{2\epsilon} \geq \frac{p}{4} \geq 32n \log \left( 1 + \frac{44}{3} \sqrt{\frac{2}{\epsilon}} \right) \geq 32n \log \left( 1 + \frac{44}{3} \sqrt{\frac{p\|\Sigma\|}{rk(\Sigma)}} \right).$$

Hence, $k^* \leq \tau \log(1/\epsilon) < n$. Moreover, $r_k^* / \lambda_k^* = \Theta(p\epsilon) = \Theta(n \log(1/\epsilon))$ so the threshold $t$ for the SNR ratio is $t = \Omega(1/\epsilon \log(1/\epsilon))$. This threshold therefore grow to infinity if $\epsilon \to 0$. As $\text{Tr}(\Sigma) \leq pe + \tau$ and the parameter $v \gtrsim pe/(n\epsilon) \gtrsim 1$, Corollary 1 shows in this example that, if $\|\beta^*\|^2 / \sigma^2 \geq 1 / (\epsilon \log(1/\epsilon))$, with probability larger than $1 - e^{-\zeta n}$,

$$\|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|^2 \lesssim \|\beta^*\|^2 \frac{pe + \tau}{n} = \|\beta^*\|^2 \left( \epsilon \log(1/\epsilon) + \frac{\tau}{n} \right).$$

Our rates of convergence in this example can therefore be, up to logarithmic factors as fast as $\epsilon \lor (1/n)$, while [4, Theorem 6] gives in this setting a
rate \((1/\log(1/\epsilon)) \vee (1/n)\) that is exponentially slower. In addition, let us recall that Corollary 1 here shows that the rate \(\epsilon \vee (1/n)\) holds with probability \(1 - e^{-\zeta n}\) while [4, Theorem 6] only shows that the logarithmically slower rate \((1/\log(1/\epsilon)) \vee (1/n)\) holds with constant probability.

3.2. Extension to other loss functions

The purpose of this section is to show that the analysis developed to prove the main theorem can be easily extended and that the excess risk of \(\hat{\beta}\) with respect to other loss functions can be controlled with the same arguments. To illustrate this general principle, we consider two losses, namely, the Huber and absolute losses. Both losses have been used repeatedly in robust statistics. Formally, let \(\ell : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^+\) denote one of the following convex loss function:

- The **Huber loss** is defined, for any \(u, y \in \mathbb{R}\), by
  \[
  \ell(u, y) = \varphi_H(u - y), \quad \text{where} \quad \varphi_H(u) = \begin{cases} 
  \frac{1}{2}u^2 & \text{if } |u| \leq \delta \\
  \delta |u| - \delta^2/2 & \text{if } |u| > \delta
  \end{cases}
  \]

  Notice that \(\varphi_H\) is \(\delta\)-Lipschitz.

- The **absolute loss** is defined, for any \(u, y \in \mathbb{R}\), by \(\ell(u, y) = \varphi_A(u - y)\), where \(\varphi_A(u) = |u|\) is 1-Lipschitz.

For both losses, recall that, for any \(\beta \in \mathbb{R}^p\), by \(PL_\hat{\beta} = P[\ell_\hat{\beta} - \ell_{\beta^*}]\), with \(\ell_{\beta}(x, y) = \ell(\langle x, \beta \rangle, y)\).

**Theorem 2.** There exist absolute constants \(\zeta, c, c_2\) such that the following holds. Let \(k^*, \rho, v, r^*\) be defined as in Theorem 1. If \(k^* \leq cn\), then

- if \(\ell\) is the Huber loss with \(\delta = c_2 \sigma\), with probability larger \(1 - 10e^{-(\zeta \wedge v)n}\),
  \[PL_\hat{\beta} \leq c(r^*)^2\]

- if \(\ell\) is the absolute loss, with probability larger \(1 - 8e^{-(\zeta \wedge v)n}\),
  \[PL_\hat{\beta} \leq cr^*\]

**Remark 1.** Theorem 2 is proved in Section 4.3. It shows that the excess risk for the Huber loss is of the same order as the one for the square loss. It is the square root of these rate for the absolute loss. Both results are expected as the same phenomenon appear in small dimension also, see for example [14].

4. Proofs of the main results

The remaining of the paper is devoted to the proofs of the main results. Section 4.1 (resp. 4.2) shows the estimation bound (resp. the prediction bounds) in Theorem 1.
4.1. 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 analysis of the prediction risk of $\hat{\beta}$. This approach is now classical in statistical learning, it has been applied successfully, for example, in [20, 25, 26, 27].

**Theorem 3.** There exist absolute constants $c$ and $\zeta$ such that the following holds. Let $k^*$, $v$ and $\rho$ be defined as in Theorem 1. If $k^* \leq cn$, the estimator $\hat{\beta}$ defined in Equation (2) satisfies

$$P(\|\hat{\beta} - \beta^*\|_2 \leq \rho) \geq 1 - 4 \exp(-v \land 1/5)n) \ .$$  

(7)

**Proof of Theorem 3.** The proof starts with the following lemma.

**Lemma 1.** With probability conditionally on $X$ larger than $1 - e^{-n/2}$,

$$\|\hat{\beta} - \beta^\ast\|_2 \leq \|\beta^\ast\|_2 + 2\sigma \sqrt{\frac{n}{\sigma(X)}} .$$  

(8)

**Proof of Lemma 1.** Classical results of linear algebra show that

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

where $X^+$ denotes the Moore-Penrose pseudo-inverse of $X$. Therefore,

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

(9)

where the last inequality follows from the triangular inequality and the fact that $X^+ X - I_p$ is the projection matrix onto the null-space of $X$. Since $\|X^+ \xi\|_2 \leq \|X^+\| \|\xi\|_2$, the function $\xi \mapsto \|X^+ \xi\|_2$ is $\|X^+\|$-Lipschitz with respect to the Euclidean norm. From Borell’s Gaussian concentration inequality, with probability conditionally on $X$ larger than $1 - \exp(-n/2)$,

$$\|X^+ \xi\|_2 \leq \mathbb{E}[\|X^+ \xi\|_2 | X] + \sigma \|X^+\| \sqrt{n} .$$  

(10)

Since rank($X$) $\leq n$, $\|X^+\| \leq \sigma_n^{-1}(X)$. Similarly, rank(($(X^+)^T X^+$) $\leq$ rank($X^+$) $\leq n$. Therefore, writing $\mathbb{E}[\cdot]$ for $\mathbb{E}[\cdot|X]$,

$$\mathbb{E}[\|X^+ \xi\|_2] \leq (\mathbb{E}[\|X^+ \xi\|_2^2])^{1/2}$$

$$= (\mathbb{E}[\xi^T (X^+)^T X^+ \xi])^{1/2} = \sigma((\mathbb{E}[(X^+)^T X^+])^{1/2}$$

$$= \sigma(\sum_{i=1}^n \lambda_i ((X^+)^T X^+))^{1/2} = \sigma(\sum_{i=1}^n \sigma_i^2(X^+))^{1/2}$$

$$= \sigma(\sum_{i=1}^n \sigma_i^{-2}(X))^{1/2} \leq \sigma \sqrt{\frac{n}{\sigma(X)}} .$$

Plugging (10) and this bound on $\mathbb{E}[\|X^+ \xi\|_2 | X]$ into (9) concludes the proof. $\blacksquare$
Lemma 1 provides a random bound on the estimation error of $\hat{\beta}$. To prove Theorem 3, it remains to bound from below, with high probability, the smallest eigenvalue $\sigma_n^2(X)$ of $XX^T$. This control is obtained in the following lemma.

**Lemma 2.** With probability larger than $1 - 2\exp(-p/18) - \exp(-nv)$, we have

$$\sigma_n(X) \geq \sqrt{\frac{r_{k^*}(\Sigma)}{8}}.
$$

**Proof.** The matrix $X^T$ 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 $\sigma_n(X) = \sigma_n(X^T)$ is distributed as $\sigma_n(\Sigma^{1/2}Gx)$. From the Courant-Fischer-Weyl min-max principle, we have

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

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

$$\sum_{i=1}^p \lambda_i(\Sigma)g_i^2 \geq \sum_{i=k^*}^p \lambda_i(\Sigma)g_i^2.$$  

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

$$\|\Lambda^{1/2}Gx\|_2^2 \geq \frac{1}{2}r_{k^*}(\Sigma). \tag{11}$$

Equation (11) holds for any fixed $x$ in the unit sphere $S^{n-1}$. To obtain uniform deviations, let us introduce an $\epsilon$-net $\Gamma_\epsilon$ of $S^{n-1}$. For any $x \in S^{n-1}$, there exists $y \in \Gamma_\epsilon$ such that $\|x - y\|_2 \leq \epsilon$. Thus,

$$\|\Sigma^{1/2}Gx\|_2 \geq \|\Sigma^{1/2}Gy\|_2 - \|\Sigma^{1/2}G(x - y)\|_2 \geq \|\Sigma^{1/2}Gy\|_2 - \epsilon\|\Sigma^{1/2}G\|.$$  

Since the operator norm is sub-multiplicative, $\|\Sigma^{1/2}G\| \leq \sqrt{\|\Sigma\|\|G\|}$. To bound the operator norm $\|G\|$, we use the following result.

**Theorem 4.** [31][Theorem 5.35]. Let $p \geq n$ and let $G$ denote a $p \times n$ matrix with independent standard Gaussian entries. For every $0 < \delta \leq 1$, with probability at least $1 - \delta$:

$$\sqrt{p} - \sqrt{n} - \sqrt{2\log(2/\delta)} \leq \sigma_{\min}(G) \leq \sigma_1(G) \leq \sqrt{p} + \sqrt{n} + \sqrt{2\log(2/\delta)}. \tag{12}$$
From Theorem 4, with probability larger that $1 - 2 \exp(-p/18)$,

$$\|G\| \leq \sqrt{p} + \sqrt{n} + \sqrt{\frac{2p}{18}} \leq \sqrt{p \left(1 + \frac{1}{2} + \frac{1}{3}\right)} = \frac{11\sqrt{p}}{6}.$$ 

It follows that

$$\min_{x \in S^{n-1}} \|\Sigma^{1/2}Gx\|_2 \geq \min_{y \in \Gamma_\varepsilon} \|\Sigma^{1/2}Gy\|_2 - \frac{11\sqrt{p}}{6} \varepsilon \sqrt{p\|\Sigma\|}.$$  \hspace{1cm} (13)

Hence, for

$$\varepsilon = \frac{6}{44} \sqrt{\frac{r_{k^*}(\Sigma)}{p\|\Sigma\|}},$$

we have

$$\min_{x \in S^{n-1}} \|\Sigma^{1/2}Gx\|_2 \geq \min_{y \in \Gamma_\varepsilon} \|\Sigma^{1/2}Gy\|_2 - \sqrt{\frac{r_{k^*}(\Sigma)}{4}}.$$  \hspace{1cm} (14)

Taking a union bound in (11), we get that, for this value of $\varepsilon$, with probability at least $1 - \exp \left(-2nv + \log(|\Gamma_\varepsilon|)\right)$,

$$\min_{x \in S^{n-1}} \|\Sigma^{1/2}Gx\|_2 \geq \sqrt{r_{k^*}(\Sigma) \left(\frac{1}{\sqrt{2}} \frac{1}{4}\right)} \geq \sqrt{\frac{r_{k^*}(\Sigma)}{8}}.$$ 

A standard volume argument shows that, for every $\varepsilon > 0$, $|\Gamma_\varepsilon| \leq (1 + 2/\varepsilon)^n$. Therefore, the probability estimate is bounded from below by

$$1 - \exp \left(-2nv + n \log \left(1 + \frac{44}{3} \sqrt{\frac{p\|\Sigma\|}{r_{k^*}(\Sigma)}}\right)\right).$$

By definition of $k^*$, this probability is bounded from below by

$$1 - \exp(-nv).$$

This concludes the proof of Lemma 2.

Theorem 3 then follows directly from Lemmas 1 and 2. \hfill \blacksquare

4.2. Proof of the prediction bound in Theorem 1

Let $\beta^* + B(\rho) = \{\beta \in \mathbb{R}^p : \|\beta - \beta^*\|_2 \leq \rho\}$. Let $P_n \mathcal{L}_\beta := n^{-1} \sum_{i=1}^n (\ell_\beta(x_i, y_i) - \ell_{\beta^*}(x_i, y_i))$ denote the empirical excess-risk. The proof starts with the following elementary result.
**Lemma 3.** With probability larger than $1 - \exp(-n/16)$, $P_n \mathcal{L}_\beta \leq -(1/2)\sigma^2$. Moreover, for any $r^*$, let $\Omega_{r^*,\rho}$ denote the following event

$$\Omega_{r^*,\rho} = \{\forall \beta \in \mathbb{R}^p \text{ such that } \beta - \beta^* \in B(\rho) \setminus B_\Sigma(r^*), P_n \mathcal{L}_\beta > -(1/2)\sigma^2\} .$$

On the event

$$\Omega_{r^*,\beta} \cap \{\hat{\beta} - \beta^* \in B(\rho)\} \cap \{P_n \mathcal{L}_\beta \leq -(1/2)\sigma^2\} ,$$

$\hat{\beta} - \beta \in B_\Sigma(r^*)$, that is

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

**Proof.** Since $\hat{\beta} \in H_n$, $\langle x_i, \hat{\beta} \rangle = y_i$ for any $i \in \{1, \ldots, n\}$, so $P_n \ell_{\hat{\beta}} = 0$ and

$$P_n \mathcal{L}_{\hat{\beta}} = P_n(\ell_{\hat{\beta}} - \ell_{\beta^*}) = -P_n \ell_{\beta^*} = -\frac{1}{n} \sum_{i=1}^n \xi_i^2 .$$

Since $\xi_i \sim \mathcal{N}(0, \sigma^2)$, from Proposition 1, $\sum_{i=1}^n \xi_i^2$ is sub-exponential with parameters $(2\sigma\sqrt{n}, 4\sigma^2)$ and from Proposition 2, with probability larger than $1 - \exp(-n/16)$,

$$P_n \mathcal{L}_{\hat{\beta}} = -\frac{1}{n} \sum_{i=1}^n \xi_i^2 \leq -(1/2)\sigma^2 . \tag{15}$$

On $\Omega_{r^*,\rho}$ all $\beta$ such that $\|\beta - \beta^*\|_2 \leq \rho$ and $\|\beta - \beta^*\|_\Sigma > r^*$ satisfy $P_n \mathcal{L}_\beta > -(1/2)\sigma^2$. Therefore, on $\Omega_{r^*,\rho}$, if $\|\hat{\beta} - \beta^*\|_2 \leq \rho$ and $P_n \mathcal{L}_{\hat{\beta}} \leq -(1/2)\sigma^2$, $\hat{\beta}$ cannot satisfy $\|\hat{\beta} - \beta^*\|_\Sigma > r^*$. Hence,

$$\{\hat{\beta} - \beta^* \in B_\Sigma(r^*)\} \supset \Omega_{r^*,\rho} \cap \{\hat{\beta} - \beta^* \in B(\rho)\} \cap \{P_n \mathcal{L}_\beta \leq -(1/2)\sigma^2\} .$$

By Lemma 3, to bound the excess risk of $\hat{\beta}$, it is sufficient to show that $r^*$ defined in (6) is such that, with high probability

$$\inf_{\beta: \beta - \beta^* \in B(\rho) \setminus B_\Sigma(r^*)} \{P_n \mathcal{L}_\beta\} > -(1/2)\sigma^2 . \tag{16}$$

**Theorem 5.** There exists an absolute constant $\zeta$ such that, with probability larger than $1 - 2e^{-\zeta n}$,

$$\inf_{\beta: \beta - \beta^* \in B(\rho) \setminus B_\Sigma(r^*)} \{P_n \mathcal{L}_\beta\} > -(1/2)\sigma^2 ,$$

where $r^*$ is the complexity parameter defined in (6).

By Lemma 3 and Theorem 3, this means that, with probability larger than $1 - 6e^{-\zeta n} - e^{-\nu n}$,

$$\|\hat{\beta} - \beta^*\|_2 \leq \rho, \quad \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \leq r^* . \tag{17}$$
Proof of Theorem 5. Let \( \beta \in \beta^* + B(\rho) \setminus B(\rho) \) and denote by \( r = \|\Sigma^{1/2} (\beta - \beta^*)\|_2 \), so \( r > r^* \) and

\[
\beta - \beta^* \in H_{r, \rho} = B(\rho) \cap S_{\Sigma}(r) .
\]

Recall that, for any \( \beta \in \mathbb{R}^p \), as

\[
\langle x_i, \beta \rangle - y_i = \langle x_i, \beta - \beta^* \rangle + \langle x_i, \beta^* \rangle - y_i = \langle x_i, \beta - \beta^* \rangle - \xi_i ,
\]

we have

\[
P_n L_{\beta} = \frac{1}{n} \sum_{i=1}^{n} (\langle x_i, \beta \rangle - y_i)^2 - \langle x_i, \beta^* \rangle^2 - 2 \frac{1}{n} \sum_{i=1}^{n} \xi_i \langle x_i, \beta - \beta^* \rangle . \tag{18}
\]

Write now \( \alpha = r^*/r \in (0, 1) \) and \( \beta_0 = \beta^* + \alpha (\beta - \beta^*) \), so

\[
P_n L_{\beta} = \alpha^{-2} \frac{1}{n} \sum_{i=1}^{n} (\langle x_i, \beta_0 - \beta^* \rangle)^2 - \alpha^{-1} \frac{2}{n} \sum_{i=1}^{n} \xi_i \langle x_i, \beta_0 - \beta^* \rangle . \tag{19}
\]

By definition, \( \|\Sigma^{1/2} (\beta_0 - \beta^*)\|_2 = r^* \) and \( \|\beta_0 - \beta^*\| \leq \alpha \rho \), that is, \( \beta_0 - \beta^* \in H_{r^*, \alpha \rho} = S_{\Sigma}(r^*) \cap B(\alpha \rho) \). Define then

\[
Q_{r^*, \rho} = \sup_{\beta - \beta^* \in H_{r^*, \rho}} \left[ \frac{1}{n} \sum_{i=1}^{n} \langle x_i, \beta - \beta^* \rangle^2 - \mathbb{E} \langle x_i, \beta - \beta^* \rangle^2 \right] ,
\]

\[
M_{r^*, \rho} = \sup_{\beta - \beta^* \in H_{r^*, \rho}} \left[ \frac{2}{n} \sum_{i=1}^{n} \xi_i \langle x_i, \beta - \beta^* \rangle \right] .
\]

By (19), we have thus

\[
\inf_{\beta \in \beta^* + H_{r^*, \rho}} P_n L_{\beta} \geq \alpha^{-2} \left[ (r^*)^2 - Q_{r^*, \alpha \rho} \right] - 2 M_{r^*, \alpha \rho} \alpha^{-1}
\]

\[
\geq \alpha^{-2} \left[ (r^*)^2 - Q_{r^*, \rho} \right] - 2 M_{r^*, \rho} \alpha^{-1} . \tag{20}
\]

It remains to bound the quadratic process \( Q_{r^*, \rho} \) and the multiplier process \( M_{r^*, \rho} \). This control is based on the Gaussian width of the sets \( H_{r^*, \rho} \). Recall that the Gaussian width of a subset \( H \subset \mathbb{R}^p \) is defined by

\[
w^*(H) = \mathbb{E} \left[ \sup_{h \in H} \langle G, h \rangle \right] , \quad \text{where} \quad G \sim \mathcal{N}(0, I_p) .
\]

The useful controls are provided in the following lemma, whose proof is postponed to Section A.2.
Lemma 4. Let $r, \rho \geq 0$ and $\delta, \eta \in (0, 1)$. There exists an absolute constant $c$ such that, with probability larger than $1 - \delta$,

$$Q_{r, \rho} \leq c\left[C^2_{r, \rho} + rC_{r, \rho} + r^2(D_{\delta, n} \vee D^2_{\delta, n})\right],$$

where the complexity $C_{r, \rho} = w^*(\Sigma^{1/2}H_{r, \rho})/\sqrt{n}$ and $D_{\delta, n} = \sqrt{\log(1/\delta)/n}$. Moreover, there exists another absolute constant $c$ such that, with probability larger than $1 - \eta$,

$$M_{r, \rho} \leq c\sigma\left[C_{r, \rho} + rD_{\eta, n}\right].$$

We apply Lemma 4 with $\eta = \delta = e^{-\zeta^2n}$, $r = r^*$ and $\rho$. We have $D^2_{\delta, n} \leq D_{\delta, n} = \zeta < 1$. It shows that $P(\Omega^*) \geq 1 - 2e^{-\zeta^2n}$, where

$$\Omega^* = \{Q_{r^*, \rho} \leq c\left[C^2_{r^*, \rho} + r^*C_{r^*, \rho} + \zeta(r^*)^2\right] \cap \{M_{r^*, \rho} \leq c\sigma[C_{r^*, \rho} + r^*\zeta]\}.$$

Moreover, from Equation (20) and the fact that $\alpha = r^*/r$, on $\Omega^*$

$$\inf_{\beta \in \beta^* + H_{r, \rho}} P_n\mathcal{L}_\beta \geq r^2(1 - c\zeta) - c\left(\frac{C_{r^*, \rho}}{\alpha^2} + r\frac{C_{r^*, \rho}}{\alpha} - 2c\sigma\frac{C_{r^*, \rho}}{\alpha} + r\zeta\right). \quad (21)$$

It remains to bound the Gaussian width $w^*(\Sigma^{1/2}H_{r, \rho})$ to bound the complexity $C_{r^*, \rho}$. This control is provided in the following lemma, whose proof is provided in Section A.3.

Lemma 5. Let $r, \rho \geq 0$. Then,

$$w^*(\Sigma^{1/2}H_{r, \rho}) = \sqrt{2W_{r, \rho}}$$

where

$$W_{r, \rho} = \sum_{i=1}^{p} r^2 \wedge \lambda_i(\Sigma)^2 .$$

From Lemma 5,

$$w^*(\Sigma^{-1/2}H_{r^*, \rho}) \leq c\sqrt{W_{r^*, \rho}} .$$

The choice of $r^*$ ensures that

$$W_{r^*, \rho} \leq n(\zeta r^*)^2 \quad \text{so} \quad \frac{C_{r^*, \rho}}{\alpha} \leq \zeta r^* .$$

Plugging this inequality into (21) shows that, on $\Omega^*$,

$$\inf_{\beta \in \beta^* + H_{r, \rho}} P_n\mathcal{L}_\beta \geq r^2(1 - 3c\zeta) - 4c\zeta r^* .$$

The inequality $ab \leq (a^2 + b^2)/2$ with $a = 4cr\zeta$ and $b = \sigma$ shows that, on $\Omega^*$,

$$\inf_{\beta \in \beta^* + H_{r, \rho}} P_n\mathcal{L}_\beta \geq r^2(1 - 3c\zeta - 8c^2\zeta^2) - \frac{\sigma^2}{2} .$$

Choosing $\zeta$ sufficiently small concludes the proof. □
4.3. Proof of Theorem 2

The proof is based on the following lemma, whose proof can be found in [2] and [13] for example.

**Lemma 6.** Assume that \( \ell(u, y) = \rho(u - y) \), where \( \rho \) is \( L \)-Lipschitz. There exists an absolute constant \( c \) such that, for any positive \( r, \rho \), with probability larger than \( 1 - \eta \),

\[
\sup_{\beta \in H_{r, \rho}} |(P_n - P)(\ell_\beta - \ell_{\beta^*})| \leq \frac{cL}{\sqrt{n}} (w^*(H_{r, \rho}) + \sqrt{\log(1/\eta)r}) .
\]

From Theorem 1, with probability larger than \( 1 - 7e^{-\zeta n} \), \( \|\hat{\beta} - \beta^*\| \leq \rho \) and \( \|\Sigma^{1/2}(\hat{\beta} - \beta^*)\|_2 \leq r^* \). Consequently, from Lemmas 5 and 6, with probability larger than \( 1 - \eta - 7e^{-\zeta n} \),

\[
P_{\beta} \leq P_{n\ell_\beta} + \frac{cL}{\sqrt{n}} \left( \sum_{i=1}^{p} (r^*)^2 \wedge \lambda_i(\Sigma) \rho^2 \right)^{1/2} + cL r^* \sqrt{\frac{\log(1/\eta)}{n}}.
\]

By definition of \( r^* \) this implies that

\[
P_{L_\beta} \leq P_{nL_\beta} + cL r^* \left( \zeta + \sqrt{\frac{\log(1/\eta)}{n}} \right).
\]

For the absolute loss function \( P_nL_{\hat{\beta}} \leq 0 \) and \( L = 1 \), so the proof is complete by taking \( \eta = e^{-n} \).

For the Huber loss function, \( L = c_2 \sigma \) and \( P_nL_{\hat{\beta}} = -P_n\ell_{\hat{\beta}}^* = -(1/n) \sum_{i=1}^{p} \rho H(\xi_i) \).

Moreover,

\[
P_n\ell_{\hat{\beta}}^* \geq \frac{1}{n} \sum_{i=1}^{p} \xi_i^2 \mathbf{1}\{|\xi_i| \leq c_2 \sigma\} = \frac{1}{n} \sum_{i=1}^{p} \xi_i^2 - \frac{1}{n} \sum_{i=1}^{p} \xi_i^2 \mathbf{1}\{|\xi_i| \geq c_2 \sigma\}.
\]

By (15), with probability larger than \( 1 - \exp(-n/16) \), \( (1/n) \sum_{i=1}^{p} \xi_i^2 \geq \sigma^2/2 \).

Similar arguments show that there exists an absolute constant \( \zeta \) such that, with probability \( 1 - \exp(-\zeta^2 n) \),

\[
\frac{1}{n} \sum_{i=1}^{p} \xi_i^2 \mathbf{1}\{|\xi_i| \geq c_2 \sigma\} \leq \sigma^2/6 + E[\xi_i^2 \mathbf{1}\{|\xi| \geq c_2 \sigma\}] .
\]

Moreover, from Cauchy-Schwarz and Markov inequalities,

\[
E[\xi_i^2 \mathbf{1}\{|\xi| \geq c_2 \sigma\}] \leq \sqrt{3} \sigma^2 \sqrt{P(|\xi| \geq c_2 \sigma)} \leq \frac{\sqrt{3} \sigma^2}{c_2} .
\]

It follows that, with probability \( 1 - 2e^{-\zeta n} \), if \( c_2 = 6\sqrt{3} \),

\[
P_nL_{\hat{\beta}} = -P_n\ell_{\hat{\beta}}^* \leq -\sigma^2 \left( \frac{1}{2} - \frac{1}{6} - \frac{\sqrt{3}}{c_2} \right) = -\frac{\sigma^2}{6} .
\]
Plugging this estimate into (22) yields, with probability $1 - 10e^{-(\zeta \wedge r) n}$,

$$PL_{\hat{\beta}} \leq -\frac{\sigma^2}{12} + 2\zeta r^*.$$  

The proof is complete since

$$2\zeta r^* = 2(\sqrt{12}\zeta r^*) \frac{\sigma}{\sqrt{12}} \leq 12\zeta^2 (r^*)^2 + \frac{\sigma^2}{12} .$$

Appendix A: Supplementary material

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

The following definition and propositions can be found in [32].

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

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

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

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

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

A.2. Proof Lemma 4

The proof of the control of the quadratic process follows from [16, Theorem 5.5] and the majorizing measure theorem (see [28, Theorem 2.4.1]).

Let $(X_t)_{t \in T}$ be a stochastic process indexed by a set $T$ of $\bar{n}$-tuples $t = (t_1, \ldots, t_n)$. Let us assume that the random variables $X_{t_i}: \Omega_t \mapsto \mathbb{R}$ are sub-Gaussian.

For every $t \in T$, let

$$A_t = \frac{1}{n} \sum_{i=1}^n (X_{t_i}^2 - \mathbb{E}X_{t_i}^2) . \quad (25)$$

Define on $T$ the pseudo-distance $d_{\psi_2}$, by

$$d_{\psi_2}(t, s) = \max_{i=1,\ldots,n} \|X_{t_i} - X_{s_i}\|_{\psi_2} , \quad (26)$$

where, for any real random variable $X$, $\|X\|_{\psi_2} = \inf\{C > 0 : \mathbb{E}\exp(|X|^2 / C^2) \leq 2\}$. The radius associated to $T$ is defined as

$$\Delta_{\psi_2}(T) = \sup_{t \in T} \max_{i=1,\ldots,n} \|X_{t_i}\|_{\psi_2} . \quad (27)$$
Theorem 6 (Theorem 5.5 in [16]). Let \((A_t)_{t \in T}\) be the process of averages defined in (25). There exists an absolute constant \(c > 0\) such that, for any \(\delta\) in \((0, 1)\), with probability larger than 1 – \(\delta\),

\[
\sup_{t \in T} A_t \leq c \left[ \frac{\gamma_2^2(T, d_{\psi_2})}{n} + \Delta_{\psi_2}(T) \frac{\gamma_2(T, d_{\psi_2})}{\sqrt{n}} + K \frac{\log(1/\delta)}{n} + M \sqrt{\frac{\log(1/\delta)}{n}} \right],
\]

(28)

where the definition of \(\gamma_2\) can be found in [28, Definition 2.2.19],

\[ K = \sup_{t \in T} \max_{i=1,\ldots,n} \|X_{t_i}\|_{\psi_2}^2 \quad \text{and} \quad M = \sup_{t \in T} \left( \frac{1}{n} \sum_{i=1}^{n} \|X_{t_i}\|_{\psi_2}^4 \right)^{1/2}. \]

To apply Theorem 6 to bound \(Q_{r, \rho}\), let \(T = \{(x_1, \beta), \ldots, (x_n, \beta), \beta \in H_{r, \rho}\}\) and, for any \(t = (\langle x_1, \beta \rangle, \ldots, \langle x_n, \beta \rangle) \in T\), let

\[ X_{t_i} = \langle x_i, \beta \rangle, \quad \text{so} \quad Q_{r, \rho} = \sup_{t \in T} A_t. \]

For any \(i = 1, \ldots, n\), \(X_{t_i} = \langle x_i, \beta \rangle \sim \mathcal{N}(0, \|\Sigma_2^{1/2} \beta\|^2_2) = \mathcal{N}(0, r^2)\). Therefore, \(\|X_{t_i}\|_{\psi_2} = r\), for any \(t \in T\) and any \(i = \{1, \ldots, n\}\), so \(\Delta_{\psi_2}(T) = K = M = r^2\). Moreover, in our case \(d_{\psi_2} = \|\cdot\|_\Sigma\) and from the majorizing measure theorem, see [28, Theorem 2.4.1], there exists an absolute constant \(c > 0\) such that \(\gamma_2(T, d_{\psi_2}) \leq w^*(\Sigma_2^{1/2} H_{r, \rho})\), so, by Theorem 6, with probability 1 – \(\delta\)

\[ Q_{r, \rho} \leq c [C_{r, \rho} + r C_{r, \rho} + r^2 (D_{\delta, \rho} \vee D_{\delta, \eta}^2)]. \]

Let us turn to the control of the multiplier process \(M_{r, \rho}\). Since the noise \(\xi\) is Gaussian with variance \(\sigma^2\), independent of \(x\), by [26, Corollary 1.10], there exists an absolute constant \(c\) such that, for any \(\delta\) in \((0, 1)\), with probability larger than 1 – \(\delta\),

\[ n M_{r, \rho} \leq c \sqrt{n} \sigma w^*(\Sigma_2^{1/2} H_{r, \rho}) + r \sqrt{\log(1/\delta)}. \]

A.3. Proof of Lemma 5

\[ w^*(\Sigma_2^{1/2} H_{r, \rho}) = \mathbb{E} \sup_{t \in \Sigma_2^{1/2} H_{r, \rho}} \langle G, t \rangle, \]

where \(G \sim \mathcal{N}(0, I_\rho)\), and

\[ \Sigma_2^{1/2} H_{r, \rho} = \{\Sigma_2^{1/2} t \in \mathbb{R}^p : \|t\| \leq \rho, \|\Sigma_2^{1/2} t\|_2 = r\} = \{t \in \mathbb{R}^p : \|\Sigma^{-1/2} t\|_2 \leq \rho, \|t\|_2 = r\} = \left\{ t \in \mathbb{R}^p : \sum_{i=1}^{n} \frac{t_i^2}{\lambda_i(\Sigma) \rho^2} \leq 1, \sum_{i=1}^{n} \frac{t_i^2}{\rho^2} \leq 1 \right\} \subset \left\{ t \in \mathbb{R}^p : \sum_{i=1}^{n} \frac{t_i^2}{\lambda_i(\Sigma) \rho^2 \wedge \rho^2} \leq 2 \right\}. \]
The Gaussian mean-width of an ellipsoid is given by [28, Proposition 2.5.1] and it follows that
\[
w^*(\Sigma^{1/2}H_{r,\rho}) \leq \sqrt{2}(\sum_{i=1}^{p} \lambda_i(\Sigma)\rho^2 \wedge r^2)^{1/2}.
\]

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] Pierre Alquier, Vincent Cottet, Guillaume Lecué, et al., *Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions*, The Annals of Statistics 47 (2019), no. 4, 2117–2144.
[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] Florentina Bunea, Seth Strimas-Mackey, and Marten Wegkamp, *Interpolation under latent factor regression models*, arXiv preprint arXiv:2002.02525 (2020).
[12] George Casella, *Minimax ridge regression estimation*, The Annals of Statistics (1980), 1036–1056.
[13] Geoffrey Chinot, ERM and RERM are optimal estimators for regression problems when malicious outliers corrupt the labels, arXiv preprint arXiv:1910.10923 (2019).
[14] Geoffrey Chinot, Guillaume Lecué, and Matthieu Lerasle, Robust statistical learning with lipschitz and convex loss functions, Probability Theory and Related Fields (2019).
[15] Christine De Mol, Ernesto De Vito, and Lorenzo Rosasco, Elastic-net regularization in learning theory, Journal of Complexity 25 (2009), no. 2, 201–230.
[16] Sjoerd Dirksen et al., Tail bounds via generic chaining, Electronic Journal of Probability 20 (2015).
[17] Vitaly Feldman, Does learning require memorization? a short tale about a long tail, arXiv preprint arXiv:1906.05271 (2019).
[18] Arthur E Hoerl and Robert W Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1970), no. 1, 55–67.
[19] 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.
[20] Vladimir Kolchinskii 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
[21] Guillaume Lecué and Shahar Mendelson, Learning subgaussian classes: Upper and minimax bounds, arXiv preprint arXiv:1305.4825 (2013).
[22] 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.
[23] Tengyuan Liang and Alexander Rakhlin, Just interpolate: Kernel" ridgeless" regression can generalize, arXiv preprint arXiv:1808.00387 (2018).
[24] Song Mei and Andrea Montanari, The generalization error of random features regression: Precise asymptotics and double descent curve, arXiv preprint arXiv:1908.05355 (2019).
[25] Shahar Mendelson, Learning without concentration, Conference on Learning Theory, 2014, pp. 25–39.
[26] ______, Upper bounds on product and multiplier empirical processes, Stochastic Processes and their Applications 126 (2016), no. 12, 3652–3680.
[27] ______, On multiplier processes under weak moment assumptions, Geometric Aspects of Functional Analysis, Springer, 2017, pp. 301–318.
[28] Michel Talagrand, Upper and lower bounds for stochastic processes: modern methods and classical problems, vol. 60, Springer Science & Business Media, 2014.
[29] Robert Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1996), no. 1, 267–288.
[30] 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.
[31] Roman Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, arXiv preprint arXiv:1011.3027 (2010).

[32] Martin J Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48, Cambridge University Press, 2019.

[33] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, *Understanding deep learning requires rethinking generalization*, arXiv preprint arXiv:1611.03530 (2016).

[34] 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.