Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results

Noureddine El Karoui
Department of Statistics, UC Berkeley*

November 11th, 2013

Abstract

We study the behavior of high-dimensional robust regression estimators in the asymptotic regime where $p/n$ tends to a finite non-zero limit. More specifically, we study ridge-regularized estimators, i.e

$$
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i - \mathbf{X}_i'\beta) + \frac{\tau}{2} \|\beta\|^2.
$$

In a recently published paper, we had developed with collaborators probabilistic heuristics to understand the asymptotic behavior of $\hat{\beta}$. We give here a rigorous proof, properly justifying all the arguments we had given in that paper. Our proof is based on the probabilistic heuristics we had developed, and hence ideas from random matrix theory, measure concentration and convex analysis.

While most the work is done for $\tau > 0$, we show that under some extra assumptions on $\rho$, it is possible to recover the case $\tau = 0$ as a limiting case.

We require that the $\mathbf{X}_i$’s be i.i.d with independent entries, but our proof handles the case where these entries are not Gaussian.

A 2-week old paper of Donoho and Montanari studied a similar problem by a different method and with a different point of view. At this point, their interesting approach requires Gaussianity of the design matrix.

1 Introduction

Robust regression estimators, also known as regression $M$-estimates, have been of interest in Statistics for at least the last five decades. They are natural extensions of the least-squares problem: namely we estimate a regression vector by solving the optimization problem

$$
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i - \mathbf{X}_i'\beta).
$$

(1)

Here, $\mathbf{X}_i \in \mathbb{R}^p$ is a vector of predictors and $Y_i \in \mathbb{R}$ is a scalar response. $\rho$ is a function from $\mathbb{R}$ to $\mathbb{R}$. Typically once assumes that there is a linear relationship between $\mathbf{X}_i$ and $Y_i$, i.e

$$
Y_i = \mathbf{X}_i'\beta_0 + \epsilon_i,
$$

where $\epsilon_i$ are considered to be unknown random errors, and $\beta_0$ is an unknown fixed vector one wishes to estimate. The $n \times p$ matrix $\mathbf{X}$ whose $i$-th row is $\mathbf{X}_i$ is called the design matrix.

Huber’s papers from the 1970’s (Huber (1972), Huber (1973)) contain a number of very interesting results, including limiting behavior for $\hat{\beta}$ as $n \to \infty$ when $p$ is held fixed. Huber also raised the question

* I would like to thank Peter Bickel, Bin Yu and Derek Bean for many interesting discussions on high-dimensional robust regression estimators. I am especially grateful to Peter Bickel for many fascinating discussions that greatly improved my understanding of this topic. Support from NSF grant DMS-0847647 (CAREER) is gratefully acknowledged. **AMS 2010 MSC:** Primary: 62E20. Secondary: 60F99 **Key words and Phrases:** high-dimensional inference, random matrix theory, concentration of measure, proximal mapping, regression M-estimates, robust regression. **Contact:** nkaroui@berkeley.edu
of understanding the behavior of the estimators when \( p \) is large and obtained partial results in the least-squares case. Further interesting contributions happened in the mid to late 80’s with work of Portnoy (Portnoy (1984), Portnoy (1985), Portnoy (1987)) and Mammen (Mammen (1989)). In these studies, the authors studied the behavior of regression M-estimates when \( p \) and \( n \) are both large, but \( p/n \to 0 \) at various rates. Some of the papers refer to fixed design (i.e \( X \) is non-random and the only source of randomness in the problem are \( \epsilon_i \)’s), others treat the random design case (i.e both \( X \) and \( \epsilon_i \) are random).

A central result of Huber (see e.g Huber and Ronchetti (2009)) is that when \( p \) is held fixed, and \( \epsilon_i \)’s are i.i.d, the optimal \( \rho \) one can use is \(-\log f_\epsilon\), where \( f_\epsilon \) is the density of the errors - at least when one measures quality of the estimator by the size of \( \text{cov}(\hat{\beta}) \). In El Karoui et al. (2012), El Karoui et al. (2013) a group of us looked at corresponding questions in the high-dimensional setting where \( p/n \) is not small and found the situation to be very different. Indeed, it was clear that one could do better than using \(-\log f_\epsilon\). In El Karoui et al. (2012), we proposed a probabilistic heuristic to understand the behavior of \( \hat{\beta} \) and verified the quality of its predictions on several simulations and computations. Our heuristic led to the formulation of a natural variational problem, which we solved in Bean et al. (2013). Interestingly, the solution of the variational problem depends in general on \( p/n \), i.e the dimensionality of the problem. (El Karoui et al. (2012) is the long form of the paper El Karoui et al. (2013), which is very short due to page-limit requirements.)

Our heuristic is based on random matrix and concentration of measure ideas. We prove in this paper that these ideas can be used rigorously and indeed, under various assumptions, rigorously justify the claims made in El Karoui et al. (2012) and El Karoui et al. (2013).

The assumptions under which we operate for the design matrix reflect the central role played by the concentration of measure phenomenon (Ledoux (2001)) in this problem.

A couple of weeks ago, Donoho and Montanari (Donoho and Montanari (2013)) announced a proof of some of the results explained in El Karoui et al. (2013) under the assumption that the design matrix is full of i.i.d Gaussian random variables (i.e \( X_i \)’s are independent with i.i.d Gaussian entries). Their proof uses different ideas than ours - it is based on the technology of rigorous analysis of approximate message passing algorithms (see Donoho et al. (2009) and Bayati and Montanari (2012)).

By working under concentration assumptions, we are able to show a number of the same results without requiring i.i.d-ness of the entries of the vectors \( X_i \)’s. However, to prove the main result, we still need the \( X_i \)’s to have i.i.d entries, but they do not need to be Gaussian. Donoho and Montanari also make interesting connections with rigorous work in statistical physics, namely to the so-called Shcherbina-Tirozzi model (Shcherbina and Tirozzi (2003) and Talagrand (2003)) and other heuristic approaches based on approximate message passing (Rangan (2011)).

Our proof also makes rigorous the probabilistic heuristics that were developed in El Karoui et al. (2013). Our point of view is that the properties of \( \hat{\beta} \) defined in Equation (1) via connections to random matrix theory. As such, our proof relies heavily on leave-one-out, martingale and concentration of measure ideas, as some of our previous work (see e.g El Karoui (2009)) did in establishing these connections. Leave-one-out ideas seem to be known in Physics under the name “cavity method”, so our general approach falls broadly in that category. A number of the tools we use are commonly used in the spectral analysis of large random matrices via the Stieltjes transform method (see Marčenko and Pastur (1967), Wachter (1978), Silverstein (1995)).

1.1 Focus of the paper

We focus on the problem of understanding

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i - X'_i \beta) + \frac{\tau}{2} \|\beta\|^2
\]  

(2)

where \( \tau > 0 \). We will see later (see Section 6) that under certain conditions on \( \rho \) the understanding of \( \hat{\beta} \) for various \( \tau \)’s will lead us to rigorous understanding of \( \hat{\beta} \) when \( \tau = 0 \).

Different parts of the proof require different assumptions. So we label the assumptions accordingly.

For the first part of the proof (i.e “leave-one-observation-out”), we work under the following assumptions:
\begin{itemize}
  \item **O1**: \(p/n\) has a finite non-zero limit.
  \item **O2**: \(\rho\) is twice differentiable, convex and non-linear. \(\psi = \rho'\). Note that \(\psi' \geq 0\) since \(\rho\) is convex. We assume that \(\rho \geq 0\) and \(\rho(0) = 0\). Note that this implies that \(\text{sign}(\psi(x)) = \text{sign}(x)\).
  \item **O3**: \(\psi(|x|) = O(|x|^m)\) at infinity for some \(m\). Furthermore, \(\psi'\) is \(L(u)\)-Lipschitz on \((-|u|, |u|)\), where \(L(|u|) \leq K|u|^m\) as \(|u| \to \infty\). Note that this implies that \(\rho\) grows at most polynomially at \(\infty\).
  \item **O4**: \(X_i\)'s are independent and identically distributed. Furthermore, for any 1-Lipschitz convex function \(F\), \(P(|F(X_i) - m_F| > t) \leq C_n \exp(-c_n t^2)\), \(C_n\) and \(c_n\) can vary with \(n\). For simplicity, we assume that \(c_n = O(1/(\log(n))^{\alpha/2})\) for some \(\alpha \geq 0\). \(X_i\)'s have mean 0 and \(\text{cov}(X_i) = \text{Id}_p\).
  \item **O5**: \(\{X_i\}_{i=1}^n\) are independent of \(\{\epsilon_i\}_{i=1}^n\).
  \item **O6**: for any fixed \(k\), \(\frac{1}{n} \sum_{i=1}^n \mathbb{E}(\psi^{2k}(\epsilon_i))\) remains uniformly bounded in \(p\) and \(n\), as both grow to infinity.
  \item **O7**: \(\text{sup}_{1 \leq i \leq n} |\epsilon_i| \triangleq \mathcal{E}_n = O((\log n)^\beta)\) and \(\epsilon_i\)'s are independent.
\end{itemize}

For the second part of the proof (i.e “leave-one-predictor-out”), we need all the previous assumptions and

- **P1**: \(X_i\)'s have i.i.d entries.

We note that according to Corollary 4.10 and the discussion that follows in Ledoux (2001), Assumptions O4 and P1 are compatible. O4 is for instance satisfied if the entries of \(X_i\)'s are bounded by \(O((\log n)^{\alpha/2})\). Another example is the case of \(X_1 \sim \mathcal{N}(0, \text{Id}_p)\).

For the last part of the proof, when we combine everything together, we will need the following assumptions on top of all the others:

- **F1**: the \(\epsilon_i\)'s have identical distribution and for any \(r > 0\), if \(Z \sim \mathcal{N}(0, 1)\), independent of \(\epsilon_i\), \(\epsilon_i + rZ\) has a density \(f\) which is increasing on \((-\infty, 0)\) and decreasing on \((0, \infty)\). Furthermore, \(\lim_{|t| \to \infty} t f(t) = 0\).

- **F2**: For any fixed \(k\), \(\mathbb{E}(|\epsilon_i|^k) < \infty\).

We refer the reader to Lemma C-1 and the discussion immediately following it for examples of such densities. We note that symmetric (around 0) log-concave densities will for instance satisfy all the assumptions we made about the \(\epsilon_i\)'s. See Karlin (1968) and Ibragimov (1956) for instance.

The aim of the paper is to prove the following theorem:

**Theorem 1.1.** Consider \(\hat{\beta}\) defined in Equation (2) and assume that \(\tau > 0\) is given. Under Assumptions O1-O7, P1 and F1-F2, we have: as \(p, n\) tend to infinity while \(p/n \to \kappa \in (0, \infty)\), \(\var{\|\hat{\beta}\|} \to 0\).

Furthermore, if \(\hat{\epsilon}_t = \epsilon + r_t(\kappa)Z\), where \(\epsilon\) has the same distribution as \(\epsilon_i\)'s and \(Z\) is a \(\mathcal{N}(0, 1)\) random variable independent of \(\epsilon\), we have: \(\|\hat{\beta}\| \to r_t(\kappa)\) and there exists a constant \(c_t(\kappa)\) such that

\[
\begin{cases}
  \mathbb{E}\left((\text{prox}_{c_t(\kappa)}(\rho))(\hat{\epsilon}_t))^2\right) = 1 - \kappa + \tau c_t(\kappa) \\
  \kappa r_t^2(\kappa) = \mathbb{E}\left((\hat{\epsilon}_t - \text{prox}_{c_t(\kappa)}(\rho))(\hat{\epsilon}_t))^2\right) 
\end{cases}
\]

(3)

We use the notation \(\text{prox}_{c}(\rho)\) to denote the proximal mapping of the function \(c \rho\). This notion was introduced in Moreau (1965). We recall that

\[
\text{prox}_{c}(\rho)(x) = \arg\min_{y \in \mathbb{R}}(c \rho(y) + \frac{1}{2}(x - y)^2), \text{ or equivalently,}

\text{prox}_{\psi}(\rho)(x) = (\text{Id} + c \psi)^{-1}(x). 
\]

The proximal mapping is an important notion in convex analysis and convex optimization (see for instance Beck and Teboullle (2010) for a nice review of analytic properties and an introduction to proximal gradient algorithms). We note that even when \(\rho\) is not differentiable, \(\text{prox}_{c}(\rho)(x)\) is a well-defined function.

As explained in Bean et al. (2013), the previous system can be reformulated in terms of \(\text{prox}_{f}((c(\kappa)\rho)^*)\), where \(f^*\) represents the Fenchel-Legendre dual of \(f\).
Remarks on the assumptions

In the context of robust statistics, where regression M-estimates are commonly used, \( \rho \) is often taken to grow linearly or square-roots of quadratic forms. However, as documented in Ledoux (2001), a large number of natural or "reasonable" distributions satisfy the O4 assumptions. Our choice of having a potentially varying \( c_n \) is motivated by the idea that we could, for instance, relax an assumption of boundedness of the entries of \( X_i \)'s - that guarantees that O4 is satisfied when \( X_i \) has i.i.d entries - and replace it by an assumption concerning the moments of \( X_i \)'s: this is what we did for instance in El Karoui (2009) through a truncation of triangular arrays argument. We also refer the interested reader to that paper for a short list of distributions satisfying O4. Finally, we could replace the \( \exp(-c_n t^2) \) upper bound in O4 by \( \exp(-c_n t^\alpha) \) for some fixed \( \alpha > 0 \) and it seems that all our arguments would go through. We chose not to do work under these more general assumptions because it would involve extra book-keeping and does not enlarge the set of distributions we can consider enough to justify this extra technical cost.

Our assumption that \( 1/c_n \) increases like a power of \( \log(n) \) at most is quite restrictive when it comes to bounded random variables - but is of course satisfied by e.g Gaussian random variables where \( c_n \) is a constant independent on \( n \) - and motivated by simplifying the book-keeping needed in our proof. Having \( 1/c_n \) grow like \( n^\gamma \) for a small \( \gamma \) should be feasible - with \( \gamma \) depending on \( m \) and \( m_1 \). In the first part of the proof we keep track of the impact of \( c_n \) to show this aspect of the problem.

Statistically, regression M-estimates are quite widely used. But in the random design case studied here, they are known to have somewhat undesirable properties (Baranchik (1973), Stein (1960)) even in very simple situations. We do not dwell more on these otherwise interesting issues, since they are a bit tangential to the main aim of this particular paper, which is to give a rigorous justification of the heuristic manipulations made in El Karoui et al. (2013).

Notations

We will repeatedly use the following notations: polyLog\((n)\) is used to replace a power of \( \log(n) \); \( \lambda_{\text{max}}(M) \) denotes the largest eigenvalue of the matrix \( M \); \( |||M|||_2 \) denotes the largest singular value of \( M \). We call \( \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n X_i X_i' \) the usual sample covariance matrix of the \( X_i \)'s. We say that \( X \preceq Y \in L_k \) if \( \text{E} (|X|^k) \leq \text{E} (|Y|^k) \). We use the notation \( u_n \preceq v_n \) to say that there exists a constant \( K \) independent of \( n \) such that \( u_n \leq K v_n \) for all \( n \). We use the usual statistical notation \( \beta_{(i)} \) to denote the regression vector we obtain when we do not use the pair \( (X_i, Y_i) \) in our optimization problem. We will also use the notation \( X_{(i)} \) to denote \( \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\} \). We use the notation \( (a, b) \) for either the interval \( (a, b) \) or the interval \( [b, a] \): in several situations, we will have to localize quantities in intervals using two values \( a \) and \( b \) but we will not know whether \( a < b \) or \( b > a \). We denote by \( X \) the \( n \times p \) design matrix from \( i \)-th row is \( X_i' \).

Remarks

Note that under our assumptions on \( \rho \), \( \hat{\beta} \) is defined as the solution of

\[
f(\hat{\beta}) = 0 \quad \text{with} \quad f(\beta) = \frac{1}{n} \sum_{i=1}^n -X_i \psi(\epsilon_i - X_i' \beta) + \tau \beta.
\]
We call
\[
F(\beta) = \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i - X'_i \beta) + \frac{\tau}{2} \| \beta \|^2 .
\] (6)

We call \( R_i = \epsilon_i - X'_i \beta \) (i.e the residuals), \( S = \frac{1}{n} \sum_{i=1}^{n} \psi'(R_i) X_i X'_i \) and \( c_\tau = \frac{1}{n} \text{trace} (S + \tau \text{Id})^{-1}. \)

2 Preliminaries

2.1 General remarks

**Proposition 2.1.** Let \( \beta_1 \) and \( \beta_2 \) be two vectors in \( \mathbb{R}^p \). Then
\[
\| \beta_1 - \beta_2 \| \leq \frac{1}{\tau} \| f(\beta_1) - f(\beta_2) \|. \tag{7}
\]

When \( \rho \) is strongly convex with modulus of convexity \( C \), we also have
\[
\| \beta_1 - \beta_2 \| \leq \frac{1}{C \lambda_{\text{min}}(\Sigma) + \tau} \| f(\beta_1) - f(\beta_2) \| .
\]

**Proof.** Let \( \beta_1 \) and \( \beta_2 \) be two vectors in \( \mathbb{R}^p \). We have
\[
f(\beta_1) - f(\beta_2) = \tau(\beta_1 - \beta_2) + \frac{1}{n} \sum_{i=1}^{n} X_i \left[ \psi(\epsilon_i - X'_i \beta_2) - \psi(\epsilon_i - X'_i \beta_1) \right].
\]

We can use the mean value theorem to write
\[
\psi(\epsilon_i - X'_i \beta_2) - \psi(\epsilon_i - X'_i \beta_1) = \psi'(\gamma_{\epsilon_i, X'_i \beta_1, X'_i \beta_2}) X'_i (\beta_1 - \beta_2),
\]

where \( \gamma_{\epsilon_i, X'_i \beta_1, X'_i \beta_2} \) is in the interval \( (\epsilon_i - X'_i \beta_1, \epsilon_i - X'_i \beta_2) \) - we do not care about the order of the endpoint in our notation.

We therefore have
\[
f(\beta_1) - f(\beta_2) = \tau(\beta_1 - \beta_2) + \frac{1}{n} \sum_{i=1}^{n} \psi'(\gamma_{\epsilon_i, X'_i \beta_1, X'_i \beta_2}) X_i X'_i (\beta_1 - \beta_2),
\]
which we write
\[
f(\beta_1) - f(\beta_2) = (S_{\beta_1, \beta_2} + \tau \text{Id}_p)(\beta_1 - \beta_2), \tag{8}
\]

where
\[
S_{\beta_1, \beta_2} = \frac{1}{n} \sum_{i=1}^{n} \psi'(\gamma_{\epsilon_i, X'_i \beta_1, X'_i \beta_2}) X_i X'_i .
\]

We therefore have
\[
\beta_1 - \beta_2 = (S_{\beta_1, \beta_2} + \tau \text{Id}_p)^{-1} (f(\beta_1) - f(\beta_2)).
\]

Since \( \rho \) is convex, \( \psi' = \rho'' \) is non-negative and \( S_{\beta_1, \beta_2} \) is positive semi-definite. In the semi-definite order, we have \( S_{\beta_1, \beta_2} + \tau \text{Id} \succeq \tau \text{Id} \). When \( \rho \) is strongly convex with modulus \( C \), we have \( \psi'(x) \geq C \) (see Theorem 4.3.1 in Hiriart-Urruty and Lemaréchal (2001)) and therefore, \( S_{\beta_1, \beta_2} + \tau \text{Id} \succeq C \Sigma + \tau \text{Id} \succeq (C \lambda_{\text{min}}(\Sigma) + \tau) \text{Id}_p \). In particular,
\[
\| \beta_1 - \beta_2 \| \leq \frac{1}{\tau} \| f(\beta_1) - f(\beta_2) \| .
\]

In the strongly convex case, we have
\[
\| \beta_1 - \beta_2 \| \leq \frac{1}{C \lambda_{\text{min}}(\Sigma) + \tau} \| f(\beta_1) - f(\beta_2) \| .
\]

\( \square \)
In the proof of Proposition 2.1, it is clear that all we need is that “enough” $\psi' (\gamma^*_i, X_i^* X_i')$’s are greater than a constant $C$. More precisely, let us call $N = \text{Card} \left\{ i : \psi' (\gamma^*_i, X_i^* X_i') \geq C \right\}$ and let us call $I$ the corresponding set of indices. Results similar to that of Proposition 2.1 then hold, with $\hat{\Sigma}$ being replaced by $\hat{\Sigma}_I = \frac{1}{n} \sum_{i \in I} X_i X_i'$. This could perhaps be used in certain situations to move away from strong convexity assumptions when we deal with the un-penalized (i.e $\tau = 0$) case. See Section 6 for more details about this question. Strong convexity is a very strong (and somewhat undesirable) requirement on $\rho$ for many applications in Statistics.

Proposition 2.1 yields the following lemma.

**Lemma 2.1.** For any $\beta_1$, 
\[ \| \hat{\beta} - \beta_1 \| \leq \frac{1}{\tau} \| f(\beta_1) \| . \]

The lemma is a simple consequence of Equation (7) since by definition $f(\hat{\beta}) = 0$.

In the following, we will strive to find approximations of $\hat{\beta}$. We will therefore use Lemma 2.1 repeatedly.

### 2.2 Boundedness of $\| \hat{\beta} \|$ 

We have the following lemma.

**Lemma 2.2.** Let us call $W_n = \frac{1}{n} \sum_{i=1}^{n} X_i \psi(\epsilon_i)$. We have 
\[ \| \hat{\beta} \| \leq \frac{1}{\tau} \| W_n \| . \]

In particular, when $X_i$ are independent and have covariance $\text{Id}_p$, 
\[ E \left( \| \hat{\beta} \|^2 \right) \leq \frac{1}{\tau^2} \frac{1}{n} \sum_{i=1}^{n} E \left( \psi^2(\epsilon_i) \right) . \]  

A similar result holds in $L_{2k}$ - provided the entries of $X_i$ has cumulants of order $2k$. This is automatically satisfied under our assumptions.

This guarantees that $\| \hat{\beta} \|$ is bounded in $L_{2k}$ provided $\frac{1}{n} \sum_{i=1}^{n} E \left( |\psi(\epsilon_i)|^{2k} \right)$ is bounded. If this latter quantity is polyLog($n$) so is $E \left( \| \hat{\beta} \|^{2k} \right)$.

We also have
\[ \| \hat{\beta} \| \leq \sqrt{\frac{2}{\tau}} \left( \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i) \right) , \]  

and hence
\[ E \left( \| \hat{\beta} \|^{2k} \right) \leq \frac{2^k}{\tau^k} E \left( \left[ \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i) \right]^k \right) . \]

Though from a probabilistic point of view our various bounds might look interchangeable, it is important to have both from the point of view of statistical applications. Indeed, in robust regression, where $\epsilon_i$’s can have heavy tails, one would typically used bounded $\psi$ functions (for instance the Huber functions or smoothed version of the Huber functions - see Huber and Ronchetti (2009), p. 84, Equation (4.51) for a definition of the exponential of the Huber functions). The bound based on Equation (9) is then particularly helpful.

**Proof.** The first inequality follows easily from taking $\beta_1 = 0$ in Lemma 2.1 and realizing that $W_n = f(0)$. The second inequality follows from the fact that, if $e$ is an $n$-dimensional vector with entries all equal to 1, $W_n = X'D_\psi e/n$, where $X$ is $n \times p$ and $D_\psi$ is a diagonal matrix whose $(i,i)$ entry is $\psi(\epsilon_i)$. Hence,
\[ W_n^2 = \frac{1}{n^2} e'D_\psi XX' D_\psi e , \]
and therefore, \( E(W_n^2) = \frac{p}{n^2} \sum_{i=1}^n E(\psi^2(\epsilon_i)) \), since \( E(X'X) = p \text{Id}_n \) and \( \{\epsilon_i\}_{i=1}^n \) is independent of \( \{X_i\}_{i=1}^n \).

For the \( L_{2k} \) bound, can use \( E(\|W_n\|^{2k}) \leq p^{k-1} \sum_{j=1}^p E(W_{n,j}^{2k}) \), because for \( \alpha_i > 0 \), \( (\sum_{i=1}^n \alpha_i)^k \leq p^{k-1} \sum \alpha_i^k \) by convexity.

Let us work temporarily conditional on \( \epsilon_i \). We control \( E(W_n^{2k}(j)) \) through the use of cumulants since \( W_n(j) = \sum_{i=1}^n X_i(j) \psi(\epsilon_i)/n \), so the \( 2k \)-th cumulant of \( W_n(j) \) is \( \sum_{i=1}^n \psi^{2k}(\epsilon_i)/n^{2k} \kappa_{2k}(X_i(j)) \). These cumulants are all of order \( n^{1-2k} \), if \( \sum_\psi^{2k}(\epsilon_i)/n = O(1) \). By the classical connection between moments and cumulants, we see that \( E(W_n^{2k}(j)) = O(n^{-k}) \) if \( \frac{1}{n} \sum_{i=1}^n E(\psi^{2k}(\epsilon_i)) \). Hence, \( E(\|W_n\|^{2k}) = O(p^{k-1}n^{-k}) = O(1) \).

The proof of Equation (10) simply follows from observing that

\[
\frac{\tau}{2} \|\hat{\beta}\|^2 \leq \frac{1}{n} \sum_{i=1}^n \rho(\epsilon_i - X_i\hat{\beta}) + \frac{\tau}{2} \|\hat{\beta}\|^2 \leq \frac{1}{n} \sum_{i=1}^n \rho(\epsilon_i) .
\]

Indeed, since, according to Equation (6),

\[
\hat{\beta} = \text{argmin}_\beta F(\beta) ,
\]

we also have

\[
F(\hat{\beta}) \leq F(0) = \frac{1}{n} \sum_{i=1}^n \rho(\epsilon_i) ,
\]

and the result follows immediately.

\( \square \)

## 3 Approximating \( \hat{\beta} \) by \( \hat{\beta}_{(i)} \): leave-one-observation-out

We consider the situation where we leave one observation out. We call

\[
\hat{r}_{j,(i)} = \epsilon_j - X_j\hat{\beta}_{(i)} \quad \text{and} \quad S_i = \frac{1}{n} \sum_{j \neq i} \psi'(\hat{r}_{j,(i)}) X_j X_j' .
\]

We also call

\[
f_i(\beta) = -\frac{1}{n} \sum_{j \neq i} X_j \psi(\epsilon_j - X_j'\beta) + \tau \beta .
\]

We call \( \hat{\beta}_{(i)} \) the solution of \( f_i(\hat{\beta}_{(i)}) = 0 \) and call it the leave-one-out estimate.

Let us consider

\[
\tilde{\beta}_i = \hat{\beta}_{(i)} + \frac{1}{n} (S_i + \tau \text{Id})^{-1} X_i \psi(\text{prox}_{c_i}(\rho)(\hat{r}_{i,(i)})) \triangleq \hat{\beta}_{(i)} + \eta_i , \quad (11)
\]

where

\[
c_i = \frac{1}{n} X_i'(S_i + \tau \text{Id})^{-1} X_i , \quad (12)
\]

\[
\eta_i = \frac{1}{n} (S_i + \tau \text{Id})^{-1} X_i \psi(\text{prox}_{c_i}(\rho)(\hat{r}_{i,(i)})) . \quad (13)
\]

All these approximations are “very natural” in light of the probabilistic heuristics we derived for this problem in El Karoui et al. (2013) - so we refer the reader to that paper for explanations about why we choose to introduce these quantities. One of the aim of the paper is to show that these heuristics are valid and indeed open up the horizon to rigorous proofs.

The aim of the work that follows is to show that \( \tilde{\beta} \) can be very well approximated by \( \tilde{\beta}_i \). In Corollary 3.1, we show that the approximation is accurate to order \( \text{polyLog}(n)/n \) in Euclidian norm, if for instance \( 1/c_n = \text{polyLog}(n) \). We refer the reader to Corollary 3.1 for full details.
3.1 Deterministic bounds

Proposition 3.1. We have
\[ \| \hat{\beta} - \tilde{\beta}_i \| \leq \frac{1}{\tau} \| R_i \| , \]  
where
\[ R_i = \frac{1}{n} \sum_{j \neq i} \left[ \psi'(\gamma^*(X_j, \hat{\beta}_i(i), \eta_i)) - \psi'(\hat{r}_{j,(i)}) \right] X_jX'_j\eta_i , \]  
and \( \gamma^*(X_j, \hat{\beta}_i(i), \eta_i) \) is in the ("unordered") interval \( (\tilde{r}_{j,(i)}, \hat{r}_{j,(i)} - X'_j\eta_i) \).

Proof. We have of course,
\[ f(\hat{\beta}_i) = f(\tilde{\beta}_i) - f_i(\hat{\beta}_i(i)) = -\frac{1}{n}X_i\psi(\epsilon_i - X'_i\hat{\beta}_i) + \frac{1}{n} \sum_{j \neq i} X_j \left[ \psi(\epsilon_j - X'_j\hat{\beta}_i(i)) \right] + \psi(\epsilon_j - X'_j(\hat{\beta}_i(i) + \eta_i)) \right] + \tau \eta_i . \]

By the mean-value theorem, we also have
\[ \psi(\epsilon_j - X'_j\hat{\beta}_i(i)) - \psi(\epsilon_j - X'_j\hat{\beta}_i(i)) = \psi'(\hat{r}_{j,(i)}) \right] X'_j\eta_i , \]
where \( \gamma^*(X_j, \hat{\beta}_i(i), \eta_i) \) is in the ("unordered") interval \( (\tilde{r}_{j,(i)}, \hat{r}_{j,(i)} - X'_j\eta_i) \).

Hence, if \( R_i \) is the quantity defined in Equation (15),
\[ \frac{1}{n} \sum_{j \neq i} X_j \left[ \psi(\epsilon_j - X'_j\hat{\beta}_i(i)) - \psi(\epsilon_j - X'_j(\hat{\beta}_i(i) + \eta_i)) \right] = \frac{1}{n} \sum_{j \neq i} \psi'(\hat{r}_{j,(i)})X_jX'_j\eta_i + R_i , \]
\[ = S_i\eta_i + R_i . \]

In light of the previous simplifications, we have
\[ f(\hat{\beta}_i) = -\frac{1}{n}X_i\psi(\epsilon_i - X'_i\hat{\beta}_i) + (S_i + \tau \text{Id})\eta_i + R_i . \]

Since by definition, \( \eta_i = \frac{1}{n}(S_i + \tau \text{Id})^{-1}X_i\psi(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) \),
\[ (S_i + \tau \text{Id})\eta_i = \frac{1}{n}X_i\psi(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) . \]

In other respects,
\[ \epsilon_i - X'_i\hat{\beta}_i = \tilde{r}_{i,(i)} - \epsilon_i\psi(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) . \]

When \( \psi \) is differentiable, \( x - c_i\psi(\text{prox}_{c_i}(\rho)(x)) = \text{prox}_{c_i}(\rho)(x) \) almost by definition of the proximal mapping (see Lemma A-1 and its proof). Therefore,
\[ -\frac{1}{n}X_i\psi(\epsilon_i - X'_i\hat{\beta}_i) + (S_i + \tau \text{Id})\eta_i = \frac{1}{n}X_i \left[ -\psi(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) + \psi(\text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})) \right] = 0 . \]

We conclude that
\[ f(\hat{\beta}_i) = R_i . \]

Applying Lemma 2.1, we see that
\[ \| \hat{\beta} - \tilde{\beta}_i \| \leq \frac{1}{\tau} \| R_i \| . \]

\[ \square \]
3.1.1 On $\mathcal{R}_i$

Lemma 3.1. We have
\[
\|\eta_i\| \leq \frac{1}{\sqrt{nT}} \frac{\|\bar{X}_i\|}{\sqrt{n}} \left[ |\psi(\tilde{r}_{i,i})| \wedge \frac{|\bar{r}_{i,i}|}{c_i} \right],
\]
and
\[
\|\mathcal{R}_i\| \leq |||\hat{\Sigma}|||_2 \sup_{j \neq i} \left| \psi'(\gamma^*(X_j, \tilde{\beta}_{i}, \eta_i)) - \psi'(\tilde{r}_{j,i}) \right| \frac{1}{\sqrt{nT}} \frac{\|\bar{X}_i\|}{\sqrt{n}} \left[ |\psi(\tilde{r}_{i,i})| \wedge |\bar{r}_{i,i}|/c_i \right].
\]

Proof. We have
\[
\mathcal{R}_i = \frac{1}{n} \sum_{j \neq i} \left[ \psi'(\gamma^*(X_j, \tilde{\beta}_{i}, \eta_i)) - \psi'(\tilde{r}_{j,i}) \right] X_j X'_j \eta_i.
\]

Of course, $\mathcal{S} = \frac{1}{n} \sum_{j \neq i} \left[ \psi'(\gamma^*(X_j, \tilde{\beta}_{i}, \eta_i)) - \psi'(\tilde{r}_{j,i}) \right] X_j X'_j$ can be written $\mathcal{S} = \frac{1}{n} X' DX$, where $D$ is a diagonal matrix with $(j, j)$ entry $\left[ \psi'(\gamma^*(X_j, \tilde{\beta}_{i}, \eta_i)) - \psi'(\tilde{r}_{j,i}) \right]$ and $(i, i)$ entry $0$.

Using the fact that $||| \cdot |||_2$ is a matrix norm, we see that $|||\mathcal{S}|||_2 \leq |||\hat{\Sigma}|||_2 |||D|||_2$. This implies that
\[
|||\mathcal{R}_i||| \leq |||\hat{\Sigma}|||_2 \sup_{j \neq i} \left| \psi'(\gamma^*(X_j, \tilde{\beta}_{i}, \eta_i)) - \psi'(\tilde{r}_{j,i}) \right| \|\eta_i\|,
\]
where $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X'_i$ is the usual sample covariance matrix.

We note that
\[
\|\eta_i\| \leq \frac{1}{\sqrt{nT}} \frac{\|\bar{X}_i\|}{\sqrt{n}} |\psi(\prox_{c_i}(\rho)(\tilde{r}_{i,i}))|.
\]

Using Lemma A-1, we see that
\[
|\psi(\prox_{c_i}(\rho)(\tilde{r}_{i,i}))| \leq |\psi(\tilde{r}_{i,i})| \wedge \frac{|\bar{r}_{i,i}|}{c_i}.
\]

The lemma is shown. \qed

3.1.2 On $\gamma^*(X_j, \tilde{\beta}_{i}, \eta_i)$ and related quantities

We now show how to control $\frac{1}{\sqrt{n}} \sup_{j \neq i} \left| \psi'(\gamma^*(X_j, \tilde{\beta}_{i}, \eta_i)) - \psi'(\tilde{r}_{j,i}) \right|

Lemma 3.2. Let us call
\[
\mathcal{B}_n(i) = \sup_{j \neq i} \left[ |\epsilon_j - X'_j \tilde{\beta}_{i}| + |X'_j \eta_i| \right].
\]

Suppose, as in our assumptions, that $\psi'$ is $L(\mathcal{B}_n(i))$ Lipschitz on $(-\mathcal{B}_n(i), \mathcal{B}_n(i))$. Then,
\[
\sup_{j \neq i} \left| \psi'(\gamma^*(X_j, \tilde{\beta}_{i}, \eta_i)) - \psi'(\tilde{r}_{j,i}) \right| \leq L(\mathcal{B}_n(i)) \sup_{j \neq i} |X'_j \eta_i|.
\]

It follows that
\[
|||\mathcal{R}_i||| \leq \sup_{j \neq i} |X'_j \eta_i| \frac{L(\mathcal{B}_n(i))}{\sqrt{nT}} \frac{\|\bar{X}_i\|}{\sqrt{n}} |||\hat{\Sigma}|||_2 \left[ |\psi(\tilde{r}_{i,i})| \wedge |\bar{r}_{i,i}|/c_i \right].
\]

We note that we could replace the assumption concerning the Lipschitz property of $\psi'$ on $(-\mathcal{B}_n(i), \mathcal{B}_n(i))$ by saying that $\psi'$ has modulus of continuity $\omega_n$ when restricted to this interval and putting growth condition on this modulus. We chose not to do this to simplify the exposition.

Proof. By definition, we have
\[
|\gamma^*(X_j, \tilde{\beta}_{i}, \eta_i) - \tilde{r}_{j,i}| \leq |X'_j \eta_i|.
\]

Of course,
\[
X'_j \eta_i = \psi(\prox_{c_i}(\rho)(\tilde{r}_{i,i})) \frac{1}{n} X'_j (S_i + \tau I_d)^{-1} X_i.
\]
Therefore,

$$\sup_{j \neq i} |\gamma' (X_j, \tilde{\beta}(i), \eta_i)| \leq \sup_{j \neq i} \left[ |\epsilon_j - X_j' \tilde{\beta}(i) + |X_j' \eta_i| \right]$$

We call

$$\mathcal{B}_n(i) = \sup_{j \neq i} \left[ |\epsilon_j - X_j' \tilde{\beta}(i) + |X_j' \eta_i| \right].$$

Since, by assumption, \( \psi' \) is \( L(B_n(i)) \)-Lipschitz on \((-\mathcal{B}_n(i), \mathcal{B}_n(i))\). Then

$$\sup_{j \neq i} \left| \psi' (\gamma (X_j, \tilde{\beta}(i), \eta_i)) - \psi' (\tilde{\gamma}_j(i)) \right| \leq L(B_n(i)) \sup_{j \neq i} |X_j' \eta_i|.$$  

The bound for \( ||\mathcal{R}_i|| \) follows immediately.

\[ \square \]

### 3.2 Probabilistic aspects

We can rewrite the bound on \( ||\mathcal{R}_i|| \) as

$$||\mathcal{R}_i|| \leq \sup_{j \neq i} \frac{|X_j'(S_i + \tau \text{Id})^{-1} X_i|}{n} \frac{L(B_n(i)) ||X_i||}{\sqrt{n} \tau} \sqrt{\frac{||\Sigma||}{2} \left( [||\psi(\bar{\gamma}_i(i))| \land | \bar{\gamma}_i(i) | / c_i] |\text{prox}_{c_i}(\rho)(\bar{\gamma}_i(i))| \right)}. $$

The bound on \( ||\mathcal{R}_i|| \) is encouraging since it shows that we can control \( ||\tilde{\beta} - \beta|| \) in \( L_k \) provided we can control each terms in the product in \( L_5k \); indeed, for a product of \( k \) random variables \( \{W_j\}_{j=1}^k \), we have

$$\mathbb{E} \left( \prod_{j=1}^k W_j \right) \leq \prod_{j=1}^k \left[ \mathbb{E} \left( \left| W_j \right| \right)^k \right]^{1/k}$$

by Hölder’s inequality. In particular, we will later need control of

$$\mathbb{E} \left( ||\tilde{\beta} - \beta||^2 \right)$$

and will therefore require subsequent bounds to in \( L_{10} \).

#### 3.2.1 On \( \sup_{j \neq i} |X_j'(S_i + \tau \text{Id})^{-1} X_i|/n \)

We will control \( X_j'(S_i + \tau \text{Id})^{-1} X_i/n \) by appealing to Lemma B-2.

**Lemma 3.3.** Suppose \( X_i \) are independent and satisfy the concentration assumptions mentioned above. Then

$$\sup_{j \neq i} |X_j'(S_i + \tau \text{Id})^{-1} X_i|/n \leq \frac{1}{\sqrt{n}} \sup_{j \neq i} \frac{||X_i||}{\tau \sqrt{n}} \text{polyLog}(n)/c_n^{1/2}$$

in \( L_{10} \), provided we control \( \sup_{j \neq i} \frac{||X_i||}{\tau \sqrt{n}} \) in \( \sqrt{L_{20}} \).

We use the perhaps slightly unusual notation \( \sqrt{L_{20}} \) to simply say that we control \( \sqrt{\mathbb{E}(Z_{20})} \) for a random variable \( Z \).

**Proof.** Let us work conditionally on \( X(i) = \{ X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n \} \). Call \( v_{j,(i)} = (S_i + \tau \text{Id})^{-1} X_j \). The map \( F_j(X_i) = X_j'(S_i + \tau \text{Id})^{-1} X_i = X_j'v_{j,(i)} \) is Lipschitz (as a function of \( X_i \)) with Lipschitz constant

$$\sqrt{X_j'(S_i + \tau \text{Id})^{-2} X_j} \leq ||X_j||/\tau.$$  

Indeed, it is linear in \( X_i \). Therefore, using Lemma B-2, we see that

$$\frac{1}{\sqrt{n}} \sup_{j \neq i} |X_j'(S_i + \tau \text{Id})^{-1} X_i|/||X(i)|| \leq \frac{1}{\sqrt{n}} \sup_j \frac{||X_j||}{\tau \sqrt{n}} \text{polyLog}(n)/c_n + \sup_j |m_{F_j}|.$$  

with overwhelming \((X_i)\)-probability and in \( L_{10} \). Recall that in Lemma B-2, we have a choice between the mean and the median for the definition of \( m_{F_j} \). Here we choose the mean.

Since \( X_i \) has mean 0, we see that \( m_{F_j} = 0 \), so that

$$\frac{1}{\sqrt{n}} \sup_{j \neq i} |X_j'v_{j,(i)}|/||X(i)|| \leq \frac{1}{\sqrt{n}} \sup_j \frac{||X_j||}{\tau \sqrt{n}} \text{polyLog}(n)/c_n$$

with overwhelming \((X_i)\)-probability and in \( L_{10} \). We can then integrate over \( X(i) \) to get the result.  

\[ \square \]
We note that using the fact that \( X_j \rightarrow \|X_j\|/\sqrt{n} \) is \( n^{-1/2} \)-Lipschitz we see that
\[
\sup_{j \neq i} \left\| X_j \right\| / \sqrt{n} - m \|X_j\|/\sqrt{n} \leq \text{polyLog}(n)/\sqrt{nc_n} \text{ in } L_{20} .
\]
Recall that \( \text{cov} (X_i) = \text{Id}_p \). So \( m \|X_j\|/\sqrt{n} \) is of order 1 in the case we are interested in, we see that
\[
\sup_{j \neq i} \left\| X_j \right\| / \sqrt{n} = O(\sqrt{L_{20}}(1) ,
\]
provided \( nc_n \gg \text{polyLog}(n) \). This is clearly the case under our assumptions.

### 3.2.2 Control of the residuals \( R_i \) and \( \hat{r}_{i,(i)} \)

Our aim here is to show that we can control \( \sup_i |R_i| \), where \( R_i = \epsilon_i - X_j \hat{\beta} \) are the residuals from the full ridge-regression model. This will allow us to achieve control of \( B_n(i) \). As \( \hat{r}_{i,(i)} \) is much easier to understand than \( R_i \), our strategy is to relate the two.

**Lemma 3.4.** We have the deterministic bound
\[
|R_i| \leq |\hat{r}_{i,(i)}| + \frac{\|X_i\|^2}{n} \frac{1}{\tau} |\psi(\hat{r}_{i,(i)})| .
\]
Denoting by \( \mathcal{E}_n = \sup_{1 \leq i \leq n} |\epsilon_i| \), we have under our assumptions on \( \{X_i\}_{i=1}^n \),
\[
\sup_{1 \leq i \leq n} |\hat{r}_{i,(i)}| \leq \mathcal{E}_n + \|W_n\| + \frac{1}{n} \sup_{1 \leq i \leq n} \left\| X_i \right\| \psi(\mathcal{E}_n) \vee \psi(-\mathcal{E}_n) \text{polyLog}(n)/\sqrt{nc_n} \text{ in } L_k .
\]
Under the assumption that \( |\psi(x)| = O(|x|^m) \) for some fixed \( m \) at infinity, we have
\[
\sup_i |R_i| \leq K(\sup_i |\hat{r}_{i,(i)}|)^{m+1} \text{ in } L_k ,
\]
and \( \|W_n\| + \frac{1}{n} \sup_{1 \leq i \leq n} \|X_i\| \psi(\mathcal{E}_n) \vee \psi(-\mathcal{E}_n) = O_L_k(\|W_n\| + \mathcal{E}_n^{m}/\sqrt{nc_n}) .
\]

**Proof.** Recall the representation
\[
\beta_1 - \beta_2 = (S_{\beta_1,\beta_2} + \tau \text{Id}_p)^{-1} (f(\beta_1) - f(\beta_2)) .
\]
Take \( \beta_1 = \hat{\beta} \) and \( \beta_2 = \hat{\beta}_{(i)} \). Note that
\[
f(\hat{\beta}_{(i)}) = -\frac{1}{n} X_i \psi(\hat{r}_{i,(i)}) - \frac{1}{n} \sum_{j \neq i} X_i \psi(\hat{r}_{j,(i)}) + \tau \hat{\beta}_{(i)} = -\frac{1}{n} X_i \psi(\hat{r}_{i,(i)})
\]
by definition of \( \hat{\beta}_{(i)} \). Therefore,
\[
\hat{\beta} - \hat{\beta}_{(i)} = \frac{1}{n} (S_{\beta,\hat{\beta}_{(i)}} + \tau \text{Id}_p)^{-1} X_i \psi(\hat{r}_{i,(i)}) .
\]
Since \( \hat{r}_{i,(i)} - R_i = X_i'(\hat{\beta} - \hat{\beta}_{(i)}) \), we also have
\[
|R_i| \leq \left\| X_i \right\| |\hat{r}_{i,(i)}| / \sqrt{n} \leq \frac{1}{\tau} |\psi(\hat{r}_{i,(i)})| .
\]
We conclude that
\[
|R_i| \leq |\hat{r}_{i,(i)}| + \frac{\|X_i\|^2}{n} \frac{1}{\tau} |\psi(\hat{r}_{i,(i)})| .
\]
Now under assumptions, we have \( \sup_{1 \leq i \leq n} \|X_i\|^2/n - 1 = O_L_k(\text{polyLog}(n)/\sqrt{nc_n}) \), according to either Lemma B-2 or Lemma B-3. Using the fact that \( \|\hat{\beta}_{(i)}\| \leq \|W_{n,(i)}\| \) (see Lemma 2.2), the independence of \( X_i \) and \( \hat{\beta}_{(i)} \), we have, through Lemma B-2,
\[
\sup_{1 \leq i \leq n} |X_i'\hat{\beta}_{(i)}| \leq \sup_{1 \leq i \leq n} \|W_{n,(i)}\| \text{polyLog}(n)/\sqrt{nc_n} .
\]
Proposition 3.2. Under the assumption that 
\[ |\tilde{r}_{i,(i)}| \leq |\epsilon_i| + \sup_i \|X_i\| \|\psi(\epsilon_i)/n\| \text{polyLog}(n)/\sqrt{c_n} \text{ in } L_k. \]

Denoting by \( E_n = \sup_{1 \leq i \leq n} |\epsilon_i| \), we have, using the fact that \( \psi \) is increasing,
\[ \sup_{1 \leq i \leq n} |\tilde{r}_{i,(i)}| \leq E_n + \left[ \|W_n\| + n^{1/2} \sup_{1 \leq i \leq n} \frac{\|X_i\|}{n^{1/2}} \left( \psi(E_n) \vee \psi(-E_n) \right) \right] \text{polyLog}(n)/\sqrt{c_n} \text{ in } L_k, \]
for any given \( k \). We note that if \( |\psi(x)| = O(x^m) \) at \( \infty \), we have the bound \( \sup_{1 \leq i \leq n} |\tilde{r}_{i,(i)}| \leq \sup_{1 \leq i \leq n} |\tilde{r}_{i,(i)}|^{m+1} \) and therefore,
\[ \sup_{1 \leq i \leq n} |R_i| \leq \left[ E_n + \text{polyLog}(n)/\sqrt{c_n} \|W_n\| + \frac{1}{n} \sup_i \|X_i\| \|\psi(E_n) \vee \psi(-E_n)\| \right]^{m+1} \text{ in } L_k, \]
provided the bound on \( \sup_{1 \leq i \leq n} |\tilde{r}_{i,(i)}| \) holds in \( L_{nk} \). Note that this is guaranteed under our assumptions. Of course, here we are using control of \( \sup_i \|X_i\|^2/n \), which we get by controlling \( \|X_i\|/\sqrt{n} \) through concentration arguments. The fact that \( \sup_i \|X_i\|/\sqrt{n} = O_{L_k}(1) \) gives us the last statement of the lemma.

**Remark 1:** at the gist of the bound on \( \tilde{r}_{i,(i)} \) is a uniform bound on \( \|\tilde{\beta}_{(i)}\| \) in \( L_k \). If one is not concerned about having assumptions that limit the existence of moments for \( \sqrt{1/n} \sum_{i=1}^n \rho(\epsilon_i) \), one could use the bound \( \sup_i \|\tilde{\beta}_{(i)}\| \leq \sqrt{2/\tau} \sqrt{1/n} \sum_{i=1}^n \rho(\epsilon_i) \) which is immediate from Lemma 2.2. This would change slightly the appearance of our bounds on \( \sup_i |\tilde{r}_{i,(i)}| \). In particular, under our assumptions, this bound is valid.

**Remark 2:** We note that a similar result holds of course for \( \tilde{r}_{j,(i)} \). More precisely,
\[ |\tilde{r}_{j,(i)} - R_j| \leq \frac{1}{n} X_j (S_{\hat{\beta}_j,\hat{\beta}_1} + \tau \text{Id}_p)^{-1} X_i \|\psi(\tilde{r}_{i,(i)})\|, \]
and hence,
\[ |\tilde{r}_{j,(i)} - R_j| \leq \frac{\|X_j\|}{n \tau} \|X_i\| \|\psi(\tilde{r}_{i,(i)})\|. \]

Of course, this bound is very coarse and we will see that we can get a better one later.

However, this finally allows us to have the following proposition

**Proposition 3.2.** Under the assumption that \( |\psi(x)| = O(|x|^m) \), we have the bound
\[ B_n(i) \leq K \left[ E_n + \left( \|W_n\| + \frac{\epsilon_m}{\sqrt{n}} \right) \text{polyLog}(n)/\sqrt{c_n} \right]^{m+1} \text{ in } L_k, \]
where \( K \) is a constant independent of \( p \) and \( n \). When \( \|W_n\| \) and \( \frac{\epsilon_m}{\sqrt{n}} \) are bounded in \( L_k \), this bound simply becomes
\[ B_n(i) \leq K \left[ E_n \vee \text{polyLog}(n)/\sqrt{c_n} \right]^{m+1} \text{ in } L_k. \]

The same bound holds for \( \sup_i B_n(i) \) in \( L_k \).

**Proof.** The result follows easily from the fact that
\[ B_n(i) = \sup_{j \neq i} \left[ |\tilde{r}_{j,(i)}| + |X_j\| \right], \]
the fact that
\[ \sup_i \sup_{j \neq i} |\tilde{r}_{j,(i)} - R_j| \leq \sup_i \sup_{j \neq i} \frac{\|X_j\|}{n \tau} \|X_i\| |\psi(\tilde{r}_{i,(i)})|, \]
and the bounds on \( \sup_i |R_i| \) we have derived earlier. The part concerning \( \sup_i \sup_{j \neq i} |X_j\| \) is easily shown to be negligible compared to this quantity from our previous investigations concerning \( X_j (S_i + \tau \text{Id}_p)^{-1} X_i \).
3.2.3 Consequences

We have the following result. Recall that $\psi'$ is assumed to be Lipschitz with Lipschitz constant $L(u)$ on $(-|u|,|u|)$.

**Proposition 3.3.** Suppose that $|\psi(x)| = O(|x|^m)$ and $\mathcal{E}_n^m = o(\sqrt{n})$ in $L_k$. Suppose further that $L(x) \leq K|x|^m$. Then we have

$$\|\mathcal{R}_i\| \leq K\frac{\text{polyLog}(n)}{nc_n} \left( (\mathcal{E}_n \vee (c_n)^{-1/2}\text{polyLog}(n)) \right)^{2m+1} \text{ in } L_k.$$  

In particular, if $\mathcal{E}_n = \text{polyLog}(n)$ and $1/c_n = O(\text{polyLog}(n))$, we have

$$\|\mathcal{R}_i\| \leq K\frac{\text{polyLog}(n)}{n} \text{ in } L_k.$$  

Furthermore, the same bounds hold for $\sup_i \|\mathcal{R}_i\|$.

**Proof.** The proof follows by aggregating all the intermediate results we had and noticing that under our assumptions, $|||\Sigma|||_2 = O_{L_k}(c_n^{-1/2})$. This latter result follows easily from a standard $\varepsilon$-net and union bound argument for controlling $|||\Sigma|||_2$ - see e.g Talagrand (2003), Appendix A.4. We provide some details in Lemma B-5.

The statement concerning $\sup_i \|\mathcal{R}_i\|$ follows by the same method. \hfill \Box

We have the following very important corollary.

**Corollary 3.1.** Under Assumptions **O1-O7**, we have

$$\| \tilde{\beta} - \hat{\beta}_i \| = O_{L_k}(\frac{\text{polyLog}(n)}{n})$$

In particular, we have

$$\mathbb{E} \left( \| \tilde{\beta} - \hat{\beta}_i \|^2 \right) = O(\text{polyLog}(n)/n^2).$$

Also,

$$\sup_{1 \leq i \leq n} \sup_{j \neq i} |\tilde{r}_{j,(i)} - R_j| = O_{L_k}(\frac{\text{polyLog}(n)}{n^{1/2}}).$$

Finally,

$$\sup_i |R_i - \text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})| = O_{L_k}(\frac{\text{polyLog}(n)}{n^{1/2}}).$$

The only parts that may require a discussion are the ones involving the residuals. However, they follow easily from the very coarse bound

$$\sup_{j \neq i} |\tilde{r}_{j,(i)} - R_j| = \sup_{j \neq i} |X_j' (\tilde{\beta} - \hat{\beta}_i)| \leq \sup_{j \neq i} |X_j' (\tilde{\beta} - \hat{\beta}_i)| + \sup_{j \neq i} |X_j' (\hat{\beta}_i - \tilde{\beta}_i)|,$$

$$\leq \left( \sup_{1 \leq j \leq n} \frac{|X_j|}{\sqrt{n}} \right) \sqrt{n} \| \tilde{\beta} - \hat{\beta}_i \| + \sup_{j \neq i} |X_j' \eta_i|,$$

and the fact that $\left( \sup_{1 \leq j \leq n} \frac{|X_j|}{\sqrt{n}} \right) = O_{L_k}(1)$ under our assumptions. Recalling that $\| \tilde{\beta} - \hat{\beta}_i \| \leq \| \mathcal{R}_i \|$ and hence $\sup_i \| \tilde{\beta} - \hat{\beta}_i \| \leq \sup_i \| \mathcal{R}_i \|$ gives control of the first term. Control of the second term follows basically from Lemma 3.3.

Concerning the approximation of $R_i$, recall that

$$R_i = \epsilon_i - X_i' \tilde{\beta} = \epsilon_i - X_i' \tilde{\beta}_i - X_i' (\tilde{\beta} - \tilde{\beta}_i).$$

Now, given the definition of $\hat{\beta}_i$, we have

$$X_i' \hat{\beta}_i = X_i' \tilde{\beta}_i + c_i \text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)}).$$
Hence,
\[ \epsilon_i - X'_i\beta_i = \tilde{r}_{i,(i)} - c_i \text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)}) = \text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)}) , \]
where the last equality is a standard property of the proximal mapping (see Lemma A-1 if needed). So we have established that
\[ \sup_i |R_i - \text{prox}_{c_i}(\rho)(\tilde{r}_{i,(i)})| = \sup_i \left| X'_i(\beta_i - \tilde{\beta}) \right| \]
and the result follows from our previous bounds.

3.3 Asymptotically deterministic character of \( \|\tilde{\beta}\|^2 \)

**Proposition 3.4.** Under our assumptions,
\[ \text{var}\left(\|\tilde{\beta}\|^2\right) \rightarrow 0 \text{ as } n \rightarrow \infty. \]

Therefore \( \|\tilde{\beta}\|^2 \) has a deterministic equivalent in probability and in \( L_2 \).

In particular, when \( c_n = 1/\text{polyLog}(n) \), we have
\[ \text{var}\left(\|\tilde{\beta}\|^2\right) = O\left(\frac{\text{polyLog}(n)}{n}\right) . \]

**Proof.** We will use the Efron-Stein inequality to show that \( \text{var}\left(\|\tilde{\beta}\|^2\right) \) goes to 0 as \( n \rightarrow \infty \). In what follows, we assume that \( \psi(\epsilon_i) \) have enough moments for all the expectations of the type \( E\left(\|\tilde{\beta}\|^{2k}\right) \) to be bounded like \( 1/\tau^{2k} \). Note that this the content of our Lemma 2.2.

Recall that the Efron-Stein inequality Efron and Stein (1981) gives, if \( Y \) is a function of \( n \) independent random variables, and \( Y_{(i)} \) is any function of all those random variables except the \( i \)-th,
\[ \text{var}\left(Y\right) \leq \sum_{i=1}^{n} \text{var}\left(Y - Y_{(i)}\right) \leq \sum_{i=1}^{n} E\left((Y - Y_{(i)})^2\right) . \]

We first observe that
\[ E\left(\|\tilde{\beta}\|^2 - \|\tilde{\beta}_{(i)}\|^2\right)^2 \leq 2 \left[ E\left(\|\tilde{\beta}\|^2 - \|\tilde{\beta}_{i}\|^2\right)^2 + E\left(\|\tilde{\beta}_{i}\|^2 - \|\tilde{\beta}_{(i)}\|^2\right)^2 \right] . \]

Of course, using the fact that \( \tilde{\beta} = \beta - \tilde{\beta}_{i} + \tilde{\beta}_{i} \) and \( \|\beta\|^2 - \|\beta_{i}\|^2\|^2 = (\beta - \tilde{\beta}_{i})'(\beta + \tilde{\beta}_{i})^2 \), \( \|\tilde{\beta}\|^2 - \|\tilde{\beta}_{i}\|^2\|^2 = O_{L_1}(\|\beta - \beta_{i}\|^4) + O_{L_1}(\|\beta - \beta_{i}\|^4) \), by the Cauchy-Schwarz inequality, since \( E\left(\|\tilde{\beta}\|^2\right) \) exists and is bounded by \( K/\tau^2 \).

Using the results of Corollary 3.1, we see that
\[ E\left(\|\tilde{\beta}\|^2 - \|\tilde{\beta}_{i}\|^2\right)^2 = O\left(\frac{\text{polyLog}(n)}{n^2}\right) = o(n^{-1}) . \]

On the other hand, given the definition in Equation (11),
\[ \|\tilde{\beta}_{i}\|^2 - \|\tilde{\beta}_{(i)}\|^2 = 2\frac{1}{n} \beta_{(i)}(S_i + \tau\text{Id})^{-1}X_i\psi(\text{prox}_{c_i}(\tilde{r}_{i,(i)})) + \frac{1}{n} X'_i(S_i + \tau\text{Id})^{-2}X_i\psi^2(\text{prox}_{c_i}(\tilde{r}_{i,(i)})) . \]

Since \( S_i \) is independent of \( X_i \), and \( \|(S_i + \tau\text{Id})^{-1}\| \leq 1/\tau \), \( \beta_{(i)}(S_i + \tau\text{Id})^{-1}X_i = O_{L_1}(\|\beta_{(i)}\|^1/\sqrt{c_n}) \), using our concentration assumptions applied to linear forms. Therefore, we see that both terms are \( O_{L_2}(1/n^{1/2}) \) provided \( \psi(\text{prox}_{c_i}(\tilde{r}_{i,(i)})) \) has 4+ absolute moments - uniformly bounded in \( n \) - by using Hölder’s inequality. Under our assumptions, given our work on \( \tilde{r}_{i,(i)} \), the fact that the prox is a contractive mapping (Moreau (1965)) and that we assume that \( \text{sign}(\psi(x)) = \text{sign}(x) \), it is clear that this is the case. We conclude that then
\[ E\left(\|\tilde{\beta}_{i}\|^2 - \|\tilde{\beta}_{(i)}\|^2\right)^2 = O\left(\frac{1}{n^2 c_n}\right) = O\left(\frac{\text{polyLog}(n)}{n^2}\right) . \]
Taking $Y = \|\hat{\beta}\|^2$ and $Y(i) = \|\hat{\beta}(i)\|^2$ in the Efron-Stein inequality, we clearly see that
\[
\text{var} \left( \|\hat{\beta}\|^2 \right) = O\left( \frac{\text{polyLog}(n)}{n} \right) = o(1) .
\]
This shows that $\|\hat{\beta}\|$ has a deterministic equivalent in probability and in $L_2$. \qed

4 Leaving out a predictor

In El Karoui et al. (2013), we showed through probabilistic heuristics that the probabilistic properties of the entries of $\hat{\beta}$ could be understood by leaving out predictors. We now show that all the formal manipulations we did in that paper are valid under our assumptions. In that step, we do need at various points that the entries of the data vector $X_i$ be independent, whereas as we showed before, it is not important when studying what happens when we leave out an observation.

We call $V$ the $n \times (p - 1)$ matrix corresponding to the first $(p - 1)$ columns of the design matrix $X$. We call $V_i$ in $\mathbb{R}^{p-1}$ the vector corresponding to the first $p - 1$ entries of $X_i$, i.e $V_i' = (X_i(1), \ldots, X_i(p-1))$.

Let us call $\hat{\gamma}$ the solution of our optimization problem when $X_i(p) = 0$ for all $i$, i.e the solution we get when we solve our original problem with the design matrix $V$ instead of $X$.

The corresponding residuals are $\{r_{i,[p]}\}_{i=1}^n$. Hence, $r_{i,[p]} = \epsilon_i - V_i'\hat{\gamma}$. We call
\[
\begin{align*}
    u_p &= \frac{1}{n} \sum_{i=1}^n \psi'(r_{i,[p]})V_iX(p) , \\
    \mathcal{G}_p &= \frac{1}{n} \sum_{i=1}^n \psi'(r_{i,[p]})V_iV_i' .
\end{align*}
\]

Note that $\mathcal{G}_p$ is $(p - 1) \times (p - 1)$. We call
\[
\xi_n \triangleq \frac{1}{n} \sum_{i=1}^n X_i^2(p)\psi'(r_{i,[p]}) - u_p'(\mathcal{G}_p + \tau \text{Id})^{-1}u_p ,
\]
and
\[
N_p \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i(p)\psi(r_{i,[p]}) .
\]

We consider
\[
b_p \triangleq \frac{1}{\sqrt{n}} \frac{N_p}{\tau + \xi_n} .
\]

We will show later, in Subsubsection 4.2.2 that $\xi_n \geq 0$. Note that when $\xi_n > 0$, we have
\[
b_p = \frac{1}{n} \sum_{i=1}^n X_i(p)\psi'(r_{i,[p]}) - \tau b_p}{\frac{1}{n} \sum_{i=1}^n X_i^2(p)\psi'(r_{i,[p]}) - u_p'(\mathcal{G}_p + \tau \text{Id})^{-1}u_p} = \frac{n^{-1/2}N_p - \tau b_p}{\xi_n} .
\]

We call
\[
\tilde{b} = \begin{bmatrix} \hat{\gamma} \\ b_p \end{bmatrix} + \begin{bmatrix} \mathcal{G}_p & \tau \text{Id} \\ \text{Id} \end{bmatrix}^{-1}u_p .
\]

The aim of our work is to establish Corollary 4.1, which shows that $\tilde{b}$ is a $\sqrt{n}$-consistent approximation of $\hat{\beta}$ - in the sense of Euclidian norm. Because the last coordinate of $b$ has a reasonably simple probabilistic structure and our approximations are sufficiently good, we will be able to transfer our insights about this coordinate to $\hat{\beta}_p$.

Once again, the approximating quantities we consider are “very natural” in light of our work in El Karoui et al. (2013).
4.1 Deterministic aspects

Proposition 4.1. We have
\[ \|\hat{b} - \tilde{b}\| \leq \frac{1}{\tau} |b_p| \sup_{1 \leq i \leq n} \|d_{i,p}\| \|\hat{\Sigma}\|_2 \sqrt{\|(\mathbf{G}_p + \tau \Id)^{-1} u_p\|^2 + 1}. \] (23)

where \( d_{i,p} = [\psi'(\gamma_{i,p}^*) - \psi'(r_{i,[p]})] \) and \( \gamma_{i,p}^* \) is in the interval \((\epsilon_i - V_i\hat{\gamma}, \epsilon_i - X_i\tilde{b})\).

Furthermore, \( \|(\mathbf{G}_p + \tau \Id)^{-1} u_p\|^2 \leq \frac{1}{n} \sum_{i=1}^{n} X_i^2(p) \psi'(r_{i,[p]}) \).

As we saw in Equation (8), we have
\[ \|\hat{b} - \tilde{b}\| \leq \frac{1}{\tau} \|f(\tilde{b})\|, \]
where
\[ f(\tilde{b}) = -\frac{1}{n} \sum_{i=1}^{n} X_i \psi(\epsilon_i - X_i\tilde{b}) + \tau\tilde{b}. \]

We note furthermore that
\[ g(\hat{\gamma}) \triangleq -\frac{1}{n} \sum_{i=1}^{n} V_i \psi(\epsilon_i - V_i\hat{\gamma}) + \tau\hat{\gamma} = 0_{p-1}. \]

The strategy of the proof is to control \( f(\tilde{b}) \) by approximating it by \( g(\hat{\gamma}) \).

Proof. a) Work on the first \((p-1)\) coordinates of \( f(\tilde{b}) \)

We call \( f_{p-1}(\beta) \) the first \( p-1 \) coordinates of \( f(\beta) \). We call \( \hat{\gamma}_{ext} \) the \( p \)-dimensional vector whose first \( p-1 \) coordinates are \( \hat{\gamma} \) and last coordinate is 0, i.e
\[ \hat{\gamma}_{ext} = \begin{bmatrix} \hat{\gamma} \\ 0 \end{bmatrix}. \]

For a vector \( v \), we use the notation \( v_{comp,k} \) to denote the \( p-1 \) dimensional vector consisting of all the coordinates of \( v \) except the \( k \)-th.

Clearly,
\[ f_{p-1}(\tilde{b}) = f_{p-1}(\beta) - g(\hat{\gamma}) = -\frac{1}{n} \sum_{i=1}^{n} V_i \left[ \psi(\epsilon_i - X_i\tilde{b}) - \psi(\epsilon_i - V_i\hat{\gamma}) \right] + \tau(\tilde{b}_{comp,p} - \hat{\gamma}). \]

We can write by using the mean value theorem
\[ \psi(\epsilon_i - X_i\tilde{b}) - \psi(\epsilon_i - V_i\hat{\gamma}) = \psi'(r_{i,[p]}) X_i'(\hat{\gamma}_{ext} - \tilde{b}) + [\psi'(\gamma_{i,p}^*) - \psi'(r_{i,[p]})] X_i'(\hat{\gamma}_{ext} - \tilde{b}). \]

Let us call
\[ d_{i,p} = [\psi'(\gamma_{i,p}^*) - \psi'(r_{i,[p]})], \]
\[ \delta_{i,p} = [\psi'(\gamma_{i,p}^*) - \psi'(r_{i,[p]})] X_i'(\hat{\gamma}_{ext} - \tilde{b}), \]
\[ R_p = -\frac{1}{n} \sum_{i=1}^{n} d_{i,p} V_i X_i'(\hat{\gamma}_{ext} - \tilde{b}). \]

We have with this notation
\[ f_{p-1}(\tilde{b}) = -\frac{1}{n} \sum_{i=1}^{n} \psi'(r_{i,[p]}) V_i X_i'(\hat{\gamma}_{ext} - \tilde{b}) + \tau(\tilde{b}_{comp,p} - \hat{\gamma}) + R_p \triangleq A_p + R_p. \]

We note that by definition,
\[ \hat{\gamma}_{ext} - \tilde{b} = b_p \begin{bmatrix} (\mathbf{G}_p + \tau \Id)^{-1} u_p \\ -1 \end{bmatrix}, \]
\[ \tilde{b}_{comp,p} - \hat{\gamma} = -b_p (\mathbf{G}_p + \tau \Id)^{-1} u_p. \]
Therefore,

\[ A_p = -b_p \left( \frac{1}{n} \sum_{i=1}^{n} \psi'(r_i,p) V_i [V_i' (\mathcal{G}_p + \tau \text{Id})^{-1} u_p - X_i(p)] \right) + \tau (-b_p (\mathcal{G}_p + \tau \text{Id})^{-1} u_p) . \]

Recalling the definition of \( \mathcal{G}_p \) and \( u_p \), we see that

\[ A_p = -b_p (\mathcal{G}_p (\mathcal{G}_p + \tau \text{Id})^{-1} u_p - u_p + \tau (\mathcal{G}_p + \tau \text{Id})^{-1} u_p) = 0_{p-1} , \]

since \( \mathcal{G}_p (\mathcal{G}_p + \tau \text{Id})^{-1} + \tau (\mathcal{G}_p + \tau \text{Id})^{-1} = \text{Id} . \)

We conclude that

\[ f_{p-1}(\tilde{b}) = R_p . \]

b) Work on the last coordinate of \( f(\tilde{b}) \)
We call \( [f(\tilde{b})]_p \) the last coordinate of \( f(\tilde{b}) \). We recall the representation

\[ \psi(\epsilon_i - X\hat{b}_i) - \psi(\epsilon_i - V\hat{\gamma}_i) = \psi'(r_i,p) X_i'(\hat{\gamma}_{\text{ext}} - \tilde{b}) + [\psi'(\gamma^*_i,p) - \psi'(r_i,p)] X_i'(\hat{\gamma}_{\text{ext}} - \tilde{b}) \]

and call

\[ \delta_{i,p} = [\psi'(\gamma^*_i,p) - \psi'(r_i,p)] X_i'(\hat{\gamma}_{\text{ext}} - \tilde{b}) . \]

Clearly,

\[ \psi(\epsilon_i - X\hat{b}_i) = \psi(r_i,p) + \psi'(r_i,p) X_i'(\hat{\gamma}_{\text{ext}} - \tilde{b}) + \delta_{i,p} , \]

\[ = \psi(r_i,p) + \psi'(r_i,p) b_p [V_i'(\mathcal{G}_p + \tau \text{Id})^{-1} u_p - X_i(p)] + \delta_{i,p} . \]

We therefore see that

\[ [f(\tilde{b})]_p + \frac{1}{n} \sum_{i=1}^{n} X_i(p) \delta_{i,p} = -\frac{1}{n} \sum_{i=1}^{n} X_i(p) \left( \psi(r_i,p) + \psi'(r_i,p) b_p [V_i'(\mathcal{G}_p + \tau \text{Id})^{-1} u_p - X_i(p)] \right) + \tau \tilde{b}_p , \]

\[ = -\frac{1}{n} \sum_{i=1}^{n} X_i(p) \psi(r_i,p) - b_p u'_p (\mathcal{G}_p + \tau \text{Id})^{-1} u_p + b_p \frac{1}{n} \sum_{i=1}^{n} \psi'(r_i,p) X_i^2(p) + \tau b_p , \]

\[ = -\left[ \frac{1}{n} \sum_{i=1}^{n} X_i(p) \psi(r_i,p) - \tau b_p \right] + b_p \left( \frac{1}{n} \sum_{i=1}^{n} \psi'(r_i,p) X_i^2(p) - u'_p (\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right) , \]

\[ = 0 . \]

We conclude that

\[ [f(\tilde{b})]_p = -\frac{1}{n} \sum_{i=1}^{n} X_i(p) \delta_{i,p} = -\frac{1}{n} \sum_{i=1}^{n} d_{i,p} X_i(p) X_i'(\hat{\gamma}_{\text{ext}} - \tilde{b}) . \]

Representation of \( f(\tilde{b}) \)

Aggregating all the results we have obtained so far, we see that

\[ f(\tilde{b}) = \left( -\frac{1}{n} \sum_{i=1}^{n} d_{i,p} X_i X_i' \right) (\hat{\gamma}_{\text{ext}} - \tilde{b}) , \]

\[ = b_p \left( \frac{1}{n} \sum_{i=1}^{n} d_{i,p} X_i X_i' \right) \left[ (\mathcal{G} + \tau \text{Id})^{-1} u_p \right] . \]

We conclude immediately that

\[ \| f(\tilde{b}) \| \leq |b_p| \sup_{1 \leq i \leq n} |d_{i,p}| \| \Sigma \|_2 \sqrt{\| (\mathcal{G} + \tau \text{Id})^{-1} u_p \|^2 + 1} . \quad (24) \]
Calling $D_{\psi'(r_{i,[p]})}$ the diagonal matrix with $(i, i)$ entry $\psi'(r_{i,[p]})$, we see that
\[
\frac{1}{n}V' D_{\psi'(r_{i,[p]})} X(p).
\]
Therefore,
\[
\|(\mathcal{S} + \tau \text{Id})^{-1}u_p\|^2 = \frac{X(p)^r D_{\psi'(r_{i,[p]})}^1/2}{\sqrt{n}} \left( \frac{V' D_{\psi'(r_{i,[p]})}^1/2}{n} + \tau \text{Id} \right)^{-1} \frac{V' D_{\psi'(r_{i,[p]})}^1/2}{\sqrt{n}} \frac{X(p)}{n}.
\]
Clearly,
\[
\frac{D_{\psi'(r_{i,[p]})}^1/2}{\sqrt{n}} \left( \frac{V' D_{\psi'(r_{i,[p]})}^1/2}{n} + \tau \text{Id} \right)^{-1} \frac{V' D_{\psi'(r_{i,[p]})}^1/2}{\sqrt{n}} \preceq \text{Id}.
\]
So we have
\[
\|(\mathcal{S} + \tau \text{Id})^{-1}u_p\|^2 \leq \frac{1}{n}X(p)^r D_{\psi'(r_{i,[p]})} X(p) = \frac{1}{n} \sum_{i=1}^n X_i X_i^2(p) \psi'(r_{i,[p]}) = (25).
\]

4.2 Probabilistic aspects

From now on, we assume that $X(p)$, the $p$-th column of the design matrix, is independent of $\{V_i, \epsilon_i\}_{i=1}^n$.

Because $r_{i,[p]}$ are the residuals from a “full model” with $p - 1$ predictors, the analysis done above concerning the $R_t$ - see Lemma 3.4 - applies and will allow us to control $\max_{1 \leq i \leq n} |\psi'(r_{i,[p]})|^2$. (Note that the distribution of the errors is the same whether we use $p$ or $p - 1$ predictors because we assume in the regression model that $\beta_0 = 0$ - the study of ridge-regularized robust regression would require an adjustment in the non-null case where $\beta_0 \neq 0$, but since we limit ourselves to the null case, no such adjustment is needed.)

In light of Lemma 3.4 and using independence of $X_i(p)$’s and $r_{i,[p]}$, it is clear that the upper bound in Equation (25) is $O_{L_k}(\text{polyLog}(n))$.

Hence,
\[
\|(\mathcal{S}_p + \tau \text{Id})^{-1}u_p\|^2 = O_{L_k}(\text{polyLog}(n))
\]
This guarantees that
\[
\|(\mathcal{S}_p + \tau \text{Id})^{-1}u_p\|^2 \leq 1 + \|(\mathcal{S}_p + \tau \text{Id})^{-1}u_p\|^2 = O_{L_k}(\text{polyLog}(n)).
\]
We conclude, using Equation (24), that
\[
\tilde{f}(\tilde{b}) \leq K\text{polyLog}(n)|\mathbf{b}_p| \sup_{1 \leq i \leq n} |d_{i,p}| ||\hat{\Sigma}||_2 \text{ in } L_k.
\]
At a high level, we expect $\sup_{1 \leq i \leq n} |d_{i,p}|$ to be small, even compared to $\max_{1 \leq i \leq n} |\psi'(r_{i,[p]})|$ which should give us that
\[
\tilde{f}(\tilde{b}) = o_{L_k}(\text{polyLog}(n)|\mathbf{b}_p|).
\]
We now show that this latter quantity is small.

4.2.1 On $\mathbf{b}_p$

We recall the notations
\[
N_p = \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(r_{i,[p]}) X_i(p),
\]
\[
\xi_n = \frac{1}{n} \sum_{i=1}^n \psi'(r_{i,[p]}) X_i^2(p) - u_p^r (\mathcal{S}_p + \tau \text{Id})^{-1}u_p.
\]
Under our assumptions, we have $E(X_i) = 0$ and $\text{cov}(X_i) = \text{Id}_p$, and hence $E(X_i^2(p)) = 1$. Recall that since we assume that $X(p)$ is independent of $\{V_i, \epsilon_i\}_{i=1}^n$, $X(p)$ is independent of $\{r_{i,[p]}\}_{i=1}^n$. 

18
Proposition 4.2. We have

$$|b_p| \leq \frac{1}{\sqrt{n\tau}} |N_p|.$$ 

Furthermore, under our assumptions, $N_p = O_{L_k}(\text{polyLog}(n))$ and therefore

$$b_p = O_{L_k}(\text{polyLog}(n)n^{-1/2}).$$

Proof. From the definition of $b_p$, we see that, when $\xi_n \neq 0$

$$b_p = \frac{1}{\sqrt{n\tau + \xi_n}} N_p.$$ 

We will see later, in Subsubsection 4.2.2, that $\xi_n \geq 0$. It immediately then follows that

$$|b_p| \leq \frac{1}{\sqrt{n\tau}} |N_p|.$$ 

Using independence of $X(p)$ and $\{V_i, \epsilon_i\}_{i=1}^n$, we have

$$\mathbb{E} (N_p^2) = \frac{1}{n} \sum_{i=1}^n \mathbb{E} (X_i^2(p)) \mathbb{E} (\psi^2(r_{i,[p]})).$$

whether the right-hand side is finite or not.

Since $r_{i,[p]}$ are the residuals of the full model with $p - 1$ predictors, our previous analyses show that $N_p$ has as many moments as we need and $N_p = O_{L_k}(\text{polyLog}(n))$. (Indeed, it suffices to apply reasoning similar to the arguments given in Lemma 2.2 for the control of the moments and our bounds on $r_{i,[p]}$ and therefore on $\psi(r_{i,[p]})$)

We therefore have

$$|b_p| \leq \frac{1}{\sqrt{n\tau}} O_{L_k}(\text{polyLog}(n)).$$

\qed

4.2.2 On $\xi_n$

Let us write $\xi_n$ in matrix form: denoting by $X(p)$ the last column of the design matrix $X$, we have

$$\xi_n = \frac{1}{n} X(p)'D_{\psi'(r_{i,[p]})}^{1/2}MD_{\psi'(r_{i,[p]})}^{1/2}X(p),$$

where

$$M = \text{Id}_n - \frac{D_{\psi'(r_{i,[p]})}^{1/2}V}{\sqrt{n}} \left( \frac{1}{n} V' D_{\psi'(r_{i,[p]})} V + \tau \text{Id} \right)^{-1} V' D_{\psi'(r_{i,[p]})}^{1/2}. \quad (27)$$

Lemma 4.1. We have

$$\xi_n \geq 0.$$ 

Furthermore,

$$|\xi_n - \frac{1}{n} \text{trace} \left( D_{\psi'(r_{i,[p]})}^{1/2}MD_{\psi'(r_{i,[p]})}^{1/2} \right) | = O_{L_k}(\sup_{1 \leq i \leq n} \psi'(r_{i,[p]})/(\sqrt{nc_n})).$$

Proof. Let us first focus on

$$M = \text{Id}_n - \frac{1}{n} D_{\psi'(r_{i,[p]})}^{1/2}V \left( \frac{V' D_{\psi'(r_{i,[p]})} V}{n} + \tau \text{Id} \right)^{-1} V' D_{\psi'(r_{i,[p]})}^{1/2}.$$ 

When $\tau > 0$, it is clear that all the eigenvalues of $M$ are strictly positive, i.e $M$ is positive definite. Indeed, if the singular values of $n^{-1/2} D_{\psi'(r_{i,[p]})}^{1/2}V$ are denoted by $\sigma_1$, the eigenvalues of $M$ are $\tau/(\sigma_1^2 + \tau).$
Therefore, since \( \xi_n = \frac{1}{n} \psi' M v \) with \( v = D^{1/2}_{\psi'(r, [p])} X(p) \), \( \xi_n \geq 0 \).

As we have seen above, \( M \) has eigenvalues between 0 and 1. Therefore,

\[
0 < D^{1/2}_{\psi'(r, [p])} M D^{1/2}_{\psi'(r, [p])} < D_{\psi'(r, [p])}.
\]

The matrix \( M \) is independent of \( X(p) \). \( D_{\psi'(r, [p])} \) is also independent of \( X(p) \).

We can now appeal to Lemma B-3 to obtain

\[
\left| \frac{1}{n} X(p)' D^{1/2}_{\psi'(r, [p])} M D^{1/2}_{\psi'(r, [p])} X(p) - \frac{1}{n} \text{trace} \left(D^{1/2}_{\psi'(r, [p])} M D^{1/2}_{\psi'(r, [p])}\right) \right| = O_L(\frac{1}{\sqrt{n c_n}} \sup_i \psi'(r_i, [p])).
\]

\[\square\]

\section*{About \( \frac{1}{n} \text{trace}\left(D^{1/2}_{\psi'(r, [p])} M D^{1/2}_{\psi'(r, [p])}\right) \)}

\begin{lemma}
Let us call \( \mathcal{G}_p = \frac{1}{n} \sum_{i=1}^{n} \psi'(r_i, [p]) V_i V_i' \) and \( \mathcal{G}_p(i) = \mathcal{G}_p - \frac{1}{n} \psi'(r_i, [p]) V_i V_i' \). Let us also call

\[
c_{\tau, p} = \frac{1}{n} \text{trace} \left( (\mathcal{G}_p + \tau \text{Id})^{-1} \right),
\]

\[
\eta_i = \frac{1}{n} V_i' (\mathcal{G}_p(i) + \tau \text{Id})^{-1} V_i - c_{\tau, p}.
\]

Then we have

\[
\left| \frac{1}{n} \text{trace} (\text{Id}_n - M) - \left( \frac{1}{n} \text{trace} \left(D^{1/2}_{\psi'(r, [p])} M D^{1/2}_{\psi'(r, [p])}\right) \right) c_{\tau, p} \right| \leq \left[ \sup_i |\eta_i| \right] \frac{1}{n} \sum_i \psi'(r_i, [p]). \tag{29}
\]

We also have

\[
\frac{1}{n} \text{trace} (\text{Id}_n - M) = \frac{p}{n} - \tau c_{\tau, p}.
\]

\begin{proof}
We call \( d_{i,i} = \psi'(r_i, [p])/n \). Of course, by using the Sherman-Morrison-Woodbury formula (see e.g. Horn and Johnson (1990), p.19),

\[
M_{i,i} = 1 - d_{i,i} V_i' D^{1/2}_{\psi'(r, [p])} V_i/\tau + \tau \text{Id})^{-1} V_i,
\]

\[
= 1 - d_{i,i} V_i' (\mathcal{G}_p(i) + \tau \text{Id})^{-1} V_i,
\]

\[
= \frac{1}{1 + d_{i,i} V_i' (\mathcal{G}_p(i) + \tau \text{Id})^{-1} V_i}.
\]

Recall that we are interested in \( \frac{1}{n} \sum_i \psi'(r_i, [p]) M_{i,i} = \frac{1}{n} \text{trace} \left(D^{1/2}_{\psi'(r, [p])} M D^{1/2}_{\psi'(r, [p])}\right) \). Note that

\[
\text{trace} (\text{Id}_n - M) = \text{trace} (\mathcal{G}_p(\mathcal{G}_p + \tau \text{Id})^{-1}) = p - \tau \text{trace} ((\mathcal{G}_p + \tau \text{Id})^{-1}) = p - n \tau c_{\tau, p}.
\]

On the other hand,

\[
\text{trace} (\text{Id}_n - M) = \sum_i (1 - M_{i,i}) = \sum_i d_{i,i} \frac{V_i' (\mathcal{G}_p(i) + \tau \text{Id})^{-1} V_i}{1 + d_{i,i} V_i' (\mathcal{G}_p(i) + \tau \text{Id})^{-1} V_i}.
\]

With our definitions, we have

\[
\frac{1}{n} \text{trace} (\text{Id}_n - M) = \left( \frac{1}{n} \sum_i \psi'(r_i, [p]) M_{i,i} \right) c_{\tau, p} + \frac{1}{n} \sum_i \psi'(r_i, [p]) \frac{\eta_i}{1 + d_{i,i} V_i' (\mathcal{G}_p(i) + \tau \text{Id})^{-1} V_i}.
\]

It immediately follows that

\[
\left| \frac{1}{n} \text{trace} (\text{Id}_n - M) - \left( \frac{1}{n} \sum_i \psi'(r_i, [p]) M_{i,i} \right) c_{\tau, p} \right| \leq \left[ \sup_i |\eta_i| \right] \frac{1}{n} \sum_i \psi'(r_i, [p]),
\]

as announced.
\[\square\]
Controlling $\eta_i$

Lemma 4.3. Suppose we can find $\{r_{j,[p]}^{(i)}\}_{j \neq i}$ independent of $V_i$ such that $\sup_{j \neq i} |r_{j,[p]}^{(i)} - r_{j,[p]}| \leq \delta_n(i)$. Leaving out $V_i$ from a regression comes of course to mind and the work of the first section will apply.

Suppose further that we can find $K_n$ such that

$$\sup_{i} \sup_{j \neq i} |\psi'(r_{j,[p]}^{(i)}) - \psi'(r_{j,[p]})| \leq K_n$$

Then

$$\sup_i |\eta_i| = O_L\left( \frac{1}{\tau^2} K_n \|\hat{\Sigma}\|_2 + \frac{\polyLog(n)}{\sqrt{n}c_n} \right), \quad (31)$$

provided $K_n$ has $3k$ uniformly bounded moments.

Proof. We call

$$AM_{i,p} = \frac{1}{n} \sum_{j \neq i} \psi'(r_{j,[p]}^{(i)}) V_j V_j'$$

Then, using for instance the first resolvent identity, i.e $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, we see that

$$\|\|(\Theta_p(i) + \tau Id)^{-1} - (AM_{i,p} + \tau Id)^{-1}\|_2 \leq \frac{1}{\tau^2} K_n \|\hat{\Sigma}\|_2.$$

In particular,

$$\left| \frac{1}{n} V_i' (\Theta_p(i) + \tau Id)^{-1} V_i - \frac{1}{n} V_i' (AM_{i,p} + \tau Id)^{-1} V_i \right| \leq \frac{\|V_i\|^2}{n} \frac{1}{\tau^2} K_n \|\hat{\Sigma}\|_2.$$

However, since $AM_{i,p}$ is independent of $V_i$, we can use Lemma B-3 and see that

$$\sup_{1 \leq i \leq n} \left| \frac{1}{n} V_i' (AM_{i,p} + \tau Id)^{-1} V_i - \frac{1}{n} \trace ((AM_{i,p} + \tau Id)^{-1}) \right| = O_L\left( \frac{\polyLog(n)}{\sqrt{n}c_n} \right),$$

by using the fact that $\lambda^\max((AM_{i,p} + \tau Id)^{-1}) \leq \frac{1}{\tau}$.

However, by the argument we gave above,

$$\left| \frac{1}{n} \trace ((AM_{i,p} + \tau Id)^{-1}) - \frac{1}{n} \trace ((\Theta_p(i) + \tau Id)^{-1}) \right| \leq \frac{1}{\tau^2} K_n \|\hat{\Sigma}\|_2 \frac{p}{n}.$$

We conclude that

$$\sup_{1 \leq i \leq n} \left| \frac{1}{n} V_i' (\Theta_p(i) + \tau Id)^{-1} V_i - \frac{1}{n} \trace ((\Theta_p(i) + \tau Id)^{-1}) \right| \leq \frac{1}{\tau^2} K_n \|\hat{\Sigma}\|_2 \sup_{1 \leq i \leq n} \left[ \frac{p}{n} + \frac{\|V_i\|^2}{n} \right] + \frac{\polyLog(n)}{\sqrt{n}c_n},$$

in $L_k$.

Now, it is clear that $\sup_{1 \leq i \leq n} \|V_i\|^2 / n = O_L(1)$ and finally

$$\sup_{1 \leq i \leq n} \left| \frac{1}{n} V_i' (\Theta_p(i) + \tau Id)^{-1} V_i - \frac{1}{n} \trace ((\Theta_p(i) + \tau Id)^{-1}) \right| = O_L\left( \frac{1}{\tau^2} K_n \|\hat{\Sigma}\|_2 + \frac{\polyLog(n)}{\sqrt{n}c_n} \right).$$

Control of $\frac{1}{n} \trace ((\Theta_p(i) + \tau Id)^{-1}) - \frac{1}{n} \trace ((\Theta_p(i) + \tau Id)^{-1})$ Using the Sherman-Woodbury-Morrison formula, we have

$$(\Theta_p(i) + \tau Id)^{-1} - (\Theta_p + \tau Id)^{-1} = \frac{\psi'(r_{i,[p]}^{(i)}) (\Theta_p(i) + \tau Id)^{-1} V_i V_i' (\Theta_p(i) + \tau Id)^{-1}}{1 + \frac{\psi'(r_{i,[p]}^{(i)})}{n} V_i' (\Theta_p(i) + \tau Id)^{-1} V_i}.$$ 

After taking traces, we see that

$$0 \leq \trace ((\Theta_p(i) + \tau Id)^{-1}) - \trace ((\Theta_p + \tau Id)^{-1}) \leq \frac{1}{\tau},$$

in $L_k$. 

$\}$

21
since \( V_i' (\mathcal{S}_p(i) + \tau \text{Id})^{-2} V_i \leq \frac{1}{n} V_i' (\mathcal{S}_p(i) + \tau \text{Id})^{-1} V_i \).

Therefore,
\[
0 \leq \frac{1}{n} \text{trace} \left( (\mathcal{S}_p(i) + \tau \text{Id})^{-1} \right) - \frac{1}{n} \text{trace} \left( (\mathcal{S}_p + \tau \text{Id})^{-1} \right) \leq \frac{1}{n\tau}.
\]

We conclude that
\[
\sup_{1 \leq i \leq n} |\eta_i| = O_{L_k} \left( \frac{1}{\tau^2} K_n \| \hat{\Sigma} \|_2 + \frac{\text{polyLog}(n)}{\sqrt{nc_n}} + \frac{1}{n\tau} \right),
\]
provided we can use Holder’s inequality. In effect, this requires \( K_n \) to have 3\( k \) uniformly bounded moments.

\[\square\]

### 4.2.3 Control of \( K_n \)

A natural choice for \( r^{(i)}_{j,[p]} \) defined in Lemma 4.3 is to use a leave one out estimator of \( \hat{\gamma} \). Hence, all the work done in Corollary 3.1 becomes immediately relevant.

**Lemma 4.4.** With the notations of Lemma 4.3, we have
\[
\sup_i (\delta_n(i)) = O_{L_k} \left( \frac{\text{polyLog}(n)}{n^{1/2}} \right).
\]

Therefore,
\[
K_n = O_{L_k} \left( n^{-1/2} \text{polyLog}(n) \right).
\]

**Proof.** The first statement of the Lemma is an application of Corollary 3.1 with \( R_j = r_{j,[p]} \) and \( \tilde{r}^{(i)}_{j,[p]} = r^{(i)}_{j,[p]} \).

The control of \( K_n \) follows immediately by using our assumptions on \( \psi' \) and on the growth of \( \mathcal{B}_n(i) \) and \( L(\mathcal{B}_n(i)) \) we had before, now applied to the situation with \( p - 1 \) predictors.

**Important remark:** the previous remark has important consequences for \( c_i \) defined in Equation (12): we just showed that \( \sup_i \frac{1}{n} V_i' (\mathcal{S}_p(i) + \tau \text{Id})^{-1} V_i - c_{\tau,p} \right) = O(\text{polyLog}(n)/\sqrt{n}) \). Recalling the notation
\[
c_{\tau} = \frac{1}{n} \text{trace} \left( \left( \frac{1}{n} \sum_{i=1}^{n} \psi'(R_i)X_iX_i' + \tau \text{Id}_p \right)^{-1} \right),
\]
which is the analog of \( c_{\tau,p} \) when we use all the predictors and not only \( (p - 1) \), we see that \( \sup_i |c_i - c_{\tau}| = O(n^{-1/2} \text{polyLog}(n)) \).

### 4.2.4 Control of \( \xi_n \) and \( b_p \)

We can combine all the results we have obtained so far in the following proposition.

**Proposition 4.3.** We have
\[
|c_{\tau,p}(\xi_n + \tau) - \frac{p}{n}| \leq O_{L_k} \left( \left( \sup_i \psi'(r_{i,[p]}) \right) \left( \frac{\text{polyLog}(n)}{\sqrt{nc_n}} + \frac{1}{\tau^2} K_n \| \hat{\Sigma} \|_2 + \frac{1}{n\tau} \right) \right) = O_{L_k} \left( \frac{\text{polyLog}(n)}{\sqrt{n}} \right).
\]

Furthermore, under our assumptions,
\[
\left( \frac{p}{n} \right)^2 n \mathbb{E} \left( b_p^2 \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( (c_{\tau,p} \psi(r_{i,[p]})^2 \right) + o(1).
\]

**Proof.** The proof of Equation (32) consists just in aggregating all the previous results and noticing that \( c_{\tau,p} \leq p/(n\tau) \) and therefore remains bounded.

We recall that
\[
(\tau + \xi_n) \sqrt{n} b_p \{V_i, \epsilon_i\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(r_{i,[p]} X_i(p).
\]
Therefore,
\[
c_{r,p}(\tau + \xi)_n \sqrt{n}b_p|\{V_i, \epsilon_i\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} c_{r,p}\psi(r_{i,[p]})X_i(p)
\]
Now, \(c_{r,p}\psi(r_{i,[p]})\), which depends only on \(\{V_i, \epsilon_i\}_{i=1}^{n}\) is independent of \(\{X_i(p)\}_{i=1}^{n}\).

We conclude that
\[
E \left( (c_{r,p}(\tau + \xi)_n \sqrt{n}b_p)^2 \right) = \frac{1}{n} \sum_{i=1}^{n} E \left( (c_{r,p}\psi(r_{i,[p]})^2 \right).
\]

Given the result in Equation (32), this means that
\[
\left( \frac{p}{n} \right)^2 nE (b_{p}^2) = \frac{1}{n} \sum_{i=1}^{n} E \left( (c_{r,p}\psi(r_{i,[p]})^2 \right) + o(1).
\]

4.2.5 On \(d_{i,p}\)

Recall the definition
\[
d_{i,p} = [\psi'(\gamma_{i,p}^*) - \psi'(r_{i,[p]})],
\]
where \(\gamma_{i,p}^* \in (r_{i,[p]}, r_{i,[p]} + \nu_i)\), with
\[
\nu_i = b_pX'_i \left[ (\Theta_p + \tau \text{Id})^{-1}u_p \right] = b_p\pi_i.
\]

We call \(\tilde{B}_n(i) = \sup_i |r_{i,[p]}| + \sup_i |\pi_i|\).

We have the following result.

**Proposition 4.4.** We have
\[
\sup_i |d_{i,p}| = O_L\left( \frac{\text{polyLog}(n)}{\sqrt{n}} \right).
\]

Hence,
\[
\sup_i |d_{i,p}| = O_L\left( \frac{\text{polyLog}(n)}{\sqrt{n}} \right).
\]

**Proof.** Recall the definition
\[
d_{i,p} = [\psi'(\gamma_{i,p}^*) - \psi'(r_{i,[p]})],
\]
where \(\gamma_{i,p}^* \in (r_{i,[p]}, r_{i,[p]} + \nu_i)\), with
\[
\nu_i = b_pX'_i \left[ (\Theta_p + \tau \text{Id})^{-1}u_p \right] = b_p\pi_i.
\]
Therefore,
\[
\pi_i = V_i'(\Theta_p + \tau \text{Id})^{-1}u_p - X_i(p).
\]

Recall that \(u_p = \frac{1}{n}V'D_{\psi'(r_{i,[p]})}X(p)\). According to Lemma B-2, we have
\[
\sup_i |V_i'(\Theta_p + \tau \text{Id})^{-1}u_p| = O_L\left( \frac{\text{polyLog}(n)}{c_1/2} \sup_i \|V_i'(\Theta_p + \tau \text{Id})^{-1}V_n V'D_{\psi'(r_{i,[p]})}\|^2 \right).
\]
Now,
\[
\|V_i'(\Theta_p + \tau \text{Id})^{-1}\frac{1}{n}V'D_{\psi'(r_{i,[p]})}\|^2 = \frac{1}{n}V_i'(\Theta_p + \tau \text{Id})^{-1}V_n V'D_{\psi'(r_{i,[p]})}^2 (\Theta_p + \tau \text{Id})^{-1}V_i.
\]
Since $\mathfrak{S}_p = \frac{V' D_{\psi'(r_{i,[p]})}}{n} V$, we have $\frac{V' D_{\psi'(r_{i,[p]})}}{n} V \leq |||D_{\psi'(r_{i,[p]})}|||_2 \mathfrak{S}_p$ and we conclude that

$$\frac{1}{n} V_i' (\mathfrak{S}_p + \tau \text{Id})^{-1} \frac{V' D_{\psi'(r_{i,[p]})}}{n} (\mathfrak{S}_p + \tau \text{Id})^{-1} V_i \leq |||V_i^2||| |||D_{\psi'(r_{i,[p]})}|||_2 = \frac{|||V_i^2|||}{n\tau} \sup_i \psi'(r_{i,[p]}) .$$

We also note that $\sup_i X_i(p) = O_{L_k}(\text{polyLog}(n)/\sqrt{n})$ and conclude that

$$\sup_i |\pi_i| = O_{L_k} \left( \frac{\text{polyLog}(n)}{c_n^{1/2}} \left[ 1 + \sup_i \psi'(r_{i,[p]}) \sup_i \| V_i \|^2 \right] \right) ,$$

$$= O_{L_k} \left( \frac{\text{polyLog}(n)}{c_n^{1/2}} \left[ \sup_i \psi'(r_{i,[p]}) \right] \right) .$$

Recalling that $b_p = O_{L_k}(n^{-1/2}\text{polyLog}(n))$, we finally see that

$$\sup_i \nu_i = O_{L_k} \left( \frac{\text{polyLog}(n)}{\sqrt{n}c_n^{1/2}} \left[ \sup_i \psi'(r_{i,[p]}) \right] \right) .$$

As before, we can control $\sup_i \psi'(r_{i,[p]})$ by using the work done in Proposition 3.2, since $r_{i,[p]}$ are the full residuals when we work with $p - 1$ predictors. The growth conditions we have imposed on $\psi'$ and $E_n$ therefore guarantee control of $[\sup_i \psi'(r_{i,[p]})]$ in $L_k$. Recall that $\tilde{B}_n(i) = \sup_i |r_{i,[p]}| + \sup_i |\pi_i|$. Now our assumptions guarantee that

$$\sup_i |d_{i,p}| = O_{L_k} \left( \frac{\text{polyLog}(n)}{\sqrt{n}c_n^{1/2}} L(\tilde{B}_n(i)) \left[ \psi'(-\tilde{B}_n(i)) \lor \psi'(\tilde{B}_n(i)) \right] \right) .$$

Proposition 3.2 then allows us to conclude, by giving us polyLog bounds on $\tilde{B}_n(i)$.

\[\square\]

### 4.3 Final conclusions

We finally have:

**Corollary 4.1.** Assuming that $1/c_n = O(\text{polyLog}(n))$, we have

$$||| \tilde{\beta} - \bar{b} ||| \leq \frac{1}{\tau} O_{L_k} \left( \frac{\text{polyLog}(n)}{n} \right) .$$

In particular,

$$\sqrt{n}(\tilde{\beta}_p - b_p) = O_{L_k}(\text{polyLog}(n)/\sqrt{n}) ,$$

$$\sup_i |X'_i(\tilde{\beta} - \bar{b})| = O_{L_k} \left( \frac{\text{polyLog}(n)}{\sqrt{n}} \right) ,$$

$$\sup_i |R_i - r_{i,[p]}| = O_{L_k} \left( \frac{\text{polyLog}(n)}{\sqrt{n}} \right) .$$

The corollary is just the aggregation of all of our results.

The last statement is the only one that might need an explanation. With the notations of the proof of Proposition 4.4, we have $R_i - r_{i,[p]} = X'_i(\bar{b} - \tilde{\beta}) + \nu_i$. The results in the proof of Proposition 4.4 as well as the bound on $||| \bar{b} - \tilde{\beta} |||$ give us the announced result.

We note that when the vectors $X_i$’s are i.i.d with i.i.d entries, all the coordinates play a symmetric role, so using the results of the previous corollary, Equation (33) and summing over all the coordinates, we have, asymptotically,

$$\frac{p}{n} E \left( |||\tilde{\beta}|||^2 \right) = \frac{p^2}{n} E \left( b_p^2 \right) + o(1) = \frac{1}{n} \sum_{i=1}^n E \left( (c_{r,p} \psi'(r_{i,[p]})^2 \right) + o(1) . \tag{34}$$
4.3.1 On $c_{\tau,p}$ and $c_{\tau}$

**Proposition 4.5.** We have

$$|c_{\tau} - c_{\tau,p}| = O_{L_k}(n^{-1/2} \text{polyLog}(n)).$$

*Proof.* Let us recall the notation

$$S = \frac{1}{n} \sum_{i=1}^{n} \psi'(R_i)X_iX_i'.$$

If we call $\Gamma = \frac{1}{n} \sum_{i=1}^{n} \psi'(R_i)V_iV_i'$ and $a = \frac{1}{n} \sum_{i=1}^{n} \psi'(R_i)X_i^2(p)$, we see that

$$S = \left( \begin{array}{cc} \Gamma & \nu \\ \nu & a. \end{array} \right).$$

According to Lemma C-2, we see that

$$|c_{\tau} - \frac{1}{n} \text{trace}((\Gamma + \tau \text{Id})^{-1})| \leq \frac{1}{n} \frac{1 + a/\tau}{\tau}.$$ 

It is clear that under our assumptions, $a = O_{L_k}(\text{polyLog}(n))$. It is also clear that

$$\sup_i |\psi'(R_i) - \psi'(r_i[p])| = O_{L_k}(\text{polyLog}(n)/\sqrt{n}).$$

Hence, using arguments similar to the ones we have used in the proof of Lemma 4.3, we see that

$$\left| \frac{1}{n} \text{trace}((\Gamma + \tau \text{Id})^{-1}) - \frac{1}{n} \text{trace}((\mathcal{S}_p + \tau \text{Id})^{-1}) \right| = O_{L_k}(\text{polyLog}(n)/\sqrt{n}).$$

Since $c_{\tau,p} = \frac{1}{n} \text{trace}((\mathcal{S}_p + \tau \text{Id})^{-1})$, the result we announced follows immediately. \hfill \square

In light of this result, we see that Equation (34) can be re-written

$$\frac{p}{n} \mathbb{E} \left( \|\tilde{\beta}\|^2 \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( (c_{\tau} \psi(R_i))^2 \right) + o(1) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( (c_{i} \psi(\text{prox}_{c_{\tau}}(\rho)(\tilde{r}_{i,(i)})^2)) \right) + o(1),$$

where we have used the remark we made after Lemma 4.4 that showed that $\sup_i |c_i - c_{\tau}| = O_{L_k}(n^{-1/2} \text{polyLog}(n))$. (See also Lemma A-2 and its proof where we compute the derivative of $\text{prox}_{c_{\tau}}(\rho)(x)$ with respect to $c$.)

So we finally have

$$\frac{p}{n} \mathbb{E} \left( \|\tilde{\beta}\|^2 \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( (c_{\tau} \psi(\text{prox}_{c_{\tau}}(\rho)(\tilde{r}_{i,(i)}))^2) \right) + o(1).$$

(35)

This will give us the second equation of our system. We also note that for any $x$, $c_{\tau} \psi(\text{prox}_{c_{\tau}}(\rho)(x)) = x - \text{prox}_{c_{\tau}}(\rho)(x) = \text{prox}_1((c_{\tau} \rho)^*)(x)$ - see e.g Moreau (1965). In Bean et al. (2013), we found that this formulation was nicer when further analytic manipulations where needed.

## 5 Putting things together

### 5.1 On the asymptotic distribution of $\tilde{r}_{i,(i)}$

We have the following lemma.

**Lemma 5.1.** As $n$ and $p$ tend to infinity, $\tilde{r}_{i,(i)}$ behaves like $\epsilon_i + \sqrt{\mathbb{E} \left( \|\tilde{\beta}\|^2 \right)} Z$, where $Z \sim \mathcal{N}(0,1)$, in the sense of weak convergence.

Furthermore, if $i \neq j$, $\tilde{r}_{i,(i)}$ and $\tilde{r}_{j,(j)}$ are asymptotically independent.
Proof. The only problem is of course showing that \( \hat{\beta}'(i)X_i \) is approximately \( \mathcal{N}(0, \mathbf{E}\left(\|\hat{\beta}\|^2\right)) \). Recall that \( \hat{\beta}'(i) \) is independent of \( X_i \). We assume without loss of generality that \( \|\hat{\beta}'(i)\| \) remains bounded away from 0 in our asymptotics. Note that if it is not the case \( \mathbf{E}\left(\|\hat{\beta}'(i)\|^2\right) \to 0 \) and so \( \hat{\beta}'(i)X_i \to 0 \), so the result holds.

Because \( \text{var}\left(\|\hat{\beta}\|^2\right) \to 0 \) and \( \text{var}\left(\|\hat{\beta}'(i)\|^2\right) \to 0 \), we see that

\[
\frac{\|\hat{\beta}'(i)\|}{\mathbf{E}\left(\|\hat{\beta}'(i)\|^2\right)} \to \text{in probability.}
\]

Provided that we can apply the Lindeberg-Feller theorem (see e.g Breiman (1992), p.186) conditional on \( X(i) \), we will have

\[
\frac{\hat{\beta}'(i)X_i}{\|\hat{\beta}'(i)\|} | X(i) \to \mathcal{N}(0,1).
\]

Because the limit is independent of \( \|\hat{\beta}'(i)\| \), we see that the result holds unconditionally, if we can apply the Lindeberg-Feller theorem with high \( X(i) \)-probability.

And because \( \|\hat{\beta}'(i)\|^2/\mathbf{E}\left(\|\hat{\beta}\|^2\right) \to 1 \) in probability, Slutsky’s lemma allows us to conclude that under these assumptions we have

\[
\sqrt{\mathbf{E}\left(\|\hat{\beta}\|^2\right)} \frac{\hat{\beta}'(i)X_i}{\|\hat{\beta}'(i)\|} \to \mathcal{N}(0,1).
\]

The only question we have to check is therefore to verify that we can apply the Lindeberg-Feller theorem conditionally on \( X(i) \), at least with high \( X(i) \)-probability. Recall that we have shown that

\[
\hat{\beta}_p = O_L\left(\frac{1}{\sqrt{nH}}\right).
\]

The same arguments we used apply also to \( (\hat{\beta}(i))_p \). So it is clear that

\[
\mathbf{E}\left(\|\hat{\beta}(i)\|^3\right) = O(n^{-3/2}).
\]

We conclude that \( \sum_{k=1}^n \|((\hat{\beta}(i))_k)^3 = O_L, X(i) (n^{-1/2}) \ll \|\hat{\beta}(i)\|^2 \) with high \( X(i) \)-probability, since we are in the setting where \( \|\hat{\beta}(i)\|^2 \) is bounded away from 0.

This shows the first part of the lemma.

For the second part, we use a leave-two-out approach, namely we use the approximation \( \hat{\beta}(i) = \hat{\beta}(i) - \hat{\beta}(i)(X_i) + O_{L}(\text{polyLog}(n)/(\sqrt{nH})) \) and similarly for \( \hat{\beta}(j) \). It is clear that \( \hat{\beta}(i) \) and \( \hat{\beta}(j) \) are asymptotically independent conditional on \( X(ij) \). But because their dependence on \( X(ij) \) is only through \( \|\hat{\beta}(ij)\| \), which is asymptotically deterministic, we see that \( \hat{\beta}(i) \) and \( \hat{\beta}(j) \) are asymptotically independent.

The lemma is shown. \( \Box \)

We are now in position to show that \( c_r = \frac{1}{n} \text{tr}(S + \tau \text{Id}_p)^{-1} \) is asymptotically deterministic and that the empirical distribution of the residuals \( \hat{R}_i \) is asymptotically non-random.

**Lemma 5.2.** Consider the random function

\[
g_n(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + x\psi'(\text{prox}_x(\rho)(\hat{\beta}(i)))},
\]

defined for \( x \geq 0 \). Let \( B > 0 \) be in \( \mathbb{R}_+ \). Call \( F_{\rho,B}(u) = (\psi(0) + L(|u|)|u| + BL(|u|)(|\psi(u)| + |\psi(-u)|)) \), where \( L(|u|) \) is the Lipschitz constant of \( \psi' \) on \([-|u|, |u|] \). We have, for any \( (x,y) \in \mathbb{R}_+^2 \), and \( x \leq B, y \leq B \)

\[
\sup_{(x,y)|x-y| \leq \eta, x \leq B, y \leq B} |g_n(x) - g_n(y)| \leq \eta \frac{1}{n} \sum_{i=1}^n F_{\rho,B}(\hat{\beta}(i)).
\]
In particular, we have
\[
P^* \left( \sup_{(x,y):|x-y|\leq \eta, x \leq B, y \leq B} |g_n(x) - g_n(y)| > \delta \right) \leq \frac{\eta}{\delta} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( F_{\rho,B}(\tilde{r}_{i,\epsilon}) \right).
\]

Hence, \( g_n \) is stochastically equicontinuous on \([0,B]\) for any \( B > 0 \) given, since under our assumptions \( \mathbb{E} \left( F_{\rho,B}(\tilde{r}_{i,\epsilon}) \right) \) is uniformly bounded in \( n \),

We used the notation \( P^* \) above to denote our probability and avoid a discussion of potential measure theoretic issues associated with taking a supremum over a non-countable collection of random variables. We refer the reader to e.g Pollard (1984) for more details on stochastic equicontinuity.

**Proof.** Let us consider the function
\[
h_u(x) = \frac{1}{1 + x \psi'(\operatorname{prox}_x(\rho)(u))} = \frac{\partial}{\partial u} \operatorname{prox}_x(\rho)(u).
\]

The last equality comes from Lemma A-3.

We have
\[
|h_u(x) - h_u(y)| \leq |x \psi'(\operatorname{prox}_x(\rho)(u)) - y \psi'(\operatorname{prox}_y(\rho)(u))| \wedge 1.
\]

Therefore,
\[
|h_u(x) - h_u(y)| \leq |x - y| \psi'(\operatorname{prox}_x(\rho)(u)) + y \psi'(\operatorname{prox}_x(\rho)(u)) - \psi'(\operatorname{prox}_y(\rho)(u))|.
\]

In particular, if \(|x - y| \leq \eta\), and \( x \vee y \leq B \)
\[
\sup_{y:|x-y|\leq \eta} |h_u(x) - h_u(y)| \leq \eta \psi'(\operatorname{prox}_x(\rho)(u)) + (x + \eta) \sup_y |\psi'(\operatorname{prox}_x(\rho)(u)) - \psi'(\operatorname{prox}_y(\rho)(u))|.
\]

Note

Under our assumptions, Lemma A-1 implies that, for \( y \geq 0 \), \( \sup_y |\operatorname{prox}_y(\rho)(u)| \leq |u| \). One of our assumptions is that \( \psi' \) is Lipschitz on any \([-t,t]\) with Lipschitz constant \( L(t) \). Therefore,
\[
|\psi'(\operatorname{prox}_x(\rho)(u)) - \psi'(\operatorname{prox}_y(\rho)(u))| \leq L(|u|)|\operatorname{prox}_x(\rho)(u) - \operatorname{prox}_y(\rho)(u)|.
\]

We recall that, according to Lemma A-2,
\[
\frac{\partial}{\partial x} \operatorname{prox}_x(\rho)(u) = -\frac{\psi(\operatorname{prox}_x(\rho)(u))}{1 + c \psi'(\operatorname{prox}_x(\rho)(u))}.
\]

Furthermore, since \( \psi \) is non-decreasing and changes sign at 0, we also have
\[
\sup_x |\frac{\partial}{\partial x} \operatorname{prox}_x(\rho)(u)| \leq |\psi(u)| \vee |\psi(-u)|.
\]

This naturally gives us a bound on the Lipschitz constant of the function \( x \rightarrow \operatorname{prox}_x(\rho)(u) \). We finally conclude that
\[
|\psi'(\operatorname{prox}_x(\rho)(u)) - \psi'(\operatorname{prox}_y(\rho)(u))| \leq L(|u|)\left[ |\psi(u)| \vee |\psi(-u)| \right]|x - y|.
\]

We therefore have
\[
\sup_{y:|x-y|\leq \eta} |h_u(x) - h_u(y)| \leq \eta \psi'(\operatorname{prox}_x(\rho)(u)) + (x + \eta)L(|u|)\left[ |\psi(u)| \vee |\psi(-u)| \right] \eta.
\]

Of course, \( \psi'(\operatorname{prox}_x(\rho)(u)) \leq \psi'(0) + L(|u|)|u| \), by using again the fact that \( |\operatorname{prox}_x(\rho)(u)| \leq |u| \), the fact that \( \operatorname{prox}_x(\rho)(0) = 0 \) and the fact that the Lipschitz constant of \( \psi' \) on \([-|\operatorname{prox}_x(\rho)(u)|,|\operatorname{prox}_x(\rho)(u)|] \) is less than \( L(u) \).
Therefore, if \(x + \eta \leq B\), we have

\[
\sup_{y : |x - y| \leq \eta} |h_u(x) - h_u(y)| \leq \eta \left( [\psi'(0) + L(|u|)|u|] + BL(|u|)[|\psi(u)| + |\psi(-u)|] \right).
\]

Therefore, we also have

\[
\sup_{(x,y): |x - y| \leq \eta, x \vee y \leq B} |h_u(x) - h_u(y)| \leq \eta \left( [\psi'(0) + L(|u|)|u|] + BL(|u|)[|\psi(u)| + |\psi(-u)|] \right).
\]

We denote by \(F_{\rho,B}(u) = ([\psi'(0) + L(|u|)|u|] + BL(|u|)[|\psi(u)| + |\psi(-u)|])\). This analysis shows that for \(x\) given, if \(|x - y| \leq \eta\) and \(x \vee y \leq B\), we have

\[
\sup_{(x,y): |x - y| \leq \eta, x \leq B, y \leq B} |g_n(x) - g_n(y)| \leq \eta \frac{1}{n} \sum_{i=1}^{n} F_{\rho,B}(\tilde{r}_{i,(i)}).
\]

We can now take expectations, and get the result in \(L_1\) provided \(E \left( F_{\rho,B}(\tilde{r}_{i,(i)}) \right)\) is finite and remains bounded in \(n\). However, this holds since \(F_{\rho,B}\) grows at most polynomially at \(\infty\), and \(\epsilon_i, \|\hat{\beta}_{(i)}\|\) and \(X_i\) have infinitely many moments, by Assumptions \(O4, F2\) and our work on \(\|\hat{\beta}\|\).

We have established stochastic equicontinuity of \(g_n(x)\) on \([0, B]\).

\[\square\]

Lemma 5.3. Let us call \(G_n(x) = E (g_n(x))\). For any given \(x_0 \leq B\),

\[g_n(x_0) - G_n(x_0) = o_{L_2}(1).\]

Under our assumptions, we also have

\[E^* \left( \sup_{0 \leq x \leq B} |g_n(x) - G_n(x)| \right) \to 0.\]

Proof. Asymptotic pairwise independence of \(\tilde{r}_{i,(i)}\) implies that

\[\operatorname{var}(g_n(x_0)) \to 0\]

and therefore gives the first result.

Let us pick \(\epsilon > 0\). By the stochastic equicontinuity of \(g_n\) and our \(L_1\) bound, we can find \(x_1, \ldots, x_K\), independent of \(n\), such that for all \(x \in [0, B]\), there exists \(l\) such that, when \(n\) is large enough,

\[E (|g_n(x) - g_n(x_l)|) \leq \epsilon.\]

Note that

\[|g_n(x) - G_n(x)| \leq |g_n(x) - g_n(x_l)| + |g_n(x_l) - G_n(x_l)| + |G_n(x_l) - G_n(x)|.\]

We immediately get

\[E^* \left( \sup_{0 \leq x \leq B} |g_n(x) - G_n(x)| \right) \leq 2\epsilon + E \left( \sup_{1 \leq l \leq K} |g_n(x_l) - G_n(x_l)| \right).\]

Because \(K\) is finite, the fact that for all \(l\), \(|g_n(x_l) - G_n(x_l)| \to 0\) in \(L_2\) implies that \(\sup_{1 \leq l \leq K} |g_n(x_l) - G_n(x_l)| \to 0\) in \(L_2\). In particular, if \(n\) is sufficiently large,

\[E \left( \sup_{1 \leq l \leq K} |g_n(x_l) - G_n(x_l)| \right) \leq \epsilon.\]

The lemma is shown.  \[\square\]
Lemma 5.4. Call \( c_r = \frac{1}{n} \text{trace}((S + \tau \text{Id}_p)^{-1}) \). Call as before

\[
g_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + x\psi'(\text{prox}_x(p)(\tilde{r}_{i,(i)}))}
\]

Then \( c_r \) is a near solution of

\[
\frac{p}{n} - \tau x - 1 + g_n(x) = 0 ,
\]

i.e \( \frac{p}{n} - \tau c_r - 1 + g_n(c_r) = o_{L_k}(1) \).

Asymptotically, near solutions of

\[
\delta_n(x) \triangleq \frac{p}{n} - \tau x - 1 + g_n(x) = 0 ,
\]

are close to solutions of

\[
\Delta_n(x) = \frac{p}{n} - \tau x - 1 + E(g_n)(x) = 0 .
\]

More precisely, call \( T_{n,\epsilon} = \{ x : |\Delta_n(x)| < \epsilon \} \). Note that \( T_{n,\epsilon} \subseteq (0, p/(n\tau) + \epsilon/\tau) \). For any given \( \epsilon \), as \( n \to \infty \), near solutions of \( \delta_n(x_n) = 0 \) belong to \( T_{n,\epsilon} \) with high-probability.

Our assumptions concerning the distribution of \( \epsilon_i \)'s, specifically F1, guarantee that as \( n \to \infty \), there is a unique solution to \( \Delta_n(x) = 0 \).

Hence \( c_r \) is asymptotically deterministic.

Proof. Let \( \delta_n \) be the function

\[
\delta_n(x) = \frac{p}{n} - \tau x - 1 + g_n(x) ,
\]

and \( \Delta_n(x) = E(\Delta_n(x)) \). Call \( x_n \) the solution \( \delta_n(x_n) = 0 \) and \( x_{n,0} \) the solution of \( \Delta_n(x_{n,0}) = 0 \). Since \( 0 \leq g_n \leq 1 \), we see that \( x_n \leq p/(n\tau) \), for otherwise, \( \delta_n(x) < 0 \). The same argument shows that if \( x > (p/n + \epsilon)/\tau \), \( \Delta_n(x) < -\epsilon \) and \( x \notin T_{n,\epsilon} \). Similarly, near solutions of \( \delta_n(x) = 0 \) must be less or equal to \( (p/n + \epsilon)/\tau \).

• Proof of the fact that \( c_r \) is such that \( \delta_n(c_r) = o(1) \)

An important remark is that \( c_r \) is a near solution of \( \delta_n(x) = 0 \). This follows most clearly for arguments we have developed for \( c_{r,p} \) so we start by giving details through arguments for this random variable. Recall that in the notation of Lemma 4.2, we had

\[
p/n - \tau c_{r,p} = \frac{1}{n} \text{trace}(\text{Id}_n - M) .
\]

Now, according to Equation (30),

\[
\frac{1}{n} \text{trace}(\text{Id}_n - M) = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \psi'(r_{i,[p]})} \frac{1}{n} V'_i(V_i(\mathcal{S}_p(i) + \tau \text{Id})^{-1}) .
\]

According to Lemmas 4.3 and 4.4, we have

\[
\sup \left| \frac{1}{n} V'_i(\mathcal{S}_p(i) + \tau \text{Id})^{-1} V_i - c_{r,p} \right| = O_{L_k}(\text{polyLog}(n)n^{-1/2}).
\]

Of course, when \( x \geq 0 \) and \( y \geq 0 \), \( |1/(1 + x) - 1/(1 + y)| \leq |x - y| \). Using our bounds on \( \psi'(r_{i,[p]}) \), we easily see that,

\[
p/n - \tau c_{r,p} - 1 + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + c_{r,p}\psi'(r_{i,[p]})} = O_{L_k}(n^{-1/2}\text{polyLog}(n)) .
\]

Exactly the same computations can be made with \( c_r \), so we have established that

\[
p/n - \tau c_r - 1 + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + c_r\psi'(R_i)} = O_{L_k}(n^{-1/2}\text{polyLog}(n)) .
\]
Now we have seen in Corollary 3.1 that
\[
\sup_i |R_i - \text{prox}_{c_i}(\tilde{r}_i, i)| = O_{L_k}(n^{-1/2}\text{polyLog}(n)) .
\]
Through our assumptions on \(\psi'\), this of course implies that
\[
\sup_i |\psi'(R_i) - \psi'(\text{prox}_{c_i}(\rho)(\tilde{r}_i, i))| = O_{L_k}(n^{-1/2}\text{polyLog}(n))
\]
We have furthermore noted that \(\sup_i |c_i - c_r| = O_{L_k}(n^{-1/2}\text{polyLog}(n))\) after Lemma 4.4. Using the proof of Lemma A-2, this implies that
\[
\sup_i |\psi'[\text{prox}_{c_i}(\rho)(\tilde{r}_i, i)] - \psi'[\text{prox}_{c_r}(\rho)(\tilde{r}_i, i)]| = O_{L_k}(n^{-1/2}\text{polyLog}(n))
\]
and therefore
\[
\sup_i |\psi'[R_i] - \psi'[\text{prox}_{c_r}(\rho)(\tilde{r}_i, i)]| = O_{L_k}(n^{-1/2}\text{polyLog}(n))
\]
So we have established that \(\delta_n(c_r) = O_{L_k}(n^{-1/2}\text{polyLog}(n))\).

\[\Box\]

Final details

Note that for any given \(x\), \(\delta_n(x) - \Delta_n(x) = o_P(1)\) by using Lemma 5.3. In our case, with the notation of this lemma, \(B = \frac{\rho}{(n\tau) + \eta/\tau}\), for \(\eta > 0\) given.

This implies that, for any given \(\epsilon > 0\)
\[
\sup_{x \in (0, \frac{\rho}{n\tau} + \eta/\tau)} |\delta_n(x) - \Delta_n(x)| < \epsilon ,
\]
with high-probability when \(n\) is large. Therefore, for any \(\epsilon > 0\)
\[
|\Delta_n(x_n)| \leq \epsilon
\]
with high-probability. This exactly means that \(x_n \in T_{n, \epsilon}\) with high-probability. The same argument applies for near solutions of \(\delta_n(x) = 0\), which, for any \(\epsilon > 0\) must belong to \(T_{n, \epsilon}\) as \(n \to \infty\) with high-probability. Of course, there is nothing random about \(T_{n, \epsilon}\) which is a deterministic set. Note that \(T_{n, \epsilon}\) is compact because it is bounded and closed, using the fact that \(g_n\) and \(E(g_n)\) are continuous.

If \(T_{n, 0}\) were reduced to a single point, we would have established the asymptotically deterministic character of \(c_r\).

Given our work concerning the limiting behavior of \(\tilde{r}_{i,(i)}\) and our assumptions about \(\epsilon_i\)'s, we see that Lemma C-1 applies to \(\lim_{n \to \infty} \Delta_n(x)\) under assumption \(F1\). Therefore, as \(n \to \infty\), \(T_{n, 0}\) is reduced to a point and \(c_r\) is asymptotically non-random.

As we had noted in El Karoui et al. (2012),
\[
\frac{\partial}{\partial t} \text{prox}_c(\rho)(t) = \text{prox}_c(\rho)'(t) = \frac{1}{1 + c\psi'(\text{prox}_c(\rho)(t))} .
\]
So \(\Delta_n\) can be interpreted as
\[
\Delta_n(x) = \frac{p}{n} - \tau x - 1 + \frac{1}{n} \sum_{i=1}^n E(\text{prox}_x(\rho)'(\tilde{r}_i, i)) .
\]
The fact that \(c_r\) is asymptotically arbitrarily close to the root of \(\Delta_n(x) = 0\) gives us the first equation in the system appearing in Theorem 1.1.

The second equation of the system comes from Equation (35). Theorem 1.1 is shown.
6 From the ridge-regularized to the un-regularized problem

Our original motivation in El Karoui et al. (2012) was to study the unpenalized problem, namely \( \hat{\beta} \) was defined as

\[
\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho(Y_i - X_i' \beta) .
\]

We now explain how we can derive the system in the unpenalized case from the one we have obtained in the penalized case, when \( p/n < 1 \).

We first note that when \( p < n \), and for instance the \( X_i \)'s have a continuous distribution, if \( Y_i = X_i' \beta_0 + \epsilon_i \),

\[
\hat{\beta} - \beta_0 = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i - X_i' \beta) .
\]

So to understand the error we make when using regression M-estimates, i.e the vector \( \hat{\beta} - \beta_0 \), it is enough to study the properties of our estimator in the null case where \( \beta_0 = 0 \). Of course, we have previously studied the penalized version of this particular problem.

We have the following theorem.

**Theorem 6.1.** Suppose our assumptions O1-O7, P1 and F1-F2 hold. Suppose further that \( \lim sup p/n < 1 \). Call

\[
\hat{\beta} = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i - X_i' \beta) , \quad \text{and} \quad \hat{\beta}_\tau = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i - X_i' \beta) + \tau \| \beta \|^2 / 2 .
\]

If \( \rho \) is strongly convex,

\[
\lim_{\tau \to 0} \| \hat{\beta}_\tau - \hat{\beta} \| = o_P(1) .
\]

Hence \( \| \hat{\beta} \| \) is asymptotically deterministic and can be computed via

\[
\| \hat{\beta} \| = \lim_{\tau \to 0} \| \hat{\beta}_\tau \| .
\]

**Proof.** We call \( f_\tau(\beta) = -\frac{1}{n} \sum_{i=1}^{n} X_i \psi(\epsilon_i - X_i' \beta) + \tau \beta \) and \( f(\beta) = f_0(\beta) \).

Since by definition, \( \hat{\beta} \) is such that

\[
\sum X_i \psi(\epsilon_i - X_i' \hat{\beta}) = 0 ,
\]

we see that \( f_\tau(\hat{\beta}) = \tau \hat{\beta} \). By a similar token, we see that \( f_0(\hat{\beta}_\tau) = -\tau \hat{\beta}_\tau \). If \( \rho \) is strongly convex with modulus of convexity \( C \), we see, using Proposition 2.1 that, by working with \( \nabla f_\tau \), we get

\[
\| \hat{\beta}_\tau - \hat{\beta} \| \leq \frac{\tau}{C \lambda_{\min}(\Sigma)} \| \hat{\beta} \| ,
\]

and by working with \( \nabla f_0 \) - along the same lines as in the proof of Proposition 2.1 - we get

\[
\| \hat{\beta}_\tau - \hat{\beta} \| \leq \frac{\tau}{C \lambda_{\min}(\Sigma)} \| \hat{\beta}_\tau \| .
\]

Recall that we showed in Equation (10) that

\[
\| \hat{\beta}_\tau \| \leq \sqrt{\frac{2}{\tau}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i)} .
\]
This shows that
\[ \| \hat{\beta}_\tau - \tilde{\beta} \| \leq \frac{\sqrt{2}\tau}{C\lambda_{\min}(\hat{\Sigma})} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i)} . \]

Under our assumptions, \( \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i) = O_P(1) \). Under the assumptions that, for instance, the entries of \( X_i \)'s are i.i.d with 4+\( \epsilon \) moments (which is always the case under our assumptions), it is well known that \( \lambda_{\min}(\hat{\Sigma}) \to (1 - \sqrt{\frac{2}{n}})^2 \) in probability and a.s (Bai (1999)).

We conclude that \( \| \hat{\beta}_\tau - \tilde{\beta} \| \to 0 \) in probability as \( \tau \to 0 \) under our assumptions.

Under for instance Gaussian design assumptions (i.e \( X_i \)'s have distribution \( \mathcal{N}(0, \text{Id}_p) \)), it is possible to bound \( E \left( \frac{1}{\lambda_{\min}(\hat{\Sigma})} \right) \) using essentially results in Silverstein (1985) as well as elementary but non-trivial linear algebra (see the appendix of Halko et al. (2011) for instance). This would give an approximation in \( L_2 \), provided the random variable \( \rho(\epsilon_i) \) has enough moment.

It would be possible with quite a bit of extra work to dispense with the assumption of strong convexity and move for instance to strict convexity (see Hiriart-Urruty and Lemaréchal (2001) if needed). We refer to the discussion after the proof of Proposition 2.1 for more details about this issue.

We note that convergence in probability of \( \hat{\beta} \) is enough for our confidence interval statements from Bean et al. (2013) (details in the supplementary material of that paper) to go through.

6.1 Other extensions

Moving from random vectors \( X_i \)'s like the ones we have studied to vectors of the form \( \tilde{X}_i = \lambda_i X_i \), where \( \lambda_i \) is a random variable (i.e a scalar) independent of \( X_i \) does not offer any new conceptual difficulties. Indeed our heuristic work in El Karoui et al. (2013) handled - heuristically of course - that case, so the arguments we gave here would be easy to modify. This extended class of models - which is akin to elliptical distributions in multivariate statistics (see Anderson (2003)) - is interesting because it includes distributions that do not share the geometric properties that “concentrated” random vectors share. We do not solve the elliptical problem here in complete details because of the extra notational burden involved.

Another easy extension of the work presented here is to study the weighted regression case, i.e for weights \( \{w_i\}_{i=1}^{n} \), \( \tilde{\beta} \) is defined as
\[ \tilde{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} w_i \rho(\epsilon_i - X'_i\beta) + \frac{\tau}{2} \|\beta\|^2 . \]

Once again, only minor modifications are needed to our proof - the heuristic we proposed easily handled this. More generally, working on the problem of understanding
\[ \hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_i(\epsilon_i - X'_i\beta) + \frac{\tau}{2} \|\beta\|^2 , \]

where \( \rho_i \) are potentially different functions and \( X_i \)'s are “elliptical” (as defined above) seems to be within relatively easy reach of the method developed and presented here.

Finally, we see that when \( Y_i = X'_i\beta_0 + \epsilon_i \),
\[ \hat{\beta}_\tau - \beta_0 = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho(\epsilon_i - X'_i\beta) + \frac{\tau}{2} \|\beta + \beta_0\|^2 . \]

This problem - a mild variant of the one we have studied here - should be amenable to analysis with the method we used here.
APPENDIX

A Notes on the prox

Lemma A-1. Let $\rho$ be such that $\psi$ changes sign at 0, i.e $\text{sign}(\psi(x)) = \text{sign}(x)$ for $x \neq 0$. Then,

$$\text{prox}_\rho(\rho)(0) = 0.$$ 

Furthermore,

$$|\psi(\text{prox}_\rho(\rho)(x))| \leq |\psi(x)|.$$

Also,

$$|\psi(\text{prox}_\rho(\rho)(x))| \leq |x|/c.$$

Proof. By definition, we have

$$\text{prox}_\rho(\rho)(x) + c\psi(\text{prox}_\rho(\rho)(x)) = x.$$ 

Therefore,

$$\text{prox}_\rho(\rho)(0) = -c\psi(\text{prox}_\rho(\rho)(0)).$$

Hence, if we call $y = \text{prox}_\rho(\rho)(0)$, we have $\text{sign}(y) = -\text{sign}(\psi(y))$. The assumptions on $\psi$ therefore guarantee that $y = 0$, for otherwise we would have a contradiction.

We also note that $\text{sign}(\text{prox}_\rho(\rho)(x)) = \text{sign}(x)$, since $\text{sign}[\psi(\text{prox}_\rho(\rho)(x))] = \text{sign}(\text{prox}_\rho(\rho)(x))$.

Using contractivity of the prox, we see that

$$|\text{prox}_\rho(\rho)(x)| = |\text{prox}_\rho(\rho)(x) - \text{prox}_\rho(\rho)(0)| \leq |x|.$$

Since $\rho$ is convex, we see that $\psi$ is increasing. If $x > 0$, $\text{prox}_\rho(\rho)(x) > 0$, and therefore,

$$0 \leq \psi(\text{prox}_\rho(\rho)(x)) \leq \psi(x).$$

Similarly, if $x < 0$, $x \leq \text{prox}_\rho(\rho)(x) < 0$ and therefore, $\psi(x) \leq \psi(\text{prox}_\rho(\rho)(x)) < 0$. The second statement of the lemma is shown.

The last statement is a simple consequence of the fact that $c\psi(\text{prox}_\rho(\rho)(x)) = x - \text{prox}_\rho(\rho)(x)$, from which it immediately follows that

$$|\psi(\text{prox}_\rho(\rho)(x))| \leq \frac{|x|}{c}.$$

\[\square\]

We will also need the following simple result.

Lemma A-2. Suppose $x$ is a real and $\rho$ is twice differentiable and convex. Then, for $c > 0$,

$$\frac{\partial}{\partial c} \rho(\text{prox}_\rho(\rho)(x)) = -\frac{\psi^2(\text{prox}_\rho(\rho)(x))}{1 + c\psi'(\text{prox}_\rho(\rho)(x))}.$$ 

In partial, at $x$ given $c \rightarrow \rho(\text{prox}_\rho(\rho)(x))$ is decreasing in $c$.

Proof. Using the fact that

$$\text{prox}_\rho(\rho)(x) + c\psi(\text{prox}_\rho(\rho)(x)) = x,$$

we easily see that

$$\frac{\partial}{\partial c} \text{prox}_\rho(\rho)(x) = -\frac{\psi(\text{prox}_\rho(\rho)(x))}{1 + c\psi'(\text{prox}_\rho(\rho)(x))}.$$ 

It then follows immediately that

$$\frac{\partial}{\partial c} \rho(\text{prox}_\rho(\rho)(x)) = -\frac{\psi^2(\text{prox}_\rho(\rho)(x))}{1 + c\psi'(\text{prox}_\rho(\rho)(x))}.$$ 

The denominator is positive, from which we immediately deduce that $c \rightarrow \rho(\text{prox}_\rho(\rho)(x))$ is decreasing in $c.$

\[\square\]
We also make the following observation, which was essential to finding the system of equations in El Karoui et al. (2013)

**Lemma A-3.** We have
\[
\frac{\partial}{\partial x} \text{prox}_c(\rho)(x) = \frac{1}{1 + c \psi'(\text{prox}_c(\rho)(x))}.
\]

A proof of this fact follows immediately from the well-known representation (see Moreau (1965))
\[
\text{prox}_c(\rho)(x) = (\text{Id} + c \psi)^{-1}(x).
\]

We finally make notice of the following simple fact.

**Lemma A-4.** The function \(x \to [x \psi(\text{prox}_c(\rho)(u))^2](\text{defined on } \mathbb{R}_+\) is increasing, for any \(u\).

**Proof.** Let us consider \(f_u(x) = x \psi(\text{prox}_c(\rho)(u))\). Note that \(f_u(x) = u - \text{prox}_c(\rho)(u)\). So
\[
\frac{\partial}{\partial x} f_u(x) = \frac{\psi(\text{prox}_c(\rho)(u))}{1 + x \psi'(\text{prox}_c(\rho)(u))}.
\]

Hence,
\[
\frac{\partial}{\partial x} f_u^2(x) = 2x \frac{\psi^2(\text{prox}_c(\rho)(u))}{1 + x \psi'(\text{prox}_c(\rho)(u))} \geq 0
\]
since \(x \geq 0\) and \(\psi' \geq 0\).

\[\square\]

**B **On convex Lipschitz functions of random variables

In this section, we provide a brief reminder concerning convex Lipschitz functions of random variables.

**Lemma B-1.** Suppose that \(\{X_i\} \in \mathbb{R}^p\) satisfy the following concentration property: \(\exists C_n, c_n\) such that for any \(F_i\), a convex, \(1\)-Lipschitz function of \(X_i\),
\[
P(|F_i(X_i) - m_i| \geq t) \leq C_n \exp(-c_n t^2)\text{,}
\]
where \(m_i\) is deterministic. Then if \(F_n = \sup_i |F_i(X_i) - m_i|\), we have, even when the \(X_i\)’s are dependent:

1. if \(u_n = \sqrt{\log(n)/c_n}\), \(E(F_n) \leq u_n + C_n/(2 \sqrt{c_n \log n}) = \frac{\sqrt{\log n}}{\sqrt{c_n}} (1 + C_n/(2 \log n))\). Similar bounds hold in \(L_k\) for any finite given \(k\).

2. when \(C_n \leq C\), where \(C\) is independent of \(n\), there exists \(K\), independent of \(n\) such that \(F_n/u_n \leq K\) with overwhelming probability, i.e probability asymptotically smaller than any power of \(1/n\).

In particular,
\[
F_n = O(polyLog(n)/\sqrt{c_n})
\]
in probability and any \(L_k\), \(k\) fixed and given.

We note that similar techniques can be used to extend the result to situations where we have \(P(|F_i(X_i) - m_i| \geq t) \leq C_n \exp(-c_n t^\alpha)\), with \(\alpha \neq 2\). Of course, the order of magnitudes of the bounds then change.

**Proof.** Clearly,
\[
P(F_n \geq t) \leq 1 \wedge n C_n \exp(-c_n t^2)\text{.}
\]

Hence, for any \(u \geq 0\),
\[
E(F_n^k) \leq u^k + \int_0^\infty k t^{k-1} n C_n \exp(-c_n t^2) dt.
\]

Since \(E(F_n^k) = \int_0^\infty k t^{k-1} P(F_n \geq t) dt\). Standard computations show that when \(uc_n^2\) is large, and \(k \geq 1\),
\[
\int_0^\infty t^{k-1} \exp(-c_n t^2) dt \sim O\left(\frac{u^k}{2c_n u^2 \exp(-c_n u^2)}\right).
\]
So we see that in that case, for a constant \( k \) that depends only on \( n \),
\[
\mathbb{E} \left( F_n^k \right) \leq u^k \left( 1 + K_k n C_n \exp \left( -c_n u^2 \right) \right).
\]
Taking \( u_n = \sqrt{\log n / c_n} \), we see that
\[
\mathbb{E} \left( F_n^k \right) \leq u_n^k \left( 1 + K_k \frac{C_n}{\log n} \right).
\]
We conclude that when \( C_n / \log n \) remains bounded, \( \mathbb{E} \left( F_n^k \right) / u_n^k \) remains bounded. In the case \( k = 1 \), it is easy to see that \( K_k = 1/2 \) and we do not require \( \sqrt{c_n u} \) to be large for our arguments to go through. This gives the bound announced in the Lemma.

The probabilistic bound comes simply from the fact that
\[
P(\mathcal{F}_n \geq t u_n) \leq C_n \exp \left( -\log_n (t^2 - 1) \right) \leq C_n \exp \left( -\log_n (t^2 - 1) \right).
\]
Hence,
\[
P \left( \mathcal{F}_n / u_n \geq K \right) \leq n^{-d},
\]
for any given \( d \) if \( K \) is large enough. If we allow \( K \) to grow like a power of \( \log n \), we also see that the right hand side above can be made even smaller.

\[ \square \]

We recall that we denote by \( X(i) = \{ X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n \} \). If \( I \) is a subset of \( \{ 1, \ldots, n \} \) of size \( n - 1 \), we call \( X_I \) the collection of the corresponding \( X_i \) random variables. We call \( X_{\bar{I}} \) the remaining random variable.

**Lemma B-2.** Suppose \( X_i \)'s are independent and satisfy the concentration inequalities as above. Consider the situation where \( F_{I_k} \) is a convex Lipschitz function of 1 variable, depending on \( X_{I_k} \) only and call \( L_{I_k} \) its Lipschitz constant (at \( X_{I_k} \) given)- which is assumed to be random. Call \( m_{F_{I_k}} = m_{F_{I_k}(X_{I_k}^\circ)} \) \( X_{I_k} \), \( m \) being the mean or the median. As before, call \( \mathcal{F}_n = \sup_{j=1,\ldots,n} \left| F_{I_j}(X_{I_j}^\circ) - m_{F_{I_j}} \right| \) Then \( \mathcal{F}_n = O\left( \sqrt{\log n / c_n \sup_{1 \leq j \leq n} L_{I_j}} \right) \) in probability and in \( \sqrt{L_{2k}} \), i.e there exists \( K > 0 \), independent of \( n \), such that
\[
\mathbb{E} \left( \mathcal{F}_n^k \right) \leq K \left( \sqrt{\log n / c_n} \right)^k \sqrt{\mathbb{E} \left( \sup_{1 \leq j \leq n} L_{I_j}^{2k} \right)}.
\]

Hence, \( \mathcal{F}_n \) is \( \text{polyLog}(n)/c_n^{1/2} \sup_{1 \leq j \leq n} L_{I_j} \) in \( \sqrt{L_{2k}} \).

**Proof.** We call \( \mathcal{L} = \sup_i L_i \). By Holder’s inequality, we have
\[
\mathbb{E} \left( \mathcal{F}_n^k \right) = \mathbb{E} \left( \left( \mathcal{F}_n / \mathcal{L} \right)^k / \mathcal{L}^k \right) \leq \sqrt{\mathbb{E} \left( \mathcal{F}_n^{2k} / \mathcal{L}^{2k} \right)} \sqrt{\mathbb{E} \left( \mathcal{L}^{2k} \right)}.
\]
Let us call \( \bar{F}_n = \mathcal{F}_n / \mathcal{L} \). As before,
\[
\mathbb{E} \left( \bar{F}_n^k \right) \leq u^k + \sum_{j=1}^n \int_u^\infty kx^{k-1} P(|F_{I_j}(X_{I_j}^\circ) - m_{F_{I_j}}| \geq \mathcal{L}x),
\]
\[
\leq u^k + \sum_{j=1}^n \int_u^\infty kx^{k-1} P(|F_{I_j}(X_{I_j}^\circ) - m_{F_{I_j}}| \geq L_{I_j}x),
\]
\[
= u^k + \sum_{j=1}^n \int_u^\infty kx^{k-1} \mathbb{E} \left( P(|F_{I_j}(X_{I_j}^\circ) - m_{F_{I_j}}| \geq L_{I_j}x|X_{I_j}) \right).
\]
Now our assumptions guarantee that

\[ P(|F_k(X_{I_k}^j) - m_{F_k}| \geq L_k x |X_{I_k}^j|) \leq C_n \exp(-c_n x^2), \]

since \( F_k/L_k \) is 1-Lipschitz (and independent of \( X_{I_k}^j \)). We conclude that

\[ E \left( \tilde{F}_n^k \right) \leq u^k + nC_n \int_u^\infty kx^{k-1} \exp(-c_n x^2). \]

This is exactly the same situation as we had before and the conclusion follows.

\[ \square \]

**Lemma B-3.** Suppose the assumptions of the previous Lemma are satisfied. Consider \( Q_{I_j} = \frac{1}{n} X_{I_j}^I M_{I_j} X_{I_j}^I \), where \( M \) is a random symmetric matrix depending only on \( X_{I_j} \) whose largest eigenvalue is \( \lambda_{\text{max}, I_j} \). Assume that \( \text{cov} (X_i) = \Id_p \) and \( nc_n \to \infty \). Then, we have in \( L_k \),

\[ \sup_{1 \leq j \leq n} \left| Q_{I_j} - \frac{1}{n}\text{trace} (M_{I_j}) \right| = O_k \left( \frac{\text{polyLog}(n)}{\sqrt{nc_n}} \sup_{1 \leq j \leq n} \lambda_{\text{max}, I_j} \right). \]

The same bound holds when considering a single \( Q_{I_j} \) without the \( \text{polyLog}(n) \) term.

**Proof.** Lemma B-2 applies to \( \sqrt{Q_{I_j}} \) and \( \sup_{1 \leq j \leq n} |\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}| \). The corresponding Lipschitz constant if of course \( \sqrt{\lambda_{\text{max}, I_j}} / n \).

So all we need to do is show that we can go from this control to the control of \( \sup_{1 \leq j \leq n} |Q_{I_j} - \frac{1}{n}\text{trace} (M_{I_j})| \).

Of course,

\[ |Q_{I_j} - \frac{1}{n}\text{trace} (M_{I_j})| \leq |Q_{I_j} - m_{\sqrt{Q_{I_j}}} + m_{\sqrt{Q_{I_j}}} - \frac{1}{n}\text{trace} (M_{I_j})|. \]

The idea from there is simply to use the fact that for \( a \) and \( b \) non-negative, \((a+b)^k \leq 2^{k-1}(a^k + b^k)\). Using Proposition 1.9 in Ledoux (2001), we know that

\[ |m_{\sqrt{Q_{I_j}}} - \frac{1}{n}\text{trace} (M_{I_j})| \leq \frac{C_n}{nc_n} \lambda_{\text{max}}(M_{I_j}). \]

On the other hand,

\[ |Q_{I_j} - m_{\sqrt{Q_{I_j}}} + m_{\sqrt{Q_{I_j}}} - \frac{1}{n}\text{trace} (M_{I_j})| \leq \left| \sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}} \right|^2 + 2 \left| \sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}} \right| \lambda_{\text{max}}(M_{I_j}). \]

since \( m_{\sqrt{Q_{I_j}}} \geq 0 \).

Therefore,

\[ \sup_{1 \leq j \leq n} |Q_{I_j} - m_{\sqrt{Q_{I_j}}} + m_{\sqrt{Q_{I_j}}} - \frac{1}{n}\text{trace} (M_{I_j})| \leq \sup_{1 \leq j \leq n} \left| \sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}} \right|^2 + 2 \left| \sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}} \right| \sup_{1 \leq j \leq n} m_{\sqrt{Q_{I_j}}} \lambda_{\text{max}}(M_{I_j}). \]

Lemma B-2 gives us control of \( \sup_{1 \leq j \leq n} |\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}| \) in \( L_{2k} \) and therefore control of \( \sup_{1 \leq j \leq n} |\sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}}| \) in \( L_{2k} \) with a bound of the form \( \frac{\text{polyLog}(n)}{\sqrt{nc_n}} \sup_{1 \leq j \leq n} \lambda_{\text{max}}(M_{I_j}). \)

The result will therefore be shown provided we control \( \left[ \sup_{1 \leq j \leq n} \left| \sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}} \right| \right] \left[ \sup_{1 \leq j \leq n} m_{\sqrt{Q_{I_j}}} \right] \).

By using Holder’s inequality and our control of \( \left[ \sup_{1 \leq j \leq n} \left| \sqrt{Q_{I_j}} - m_{\sqrt{Q_{I_j}}} \right| \right] \left[ \sup_{1 \leq j \leq n} m_{\sqrt{Q_{I_j}}} \right] \) in \( L_{2k} \), it is clear that the only issue remaining is control of \( \sup_{1 \leq j \leq n} m_{\sqrt{Q_{I_j}}} \) in \( L_{2k} \).

Since \( m_{\sqrt{Q_{I_j}}} = E_{X_{I_j}} \left( X_{I_j}^I M_{I_j} X_{I_j}^I / n \right) \leq \sqrt{E_{X_{I_j}} \left( X_{I_j}^I M_{I_j} X_{I_j}^I / n \right)} = \sqrt{\text{trace} (M_{I_j}) / n}, \) since \( \text{cov} (X_{I_j}) = \Id_p \), we see that

\[ \sup_{1 \leq j \leq n} m_{\sqrt{Q_{I_j}}} \leq \sqrt{p/n} \sup_{1 \leq j \leq n} \sqrt{\lambda_{\text{max}, I_j}}. \]
Therefore,
\[
\left[ \sup_{1 \leq j \leq n} \sqrt{Q_{ij}} - m \sqrt{Q_{ij}} \right] \leq K \frac{\text{polyLog}(n) \sqrt{p/n}}{\sqrt{nc_n}} \sup_{1 \leq j \leq n} \lambda_{\max, i} \text{ in } L_k,
\]
provided all the random quantities we work with have $2k$ moments.

The conclusions of the Lemma follow by recalling our assumption that $p/n$ remains bounded and using the fact that $1/c_n \geq K/\sqrt{c_n}$ in the situations we are considering, i.e $c_n$ bounded but possibly going to zero.

Lemma B-4. Suppose $S_i = \frac{1}{n} \sum_{j \neq i} d_j X_j X_j'$ where $\{d_j\}_{j=1}^n$ depends on $\{X_i\}_{i=1}^n$. Suppose we can find $\{d_{j(i)}\}_{j \neq i}$ independent of $X_i$ such that $\sup_{j \neq i} |d_j - d_{j(i)}| \leq \delta_n(i)$. Then we have
\[
\left| \frac{1}{n} X_i'(S_i + \tau \text{Id}_p)^{-1} X_i - \frac{1}{n} \text{trace} \left( (S_i + \tau \text{Id}_p)^{-1} \right) \right| = O_{L_k} \left( \delta_n(i) ||| \bar{\Sigma} |||_2 \vee \frac{1}{nc_n} \right).
\]

If the same can be done with all $i$'s and $\sup_{1 \leq i \leq n} \delta_n(i) \leq K_n$, we have
\[
\sup_i \left| \frac{1}{n} X_i'(S_i + \tau \text{Id}_p)^{-1} X_i - \frac{1}{n} \text{trace} \left( (S_i + \tau \text{Id}_p)^{-1} \right) \right| = O_{L_k} \left( (1 + \frac{n \text{polyLog}(n)}{\sqrt{nc_n}}) K_n ||| \bar{\Sigma} |||_2 \vee \frac{n \text{polyLog}(n)}{\sqrt{nc_n}} \right)
\]

Proof. We call $S_i = \frac{1}{n} \sum_{j \neq i} d_{j(i)} X_j X_j'$. Recall the first resolvent identity, i.e $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$.

Clearly, by applying this identity of $S_i + \tau \text{Id}_p$ and $\tilde{S}_i + \tau \text{Id}_p$, we have
\[
\left| \frac{1}{n} X_i'(S_i + \tau \text{Id}_p)^{-1} X_i - \frac{1}{n} X_i' (\tilde{S}_i + \tau \text{Id}_p)^{-1} X_i \right| \leq \frac{\|X_i\|^2}{n} \frac{1}{\tau} \delta_n(i) ||| \bar{\Sigma} |||_2,
\]
by using the fact that $||| \frac{1}{n} \sum_{j \neq i} (\tilde{d}_{j(i)} - d_j) X_j X_j' |||_2 \leq \sup_{j \neq i} |\tilde{d}_{j(i)} - d_j| ||| \bar{\Sigma} |||_2$.

Since $\tilde{S}_i$ is independent of $X_i$, we can apply the results from Lemma B-3 to see that
\[
\frac{1}{n} X_i'(\tilde{S}_i + \tau \text{Id}_p)^{-1} X_i - \frac{1}{n} \text{trace} \left( (\tilde{S}_i + \tau \text{Id}_p)^{-1} \right) = O_{L_k} \left( \frac{1}{\sqrt{nc_n}} \right).
\]

On the other hand, by making use again of the first resolvent identity, we see that
\[
\frac{1}{n} \text{trace} \left( (S_i + \tau \text{Id}_p)^{-1} - (\tilde{S}_i + \tau \text{Id}_p)^{-1} \right) = \frac{1}{n} \text{trace} \left( (S_i + \tau \text{Id}_p)^{-1} \frac{1}{n} \sum_{j \neq i} (\tilde{d}_{j(i)} - d_j) X_j X_j' (\tilde{S}_i + \tau \text{Id}_p)^{-1} \right),
\]
\[
= \frac{1}{n^2} \sum_{j \neq i} (\tilde{d}_{j(i)} - d_j) X_j X_j' (\tilde{S}_i + \tau \text{Id}_p)^{-1} (S_i + \tau \text{Id}_p)^{-1} X_j,
\]
\[
\leq \frac{1}{n^2} \sum_{j \neq i} |\tilde{d}_{j(i)} - d_j| \frac{\|X_j\|^2}{\tau^2},
\]
\[
\leq \delta_n(i) \frac{1}{n} \sum_{i=1}^n \frac{\|X_j\|^2}{n \tau^2}.
\]

We conclude that
\[
\left| \frac{1}{n} X_i'(S_i + \tau \text{Id}_p)^{-1} X_i - \frac{1}{n} \text{trace} \left( (S_i + \tau \text{Id}_p)^{-1} \right) \right| = O_{L_k} \left( \frac{\|X_i\|^2}{n} \frac{1}{\tau} \delta_n(i) ||| \bar{\Sigma} |||_2 + \frac{1}{\sqrt{nc_n} \tau} + \delta_n(i) \frac{1}{n} \sum_{i=1}^n \frac{\|X_j\|^2}{n \tau^2} \right).
\]

The result for the sup follows by the same technique and adjusting for taking sup in the various approximations. \qed
On the spectral norm of covariance matrices

In this subsection, we show that under our initial concentration assumptions, we can control $|||\widehat{\Sigma}|||_2$. These results are very likely known but we did not find a reference covering precisely the same question we consider. The proof is a simple adaption of the well-known $\epsilon$-net argument explained e.g in Talagrand (2003), Appendix A.4.

Lemma B-5. Suppose $X_i$’s are independent random vectors in $\mathbb{R}^p$, satisfying our concentration assumptions in $O4$, and having mean 0 and covariance $\text{Id}_p$. Let $\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_iX_i'$. Then,

$$|||\widehat{\Sigma}|||_2 = O_P(c_n^{-1/2}).$$

The results hold also in $L_k$.

Proof. We study the largest singular value, $\sigma_1$ of the matrix $X/\sqrt{n}$, where the $i$-th row of $X$ is $X_i$. Of course,

$$\sigma_1(X/\sqrt{n}) = \sup_{u,v,||u||=1,||v||=1} \frac{1}{\sqrt{n}} u'Xv.$$

Note that $u'Xv = \sum_{i=1}^{n} u_i(X_i'v)$.

Consider first the case where $c_n = 1$. Under our assumptions, $X_i'v$ are independent subGaussian random vectors, with mean 0. Note that $\text{var}(X_i'v) = 1$ if $\text{cov}(X_i) = 1$ and $||v|| = 1$. Computing the moment generating function of $u'Xv$, we see that this random variable is itself subGaussian and has variance 1. Therefore, we have for all $t$, and constants $c_1$ and $c_2$,

$$P(|u'Xv| > t) \leq c_1 \exp(-c_2t^2).$$

The $\epsilon$-net argument given in the proof of Lemma A.4.1 in Talagrand (2003) then can be applied and the conclusions of that Lemma reached. (A slight adaption is needed to handle the fact that $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^p$ but it is completely trivial and omitted). The fact that the results hold in $L_k$ is a simple consequence of the proof.

In the case where $c_n \neq 1$, we just need to note that the moment generating function of $u'Xv$ is smaller than that of a Gaussian random variable with variance $1/c_n$. The result follows immediately. \qed

C Miscellaneous results

C-1 An analytic remark

One of our assumptions concerns the existence and uniqueness of a solution of the equation $F(x) = 0$, where

$$F(x) = \frac{p}{n} - \tau x - 1 + \mathbb{E} \left( (\text{prox}_x(\rho))'(W) \right)$$

where $W$ is a random variable and $(\text{prox}_x(\rho))'(t) = \frac{\partial}{\partial t} \text{prox}_x(\rho)(t) = \frac{1}{1+x\psi'(\text{prox}_x(\rho)(t))}$.

We now show that under mild conditions on $W$ this equation has a unique solution. This guarantees that our assumptions are not terribly strong and in particular apply to problems of interest to statisticians.

Lemma C-1. Suppose that $W$ has a smooth density $f$ with $\text{sign}(f'(x)) = -\text{sign}(x)$. Suppose further that $\lim_{|t| \to \infty} tf(t) = 0$ and that $\text{sign}(\psi(x)) = \text{sign}(x)$. Then, if

$$F(x) = \frac{p}{n} - \tau x - 1 + \mathbb{E} \left( (\text{prox}_x(\rho))'(W) \right),$$

the equation $F(x) = 0$ has a unique solution.
Proof. We call
\[ G(x) \triangleq E \left( (\text{prox}_x(\rho))'(W) \right) . \]
Of course,
\[ E \left( (\text{prox}_x(\rho))'(W) \right) = \int (\text{prox}_x(\rho))'(t)f(t)dt. \]
Using contractivity of the proximal mapping (see Moreau (1965)) we see that \( \lim_{|t| \to \infty} \text{prox}_x(\rho)(t)f(t) = 0 \) under our assumptions.

Integrating the previous equation by parts, we see that
\[ E \left( (\text{prox}_x(\rho))'(W) \right) = - \int (\text{prox}_x(\rho))(t)f'(t)dt . \]
To compute \( G'(x) \), we differentiate under the integral sign (under our assumptions the conditions of Theorem 9.1 in Durrett (1996) are satisfied) to get
\[ G'(x) = \int \frac{\psi(\text{prox}_x(\rho)(t))f'(t)}{1 + x\psi'(\text{prox}_x(\rho)(t))} dt . \]
Under our assumptions, \( \text{sign}(\psi(\text{prox}_x(\rho)(t))) = \text{sign}(t) \) and \( \text{sign}(f'(t)) = -\text{sign}(t) \), so that
\[ \forall t \neq 0, \text{sign}(\psi(\text{prox}_x(\rho)(t)))f'(t)) = -1 . \]
Since the denominator of the function we integrate is positive, we conclude that
\[ G'(x) \leq 0 . \]
Since \( F'(x) = -\tau + G'(x) \), we see that \( F'(x) < 0 \). Therefore \( F \) is a decreasing function on \( \mathbb{R}_+ \). Of course, \( F(0) = p/n \) and \( \lim_{x \to \infty} F(x) = -\infty \). So we conclude that the equation \( F(x) = 0 \) has a unique root. \( \square \)

Remark: the conditions on the density of \( W \) are satisfied in many situations. For instance if \( W = \epsilon + rZ \), where \( \epsilon \) is symmetric about 0 and log-concave, \( Z \) is \( N(0, 1) \) and \( r > 0 \), it is clear that the density of \( W \) satisfies the conditions of our lemma. Similar results hold under weaker assumptions on \( \epsilon \) of course but since the paper is already a bit long, we do not dwell on these issues which are well-known in the theory of log-concave functions (see e.g Karlin (1968), Prékopa (1973) and Ibragimov (1956)).

C-2 A linear algebraic remark

We have the following lemma.

Lemma C-2. Suppose the \( p \times p \) matrix \( A \) is positive semi-definite and
\[ A = \begin{pmatrix} \Gamma & v \\ v' & a \end{pmatrix} . \]
Here \( a \in \mathbb{R} \). Let \( \tau \) be a strictly positive real. Call \( \Gamma_{\tau} = \Gamma + \tau \text{Id}_{p-1} \). Then we have
\[ \text{trace} \left( (A + \tau)^{-1} \right) = \text{trace} \left( \Gamma_{\tau}^{-1} \right) + \frac{1 + v'\Gamma_{\tau}^{-2}v}{a + \tau - v'\Gamma_{\tau}^{-1}v} . \]
In particular,
\[ \left| \text{trace} \left( (A + \tau)^{-1} \right) - \text{trace} \left( \Gamma_{\tau}^{-1} \right) \right| \leq \frac{1 + a/\tau}{\tau} . \]
Proof. The first equation is simply an application of the block inversion formula for matrices (see Horn and Johnson (1990), p.18) and the Sherman-Morrison-Woodbury formula (Horn and Johnson (1990), p.19). Suppose temporarily that \( A \) is positive definite. Then the Schur complement formula guarantees that \( a \geq v'\Gamma^{-1}v > v'\Gamma_{\tau}^{-1}v \). The fact that \( a \geq v'\Gamma_{\tau}^{-1}v \) in general is obtained by a continuity argument (change \( A \) to \( A_\epsilon = A + \epsilon \text{Id}_p \) and let \( \epsilon \) tend to 0). This implies that
\[ \frac{1}{a + \tau - v'\Gamma_{\tau}^{-1}v} \leq \frac{1}{\tau} . \]
Since \( v'\Gamma_{\tau}^{-2}v \leq \frac{1}{4}v'\Gamma_{\tau}^{-1}v \leq a/\tau \), we get the second equation. \( \square \)
References

ANDERSON, T. W. (2003). An introduction to multivariate statistical analysis. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition.

BAI, Z. D. (1999). Methodologies in spectral analysis of large-dimensional random matrices, a review. Statist. Sinica 9, 611–677. With comments by G. J. Rodgers and Jack W. Silverstein; and a rejoinder by the author.

BARANCHIK, A. J. (1973). Inadmissibility of maximum likelihood estimators in some multiple regression problems with three or more independent variables. Ann. Statist. 1, 312–321.

BAYATI, M. and MONTANARI, A. (2012). The LASSO risk for Gaussian matrices. IEEE Trans. Inform. Theory 58, 1997–2017. URL http://dx.doi.org/10.1109/TIT.2011.2174612.

BEAN, D., BICKEL, P. J., EL KAROUI, N., and YU, B. (2013). Optimal m-estimation in high-dimensional regression. Proceedings of the National Academy of Sciences 110, 14563–14568. URL http://www.pnas.org/content/110/36/14563.abstract.

BECK, A. and TEBOUILLE, M. (2010). Convex Optimization in Signal Processing and Communications, chapter Gradient-Based Algorithms with Applications in Signal Recovery Problems, pp. 33–88. Cambridge University Press.

BREIMAN, L. (1992). Probability, volume 7 of Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. Corrected reprint of the 1968 original.

DONOOHO, D. and MONTANARI, A. (2013). High dimensional robust m-estimation: Asymptotic variance via approximate message passing. arXiv:1310.7320 .

DONOOHO, D. L., MALEKI, A., and MONTANARI, A. (2009). Message-passing algorithms for compressed sensing. Proceedings of the National Academy of Sciences 106, 18914–18919. URL http://www.pnas.org/content/106/45/18914.abstract.

DURRETT, R. (1996). Probability: theory and examples. Duxbury Press, Belmont, CA, second edition.

EFRON, B. and STEIN, C. (1981). The jackknife estimate of variance. Ann. Statist. 9, 586–596.

EL KAROUI, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. The Annals of Applied Probability 19, 2362–2405.

EL KAROUI, N., BEAN, D., BICKEL, P., LIM, C., and YU, B. (2012). On robust regression with high-dimensional predictors. Technical report.

EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C., and YU, B. (2013). On robust regression with high-dimensional predictors. Proceedings of the National Academy of Sciences URL http://www.pnas.org/content/early/2013/08/15/1307842110.abstract.

HALKO, N., MARTINSSON, P. G., and TROPP, J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. SIAM Rev. 53, 217–288. URL http://dx.doi.org/10.1137/090771806.

HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (2001). Fundamentals of convex analysis. Grundlehren Text Editions. Springer-Verlag, Berlin. Abridged version of it Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and it II [ibid.; MR1295240 (95m:90002)].

HORN, R. A. and JOHNSON, C. R. (1990). Matrix analysis. Cambridge University Press, Cambridge. Corrected reprint of the 1985 original.

40
Huber, P. J. (1972). The 1972 Wald lecture. Robust statistics: A review. *Ann. Math. Statist.* **43**, 1041–1067.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition. URL http://dx.doi.org/10.1002/9780470434697.

Ibragimov, I. A. (1956). On the composition of unimodal distributions. *Teor. Veroyatnost. i Primenen.* **1**, 283–288.

Karlin, S. (1968). *Total positivity. Vol. I*. Stanford University Press, Stanford, Calif.

Ledoux, M. (2001). *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17**, 382–400. URL http://dx.doi.org/10.1214/aos/1176347023.

Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices. *Mat. Sb. (N.S.)* **72** (114), 507–536.

Moreau, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* **93**, 273–299.

Pollard, D. (1984). *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York.

Portnoy, S. (1984). Asymptotic behavior of $M$-estimators of $p$ regression parameters when $p^2/n$ is large. I. Consistency. *Ann. Statist.* **12**, 1298–1309. URL http://dx.doi.org/10.1214/aos/1176346793.

Portnoy, S. (1985). Asymptotic behavior of $M$ estimators of $p$ regression parameters when $p^2/n$ is large. II. Normal approximation. *Ann. Statist.* **13**, 1403–1417. URL http://dx.doi.org/10.1214/aos/1176349744.

Portnoy, S. (1987). A central limit theorem applicable to robust regression estimators. *J. Multivariate Anal.* **22**, 24–50. URL http://dx.doi.org/10.1016/0047-259X(87)90073-X.

Prékopa, A. (1973). On logarithmic concave measures and functions. *Acta Sci. Math. (Szeged)* **34**, 335–343.

Rangan, S. (2011). Generalized approximate message passing for estimation with random linear mixing. In *IEEE International Symp. On Information Theory (St. Petersburg)*.

Shcherbina, M. and Tirozzi, B. (2003). Rigorous solution of the Gardner problem. *Comm. Math. Phys.* **234**, 383–422. URL http://dx.doi.org/10.1007/s00220-002-0783-3.

Silverstein, J. W. (1985). The smallest eigenvalue of a large-dimensional Wishart matrix. *Ann. Probab.* **13**, 1364–1368.

Silverstein, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55**, 331–339.

Stein, C. (1960). Multiple regression. In *Contributions to probability and statistics*, pp. 424–443. Stanford Univ. Press, Stanford, Calif.

Talagrand, M. (2003). *Spin glasses: a challenge for mathematicians*, volume 46 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]*. Springer-Verlag, Berlin. Cavity and mean field models.
WACHTER, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Annals of Probability* 6, 1–18.