On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators

Noureddine El Karoui
Department of Statistics, UC Berkeley*

Abstract

We study ridge-regularized generalized robust regression estimators, i.e

$$\widehat{\beta} = \text{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho_i(Y_i - X_i'\beta) + \frac{\tau}{2}\|\beta\|^2 \ , \text{ where } Y_i = \epsilon_i + X_i'\beta_0 \ .$$

in the situation where p/n tends to a finite non-zero limit.

Our study here focuses on the situation where the errors ϵ_i's are heavy-tailed and X_i's have an "elliptical-like" distribution. Our assumptions are quite general and we do not require homoskedasticity of ϵ_i's for instance.

We obtain a characterization of the limit of $\|\widehat{\beta}-\beta_0\|$, as well as several other results, including central limit theorems for the entries of $\widehat{\beta}$.

1 Introduction

Robust regression estimators are a standard and important tool in the toolbox of modern statisticians. They were introduced in the last sixties (Relles (1968)) and important early results appeared shortly thereafter (Huber (1972, 1973)). We recall that these estimators are defined as

$$\widehat{\beta}_\rho = \text{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho(Y_i - X_i'\beta) \ , \tag{1}$$

for ρ a function chosen by the user. Here Y_i is a scalar response and X_i is a vector of predictor in $\mathbb{R}^p$. In the context we consider here, ρ will be a convex function. Naturally, one of the main reason to use these estimators instead of the standard least-squares estimator, is to increase the robustness of $\widehat{\beta}_\rho$ to outliers in e.g Y_i's. Formally, this robustness result can be seen through results of Huber (see Huber and Ronchetti (2009)), in the low-dimensional case where p is fixed. Huber showed that when $Y_i = X_i'\beta_0 + \epsilon_i$, and when ϵ_i's are i.i.d, under some mild regularity conditions, $\widehat{\beta}_\rho$ is asymptotically normal with mean β_0 and (asymptotic) covariance

$$(X'X)^{-1}\frac{\mathbf{E}\left(\psi^2(\epsilon)\right)}{[\mathbf{E}\left(\psi'(\epsilon)\right)]^2} \ , \text{ where } \psi = \rho' \ . \tag{2}$$

The question of understanding the behavior of these estimators in the high-dimensional setting where p is allowed to grow with n was raised very early on in Huber (1973) (p.802, questions b) to f)). These questions started being answered in the mid to late eighties in work of Portnoy and Mammen (e.g Portnoy

*I would like to thank Peter Bickel for many interesting discussions on this and related topics. I am grateful to the Center of Mathematical Sciences and Applications at Harvard University for hospitality while some of this work was being completed. Support from NSF grant DMS-0847647 (CAREER) is gratefully acknowledged. **AMS 2010 MSC:** Primary: 60F99; Secondary: 62E20.**Key words and Phrases :** high-dimensional inference, random matrix theory, concentration of measure, proximal mapping, regression M-estimates, robust regression. **Contact :** nkaroui@berkeley.edu

(1984), Portnoy (1985), Portnoy (1986), Portnoy (1987), Mammen (1989)). However, these papers covered the case where $p/n \to 0$ while $p \to \infty$.

In the papers El Karoui et al. (2011) and El Karoui et al. (2013), we explained (mixing, as in Huber (1973), rigorous arguments, simulations and heuristic arguments) that the case $p/n \to \kappa \in (0,1)$ yielded a qualitatively completely different picture for this class of problems. For instance, under various technical assumptions, we explained that the risk $\|\widehat{\beta}_\rho - \beta_0\|_2$ could be characterized through a system of two non-linear equations (sharing some characteristics with the one below), the distribution of the residuals could be found and was completely different of that of the ϵ_i's, by contrast with the low dimensional case. Furthermore, we showed in Bean et al. (2013) that maximum likelihood estimators were in general inefficient in high-dimension and found dimension-adaptive loss functions ρ that yielded better estimators than the ones we would have gotten by using the standard maximum likelihood estimator, i.e using $\rho = -\log f_\epsilon$, where f_ϵ is the density of the i.i.d errors ϵ_i's. (We subsequently showed in as-of-yet-unpublished El Karoui (2013) that the techniques we had proposed in El Karoui et al. (2011) could be made mathematically rigorous under various assumptions. See also the paper Donoho and Montanari (2013) that handles only the case of i.i.d Gaussian predictors, whereas El Karoui (2013) can deal with more general assumptions on the predictors. Donoho and Montanari (2013) also make interesting connection with the Scherbina-Tirrozi model in statistical physics - see Shcherbina and Tirozzi (2003) and Talagrand (2003). For other interesting results using rigorous approximate message passing techniques, see also e.g Bayati and Montanari (2012))

In the current paper, we study a generic extension of the robust regression problem involving ridge regularization. In other words, we study the statistical properties of

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_i(Y_i - X_i'\beta) + \frac{\tau}{2}\|\beta\|^2 \ , \text{ where } Y_i = \epsilon_i + X_i'\beta_0 \ .$$

We will focus in particular on the case where there is no moment restriction on ϵ_i's. Furthermore, a key element of the study will be to show that the performance of $\widehat{\beta}$ is driven by the Euclidean geometry of the set of predictor $\{X_i\}_{i=1}^n$. To do so, we will study "elliptical" models for X_i's, i.e $X_i = \lambda_i \mathcal{X}_i$, where $\mathcal{X}_i$ has for instance independent entries. We note that when λ_i is independent of $\mathcal{X}_i$ and $\mathbf{E}\left(\lambda_i^2\right) = 1$, $\text{cov}\left(X_i\right) = \text{cov}\left(\mathcal{X}_i\right)$. Hence these families of distributions for X_i's have the same covariance, but as we will see they yield estimators whose performance vary quite substantially with the distribution of λ_i's. As we explain below, the role of λ_i's is to induce a "non-spherical geometry" on the predictors; understanding the impact of λ_i's on the performance of $\widehat{\beta}$ is hence a way to understand how the geometry of the predictors affects the performance of the estimator. We note that in the low-dimensional case, when X_i's are i.i.d, $X'X/n \to \text{cov}\left(X_1\right)$ in probability under mild assumptions, and hence the result of Huber mentioned in Equation (2) shows that the limiting behavior of $\widehat{\beta}_\rho$ defined in Equation (1) is the same under "elliptical" and non-elliptical models.

Our interest in elliptical distributions stems from the fact that, as we intuited for a related problem in El Karoui et al. (2011), the behavior of quantities of the type $X_i'QX_i$ for Q deterministic is at the heart of the performance of $\widehat{\beta}$. Hence, studying elliptical distribution settings both shed light on the impact of the geometry of predictors on the performance of the estimator and allow us to put to rest potential claims of "universality" of results obtained in the Gaussian (or geometrically similar) case. We note that in statistics there is a growing body of work showing the importance of predictor geometry on various high-dimensional problems (see e.g Diaconis and Freedman (1984); Hall et al. (2005); El Karoui (2009, 2010); El Karoui and Koesters (2011)).

One main motivation for allowing ρ_i to change with i is that it might be natural to use different loss functions for different observations if we happen to have information about distributional inhomogeneities in $\{X_i, Y_i\}_{i=1}^n$. For instance, one group of observations could have errors coming from one distribution and a second group might have errors with a different distribution. Another reason is to gain information on the case of weighted regression, in which case $\rho_i = w_i\rho$. Also, this analysis can be used to justify rigorously some of the claims made in El Karoui et al. (2011). Finally, it may prove useful in some bootstrap studies (see e.g El Karoui and Purdom (2015) for examples).

In the current paper, we consider the situation where β_0 is "diffuse", i.e all of its coordinates are small and it cannot be well approximated by a sparse vector. In this situation, use of ridge/ℓ_2 penalization is

natural. The paper also answers the question, raised by other researchers in statistics, of knowing whether the techniques of El Karoui (2013) could be used in the situation we are considering here. Finally, the paper shows that some of the heuristics of Bean et al. (2012) can be rigorously justified.

When $\rho_i = \rho$ for all i, a natural question is to know whether we can find an optimal ρ, in terms of prediction error for instance, as a function of the law of ϵ_i's - in effect asking similar questions to the ones we answered by Huber (Huber and Ronchetti (2009)) in low-dimension and in Bean et al. (2013) in high-dimension. However, the constraints we impose in the current paper on both the errors (i.e we do not want them to have moments) and the functions ρ_i's make part of the argument in Bean et al. (2013) not usable and might require new ideas. So we will consider this "optimization over ρ_i's and τ" in future work, given that the current proof is already long.

The problem and setup considered in this paper are more natural in the context of robust regression than the ones studied in the as-of-yet unpublished El Karoui (2013), where the setup chosen was targeted towards problems related to suboptimality of maximum likelihood methods. However, the strategy for the proof of the results here is similar to the strategy we devised in El Karoui (2013). There are three main conceptual novelties, that create important new problems: handling ellipticity and the fact that $\beta_0 \neq 0$ requires new ideas in the second part of the proof (i.e Section D). Dealing with heavy tails and appropriate loss functions impacts the whole proof and requires many changes compared to the proof of El Karoui (2013). Conceptually, this latter part is also the most important, as it shows that all the approximations made in earlier heuristic papers are valid, even in the presence of heavy-tailed errors. This situation is of course the one where these approximations, while having clearly shown their usefulness in giving conceptual and heuristic understanding of the statistical problem, were the most mathematically "suspicious". So it is interesting to see that they can be made to work rigorously, especially since the probabilistic heuristics developed in this earlier papers allow researchers to shed light quickly on non-trivial statistical problems.

We now state our results. We believe our notations are standard but refer the reader to p.17 in case clarification is needed.

2 Results

The main focus of the paper is in understanding the properties of

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_i(Y_i - X_i'\beta) + \frac{\tau}{2}\|\beta\|^2 \ , \text{ where } Y_i = X_i'\beta_0 + \epsilon_i \ , \tag{3}$$

and $\tau > 0$. For all $1 \leq i \leq n$, we have $\epsilon_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$.

We prove four main results in the paper:

1. we characterize the ℓ_2-risk of our estimator, i.e $\|\widehat{\beta} - \beta_0\|_2$
2. we describe the behavior of the residuals $R_i = Y_i - X_i'\widehat{\beta}$ and relate them to the leave-one-out prediction error $\tilde{r}_{i,i} = Y_i - X_i'\widehat{\beta}_{(i)}$
3. we obtain an approximate update formula for $\widehat{\beta}$ when adding an observation (and show it is very accurate)
4. we provide central limit theorems for the individual coordinates of $\widehat{\beta}$.

For the sake of clarity, we provide in the main text a series of assumptions that guarantee that our results hold. However, a more detailed and less restrictive statement of our assumptions is provided in the Appendix (see p. 14 and beyond).

2.1 Preliminaries and overview of technical assumptions

We use the notation $\text{prox}(\rho)$ to denote the proximal mapping of the function ρ, which is assumed to be convex. This notion was introduced in Moreau (1965). We recall that

$$\text{prox}(c\rho)(x) = \text{argmin}_{y\in\mathbb{R}}(c\rho(y) + \frac{1}{2}(x-y)^2) \text{ , or equivalently,}$$
$$\text{prox}(c\rho)(x) = (\text{Id} + c\psi)^{-1}(x) \text{ , where } \psi = \rho'$$

We refer the reader to Moreau (1965), Beck and Teboulle (2010), or Ruszczyński (2006), Section 7.3, for more details on this operation. Note that the previous definitions imply that

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

We give examples of proximal mappings in the Appendix, Section F, p. 51.

We now state some sufficient assumptions that guarantee that all the results stated below are correct. The main proofs are in the Appendix. The proofs done in the Appendix are done at a much greater level of generality than we are about to state and various aspects of those proofs require much weaker assumptions than those we present here. We start by giving an example where all of our conditions are met.

Example Our conditions are met when

- $p/n \to \kappa \in (0, \infty)$
- ϵ_i's are i.i.d Cauchy (with median at 0)
- $X_i = \lambda_i \mathcal{X}_i$, where $\lambda_i \in \mathbb{R}$ and $\mathcal{X}_i \in \mathbb{R}^p$ are independent. λ_i's are i.i.d with bounded support; $\mathcal{X}_i$'s are i.i.d with i.i.d $\mathcal{N}(0,1)$ entries, or i.i.d entries with bounded support and mean 0 as well as variance 1. $\{\mathcal{X}_i\}_{i=1}^n$, $\{\lambda_i\}_{i=1}^n$ and $\{\epsilon_i\}_{i=1}^n$ are independent.
- β_0 is a "diffuse" vector with $\beta_0(i) = u_{i,p}/\sqrt{p}$, $0 \leq |u_{i,p}| \leq C$ and $\sum_{i=1}^p u_{i,p}^2 = 1$.
- $\rho_i = \rho$ for all i's. $\psi = \rho'$ is bounded and ψ' is Lipschitz and bounded. $\text{sign}(\psi(x)) = \text{sign}(x)$ and $\rho(x) \geq \rho(0) = 0$.

We note that this last condition is satisfied for smoothed approximation of the Huber function, where the discontinuity in ψ' at say 1 is replaced by a linear interpolation; see below for more details. Note however that the Huber function has a priori no statistical optimality properties in the context we consider.

Sufficient conditions for our results to hold

- p/n has a finite non-zero limit.
- ρ_i's are chosen from finitely many possible convex functions. If $\psi_i = \rho_i'$, $\sup_i \|\psi_i\|_\infty \leq K$, $\sup_i \|\psi_i'\|_\infty \leq K$, for some K. ψ_i' is also assumed to be Lipschitz-continuous. Also, for all $x \in \mathbb{R}$, $\text{sign}(\psi_i(x)) = \text{sign}(x)$ and $\rho_i(x) \geq \rho_i(0) = 0$.
- $X_i = \lambda_i \mathcal{X}_i$, where $\mathcal{X}_i$'s are i.i.d with independent entries. λ_i's are independent and independent of $\mathcal{X}_i$'s. The entries of $\mathcal{X}_i$'s satisfy concentration property in the sense that if G is a convex 1-Lipschitz function (with respect to Euclidean norm), $P(|G(\mathcal{X}_i) - m_G| > t) \leq C\exp(-\mathsf{c}t^2)$, for any $t > 0$, m_G being a median of $G(\mathcal{X}_i)$. We require the same assumption to hold when considering the columns of the $n \times p$ design matrix $\mathcal{X}$. $\mathcal{X}_i$'s have mean 0 and $\text{cov}\,(\mathcal{X}_i) = \text{Id}_p$. We also assume that the coordinates of $\mathcal{X}_i$ have moments of all order. Furthermore, for any given k, the k-th moment of the entries of $\mathcal{X}_i$ is assumed to be bounded independently of n and p.
- $\mathbf{E}\left(\lambda_i^2\right) = 1$, $\mathbf{E}\left(\lambda_i^4\right)$ is bounded and $\sup_{1\leq i\leq n} |\lambda_i|$ grows a most like $C(\log n)^k$ for some k. λ_i's may have different distributions, but the number of such possible distributions is finite.
- ϵ_i's are independent. They may have different distributions, but the number of such possible distributions is finite. Those distributions are assumed to have densities that are differentiable, symmetric and unimodal. Furthermore, we assume that if f_i is the density of one such distribution, $\lim_{x\to\infty} x f_i(x) = 0$. $\{\mathcal{X}_i\}_{i=1}^n$, $\{\lambda_i\}_{i=1}^n$ and $\{\epsilon_i\}_{i=1}^n$ are independent.

- $\|\beta_0\|_2$ remains bounded. Furthermore, $\|\beta_0\|_\infty = \mathrm{O}(n^{-\mathsf{e}})$, where $1/4 < \mathsf{e}$.
- The fraction of time each possible combination of functions and distributions for $(\rho_i, \mathcal{L}(\epsilon_i), \mathcal{L}(\lambda_i))$ appears in our problem has a limit as $n \to \infty$. ($\mathcal{L}(\epsilon_i)$ and $\mathcal{L}(\lambda_i)$ are the laws of ϵ_i and λ_i.)

We now state our most important results (several others are in the Appendix, where we give the proof) and our proof strategy; naturally, the two go together to provide a sketch of proof. We postpone our discussion of both the assumptions and our results to Subsection 2.3 on p. 9.

2.2 Results and proof strategy

2.2.1 Characterization of the risk of $\widehat{\beta}$

Consider $\widehat{\beta}$ defined in Equation (3) and assume that $\tau > 0$ is given, i.e does not change with p and n. Under the technical assumptions detailed in Subsection 2.1, we have :

Theorem 2.1. *As p, n tend to infinity while $p/n \to \kappa \in (0, \infty)$, $\mathrm{var}\left(\|\widehat{\beta} - \beta_0\|^2\right) \to 0$. Furthermore, $\|\widehat{\beta} - \beta_0\| \to r_\rho(\kappa)$ in probability, for $r_\rho(\kappa)$ a deterministic scalar. Call $W_i = \epsilon_i + r_\rho(\kappa)\lambda_i Z_i$, where Z_i is a $\mathcal{N}(0,1)$ random variable independent of ϵ_i and λ_i. Then there exists a constant $c_\rho(\kappa) \geq 0$ such that*

$$\begin{cases} \left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left([prox(c_\rho(\kappa)\lambda_i^2\rho_i)]'(W_i)\right)\right] & = 1 - \kappa + \tau c_\rho(\kappa) \\ \kappa\left[\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\frac{(W_i - prox(c_\rho(\kappa)\lambda_i^2\rho_i)[W_i])^2}{\lambda_i^2}\right)\right] + \tau^2\|\beta_0\|^2 c_\rho^2(\kappa) & = \kappa^2 r_\rho^2(\kappa)\ . \end{cases} \tag{4}$$

We note that

$$\frac{(x - \mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)[x])^2}{\lambda_i^2} = c_\rho^2(\kappa)\lambda_i^2\psi_i^2(\mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)[x])$$

so in case λ_i takes the value 0, we can replace the expression on the left hand side by that on the right hand side, which does not involve dividing by λ_i^2. This alternative expression also shows that there is no problem taking expectations in our equations.

The previous system can be reformulated in terms of $\mathrm{prox}((c_\rho(\kappa)\lambda_i^2\rho_i)^*)$, where f^* represents the Fenchel-Legendre dual of f. Indeed, Moreau's prox identity (Moreau (1965)) gives

$$\mathrm{prox}((c\rho)^*)(x) = x - \mathrm{prox}(c\rho)(x)\ .$$

This is partly why we chose to write the system as we did, since it can be rephrased purely in terms of $\mathrm{prox}([c_\rho(\kappa)\lambda_i^2\rho_i]^*)$, a formulation that has proven useful in previous related problems (see Bean et al. (2013)).

We note that $r_\rho(\kappa)$ and $c_\rho(\kappa)$ will in general depend on τ, but we do not index those quantities by τ to avoid cumbersome notations.

2.2.2 Organization of the proof and strategy

The proof is quite long so we now explain the main ideas and organization of the argument. Recall that if

$$F(\beta) = \frac{1}{n}\sum_{i=1}^n \rho_i(Y_i - X_i'\beta) + \frac{\tau}{2}\|\beta\|^2\ ,$$

we have

$$\widehat{\beta} = \mathrm{argmin}_{\beta\in\mathbb{R}^p} F(\beta)\ .$$

The proof is broadly divided into three steps.

First step The first idea is to relate $\widehat{\beta}$ and $\widehat{\beta}_{(i)}$, the solution of our optimization problem when the pair (X_i, Y_i) is excluded from the problem. It is reasonable to expect that adding (X_i, Y_i) will not change too much $\tilde{r}_{j,(i)} = Y_j - X_j'\widehat{\beta}_{(i)}$ when $j \neq i$, and hence that $\tilde{r}_{j,(i)} \simeq R_j = Y_j - X_j'\widehat{\beta}$ when $j \neq i$. Armed with this intuition, we can try to use a first-order Taylor expansion of $\widehat{\beta}$ around $\widehat{\beta}_{(i)}$ in the equation $\nabla F(\widehat{\beta}) = 0$ to relate the two vectors. This is what the first part of the proof does, by surmising an

approximation for $\eta_i = \widehat{\beta} - \widehat{\beta}_{(i)}$- following along the intuitive lines above but non-trivial to come up with at the precision we need. Much work is devoted to proving that this very informed guess is sufficiently accurate for our purposes. Since "the only thing we know" about $\widehat{\beta}$ is that $\nabla F(\widehat{\beta}) = 0$, we compare $\nabla F(\widehat{\beta}) - \nabla F(\widehat{\beta}_{(i)} + \eta_i)$ to do so, and show in our preliminaries (see Section B) that controlling this latter quantity is enough to control $\|\widehat{\beta} - \widehat{\beta}_{(i)} - \eta_i\|$. Once our bound for $\|\widehat{\beta} - \widehat{\beta}_{(i)} - \eta_i\|$ is established, we use it to bound $\mathbf{E}\left(|\|\widehat{\beta} - \beta_0\|^2 - \|\widehat{\beta}_{(i)} - \beta_0\|^2|^2\right)$ and use a martingale inequality to deduce a bound on $\mathrm{var}\left(\|\widehat{\beta} - \beta_0\|^2\right)$, which we show goes to zero. The corresponding results are presented in Subsubsection 2.2.3 and the detailed mathematical analysis is in Section C.

Second step The second step of the proof is to relate $\widehat{\beta}$ to another quantity $\widehat{\gamma}$, which is the solution of our optimization problem when the last column of the matrix X is excluded from the problem - see Subsubsection 2.2.4 below and Section D for detailed mathematical analysis. Call V the corresponding design matrix. In our setting, it is reasonable to expect that $r_{i,[p]} = Y_i - X_i(p)\beta_0(p) - V_i'\widehat{\gamma} \simeq Y_i - X_i'\widehat{\beta}$. A first order Taylor expansion of $\nabla F(\widehat{\beta})$ around $(\widehat{\gamma}'\ \beta_0(p))'$ and further manipulations yields an informed "guess", denoted $\tilde{b}$ below, for $\widehat{\beta}$, and in particular for $\widehat{\beta}_p$, the last coordinate of $\widehat{\beta}$. A large amount of work is devoted to proving that the quantity denoted $\mathfrak{b}_p$ below we surmised approximates $\widehat{\beta}_p$ sufficiently well for our purposes - once again by doing delicate computations on the corresponding gradients. Since $\mathfrak{b}_p$ has a reasonably nice probabilistic representation, it is possible to write $\mathbf{E}\left(\mathfrak{b}_p^2\right)$ is terms of other quantities appearing in the problem, such as $\psi_i(r_{i,[p]})$ (where $\psi_i = \rho_i'$) and a quantity $\mathsf{c}_{\tau,p}$ that is the trace of the inverse of a certain random matrix. Because $\mathfrak{b}_p$ approximates $\widehat{\beta}_p$ sufficiently well, our approximation of $\mathbf{E}\left(\mathfrak{b}_p^2\right)$ can be used to yield a good approximation of $\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|^2\right)$. However, we want the approximation of $\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|^2\right)$ to not depend on quantities that depend on p, such as $r_{i,[p]}$ and $\mathsf{c}_{\tau,p}$. Further work is needed to show that the approximation of $\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|^2\right)$ can be made in terms of $\tilde{r}_{i,(i)}$'s - which we used in the first part of the proof - and a new quantity c_τ, which is the trace of the inverse of a certain random matrix, as was $\mathsf{c}_{\tau,p}$. The resulting approximation for $\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|^2\right)$ is essentially the second equation of our system - see Proposition (2.4) for instance.

Third step The last part of the proof - see Subsubsection 2.2.5 and Section E for detailed mathematical analysis - is devoted to first showing that $\tilde{r}_{i,(i)} = Y_i - X_i'\widehat{\beta}_{(i)}$ behaves asymptotically like $\epsilon_i + \lambda_i\sqrt{\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|^2\right)}Z_i$, where $Z_i \sim \mathcal{N}(0,1)$. The work done previously in the proof is extremely useful for that. Finally, we show that c_τ is asymptotically deterministic. The characterization of c_τ is essentially the first equation of our system - see Theorem 2.6 below. After all this is established, we can state for instance central limit theorems for $\widehat{\beta}_p$ and interesting quantities that appear in our proof.

The following few subsubsections make all our intermediate results precise. Armed with the above explanation for our approach, they provide the reader with a clear overview of the arc of our proof. The detailed mathematical analysis is given in the Appendix.

2.2.3 Leave-one-observation out approximations

We call the residuals

$$R_i = Y_i - X_i'\widehat{\beta} = \epsilon_i - X_i'(\widehat{\beta} - \beta_0)\ .$$

We consider the situation where we leave the i-th observation, (X_i, ϵ_i), out. We call

$$\widehat{\beta}_{(i)} = \mathrm{argmin}_{\beta\in\mathbb{R}^p} F_i(\beta)\ ,\ \text{where } F_i(\beta) = \frac{1}{n}\sum_{j\neq i}\rho_j(\epsilon_j + X_j'\beta_0 - X_j'\beta) + \frac{\tau}{2}\|\beta\|^2\ .$$

and

$$\tilde{r}_{j,(i)} = \epsilon_j - X_j'(\widehat{\beta}_{(i)} - \beta_0) \text{ and } S_i = \frac{1}{n}\sum_{j\neq i}\psi_j'(\tilde{r}_{j,(i)})X_jX_j'\ .$$

Note that $\tilde{r}_{j,(i)}$ are simply the leave-one-out residuals (for $j \neq i$) and the leave-one-out prediction error (for $j = i$).

Let us consider

$$\widetilde{\beta}_i = \widehat{\beta}_{(i)} + \frac{1}{n}(S_i + \tau \mathrm{Id})^{-1} X_i \psi_i(\mathrm{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) \triangleq \widehat{\beta}_{(i)} + \eta_i \ ,$$

where

$$c_i = \frac{1}{n} X_i'(S_i + \tau \mathrm{Id})^{-1} X_i \ , \text{ and } \eta_i = \frac{1}{n}(S_i + \tau \mathrm{Id})^{-1} X_i \psi_i(\mathrm{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) \ .$$

We have the following theorem.

Theorem 2.2. *Under our technical assumptions, we have, for any fixed k, when τ is held fixed,*

$$\sup_{1 \leq i \leq n} \|\widehat{\beta} - \widetilde{\beta}_i\| = \mathrm{O}_{L_k}(\frac{polyLog(n)}{n}) \ .$$

Also,

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

$$\sup_{1 \leq i \leq n} |R_i - prox(c_i \rho_i)(\tilde{r}_{i,(i)})| = \mathrm{O}_{L_k}(\frac{polyLog(n)}{n^{1/2}}) \ .$$

Finally,

$$\mathrm{var}\left(\|\widehat{\beta} - \beta_0\|_2^2\right) = \mathrm{O}\left(\frac{polyLog(n)}{n}\right)$$

A stronger version of this theorem is available in the Appendix. There are two main reasons this theorem is interesting: it provides online-update formulas for $\widehat{\beta}$ through $\widetilde{\beta}_i$, with guaranteed approximation errors. Second, it relates the full residuals, whose statistical and probabilistic properties are quite complicated to the much-simpler-to-understand "leave-one-out" prediction error, $\tilde{r}_{i,(i)}$. Indeed, because X_i is independent of $\widehat{\beta}_{(i)}$ under our assumptions, the statistical properties of $\widehat{\beta}_{(i)}' X_i$ are much simpler to understand than those of $\widehat{\beta}' X_i$.

2.2.4 Leave-one-predictor out approximations

Let V be 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 $V_i' = (X_i(1), \ldots, X_i(p-1))$. We call $X(p)$ the vector in $\mathbb{R}^n$ with j-th entry $X_j(p)$, i.e the $p-$th entry of the vector X_j. When this does not create problems, we also use the standard notation $X_{j,p}$ for $X_j(p)$.

Let us call $\widehat{\gamma}$ the solution of our optimization problem when we use the design matrix V instead of X. In other words,

$$\widehat{\gamma} = \mathrm{argmin}_{\gamma \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^n \rho_i(\epsilon_i - V_i'(\gamma - \gamma_0)) + \frac{\tau}{2}\|\gamma\|^2 \ . \tag{5}$$

For stating the following results, we will rely heavily on the following definitions:

Definition. We call the corresponding residuals $\{r_{i,[p]}\}_{i=1}^n$, i.e $r_{i,[p]} = \epsilon_i + V_i'\gamma_0 - V_i'\widehat{\gamma}$. Let

$$u_p = \frac{1}{n} \sum_{i=1}^n \psi_i'(r_{i,[p]}) V_i X_i(p) \ , \mathfrak{S}_p = \frac{1}{n} \sum_{i=1}^n \psi_i'(r_{i,[p]}) V_i V_i' \ .$$

We have $u_p \in \mathbb{R}^{p-1}$ and $\mathfrak{S}_p$ is $(p-1) \times (p-1)$. We call

$$\xi_n \triangleq \frac{1}{n} \sum_{i=1}^n X_i^2(p) \psi_i'(r_{i,[p]}) - u_p'(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p \ ,$$

$$N_p \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i(p) \psi_i(r_{i,[p]}) \ .$$

We consider

$$\mathfrak{b}_p \triangleq \beta_0(p)\frac{\xi_n}{\tau+\xi_n} + \frac{1}{\sqrt{n}}\frac{N_p}{\tau+\xi_n}\ . \tag{6}$$

Note that when $\xi_n > 0$, we have

$$\mathfrak{b}_p - \beta_0(p) = \frac{\frac{1}{n}\sum_{i=1}^n X_i(p)\psi_i(r_{i,[p]}) - \tau\mathfrak{b}_p}{\frac{1}{n}\sum_{i=1}^n X_i^2(p)\psi_i'(r_{i,[p]}) - u_p'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p} = \frac{n^{-1/2}N_p - \tau\mathfrak{b}_p}{\xi_n}\ .$$

We call

$$\widetilde{b} = \begin{bmatrix}\widehat{\gamma}\\ \beta_0(p)\end{bmatrix} + [\mathfrak{b}_p - \beta_0(p)]\begin{bmatrix}-(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\\ 1\end{bmatrix}\ . \tag{7}$$

Theorem 2.3. *Under our Assumptions, we have, for any fixed $\tau > 0$,*

$$\|\widehat{\beta} - \widetilde{b}\| \le \mathrm{O}_{L_k}\left(\frac{polyLog(n)}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right)$$

In particular,

$$\begin{aligned}
\sqrt{n}(\widehat{\beta}_p - \mathfrak{b}_p) &= \mathrm{O}_{L_k}\left(\frac{polyLog(n)n^{1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right)\ ,\\
\sup_i |X_i'(\widehat{\beta}-\widetilde{b})| &= \mathrm{O}_{L_k}\left(\frac{polyLog(n)n^{1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right)\ ,\\
\sup_i |R_i - r_{i,[p]}| &= \mathrm{O}_{L_k}\left(\left[\frac{polyLog(n)}{\sqrt{n}\wedge n^{\mathsf{e}}}\right]\vee\left[\frac{polyLog(n)n^{1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right]\right)\ .
\end{aligned}$$

Let us call

$$c_\tau = \frac{1}{n}\mathrm{trace}\left((S+\tau\mathrm{Id})^{-1}\right)\ ,\ \text{where } S = \frac{1}{n}\sum_{i=1}^n \psi_i'(R_i)X_iX_i'\ .$$

We also have

Proposition 2.4. *Under our assumptions, we have*

$$\left(\frac{p}{n}\right)^2\mathbf{E}\left(\|\widehat{\beta}-\beta_0\|_2^2\right) = \frac{p}{n}\frac{1}{n}\sum_{i=1}^n\mathbf{E}\left([c_\tau\lambda_i\psi_i(prox(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)}))]^2\right) + \tau^2\|\beta_0\|^2\mathbf{E}\left(c_\tau^2\right) + \mathrm{o}(1)\ .$$

Furthermore,

$$\sup_i |c_i - \lambda_i^2 c_\tau| = \mathrm{O}_{L_k}(n^{-1/2}polyLog(n))\ .$$

2.2.5 Final steps and related results

Lemma 2.5. *Under our assumptions , as n and p tend to infinity, $\tilde{r}_{i,(i)}$ behaves like $\epsilon_i + \lambda_i\sqrt{\mathbf{E}\left(\|\widehat{\beta}-\beta_0\|^2\right)}Z_i$, where $Z_i \sim \mathcal{N}(0,1)$ is independent of ϵ_i and λ_i, in the sense of weak convergence.*

Furthermore, if $i \neq j$, $\tilde{r}_{i,(i)}$ and $\tilde{r}_{j,(j)}$ are asymptotically (pairwise) independent. The same is true for the pairs $(\tilde{r}_{i,(i)}, \lambda_i)$ and $(\tilde{r}_{j,(j)}, \lambda_j)$

Theorem 2.6. *Under our assumptions, when $p/n \to \kappa \in (0,\infty)$, $\|\widehat{\beta}-\beta_0\|^2 \to r_\rho(\kappa)$, where $r_\rho(\kappa)$ is deterministic. Call $W_i = \epsilon_i + \lambda_i r_\rho(\kappa)Z_i$, where $Z_i \sim \mathcal{N}(0,1)$ independent of ϵ_i and λ_i. Call*

$$\begin{aligned}
\mathsf{G}_n(x) &= \frac{1}{n}\sum_{i=1}^n\mathbf{E}\left(\frac{1}{1+x\lambda_i^2\psi_i'(prox(x\lambda_i^2\rho_i)(W_i))}\right)\ \textit{and}\ \mathsf{G}(x) = \lim_{n\to\infty}\mathsf{G}_n(x)\ ,\\
\mathsf{H}_n(x) &= \frac{1}{n}\sum_{i=1}^n\mathbf{E}\left([x\lambda_i\psi_i(prox(x\lambda_i^2\rho_i)(W_i))]^2\right)\ \textit{and}\ \mathsf{H}(x) = \lim_{n\to\infty}\mathsf{H}_n(x)\ .
\end{aligned}$$

Under our assumptions, $c_\tau \to c_\rho(\kappa)$ in probability, where $c_\rho(\kappa)$ is the unique solution of the equation $\mathsf{G}(x) = 1 - \kappa + \tau x$. Furthermore, $r_\rho(\kappa)$ solves

$$\kappa^2 r_\rho^2(\kappa) = \kappa \mathsf{H}(c_\rho(\kappa)) + \tau^2 \|\beta_0\|^2 c_\rho^2(\kappa) \ .$$

We note that the equation $\mathsf{G}(x) = 1-\kappa+\tau x$ translates into the first equation of our system (4). The last equation of Theorem 2.6 is the second equation of our system (4). (The fact that the limits of G_n and H_n exist simply come from our assumptions that the proportion of times each possible triplet $(\rho_i, \mathcal{L}(\epsilon_i), \mathcal{L}(\lambda_i))$ appears has a limit as $n \to \infty$.)

From this main theorem follows the following propositions.

Proposition 2.7. *$\xi_n \to \xi$ in probability, where $\xi = \kappa / c_\rho(\kappa) - \tau > 0$.*
$N_p \Longrightarrow \mathcal{N}(0, v^2)$ where

$$v^2 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{E}\left(\lambda_i^2 \psi_i^2 [prox(c_\rho(\kappa)\lambda_i^2 \rho_i)(W_i)]\right)$$

Finally, when $\beta_0(k) = \mathrm{O}(n^{-1/2})$,

$$\sqrt{n}[(\tau + \xi)\widehat{\beta}_k - \beta_0(k)\xi] \Longrightarrow \mathcal{N}(0, v^2) \ .$$

The previous result is still true if v^2 is replaced by $\hat{v}_n^2 = \frac{1}{n}\sum_{i=1}^n \lambda_i^2 \psi_i^2 [prox(c_\tau \lambda_i^2 \rho_i)(\tilde{r}_{i,(i)})]$ and ξ is replaced by $\omega_n = p/(nc_\tau) - \tau$.

We note that ω_n is computable from the data. In our setup, λ_i's are estimable using the scheme proposed in El Karoui (2010) and $\hat{v}_n^2$ can therefore also be estimated from the data. Hence, the previous proposition allows for testing the null hypothesis that $\beta_0(k) = 0$, for any $1 \leq k \leq p$.

We are also now in position to explain the behavior of the residuals.

Proposition 2.8. *When our assumptions are satisfied and we further assume that λ_i's are uniformly bounded, we have*

$$\sup_{1\leq i\leq n} |R_i - prox(\lambda_i^2 c_\rho(\kappa)\rho_i)(\tilde{r}_{i,(i)})| = \mathrm{o}_{L_k}(1) \ .$$

The behavior of the residuals is therefore qualitatively very different in this high-dimensional setting than its counterpart in the low-dimensional setting.

2.3 Discussion of assumptions and results

2.3.1 Why consider elliptical-like predictors?

The study of elliptical distributions is quite classical in multivariate statistics (see Anderson (1984)). As pointed out by various authors (see in the context of statistics and random matrix theory Diaconis and Freedman (1984); Hall et al. (2005); El Karoui (2009)), the Gaussian distribution has a very peculiar geometry in high-dimension. It is therefore important to be able to study models that break away from these geometric restrictions, which are not particularly natural from the point of view of data analysts.

Under our assumptions, in light of Lemma G.3, it is clear that

$$\sup_{1\leq i\leq n} \left| \frac{\|X_i\|^2}{p} - \lambda_i^2 \right| = \mathrm{o}(1)$$
$$\sup_{i\neq j} \left| \frac{X_i' X_j}{p} \right| = \mathrm{o}(1) \ .$$

In the Gaussian (or Gaussian-like case of i.i.d entries for X_i, with e.g bounded entries which satisfy the assumptions we stated above), $\lambda_i = 1$. Hence, Gaussian or Gaussian-like assumptions imply that predictor vectors are situated near a sphere and are nearly orthogonal. (This simple geometry is of course closely

tied to - or a manifestation of the - concentration of measure for convex 1-Lipschitz functions of those random variables.)

This is clearly not the case for elliptical predictors, though under our assumptions, $\text{cov}(X_i) = \text{Id}_p$, even in the "elliptical" case we consider in the paper. So all the models we consider have the same covariance but the corresponding datasets may have different geometric properties.

We show in the paper that the role of the distribution of λ_i's in the performance of the estimator depends on much more than its second moment, as Theorem 2.1 makes very clear. This is a situation that is similar to corresponding results in random matrix theory - see e.g El Karoui (2009); El Karoui and Koesters (2011). It is therefore clear here again that predictor geometry (as measured by λ_i) plays a key role in the performance of our estimators in high-dimension, in contrast with the low-dimensional setting (see Huber and Ronchetti (2009) - which shows that in low-dimensional robust regression, what matters is only $\text{cov}(X_i)$).

These types of studies are also interesting and we think important as they clearly show that there is little hope of statistically meaningful "universality" results derived from Gaussian design results : moving from independent Gaussian assumptions for the entries of X_i to i.i.d assumptions does not change the geometry of the predictors, which appears to be key here as our proof's reliance on concentration of quadratic forms in $\mathcal{X}_i$ makes clear. As such, while very interesting on many counts, for instance to allow discrete predictors, moving from Gaussian to i.i.d assumptions is not a very significant perturbation of the model for statistical purposes. This is why we chose to work under elliptical assumptions. See also Diaconis and Freedman (1984) for similar observations in a different statistical context.

In conclusion, the generalized elliptical models we study in this paper prove also that many models may be such that the predictors have the same covariance $\text{cov}(X_i)$ but yield very different performance when it comes to $\lim\|\widehat{\beta} - \beta_0\|$. They therefore provide a meaningful perturbation of the Gaussian assumption, give us insights into the impact of predictor geometry on the behavior of our estimators, and give us a rough idea of the subclass of models for which we can expect similar (or "universal") performance for $\widehat{\beta}$.

Examples of distribution for $\mathcal{X}_i$ satisfying our concentration assumptions Corollary 4.10 in Ledoux (2001) shows that our assumptions are satisfied if $\mathcal{X}_i$ has independent entries bounded by $1/(2\sqrt{\mathsf{c}})$. Theorem 2.7 in Ledoux (2001) shows that our assumptions are satisfied if $\mathcal{X}_i$ has independent entries with density f_k, $1 \leq k \leq p$ such that $f_k(x) = \exp(-u_k(x))$ and $u_k''(x) \geq \sqrt{c}$ for some $c > 0$. Then $\mathsf{c} = c/2$. This is in particular the case for the case where $\mathcal{X}_i$ has i.i.d $\mathcal{N}(0,1)$ entries: then $c = 1$ and $\mathsf{c} = 1/2$. We discuss briefly after Lemma G.1 in the Appendix the impact of choosing other types of concentration assumptions.

2.3.2 Non-sparse β_0: why consider ℓ_2/ridge-regularization?

In this paper, we consider the case where β_0 cannot - in general - be approximated in ℓ_2-norm by a sparse vector. This is a situation that is thought to not be uncommon in biology (see, in a slightly different context, Michael B. Eisen and Botstein (1998) and many similar references), where sparsity assumptions are often/sometimes in doubt.

In other words, if s is a sparse vector (e.g with support of size $o(p)$), we necessarily have when β_0 is diffuse (i.e all of its entries are roughly of size $p^{-1/2}$) $\|\beta_0 - \mathsf{s}\| \not\to 0$. In the situation we consider, it is in fact unclear whether any estimator can be consistent in ℓ_2 for β_0. One interesting aspect of our study is that the System (4) might allow us to optimize (at least in certain circumstances) over the functions ρ_i's we consider to get the best performing estimator in the class of ridge-regularized robust regression estimators for β_0 and hence potentially beat sparse estimators (in the same line of thought, there are of course numerous applied examples where ridge regression outperforms Lasso in terms of prediction error).

Finally, one benefit of our analysis is that we have a central limit theorem for a linear transform of $\widehat{\beta}_p$ (see Proposition 2.7), which makes testing possible. In the situations where β_0 has some large entries (of size up to $n^{-1/4-\eta}$, $\eta > 0$) and many small ones (of size $o(n^{-1/2})$), this central limit theorem and its more refined version in Proposition E.8 could help in designing better performing estimators by using scaled versions of $\widehat{\beta}_p$, which we would threshold according to the result of our test.

2.3.3 A remark on the fixed design case

We have worked in this paper with a certain class of random designs. It is not unusual to do so in robust regression studies - see the classic Portnoy (1987, 1984, 1985). In many areas of applications, it is also unclear why statisticians should limit themselves to the study of fixed designs, in particular when they do not have control over the choice of the values of the predictors, i.e they cannot design their experiments.

However, it is also interesting to understand what remains valid of our analysis in the case of fixed design. We note that our analysis gives already a few results in this direction.

In fact, since we have shown that $\text{var}\left(\|\widehat{\beta}-\beta_0\|^2\right) \to 0$, we have shown that

$$\mathbf{E}\left(\text{var}\left(\|\widehat{\beta}-\beta_0\|^2|X\right)\right) \to 0 \text{ and } \text{var}\left(\mathbf{E}\left(\|\widehat{\beta}-\beta_0\|^2|X\right)\right) \to 0 \,,$$

since

$$\text{var}\left(\|\widehat{\beta}-\beta_0\|^2\right) = \mathbf{E}\left(\text{var}\left(\|\widehat{\beta}-\beta_0\|^2|X\right)\right) + \text{var}\left(\mathbf{E}\left(\|\widehat{\beta}-\beta_0\|^2|X\right)\right) \,.$$

Therefore, with probability (over the design X) going to 1,

$$\|\widehat{\beta}-\beta_0\|^2 - r_\rho(\kappa) \to 0 \text{ in } P_{\{\epsilon_i\}_{i=1}^n}\text{-probability} \,.$$

($P_{\{\epsilon_i\}_{i=1}^n}$-probability simply refers to probability statements with respect to the random ϵ_i's, the only source of randomness if the design matrix X is assumed to be fixed.) In other words, if the design is fixed, but results from a draw a random of a $n \times p$ matrix satisfying our distributional assumptions, Theorem 2.1 applies with probability (over the choice of design matrix) going to 1.

We note that $\|\widehat{\beta}-\beta_0\|$ is an especially important quantity in terms of prediction error in our context, which why our short discussion above focused on this quantity: if we are given a new predictor vector X_{new}, we would naturally predict an unobserved response Y_{new} by $X'_{new}\widehat{\beta}$ and hence, if $Y_{new} = \epsilon_{new} + X'_{new}\beta_0$, our prediction error will be $PE_{new} = \epsilon_{new} + X'_{new}(\beta_0 - \widehat{\beta})$. (Of course, if X_{new} satisfies $\text{cov}\,(X_{new}) = \text{Id}_p$, $\mathbf{E}_{X_{new}}\left[X'_{new}(\beta_0-\widehat{\beta})\right] = \|\beta_0-\widehat{\beta}\|_2^2$.)

2.3.4 Optimization with respect to τ and ρ

Just as the classic work of Huber on robust regression started by establishing central limit theorems for the estimator of interest (as a function of ρ) and proceeded to find optimal methods in various contexts (see Huber and Ronchetti (2009)), one objective of our work is to pave the way for answering optimality questions in the setting we consider. An important first step to do so is therefore to obtain results such as Theorem 2.1.

A natural question is therefore to ask what are the optimal ρ_i's in the context we consider, where optimality might be defined in terms of minimizing $r_\rho(\kappa)$ in Theorem 2.1 or v^2 in Proposition 2.7. For an example of such a study for $r_\rho(\kappa)$ in a slightly different context, see Bean et al. (2013). Similarly, optimization over τ should be possible. We leave however these questions for future work, since they are of a more analytic nature. (We have had success in Bean et al. (2013) in the situation where $\lambda_i = 1$ and the errors are log-concave and hence not heavy-tailed, but the technique we employed in that paper does not apply readily here.)

We also note that in our context the optimal τ is in general not going to be close to 0, so the fact that our current study requires $\tau > 0$ is not a problem (see also Bean et al. (2012)).

2.3.5 Possible extensions

Less smooth ρ's and ψ's While our approach is quite general and allows us to handle designs that are far from being Gaussian, the proof presented in this paper still requires some smoothness concerning ρ_i's and ψ_i's. On the one hand, results such as the ones obtained in Bean et al. (2013) suggest that it is often the case that optimal loss functions in high-dimension are smoother than in low dimension. So the fact that we require ψ_i's to be smooth is a source of less concern that it would be in low dimension. (Note also that the classic papers Portnoy (1985); Mammen (1989) also require smoothness properties on ψ.)

Though it is unclear whether the Huber function is optimal in any sense for the problems we are looking at, and hence whether it warrants a special focus, let us discuss this function in some detail. For the sake of simplicity let us focus on the situation where the transition from quadratic to linear happens at $x = \pm 1$. Then

$$\psi(x) = \begin{cases} x & \text{if } |x| \leq 1 \\ \text{sign}(x) & \text{if } |x| \geq 1 \end{cases} .$$

So ψ is not differentiable at 1. However, it is easy to approximate this function by a function whose derivative is Lipschitz. As a matter of fact, ψ'_η such that

$$\psi'_\eta(x) = \begin{cases} 1 & \text{if } |x| \leq 1-\eta \\ \frac{1-|x|}{\eta} & \text{if } |x| \in (1-\eta, 1) \\ 0 & \text{if } |x| \geq 1 \end{cases} ,$$

is $1/\eta$-Lipschitz. Furthermore,

$$\psi_\eta(x) = \begin{cases} x & \text{if } |x| \leq 1-\eta \\ 1-\eta+\delta(x)(\eta - \frac{\delta(x)}{2}) & \text{if } |x| = 1-\eta+\delta(x), 0 \leq \delta(x) \leq \eta \\ 1-\eta+\frac{\eta^2}{2}\text{sign}(x) & \text{if } |x| \geq 1 \end{cases} ,$$

can be made to be arbitrarily close to ψ and similarly for the corresponding ρ_η, picked such that $\rho_\eta(0) = 0$.

Our results apply to ρ_η, for any $\eta > 0$. It seems quite likely that with a bit (and possibly quite a bit) of further approximation theoretic work, it should be possible to establish results similar to Theorem 2.1 for the Huber function by taking the limit of corresponding results for ρ_η with η arbitrarily small.

We note that most of our proof (in particular Sections C and D) is actually valid with functions ρ_i's that can change with n. In particular, many results hold when ψ'_i are $L_i(n)$-Lipschitz with $L_i(n) \leq Cn^\alpha$. So one strategy to handle the case of the Huber function could be to use ψ_{η_n} with $\eta_n = 1/\log(n)$ for instance and strengthen the arguments of Section E in the Appendix - in this very specific case where ψ_{η_n} has a limit - to get the Huber case as a limiting result. Because our proof is already long, we leave the details to the interested reader and might consider this problem in detail in future work.

Weighted robust regression One motivation for working on the problem at the level of generality we dealt with is that our results should allow to tackle among other things weighted robust regression. For instance if ϵ_i's or λ_i's in our model had different distributions, it would be natural to pick the corresponding ρ_i's either as completely different functions, or maybe as $\rho_i = w_i\rho$, with w_i deterministic but possibly depending on the distribution of ϵ_i's or λ_i's. In the case where ϵ_i's and λ_i's come from finitely many possible distributions, our results handle this situation.

Most of our results - i.e those of Section C and D - are true even when w_i's are allowed to take a possibly infinite set of different values. If ϵ_i's are i.i.d, λ_i's are i.i.d and w_i's are i.i.d and these three groups of random variables are independent of each other, our arguments can be made to go through without much extra difficulties. The main potential problem is in Section E, but then distributional symmetry between the R_i's on one hand and the $\tilde{r}_{i,(i)}$ on the other hand becomes helpful, as it had in El Karoui (2013). So it is very likely that our results could be extended to cover this case at relatively little technical cost.

3 Conclusion

We have studied ridge-regularized robust regression estimators in the high-dimensional context where p/n has a finite non-zero limit. Our study has highlighted the importance of the geometry of the predictors in this problem: two models with similar covariance but different predictor geometry will in general yield estimators with very different performance. We have shown this result by studying the random design case in the context of elliptical predictors and looking at the influence of the "ellipticity parameter" λ_i on our results. Importantly, this shows that no statistically meaningful "universality" results can be derived from the study of Gaussian or i.i.d-designs, since their geometry is so peculiar (i.e they are limited to the case

$\lambda_i = 1$ for all i's). The technique used in the paper seems versatile enough to be useful for several other high-dimensional M-estimation problems.

We have also obtained central limit theorems for the coordinates of $\widehat{\beta}$ that can be used for testing whether $\beta_0(k) = 0$ for any $1 \leq k \leq p$. However, our interest was mostly in the case where β_0 is diffuse, with all coordinates small but contributing to $Y_i = \epsilon_i + X_i'\beta_0$. Our results also provide a very detailed understanding of the properties of the residuals R_i.

All these results were obtained without moment requirements on the errors ϵ_i's.

Finally, our characterization of the risk of these estimators raise interesting analytic questions related to finding optimal loss functions ρ_i's in the context we consider. We plan to study these questions in the future.

APPENDIX

A Assumptions and technical elements

Recall that the focus of the paper is on understanding the properties of

$$\widehat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_i(\epsilon_i - X_i'(\beta - \beta_0)) + \frac{\tau}{2}\|\beta\|^2 \tag{A-1}$$

where $\tau > 0$. For all $1 \leq i \leq n$, we have $\epsilon_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$.

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

Most of our proof (Sections C and D) is carried out for functions $\rho_{i,n}$ that may vary with n, so our assumptions reflect this and we carry out most of our work at this level of generality. However, we do not make the dependence of $\rho_{i,n}$ on n explicit to avoid cumbersome notations. Having these results available should make future work on weighted regression or work of a more approximation-theoretic nature (for instance using a sequence $\rho_{i,n}$ to approximate a function ρ_i that is not smooth) easier. This is one of the prime motivations for working at this level of generality.

Naturally, our assumptions are more and more restrictive as the proof progresses, so the summary of assumptions we provided in the main text is obtained by going through the assumptions and simply tallying the more restrictive ones. A sketch of proof is provided in Subsubsection 2.2.2, which should be helpful in navigating the detailed proof we provide in this appendix.

Before we delve into the details of the assumptions needed for each part of the proof to work, we summarize for the convenience of the reader the assumptions we need for the whole proof to go through.

Assumptions under which the whole proof goes through

- **A1** p/n has a finite non-zero limit
- **A2** $\|\beta_0\|$ remains bounded. Furthermore, $\|\beta_0\|_\infty = \mathrm{O}(n^{-\mathsf{e}})$, for $\mathsf{e} > 1/4$
- **A3** ρ_i's are twice differentiable, and convex. If $\psi_i = \rho_i'$, we assume that $\text{sign}(\psi_i(x)) = \text{sign}(x)$ and $\rho_i \geq 0 = \rho_i(0)$. Furthermore, there exists C such that $\|\psi_i\|_\infty \leq C$, $\|\psi_i'\|_\infty \leq C$ and ψ_i' is assumed to be Lipschitz. The functions ρ_i's can be chosen among finitely many possible functions.
- **A4** $X_i = \lambda_i \mathcal{X}_i$. λ_i's are random variables with $\lambda_i \in \mathbb{R}$. $\mathcal{X}_i \in \mathbb{R}^p$ are independent and identically distributed. Their distribution is allowed to change with p and n. The entries of $\mathcal{X}_i$ are independent. Furthermore, for any 1-Lipschitz (with respect to Euclidean norm) convex function G, if $m_{G(\mathcal{X}_i)}$ is a median of $G(\mathcal{X}_i)$, for any $t > 0$, $P(|G(\mathcal{X}_i) - m_{G(\mathcal{X}_i)}| > t) \leq C_n \exp(-\mathsf{c}_n t^2)$, C_n and c_n can vary with n. For simplicity, we assume that $1/\mathsf{c}_n = \mathrm{O}(\text{polyLog}(n))$ and C_n is bounded in n. $\mathcal{X}_i$'s have mean 0 and $\text{cov}(\mathcal{X}_i) = \text{Id}_p$. We also assume that the coordinates of $\mathcal{X}_i$ have moments of all order. Furthermore, for any given k, the k-th moment of the entries of $\mathcal{X}_i$ is assumed to be bounded independently of n and p. Also, for any $1 \leq k \leq p$, the vectors $\Theta_k = (\mathcal{X}_1(k), \ldots, \mathcal{X}_n(k))$ in $\mathbb{R}^n$ satisfy: for any 1-Lipschitz (with respect to Euclidean norm) convex function G, if $m_{G(\Theta_k)}$ is a median of $G(\Theta_k)$, for any $t > 0$, $P(|G(\Theta_k) - m_{G(\Theta_k)}| > t) \leq C_n \exp(-\mathsf{c}_n t^2)$, C_n and c_n can vary with n. As above, we assume that $1/\mathsf{c}_n = \mathrm{O}(\text{polyLog}(n))$.
- **A5** λ_i's are independent of each other and $\{\mathcal{X}_i\}_{i=1}^n$. $\mathbf{E}\left(\lambda_i^2\right) = 1$ and $\mathbf{E}\left(\lambda_i^4\right) \leq C$ and $\sup_{1 \leq i \leq n} |\lambda_i| = \mathrm{O}_{L_k}(\text{polyLog}(n))$.
- **A6** ϵ_i's are independent $\{\mathcal{X}_i\}_{i=1}^n$ and $\{\lambda_i\}_{i=1}^n$ and of each other. Furthermore, for any $r \in \mathbb{R}$, if $Z \sim \mathcal{N}(0,1)$, independent of ϵ_i, $\epsilon_i + rZ$ has a (differentiable) density $f_{i,r}$ which is increasing on $(-\infty, 0)$ and decreasing on $(0, \infty)$. Finally, $\lim_{|t| \to \infty} t f_{i,r}(t) = 0$.

- **A7** λ_i's can have different distributions. Similarly, ϵ_i's can have different distributions. However, the number of choices for the triplet $(\rho_i, \mathcal{L}(\lambda_i), \mathcal{L}(\epsilon_i))$ is finite. Furthermore, the fraction of times each such triplet appears in our problem - see Equation (A-1) has a limit. ($\mathcal{L}(\epsilon_i)$ just means the law of ϵ_i.)

We note that the entries of $\mathcal{X}_i$ do not need to have the same distribution. Our condition **A4** is satisfied when $\mathcal{X}_i$ have i.i.d $\mathcal{N}(0,1)$ entries, or independent entries that are bounded by a constant and have mean 0 and variance 1. (See Ledoux (2001), Corollary 4.10.)

Condition **A7** just means that if for instance $\rho_i = \rho$ and λ_i's are i.i.d but ϵ_i's can come from 3 distributions, the fraction of ϵ_i's coming from each of these three distributions has a limit as $n \to \infty$. (This last condition mostly plays a role in guaranteeing that we can take limits in various expressions.) The simplest case is of course when $\rho_i = \rho$, λ_i's are i.i.d and ϵ_i's are i.i.d, in which case there is only one possible choice for the triplet $(\rho_i, \mathcal{L}(\lambda_i), \mathcal{L}(\epsilon_i))$.

We now state the conditions under which we carry out the proof. We state them in one place for the convenience of the reader. A discussion follows immediately after the statement of all the conditions.

First part of the proof (Section C)

For the first part of the proof (i.e "leave-one-**O**bservation-out"), we work under the following assumptions:

- **O1**: p/n has a finite non-zero limit.
- **O2**: ρ_i's are twice differentiable, convex and non-linear. $\psi_i = \rho_i'$. Note that $\psi_i' \geq 0$ since ρ_i is convex. We assume that $\text{sign}(\psi_i(x)) = \text{sign}(x)$ and $\rho_i \geq 0 = \rho_i(0)$.
- **O3**: $\sup_{x,i} |\psi_i(x)| \leq C\text{polyLog}(n)$ where C is constant. This is natural in the context of robust statistics, since it means that we allow ρ_i's to grow at most linearly at infinity. This assumption is for instance verified for Huber functions. Furthermore, ψ_i' is assumed to be $\mathsf{L}_i(n)$-Lipschitz with $\mathsf{L}_i(n) \leq Cn^\alpha$, $\alpha \geq 0$. (We here have in mind smoothed Huber functions.) We also assume that $\sup_i \|\psi_i'\|_\infty \leq C\text{polyLog}(n)$. Finally, we assume that $\frac{1}{n}\sum_{i=1}^n \|\psi_i^2\|_\infty \leq C$, where C is a constant independent of n.
- **O4**: $X_i = \lambda_i \mathcal{X}_i$. λ_i's are random variables with $\lambda_i \in \mathbb{R}$. $\mathcal{X}_i \in \mathbb{R}^p$ are independent and identically distributed. Their distribution is allowed to change with p and n. Furthermore, for any 1-Lipschitz (with respect to Euclidean norm) convex function G, if $m_{G(\mathcal{X}_i)}$ is a median of $G(\mathcal{X}_i)$, for any $t > 0$, $P(|G(\mathcal{X}_i) - m_{G(\mathcal{X}_i)}| > t) \leq C_n \exp(-\mathsf{c}_n t^2)$, C_n and c_n can vary with n. For simplicity, we assume that, $1/\mathsf{c}_n = \text{O}(\text{polyLog}(n))$ and C_n is bounded in n. $\mathcal{X}_i$'s have mean 0 and $\text{cov}(\mathcal{X}_i) = \text{Id}_p$. We also assume that the coordinates of $\mathcal{X}_i$ have moments of all order. Furthermore, for any given k, the k-th moment of the entries of $\mathcal{X}_i$ is assumed to be bounded independently of n and p. $\{\mathcal{X}_i\}_{i=1}^n$ and $\{\lambda_i\}_{i=1}^n$ are independent.
- **O5**: $\{\mathcal{X}_i\}_{i=1}^n$ and $\{\lambda_i\}_{i=1}^n$ are independent of $\{\epsilon_i\}_{i=1}^n$. ϵ_i's are independent of each other.
- **O6**: $\sup_{1\leq i\leq n} |\lambda_i| \triangleq \mathcal{L}_n = \text{O}_{L_k}((\log n)^\beta)$ and λ_i's are independent. Furthermore, $\mathbf{E}\left(\lambda_i^2\right) = 1$. (Note that this implies that $\text{cov}(X_i) = \text{cov}(\mathcal{X}_i)$.)
- **O7**: $1 - 2\alpha > 0$ and $\|\beta_0\| = \text{O}(\text{polyLog}(n))$.

Note that we do not assume that ϵ_i's have identical distributions. Assumption **O4** is satisfied for instance when $\mathcal{X}_i$ are for instance $\mathcal{N}(0, \text{Id}_p)$ or have i.i.d entries bounded by $\text{polyLog}(n)$ - see Ledoux (2001) (this reference guarantees the concentration result we require is satisfied; the moment conditions need to be checked by other methods, but this is generally much simpler, as the case of Gaussian random variables clearly shows). Importantly, note that **O4** does not require the entries of $\mathcal{X}_i$ to be independent; see Ledoux (2001) or El Karoui (2009) for examples of $\mathcal{X}_\rangle$ satisfying **O4** with dependent entries. In other respects, the assumption $\mathbf{E}\left(\lambda_i^2\right) = 1$ plays a very minor role mathematically and could be relaxed to $\mathbf{E}\left(\lambda_i^2\right)$ is uniformly bounded without problems. Statistically, it is however important as it guarantees that $\text{cov}(X_i) = \text{Id}_p$ in all the models we consider.

Second part of the proof (Section D)

For the second part of the proof (i.e "leave-one-**P**redictor-out"), we need all the previous assumptions and

- **P1**: $\mathcal{X}_i$'s have independent entries. Furthermore, for $1 \leq k \leq p$, the vectors $\Theta_k = (\mathcal{X}_1(k), \ldots, \mathcal{X}_n(k))$ in $\mathbb{R}^n$ satisfy: for any 1-Lipschitz (with respect to Euclidean norm) convex function G, if $m_{G(\Theta_k)}$ is a median of $G(\Theta_k)$, for any $t > 0$, $P(|G(\Theta_k) - m_{G(\Theta_k)}| > t) \leq C_n \exp(-\mathsf{c}_n t^2)$, C_n and c_n can vary with n. As above, we assume that $1/\mathsf{c}_n = \mathrm{O}(\mathrm{polyLog}(n))$.
- **P2**: $\frac{1}{n}\sum_{i=1}^n \|\psi_i'\|_\infty = \mathrm{O}(1)$.
- **P3**: $\|\beta_0\|_\infty = \mathrm{O}(n^{-\mathsf{e}})$, where $\mathsf{e} > 0$. Furthermore, $\|\beta_0\|_2 \leq C$, where C is a constant independent of p and n. e satisfies $\alpha + 1/4 - \mathsf{e} < 0$.
- **P4**: $1/2 - 2\alpha > 0$ and $\min(1/2, \mathsf{e}) - \alpha - 1/4 > 0$. The latter implies that $\min(1/2, \mathsf{e}) - \alpha > 0$

We note that according to Corollary 4.10 and the discussion that follows in Ledoux (2001), Assumptions **O4** and **P1** are compatible. **O4** and **P1** are for instance satisfied if the entries of $\mathcal{X}_i$'s are independent and bounded by $\mathrm{polyLog}(n)$. Another example is the case of $\mathcal{X}_i \sim \mathcal{N}(0, \mathrm{Id}_p)$, in which case c_n is a constant independent of the dimension.

Note that we do not assume that the entries of $\mathcal{X}_i$ have the same distribution.

We note that if for instance $\alpha = 1/12$ and $\mathsf{e} = 5/12$, all the conditions in **P3-P4** are satisfied. When $\alpha = 0$, they simply become $\mathsf{e} > 1/4$.

Last part of the proof (Section E)

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 ϵ_i's may have different distributions; however, they may only come from finitely many distributions. Furthermore, for any $r \in \mathbb{R}$, if $Z \sim \mathcal{N}(0,1)$, independent of ϵ_i, $\epsilon_i + rZ$ has a differentiable density $f_{i,r}$ which is increasing on $(-\infty, 0)$ and decreasing on $(0, \infty)$. Finally, $\lim_{|t|\to\infty} t f_{i,r}(t) = 0$.
- **F2**: $\frac{1}{n}\sum_{i=1}^n \|\psi_i\|_\infty = \mathrm{O}(1)$. ψ_i' has Lipschitz constant $\mathsf{L}_i(n)$. Furthermore, $\frac{1}{n}\sum_{i=1}^n \mathsf{L}_i(n)\|\psi_i\|_\infty = \mathrm{O}(1)$.
- **F3**: $\alpha < 1/6$ and $\alpha + 1/3 < 2\min(1/2, \mathsf{e})$
- **F4**: there exists C independent of n and p such that $\mathbf{E}\left(\lambda_i^4\right) \leq C$.
- **F5**: λ_i's may have different distributions, but the set of possible distributions for λ_i is finite. Similarly, ρ_i may be different functions, but the set of possible functions ρ_i may be is finite. Also, the number of distinct triplets $(\rho_i, \mathcal{L}(\epsilon_i), \mathcal{L}(\lambda_i))$ is finite. Furthermore, the proportion of each such distinct triplet has a limit as $n \to \infty$.

Condition **F3** is clearly satisfied in the case $\alpha = 1/12$ and $\mathsf{e} = 5/12$ we mentioned above. On the other hand, condition **F5** requires that $\alpha = 0$, since it prevents ρ_i to change with n. (We note that since **F5** is required only at the very end of the proof, one could probably weaken its requirements considerably if another situation that the one we investigate really called for it.)

We refer the reader to Lemma H.1 and the discussion immediately following it for examples of densities for ϵ_i's satisfying **F1**. We note that smooth symmetric (around 0) log-concave densities will for instance satisfy all the assumptions we made about the ϵ_i's. See Karlin (1968) and Ibragimov (1956) for instance. This is also the case for the Cauchy distribution (see Theorem 1.6 in Dharmadhikari and Joag-Dev (1988)). The latter is the most relevant reference here since we care about heavy-tailed ϵ_i's.

Discussion of the assumptions

Assumptions concerning the loss functions We wanted to investigate in this paper the situation where ϵ_i's have no moment restrictions, as befits "classical" robust statistics studies. As such, it is natural to assume that ψ_i's remain bounded, which is part of assumption **A3**. We think that interesting results can be found in Sections C and D: in particular for dealing with ψ_i functions that are not smooth and require approximations by $\psi_{i,n}$ functions that are smooth, but also for bootstrap studies for instance. This is why the paper handles those cases, even though we do not use fully these results in the main statements of the paper. We note that in specific cases, one would simply need to modify various arguments in Section E to handle limits of those $\psi_{i,n}$.

Assumptions concerning the predictors Assumption **O4** is a bit stronger than we will need. For instance, Sections C and D do not actually require the $\mathcal{X}_i$'s to have identical distributions; Section E would work if we assumed that $\mathcal{X}_i$'s were coming from finitely many distributions, with the proportion of $\mathcal{X}_i$'s picked from a particular distribution having a limit as $n \to \infty$. The functions G we are working with will either be linear or square-root of quadratic forms, so we could limit our assumptions to those functions. However, as documented in Ledoux (2001) and discussed briefly in the introduction, 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 $\mathcal{X}_i$'s - that guarantees that **O4** is satisfied when $\mathcal{X}_i$ has independent entries - and replace it by an assumption concerning the moments of the entries of $\mathcal{X}_i$'s and a truncation of triangular arrays argument (see for instance Yin et al. (1988)). We also refer the interested reader to El Karoui (2009) for a short list of distributions satisfying **O4**, compiled from various parts of Ledoux (2001). Finally, we could replace the $\exp(-\mathsf{c}_n t^2)$ upper bound in **O4** by $\exp(-\mathsf{c}_n t^\beta)$ for some fixed $\beta > 0$ and it seems that all our arguments would go through. We chose not to 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. Importantly, **O4** allows the entries of $\mathcal{X}_i$'s to be dependent.

To give a concrete example, let us consider the situation where the entries of $\mathcal{X}_i$ are independent and symmetric with an exponential density chosen to have variance 1. Then it is clear that $\sup_{i,j} |\mathcal{X}_i(j)| \leq K[\log(n)]^2$ almost surely as $n, p \to \infty$. Our analysis and assumptions then apply to the predictors $\mathsf{X}_i = \lambda_i s_n \Gamma_{i,n}$, with $\Gamma_{i,n} = \mathcal{X}_i 1_{\|\mathcal{X}_i\|_\infty \leq K[\log(n)]^2}$ where s_n is chosen so that $1/s_n^2 = \mathrm{var}(\Gamma_{i,n})$ (the variance of the entries of $\Gamma_{i,n}$ is not 1, since it is a truncation of $\mathcal{X}_i$ but it is easy to see that $\mathrm{var}(\Gamma_{i,n}) \to 1$). Note that $\mathsf{X}_i = X_i$ almost surely, and therefore our statistical problem is not affected. Very minor modifications to the arguments of Section E are then needed to handle s_n and show that our results go through. Naturally the same argument could be made for other (non-exponential) distributions as long as $\sup_{i,j} |\mathcal{X}_i(j)| \leq K[\log(n)]^2$. We note that our method should also be able to handle c_n such that $1/\mathsf{c}_n$ grows faster that polyLog(n) and hence deal with an even broader class of predictor distributions, but we chose not to do it to limit the book-keeping burden in an already long proof.

Notations

We will repeatedly use the following notations: $Y_i = X_i'\beta_0 + \epsilon_i$; polyLog$(n)$ is used to replace a power of $\log(n)$; $\lambda_{\max}(M)$ denotes the largest eigenvalue of the matrix M; $|||M|||_2$ denotes the largest singular value of M. We call $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$ the usual sample covariance matrix of the X_i's when X_i's are known to have mean 0. We say that $X \leq Y$ in L_k if $\mathbf{E}\left(|X|^k\right) \leq \mathbf{E}\left(|Y|^k\right)$. We write $X \overset{\mathcal{L}}{=} Y$ to say that the random variables X and Y are equal in law. We use the usual notation $\widehat{\beta}_{(i)}$ to denote the regression vector we obtain when we do not use the pair (X_i, Y_i) or (X_i, ϵ_i) in our optimization problem, a.k.a the leave-one-out estimate. 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 whose i-th row is X_i'. We write $a \wedge b$ for $\min(a, b)$ and $a \vee b$ for $\max(a, b)$. If A and B are two symmetric matrices, $A \succeq B$ means that $A - B$ is positive semi-definite, i.e A is greater than B in the positive-definite/Loewner order. The notations o_P, O_P are used with their

standard meanings, see e.g van der Vaart (1998) p.12 for definitions. For the random variable W, we use the definition $\|W\|_{L_k} = \left[\mathbf{E}\left(|W|^k\right)\right]^{1/k}$. For sequences of random variables W_n, Z_n, we use the notation $W_n = \mathrm{O}_{L_k}(Z_n)$ (resp $W_n = \mathrm{o}_{L_k}(Z_n)$) when $\|W_n\|_{L_k} = \mathrm{O}(\|Z_n\|_{L_k})$ (resp $\|W_n\|_{L_k} = \mathrm{o}(\|Z_n\|_{L_k})$). For a vector v in $\mathbb{R}^p$, $\|v\|$ is its Euclidean norm, whereas $\|v\|_\infty = \max_{1\le k\le p}|v(k)|$. For a function f from $\mathbb{R}$ to $\mathbb{R}$, $\|f\|_\infty = \sup_{x\in\mathbb{R}}|f(x)|$.

Remarks

We call

$$F(\beta) = \frac{1}{n}\sum_{i=1}^{n}\rho_i(\epsilon_i + X_i'\beta_0 - X_i'\beta) + \frac{\tau}{2}\|\beta\|^2 \ . \tag{A-2}$$

Note that under our assumptions on ρ, $\widehat{\beta}$ is defined as the solution of

$$f(\widehat{\beta}) = 0 \text{ with} \tag{A-3}$$

$$\nabla F = f(\beta) = \frac{1}{n}\sum_{i=1}^{n} -X_i\psi_i(\epsilon_i + X_i'\beta_0 - X_i'\beta) + \tau\beta \ . \tag{A-4}$$

Recall the important definitions:

Definition. We call

$$R_i = \epsilon_i + X_i'\beta_0 - X_i'\widehat{\beta} \text{ (i.e the residuals)}, \tag{A-5}$$

$$S = \frac{1}{n}\sum_{i=1}^{n}\psi_i'(R_i)X_iX_i' \ , \tag{A-6}$$

$$c_\tau = \frac{1}{n}\mathrm{trace}\left(S + \tau\mathrm{Id}\right)^{-1} \ . \tag{A-7}$$

B Preliminaries

B-1 General remarks

Proposition B.1. *Let β_1 and β_2 be two vectors in $\mathbb{R}^p$. Then, when ρ_i's are convex and twice-differentiable,*

$$\boxed{\|\beta_1 - \beta_2\| \le \frac{1}{\tau}\|f(\beta_1) - f(\beta_2)\| \ .} \tag{B-8}$$

Proof. Let β_1 and β_2 be two vectors in $\mathbb{R}^p$. We have by definition

$$f(\beta_1) - f(\beta_2) = \tau(\beta_1 - \beta_2) + \frac{1}{n}\sum_{i=1}^{n} X_i\left[\psi_i(\epsilon_i + X_i'\beta_0 - X_i'\beta_2) - \psi_i(\epsilon_i + X_i'\beta_0 - X_i'\beta_1)\right] \ .$$

We can use the mean value theorem to write

$$\psi_i(\epsilon_i + X_i'\beta_0 - X_i'\beta_2) - \psi_i(\epsilon_i + X_i'\beta_0 - X_i'\beta_1) = \psi_i'(\gamma^*_{\epsilon_i + X_i'\beta_0, X_i'\beta_1, X_i'\beta_2})X_i'(\beta_1 - \beta_2) \ ,$$

where $\gamma^*_{\epsilon_i + X_i'\beta_0, X_i'\beta_1, X_i'\beta_2}$ is in the interval $(\epsilon_i + X_i'\beta_0 - X_i'\beta_1, \epsilon_i + X_i'\beta_0 - X_i'\beta_2)$ - recall that we do not care about the order of the endpoints in our notation.

Hence,

$$f(\beta_1) - f(\beta_2) = \tau(\beta_1 - \beta_2) + \frac{1}{n}\sum_{i=1}^{n}\psi_i'(\gamma^*_{\epsilon_i + X_i'\beta_0, X_i'\beta_1, X_i'\beta_2})X_iX_i'(\beta_1 - \beta_2) \ ,$$

which we write

$$f(\beta_1) - f(\beta_2) = (\mathsf{S}_{\beta_1,\beta_2} + \tau\mathrm{Id}_p)(\beta_1 - \beta_2) \ , \tag{B-9}$$

where

$$\mathsf{S}_{\beta_1,\beta_2} = \frac{1}{n}\sum_{i=1}^{n} \psi_i'(\gamma^*_{\epsilon_i + X_i'\beta_0, X_i'\beta_1, X_i'\beta_2}) X_i X_i' \ . \tag{B-10}$$

This shows that

$$\beta_1 - \beta_2 = \left(\mathsf{S}_{\beta_1,\beta_2} + \tau \mathrm{Id}_p\right)^{-1} \left(f(\beta_1) - f(\beta_2)\right) \ .$$

Since ρ_i's are convex, $\psi_i' = \rho_i''$ is non-negative and $\mathsf{S}_{\beta_1,\beta_2}$ is positive semi-definite. In the semi-definite order, we have $\mathsf{S}_{\beta_1,\beta_2} + \tau \mathrm{Id}_p \succeq \tau \mathrm{Id}_p$. In particular,

$$\|\beta_1 - \beta_2\| \leq \frac{1}{\tau}\|f(\beta_1) - f(\beta_2)\| \ .$$

□

Proposition B.1 yields the following lemma.

Lemma B.2. *For any β_1,*

$$\|\widehat{\beta} - \beta_1\| \leq \frac{1}{\tau}\|f(\beta_1)\| \ .$$

The lemma is a simple consequence of Equation (B-8) since by definition $f(\widehat{\beta}) = 0$.

Our strategy in what follows is to come up with "good candidates" for β_1, for which we can control $f(\beta_1)$ and transfer the information we will glean about the statistical properties of β_1 to $\widehat{\beta}$ through Lemma B.2.

B-2 On $\|\widehat{\beta}\|$ and $\|\widehat{\beta} - \beta_0\|$

We show in the following lemma that $\|\widehat{\beta}\|$ and $\|\widehat{\beta} - \beta_0\|$ cannot be too large.

Lemma B.3. *Let us call $W_n(b) = \frac{1}{n}\sum_{i=1}^n X_i \psi_i(\epsilon_i + X_i'b)$, $W_n \in \mathbb{R}^p$.*

We have, if $D_{\psi_i(Y_i)}$ is the $n \times n$ diagonal matrix with (i,i)-entry $\psi_i(Y_i) = \psi_i(\epsilon_i + X_i'\beta_0)$,

$$\|\widehat{\beta}\| \leq \frac{1}{\tau}\|W_n(\beta_0)\| = \frac{1}{\tau}\sqrt{\frac{1}{n^2} 1_n' D_{\psi_i(Y_i)} X X' D_{\psi_i(Y_i)} 1_n} \ ,$$

and if $D_{\psi_i(\epsilon_i)}$ is the $n \times n$ diagonal matrix with (i,i)-entry $\psi_i(\epsilon_i)$,

$$\|\widehat{\beta} - \beta_0\| \leq \|\beta_0\| + \frac{1}{\tau}\|W_n(0)\| = \|\beta_0\| + \frac{1}{\tau}\sqrt{\frac{1}{n^2} 1_n' D_{\psi_i(\epsilon_i)} X X' D_{\psi_i(\epsilon_i)} 1_n} \ .$$

Also,

$$\|W_n(\beta_0)\|^2 \leq \frac{1_n' D^2_{\psi_i(Y_i)} 1_n}{n} |||X'X/n|||_2 \ .$$

*Therefore, under our assumptions **O1-O6**,*

$$\mathbf{E}\left(\|\widehat{\beta}\|^2\right) \leq \frac{1}{\tau^2}\frac{p}{n} C^2 \mathit{polyLog}(n) \ , \ \textit{and} \tag{B-11}$$

$$\mathbf{E}\left(\|\widehat{\beta}\|^4\right) \leq \frac{1}{\tau^4} C \mathit{polyLog}(n) \tag{B-12}$$

Similarly, for any finite k,

$$\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|_2^k\right) \leq C_k \left[\|\beta_0\|^k + \mathit{polyLog}(n)/\tau^k\right] \ .$$

In the case $k = 2$, we have the more precise bound

$$\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|_2^2\right) \leq 2\left[\|\beta_0\|^2 + \frac{p/n}{\tau^2}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\psi_i^2(\epsilon_i)\right)\right] \ .$$

Proof. The first and key inequality simply comes from applying Lemma B.2 with $\beta_1 = 0$, after noticing that $f(0) = -W_n(\beta_0)$.

The second one comes from using $\beta_1 = \beta_0$ and noticing that $f(\beta_0) = -W_n(0) + \tau\beta_0$.

We note that under our assumptions, according to Lemma G.4

$$|||X'X/n|||_2 = \mathrm{O}_{L_k}(\mathrm{polyLog}(n))$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\psi_i^2(Y_i) \leq \frac{1}{n}\sum_{i=1}^{n}\|\psi_i^2\|_\infty = \mathrm{O}(1) \ ,$$

which gives all the results about L_k bounds.

The last result about $k = 2$ follows from computing $\mathbf{E}\left(\|W_n(0)\|^2\right) = \frac{p}{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\psi_i^2(\epsilon_i)\right)$ and using the bound

$$\|\widehat{\beta} - \beta_0\|^2 \leq 2\|\beta_0\|^2 + \frac{2}{\tau^2}\|W_n(0)\|^2 \ .$$

□

C Approximating $\widehat{\beta}$ by $\widehat{\beta}_{(i)}$: leave-one-observation-out

We consider the situation where we leave the i-th observation, (X_i, ϵ_i), out. By definition,

$$\widehat{\beta}_{(i)} = \mathrm{argmin}_{\beta\in\mathbb{R}^p} F_i(\beta) \ , \text{ where } F_i(\beta) = \frac{1}{n}\sum_{j\neq i}\rho_j(\epsilon_j + X_j'\beta_0 - X_j'\beta) + \frac{\tau}{2}\|\beta\|^2 \ .$$

We call

$$\tilde{r}_{j,(i)} = \epsilon_j - X_j'(\widehat{\beta}_{(i)} - \beta_0) \text{ and } S_i = \frac{1}{n}\sum_{j\neq i}\psi_j'(\tilde{r}_{j,(i)})X_jX_j' \ .$$

We also call

$$f_i(\beta) = -\frac{1}{n}\sum_{j\neq i}X_j\psi_j(\epsilon_j + X_j'\beta_0 - X_j'\beta) + \tau\beta = f(\beta) + \frac{1}{n}X_i\psi_i(\epsilon_i - X_i'(\beta - \beta_0)) \ .$$

We have of course

$$f_i(\widehat{\beta}_{(i)}) = 0 \ .$$

Let us consider

$$\widetilde{\beta}_i = \widehat{\beta}_{(i)} + \frac{1}{n}(S_i + \tau\mathrm{Id})^{-1}X_i\psi_i(\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})) \triangleq \widehat{\beta}_{(i)} + \eta_i \ , \tag{C-13}$$

where

$$c_i = \frac{1}{n}X_i'(S_i + \tau\mathrm{Id})^{-1}X_i \ , \text{ and} \tag{C-14}$$

$$\eta_i = \frac{1}{n}(S_i + \tau\mathrm{Id})^{-1}X_i\psi_i(\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})) \ . \tag{C-15}$$

These definitions and the approximations they will imply can be understood in light of the probabilistic heuristics we derived for a related problem in El Karoui et al. (2011) and El Karoui et al. (2013). The interested reader is referred to those papers - where we made a large effort to explain our intuitive ideas - for more information and intuition; given page limit requirements, we do not give a complete heuristic derivation of our results and refer the reader to Subsubsection 2.2.2 for a detailed explanation of our strategy. (A somewhat long heuristic derivation would also likely not be appropriate for submission to PTRF.) We note however that the rigorous proof requires refinements over the intuitive ideas. Those aspects are of a more technical nature and become apparent only through the analysis that we present here.

One of our aims is to show Theorem C.6 below, which shows that we can very accurately approximate $\widehat{\beta}$ by $\widetilde{\beta}_i$. Note that the statistical properties of $\widetilde{\beta}_i$ are easier to understand that those of $\widehat{\beta}$; our high-quality approximations will allow us to transfer our understanding of $\widetilde{\beta}_i$ to $\widehat{\beta}$.

C-1 Deterministic bounds

Proposition C.1. *We have, with $\widetilde{\beta}_i$ defined in Equation* (C-13)*,*

$$\|\widehat{\beta} - \widetilde{\beta}_i\| \leq \frac{1}{\tau}\|\mathcal{R}_i\| \,, \tag{C-16}$$

where

$$\mathcal{R}_i = \frac{1}{n}\sum_{j\neq i}\left[\psi_j'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi_j'(\tilde{r}_{j,(i)})\right] X_j X_j' \eta_i \,, \tag{C-17}$$

and $\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)$ is in the ("unordered") interval $(\tilde{r}_{j,(i)}, \tilde{r}_{j,(i)} - X_j'\eta_i) = (\epsilon_j + X_j'\beta_0 - X_j'\widehat{\beta}_{(i)}, \epsilon_j + X_j'\beta_0 - X_j'\widetilde{\beta}_i)$.

Proof. The proof and strategy are similar to the corresponding ones in El Karoui (2013). However, since there are delicate cancellations in the argument, we give all the details of the argument.

We recall that $Y_i = \epsilon_i + X_i'\beta_0$.

Since $f_i(\widehat{\beta}_{(i)}) = 0$, and $\widetilde{\beta}_i = \widehat{\beta}_{(i)} + \eta_i$,

$$f(\widetilde{\beta}_i) = f(\widetilde{\beta}_i) - f_i(\widehat{\beta}_{(i)}) = -\frac{1}{n}X_i\psi_i(Y_i - X_i'\widetilde{\beta}_i) + \frac{1}{n}\sum_{j\neq i} X_j\left[\psi_j(Y_j - X_j'\widehat{\beta}_{(i)}) - \psi_j(Y_j - X_j'(\widehat{\beta}_{(i)} + \eta_i))\right] + \tau\eta_i \,.$$

By the mean-value theorem, we also have

$$\psi_j(Y_j - X_j'\widehat{\beta}_{(i)}) - \psi_j(Y_j - X_j'(\widehat{\beta}_{(i)} + \eta_i)) = \psi_j'(\tilde{r}_{j,(i)})X_j'\eta_i + \left[\psi_j'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi_j'(\tilde{r}_{j,(i)})\right] X_j'\eta_i \,,$$

where $\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)$ is in the ("unordered") interval $(Y_j - X_j'\widehat{\beta}_{(i)}, Y_j - X_j'(\widehat{\beta}_{(i)} + \eta_i))$, i.e $(\tilde{r}_{j,(i)}, \tilde{r}_{j,(i)} - X_j'\eta_i)$.

Hence, if $\mathcal{R}_i$ is the quantity defined in Equation (C-17),

$$\begin{aligned}\frac{1}{n}\sum_{j\neq i} X_j\left[\psi_j(Y_j - X_j'\widehat{\beta}_{(i)}) - \psi_j(Y_j - X_j'(\widehat{\beta}_{(i)} + \eta_i))\right] &= \frac{1}{n}\sum_{j\neq i}\psi_j'(\tilde{r}_{j,(i)})X_jX_j'\eta_i + \mathcal{R}_i \,,\\ &= S_i\eta_i + \mathcal{R}_i \,.\end{aligned}$$

In light of the previous simplifications, we have, using

$$f(\beta) = f_i(\beta) - \frac{1}{n}X_i\psi_i(Y_i - X_i'\beta) \text{ and } f_i(\widehat{\beta}_{(i)}) = 0 \,,$$

the equality

$$f(\widetilde{\beta}_i) = -\frac{1}{n}X_i\psi_i(Y_i - X_i'\widetilde{\beta}_i) + (S_i + \tau\mathrm{Id})\eta_i + \mathcal{R}_i \,.$$

Since by definition, $\eta_i = \frac{1}{n}(S_i + \tau\mathrm{Id})^{-1}X_i\psi_i(\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)}))$,

$$(S_i + \tau\mathrm{Id})\eta_i = \frac{1}{n}X_i\psi_i(\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})) \,.$$

In other respects,

$$Y_i - X_i'\widetilde{\beta}_i = \tilde{r}_{i,(i)} - c_i\psi_i(\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})) \,.$$

When ρ is differentiable, $x - c\psi(\mathrm{prox}(c\rho)(x)) = \mathrm{prox}(c\rho)(x)$ almost by definition of the proximal mapping (see Lemma F.1). Therefore, $Y_i - X_i'\widetilde{\beta}_i = \mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})$ and

$$-\frac{1}{n}X_i\psi_i(Y_i - X_i'\widetilde{\beta}_i) + (S_i + \tau\mathrm{Id})\eta_i = \frac{1}{n}X_i\left[-\psi_i(\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})) + \psi_i(\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)}))\right] = 0.$$

We conclude that

$$f(\widetilde{\beta}_i) = \mathcal{R}_i \,.$$

Applying Lemma B.2, we see that

$$\|\widehat{\beta} - \widetilde{\beta}_i\| \leq \frac{1}{\tau}\|\mathcal{R}_i\| \,.$$

□

C-1.1 On $\mathcal{R}_i$

Clearly, controlling $\mathcal{R}_i$ is the key to controlling $\|\widehat{\beta} - \widetilde{\beta}_i\|$, so we need to develop insights into $\mathcal{R}_i$.

Lemma C.2. *We have*

$$\|\eta_i\| \leq \frac{1}{\sqrt{n}\tau} \frac{\|X_i\|}{\sqrt{n}} |\psi_i(\tilde{r}_{i,(i)})| \,, \tag{C-18}$$

and

$$\|\mathcal{R}_i\| \leq |||\widehat{\Sigma}|||_2 \sup_{j \neq i} \left| \psi_j'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi_j'(\tilde{r}_{j,(i)}) \right| \frac{1}{\sqrt{n}\tau} \frac{\|X_i\|}{\sqrt{n}} |\psi_i(\tilde{r}_{i,(i)})| \,. \tag{C-19}$$

We note that under our assumptions, we have

$$\left|\psi_i(\tilde{r}_{i,(i)})\right| \leq \|\psi_i\|_\infty \leq C\text{polyLog}(n)$$

and, using Lemma G.1

$$\sup_i \frac{\|X_i\|}{\sqrt{n}} = \mathrm{O}_{L_k}(\sup_i |\lambda_i|) \,.$$

Proof. This proof is essentially obvious. We refer the reader to a corresponding one given in El Karoui (2013) in case details are needed. □

C-1.2 On $\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)$ and related quantities

We now show how to control $\frac{1}{\sqrt{n}} \sup_{j \neq i} \left| \psi_j'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi_j'(\tilde{r}_{j,(i)}) \right|$, which is essential for turning Equation (C-16) into a useful bound.

Lemma C.3. *Suppose, as in our assumption **O3**, that ψ_i' is $\mathsf{L}_i(n)$-Lipschitz. Then,*

$$\sup_{j \neq i} \left| \psi_j'(\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i)) - \psi_j'(\tilde{r}_{j,(i)}) \right| \leq \left[\sup_{1 \leq i \leq n} \mathsf{L}_i(n) \right] \sup_{j \neq i} |X_j' \eta_i| \,.$$

Proof. By definition, we have

$$|\gamma^*(X_j, \widehat{\beta}_{(i)}, \eta_i) - \tilde{r}_{j,(i)}| \leq |X_j' \eta_i| \,.$$

The bound follows immediately, using the fact that ψ_i' is $\mathsf{L}_i(n)$-Lipschitz.

□

C-2 Stochastic aspects

Recall that we have by definition

$$X_j' \eta_i = \psi_i(\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) \frac{1}{n} X_j' (S_i + \tau \text{Id}_p)^{-1} X_i \,.$$

We can therefore bound $\|\mathcal{R}_i\|$ by

$$\|\mathcal{R}_i\| \leq \left[\sup_{j \neq i} \frac{|X_j'(S_i + \tau \text{Id}_p)^{-1} X_i|}{n} \right] \frac{\sup_{1 \leq i \leq n} \mathsf{L}_i(n)}{\sqrt{n}\tau} \frac{\|X_i\|}{\sqrt{n}} |||\widehat{\Sigma}|||_2 \left(|\psi_i(\tilde{r}_{i,(i)})| |\psi_i(\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)}))| \right) \,.$$

Therefore, we also have

$$\|\mathcal{R}_i\| \leq \left[\sup_{j \neq i} \frac{|X_j'(S_i + \tau \text{Id}_p)^{-1} X_i|}{n} \right] \frac{\mathsf{L}_i(n)}{\sqrt{n}\tau} \frac{\|X_i\|}{\sqrt{n}} |||\widehat{\Sigma}|||_2 \|\psi_i\|_\infty^2 \,.$$

This bound on $\|\mathcal{R}_i\|$ shows that we can control $\|\widehat{\beta} - \widetilde{\beta}_i\|$ in L_k provided we can control each terms in the above product in L_{3k}, by appealing to Holder's inequality and Proposition C.1.

We now turn our attention to the various elements of the bound on $\|\mathcal{R}_i\|$ and show that we can control them under our assumptions.

C-2.1 On $\sup_{j\neq i}|X_j'(S_i+\tau\mathrm{Id})^{-1}X_i/n|$

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

Lemma C.4. *Suppose X_i's are independent and satisfy Assumption **O4**; suppose λ_i's satisfy **O6**. Then*

$$\sup_{j\neq i}|X_j'(S_i+\tau\mathrm{Id})^{-1}X_i/n| \leq \frac{1}{\sqrt{n}}\sup_{j\neq i}\frac{\|\mathcal{X}_j\|}{\tau\sqrt{n}}\mathit{polyLog}(n)$$

*in L_k, for any finite k. Note also that under our Assumption **O4**, for any finite k,*

$$\sup_{j\neq i}\left|\|\mathcal{X}_j\|/\sqrt{n}\right| = \mathrm{O}_{L_k}(1)\ .$$

Proof. The proof follows from that of Lemma 2.3 in El Karoui (2013). Indeed,

$$|X_j'(S_i+\tau\mathrm{Id})^{-1}X_i/n| = |\lambda_i\lambda_j||\mathcal{X}_j'(S_i+\tau\mathrm{Id})^{-1}\mathcal{X}_i/n|\ .$$

The proof of Lemma 2.3 in El Karoui (2013) shows that

$$\sup_{j\neq i}|\mathcal{X}_j'(S_i+\tau\mathrm{Id})^{-1}\mathcal{X}_i/n| \leq \frac{1}{\sqrt{n}}\sup_{j\neq i}\frac{\|\mathcal{X}_j\|}{\tau\sqrt{n}}\mathrm{polyLog}(n)/\mathsf{c}_n^{1/2}$$

in L_k, when $\sup_{j\neq i}\frac{\|\mathcal{X}_j\|}{\tau\sqrt{n}} = \mathrm{O}_{L_{20}}(1)$; this latter result is shown in El Karoui (2013) (see the discussion after Lemma 2.3 or Lemma G.1 in the current Appendix).

Now our assumptions **O6** concerning $\sup_i|\lambda_i| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n))$ guarantee that the bounds we announced are valid. □

C-2.2 Consequences

We have the following result. Recall that ψ_i' is assumed to be Lipschitz with Lipschitz constant $\mathsf{L}_i(n)$.

Proposition C.5. *Under Assumptions **O1-O6**, we have*

$$\|\mathcal{R}_i\| = \mathrm{O}_{L_k}\left(\frac{[\sup_{1\leq i\leq n}\mathsf{L}_i(n)]\|\psi_i\|_\infty^2}{n\tau}\mathit{polyLog}(n)\right)\ .$$

Furthermore, the same bound hold for $\sup_{1\leq i\leq n}\|\mathcal{R}_i\|$ with $\sup_{1\leq i\leq n}\|\psi_i\|_\infty^2$ (instead of $\|\psi_i\|_\infty^2$) in the right-hand side.

Proof. The proof follows by aggregating all the intermediate results we had, using Holder's inequality and noticing that under our assumptions, $|||\widehat{\Sigma}|||_2 = \mathrm{O}_{L_k}(\sup_i\lambda_i^2\mathsf{c}_n^{-1}) = \mathrm{O}_{L_k}(\mathrm{polyLog}(n))$. This latter result is shown in Lemma G.4.

The statement concerning $\sup_{1\leq i\leq n}\|\mathcal{R}_i\|$ follows by the same arguments. □

We can now prove and state the following result, which relates residuals to leave-one-out-prediction errors and give a way to do online update from $\widehat{\beta}_{(i)}$ to $\widehat{\beta}$.

We recall that $\widetilde{\beta}_i$ is defined in Equation (C-13) on p.20.

Theorem C.6. *Under Assumptions **O1-O7**, we have, for any fixed k, when τ is held fixed and $\mathsf{L}_i(n)\leq Cn^\alpha$,*

$$\sup_{1\leq i\leq n}\|\widehat{\beta}-\widetilde{\beta}_i\| = \mathrm{O}_{L_k}(\frac{\mathit{polyLog}(n)}{n^{1-\alpha}})\ .$$

In particular, we have

$$\forall 1\leq i\leq n\ ,\mathbf{E}\left(\|\widehat{\beta}-\widetilde{\beta}_i\|^2\right) = \mathrm{O}(\mathit{polyLog}(n)/n^{2-2\alpha})\ .$$

Also,

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

Finally,

$$\sup_{1\leq i\leq n} |R_i - prox(c_i\rho_i)(\tilde{r}_{i,(i)})| = \mathrm{O}_{L_k}(\frac{polyLog(n)}{n^{1/2-\alpha}}) \ . \tag{C-20}$$

We note that we could state a slightly finer result involving $\mathsf{L}_i(n)$ and various powers of $\|\psi_i\|_\infty$. However, we will not need such fine results in what follows, so we opt for slightly coarser but easier-to-state statements.

Proof. The first two results simply follow from our work on $\|\mathcal{R}_i\|$.

The third result follows from the coarse bound

$$\begin{aligned}\sup_{j\neq i} |\tilde{r}_{j,(i)} - R_j| = \sup_{j\neq i} \left|X_j'(\widehat{\beta} - \widehat{\beta}_{(i)})\right| &\leq \sup_{j\neq i} \left|X_j'(\widehat{\beta} - \widetilde{\beta}_i)\right| + \sup_{j\neq i} |X_j'(\widetilde{\beta}_i - \widehat{\beta}_{(i)})| \ , \\ &\leq \left(\sup_{1\leq j\leq n} \frac{\|X_j\|}{\sqrt{n}}\right) \sqrt{n}\|\widehat{\beta} - \widetilde{\beta}_i\| + \sup_{j\neq i} |X_j'\eta_i| \ ,\end{aligned}$$

and the fact that $\left(\sup_{1\leq j\leq n} \frac{\|X_j\|}{\sqrt{n}}\right) = \mathrm{O}_{L_k}(\mathrm{polyLog}(n))$ under our assumptions. Our results on $\|\widehat{\beta} - \widetilde{\beta}_i\|$ give control of the first term. Control of the second term follows from Lemma C.4 and the assumption that ψ_i is bounded by $C\mathrm{polyLog}(n)$.

Let us now turn to the final result, i.e the approximation of the residual R_i by a non-linear function of the leave-one-out prediction error $\tilde{r}_{i,(i)}$. Recall that

$$R_i = \epsilon_i + X_i'\beta_0 - X_i'\widehat{\beta} = \epsilon_i + X_i'\beta_0 - X_i'\widetilde{\beta}_i - X_i'(\widehat{\beta} - \widetilde{\beta}_i) \ .$$

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

$$X_i'\widetilde{\beta}_i = X_i'\widehat{\beta}_{(i)} + c_i\psi_i[\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})] \ .$$

Hence, since almost by definition, if $y = \mathrm{prox}(c\rho)(x)$, $y + c\psi(y) = x$, we get

$$\epsilon_i + X_i'\beta_0 - X_i'\widetilde{\beta}_i = \tilde{r}_{i,(i)} - c_i\psi_i[\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})] = \mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)}) \ .$$

So we have established that

$$\sup_i \left|R_i - \mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})\right| = \sup_i \left|X_i'(\widetilde{\beta}_i - \widehat{\beta})\right|$$

and the result follows from our previous bounds. □

C-3 On the limiting variance of $\|\widehat{\beta}\|^2$ and $\|\widehat{\beta} - \beta_0\|^2$

An interesting consequence of our leave-one-observation-out work is that we can use the ideas and approximations developed above to show that $\|\widehat{\beta} - \beta_0\|$ and $\|\widehat{\beta}\|$ are asymptotically deterministic (in other words, they can be approximated asymptotically by deterministic sequences).

Proposition C.7. *Under our assumptions **O1-O7**,*

$$\mathrm{var}\left(\|\widehat{\beta}\|^2\right) \to 0 \ \textit{as} \ n \to \infty \ .$$

Therefore $\|\widehat{\beta}\|^2$ has a deterministic equivalent in probability and in L_2.

More precisely, we have

$$\mathrm{var}\left(\|\widehat{\beta}\|^2\right) = \mathrm{O}(\frac{polyLog(n)}{n^{1-2\alpha}}) \ .$$

The same type of results are true for $\mathrm{var}\left(\|\widehat{\beta} - \beta_0\|_2^2\right)$ *provided* $\|\beta_0\| = \mathrm{O}(polyLog(n))$.

Proof. We use the Efron-Stein inequality Efron and Stein (1981) to show that $\text{var}\left(\|\widehat{\beta}\|^2\right)$ goes to 0 as $n \to \infty$. In what follows, we rely on our approximations and our assumptions to have enough moments for all the expectations of the type $\mathbf{E}\left(\|\widehat{\beta}\|^{2k}\right)$ to be bounded like $\text{polyLog}(n)/\tau^{2k}$. Note that this the content of our Lemma B.3.

Recall the Efron-Stein inequality (Efron and Stein (1981)) : if W is a function of n independent random variables, and $W_{(i)}$ is any function of all those random variables except the i-th,

$$\text{var}\,(W) \leq \sum_{i=1}^{n} \text{var}\left(W - W_{(i)}\right) \leq \sum_{i=1}^{n} \mathbf{E}\left((W - W_{(i)})^2\right) \ .$$

In our arguments below, $\|\widehat{\beta}\|^2$ plays the role of W and $\|\widehat{\beta}_{(i)}\|^2$ plays the role of $W_{(i)}$.

We first observe that

$$\mathbf{E}\left(|\|\widehat{\beta}\|^2 - \|\widehat{\beta}_{(i)}\|^2|^2\right) \leq 2\left[\mathbf{E}\left(|\|\widehat{\beta}\|^2 - \|\widetilde{\beta}_i\|^2|^2\right) + \mathbf{E}\left(|\|\widetilde{\beta}_i\|^2 - \|\widehat{\beta}_{(i)}\|^2|^2\right)\right] \ .$$

Of course, using the fact that $\widehat{\beta} = \widehat{\beta} - \widetilde{\beta}_i + \widetilde{\beta}_i$ and $|\|\widehat{\beta}\|^2 - \|\widetilde{\beta}_i\|^2|^2 = [(\widehat{\beta} - \widetilde{\beta}_i)'(\widehat{\beta} + \widetilde{\beta}_i)]^2$, and hence $(\widehat{\beta} - \widetilde{\beta}_i)'(\widehat{\beta} + \widetilde{\beta}_i) = 2(\widehat{\beta} - \widetilde{\beta}_i)'\widehat{\beta} - \|\widehat{\beta} - \widetilde{\beta}_i\|^2$, we have

$$|\|\widehat{\beta}\|^2 - \|\widetilde{\beta}_i\|^2|^2 = \text{O}_{L_1}(\|\widehat{\beta} - \widetilde{\beta}_i\|^4) + \sqrt{\text{O}_{L_1}(\text{polyLog}(n)\|\widehat{\beta} - \widetilde{\beta}_i\|^4)} \ ,$$

by the Cauchy-Schwarz inequality, since $\mathbf{E}\left(\|\widehat{\beta}\|^k\right)$ exists and is bounded by $K\text{polyLog}(n)/\tau^k$.

Using the results of Theorem C.6, we see that

$$\mathbf{E}\left(|\|\widehat{\beta}\|^2 - \|\widetilde{\beta}_i\|^2|^2\right) = \text{O}(\frac{\text{polyLog}(n)}{n^{2-2\alpha}}) = \text{o}(n^{-1}) \ ,$$

provided $\alpha < 1/2$.

On the other hand, given the definition in Equation (C-13),

$$\|\widetilde{\beta}_i\|^2 - \|\widehat{\beta}_{(i)}\|^2 = 2\frac{1}{n}\widehat{\beta}'_{(i)}(S_i + \tau\text{Id})^{-1}X_i\psi_i(\text{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})) + \frac{1}{n^2}X_i'(S_i + \tau\text{Id})^{-2}X_i\psi_i^2(\text{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})) \ .$$

Since $\widehat{\beta}_{(i)}$ and S_i are independent of X_i, and $|||(S_i+\tau\text{Id})^{-1}|||_2 \leq 1/\tau$, $\widehat{\beta}'_{(i)}(S_i+\tau\text{Id})^{-1}X_i = \text{O}_{L_2}(|\lambda_i|\|\widehat{\beta}_{(i)}\|/\mathsf{c}_n^{1/2})$, using our assumptions **O4** on $\mathcal{X}_i$ applied to linear forms. Recall also that $\sup_i\|\psi_i\|_\infty = \text{O}(\text{polyLog}(n))$. Therefore, we see that both terms are $\text{O}_{L_2}(\text{polyLog}(n)/n\mathsf{c}_n^{1/2})$.

We conclude that then

$$\mathbf{E}\left(\left|\|\widetilde{\beta}_i\|^2 - \|\widehat{\beta}_{(i)}\|^2\right|^2\right) = \text{O}(\frac{\text{polyLog}(n)}{n^2}) \ .$$

Taking $W = \|\widehat{\beta}\|^2$ and $W_{(i)} = \|\widehat{\beta}_{(i)}\|^2$ in the Efron-Stein inequality, we clearly see that

$$\text{var}\left(\|\widehat{\beta}\|^2\right) = \text{O}\left(\frac{\text{polyLog}(n)}{n^{1-2\alpha}}\right) = \text{o}(1) \ .$$

This shows that $\|\widehat{\beta}\|^2$ has a deterministic equivalent in probability and in L_2.

• About $\|\widehat{\beta}-\beta_0\|$. The results are obtained in a similar fashion using our bound on $\|\widehat{\beta}-\beta_0\|$ and replacing everywhere in the arguments $\|\widehat{\beta}\|$ by $\|\widehat{\beta}-\beta_0\|$, and $\|\widetilde{\beta}_i\|$ by $\|\widetilde{\beta}_i-\beta_0\|$. The condition $\|\beta_0\| = \text{O}(\text{polyLog}(n))$ plays a role to guarantee in Lemma B.3 that $\|\widehat{\beta} - \beta_0\| = \text{O}_{L_k}(\text{polyLog}(n))$. □

D Leaving out a predictor

In this second step of the proof, we do need at various points that the entries of the vector $\mathcal{X}_i$ be independent, whereas as we showed before, it is not important when studying what happens when we leave out an observation.

Let V be 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 $V_i' = (X_i(1), \ldots, X_i(p-1))$. We call $X(p)$ the vector in $\mathbb{R}^n$ with j-th entry $X_j(p)$, i.e the $p-$th entry of the vector X_j. When this does not create problems, we also use the standard notation $X_{j,p}$ for $X_j(p)$.

Let us call $\widehat{\gamma}$ the solution of

$$\widehat{\gamma} = \text{argmin}_{\gamma \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \rho_i(\epsilon_i - V_i'(\gamma - \gamma_0)) + \frac{\tau}{2}\|\gamma\|^2 \ . \tag{D-21}$$

Note that $\begin{pmatrix} \widehat{\gamma} \\ 0 \end{pmatrix}$ is the solution of the original optimization problem (3) when $X_i(p)$ is replaced by 0.

In this part of the paper, we will rely heavily on the following definitions:

Definition. We call the residuals corresponding to this optimization problem $\{r_{i,[p]}\}_{i=1}^n$, in other words

$$r_{i,[p]} = \epsilon_i + V_i'\gamma_0 - V_i'\widehat{\gamma} \ .$$

We call

$$u_p = \frac{1}{n} \sum_{i=1}^{n} \psi_i'(r_{i,[p]}) V_i X_i(p) \ , \text{ and } \mathfrak{S}_p = \frac{1}{n} \sum_{i=1}^{n} \psi_i'(r_{i,[p]}) V_i V_i' \ .$$

Note that $u_p \in \mathbb{R}^{p-1}$ and $\mathfrak{S}_p$ is $(p-1) \times (p-1)$. We call

$$\xi_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i^2(p) \psi_i'(r_{i,[p]}) - u_p'(\mathfrak{S}_p + \tau \text{Id})^{-1} u_p \ , \tag{D-22}$$

and

$$N_p \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(p) \psi_i(r_{i,[p]}) \ . \tag{D-23}$$

We will show later, in Subsubsection D-2.2 that $\xi_n \geq 0$. However, we will use this information from the beginning and there are no circular arguments.

We consider

$$\mathfrak{b}_p \triangleq \beta_0(p) \frac{\xi_n}{\tau + \xi_n} + \frac{1}{\sqrt{n}} \frac{N_p}{\tau + \xi_n} \ . \tag{D-24}$$

Note that when $\xi_n > 0$, we have

$$\mathfrak{b}_p - \beta_0(p) = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i(p)\psi_i(r_{i,[p]}) - \tau \mathfrak{b}_p}{\frac{1}{n}\sum_{i=1}^{n} X_i^2(p)\psi_i'(r_{i,[p]}) - u_p'(\mathfrak{S}_p + \tau \text{Id})^{-1} u_p} = \frac{n^{-1/2} N_p - \tau \mathfrak{b}_p}{\xi_n} \ .$$

We call

$$\widetilde{b} = \begin{bmatrix} \widehat{\gamma} \\ \beta_0(p) \end{bmatrix} + [\mathfrak{b}_p - \beta_0(p)] \begin{bmatrix} -(\mathfrak{S}_p + \tau \text{Id})^{-1} u_p \\ 1 \end{bmatrix} \ . \tag{D-25}$$

The aim of our work in the second part of this proof is to establish Theorem D.10 on p.37, which shows that $\|\widetilde{b} - \widehat{\beta}\| = \text{O}(\text{polyLog}(n)/n)$ in L_k. Because the last coordinate of $\widetilde{b}$, $\mathfrak{b}_p$, has a reasonably simple probabilistic structure and our approximations are sufficiently good, we will be able to transfer our insights about this coordinate to $\widehat{\beta}_p$, the last coordinate of $\widehat{\beta}$. This is also true when considering $\sqrt{n}(\mathfrak{b}_p - \widehat{\beta}_p)$, so our approximations will be interesting at that scale, too.

The approach and approximating quantities we choose - as well as the intuition behind those choices - by using variants of the ideas discussed in our work in El Karoui et al. (2011), Bean et al. (2012) and El Karoui et al. (2013). (Once again, we do not give here a heuritic derivation as it seems inappropriate for a submission to this journal.)

D-1 Deterministic aspects

Proposition D.1. *We have*

$$\|\widehat{\beta}-\widetilde{b}\| \leq \frac{1}{\tau}|\mathfrak{b}_p-\beta_0(p)| \sup_{1\leq i\leq n}|\mathsf{d}_{i,p}|\,|||\widehat{\Sigma}|||_2\sqrt{\|(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\|^2+1}\,. \tag{D-26}$$

where $\mathsf{d}_{i,p}=[\psi_i'(\gamma_{i,p}^*)-\psi_i'(r_{i,[p]})]$ *and* $\gamma_{i,p}^*$ *is in the interval* $(\epsilon_i+V_i'\gamma_0-V_i'\widehat{\gamma},\epsilon_i+X_i'\beta_0-X_i'\widetilde{b})$.

Furthermore,

$$\|(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\|^2 \leq \frac{1}{n\tau}\sum_{i=1}^n X_i^2(p)\psi_i'(r_{i,[p]}) = \frac{1}{n\tau}\sum_{i=1}^n \lambda_i^2\psi_i'(r_{i,[p]})\mathcal{X}_i^2(p)\,. \tag{D-27}$$

As we saw in Equation (B-8) and Lemma B.2, we have

$$\|\widehat{\beta}-\widetilde{b}\| \leq \frac{1}{\tau}\|f(\widetilde{b})\|\,,$$

where

$$f(\widetilde{b}) = -\frac{1}{n}\sum_{i=1}^n X_i\psi_i(\epsilon_i+X_i'\beta_0-X_i'\widetilde{b})+\tau\widetilde{b}\,.$$

We note furthermore that, by definition of $\widehat{\gamma}$,

$$g(\widehat{\gamma}) \triangleq -\frac{1}{n}\sum_{i=1}^n V_i\psi_i(\epsilon_i+V_i'\gamma_0-V_i'\widehat{\gamma})+\tau\widehat{\gamma}=0_{p-1}\,.$$

The strategy of the proof is to control $f(\widetilde{b})$ by using $g(\widehat{\gamma})$ to create good approximations and then recalling that $g(\widehat{\gamma})=0_{p-1}$.

Proof. The proof strategy and ideas are tied to the technique developed in El Karoui (2013); however, because there are a number of delicate cancellations in the argument, we give it in full details. (Naturally, coming up with good approximating quantities required much work.)

a) Work on the first $(p-1)$ **coordinates of** $f(\widetilde{b})$

We call $\mathsf{f}_{p-1}(\beta)$ the first $p-1$ coordinates of $f(\beta)$. We call $\widehat{\gamma}_{ext}$ the p-dimensional vector whose first $p-1$ coordinates are $\widehat{\gamma}$ and last coordinate is $\beta_0(p)$, i.e

$$\widehat{\gamma}_{ext}=\begin{bmatrix}\widehat{\gamma}\\ \beta_0(p)\end{bmatrix}\,.$$

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,

$$\mathsf{f}_{p-1}(\widetilde{b}) = \mathsf{f}_{p-1}(\widetilde{b})-g(\widehat{\gamma}) = -\frac{1}{n}\sum_{i=1}^n V_i\left[\psi_i(\epsilon_i+X_i'\beta_0-X_i'\widetilde{b})-\psi_i(\epsilon_i+V_i'\gamma_0-V_i'\widehat{\gamma})\right]+\tau(\widetilde{b}_{comp,p}-\widehat{\gamma})\,.$$

We can write by using the mean value theorem, for $\gamma_{i,p}^*$ in the interval $(\epsilon_i-V_i'(\widehat{\gamma}-\gamma_0),\epsilon_i-X_i'(\widetilde{b}-\beta_0))$,

$$\begin{aligned}\psi_i(\epsilon_i+X_i'\beta_0-X_i'\widetilde{b})-\psi_i(\epsilon_i+V_i'\gamma_0-V_i'\widehat{\gamma}) &= \psi_i'(\gamma_{i,p}^*)X_i'(\widehat{\gamma}_{ext}-\widetilde{b})\,,\\ &= \psi_i'(r_{i,[p]})X_i'(\widehat{\gamma}_{ext}-\widetilde{b})+[\psi_i'(\gamma_{i,p}^*)-\psi_i'(r_{i,[p]})]X_i'(\widehat{\gamma}_{ext}-\widetilde{b})\,.\end{aligned}$$

Let us call

$$\begin{aligned}\mathsf{d}_{i,p} &= [\psi_i'(\gamma_{i,p}^*)-\psi_i'(r_{i,[p]})]\,,\\ \delta_{i,p} &= [\psi_i'(\gamma_{i,p}^*)-\psi_i'(r_{i,[p]})]X_i'(\widehat{\gamma}_{ext}-\widetilde{b})\,,\\ \mathsf{R}_p &= -\frac{1}{n}\sum_{i=1}^n \mathsf{d}_{i,p}V_iX_i'(\widehat{\gamma}_{ext}-\widetilde{b})\,.\end{aligned}$$

We have with this notation

$$\mathsf{f}_{p-1}(\widetilde{b}) = -\frac{1}{n}\sum_{i=1}^{n}\psi_i'(r_{i,[p]})V_iX_i'(\widehat{\gamma}_{ext}-\widetilde{b}) + \tau(\widetilde{b}_{comp,p}-\widehat{\gamma}) + \mathsf{R}_p \triangleq \mathsf{A}_p + \mathsf{R}_p \ .$$

We note that by definition,

$$\widehat{\gamma}_{ext}-\widetilde{b} = [\mathfrak{b}_p - \beta_0(p)]\begin{bmatrix}(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\\ -1\end{bmatrix} \ ,$$

$$\widetilde{b}_{comp,p}-\widehat{\gamma} = -[\mathfrak{b}_p-\beta_0(p)](\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p \ .$$

Therefore, $X_i'(\widehat{\gamma}_{ext}-\widetilde{b}) = [\mathfrak{b}_p-\beta_0(p)]\left[V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p - X_i(p)\right]$, and

$$\mathsf{A}_p = -(\mathfrak{b}_p-\beta_0(p))\left(\frac{1}{n}\sum_{i=1}^{n}\psi_i'(r_{i,[p]})V_i\left[V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p - X_i(p)\right] + \tau(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\right) \ .$$

Recalling the definition of $\mathfrak{S}_p$ and u_p, we see that

$$\mathsf{A}_p = -(\mathfrak{b}_p-\beta_0(p))\left(\mathfrak{S}_p(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p - u_p + \tau(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\right) = 0_{p-1} \ ,$$

since $\mathfrak{S}_p(\mathfrak{S}_p+\tau\mathrm{Id})^{-1} + \tau(\mathfrak{S}_p+\tau\mathrm{Id})^{-1} = \mathrm{Id}$.

We conclude that

$$\boxed{\mathsf{f}_{p-1}(\widetilde{b}) = \mathsf{R}_p \ .}$$

b) Work on the last coordinate of $f(\widetilde{b})$

We call $[f(\widetilde{b})]_p$ the last coordinate of $f(\widetilde{b})$. We have shown above that

$$\psi_i(\epsilon_i + X_i'\beta_0 - X_i'\widetilde{b}) - \psi_i(\epsilon_i + V_i'\gamma_0 - V_i'\widehat{\gamma}) = \psi_i'(r_{i,[p]})X_i'(\widehat{\gamma}_{ext}-\widetilde{b}) + [\psi_i'(\gamma_{i,p}^*) - \psi_i'(r_{i,[p]})]X_i'(\widehat{\gamma}_{ext}-\widetilde{b}).$$

Recall the notation

$$\delta_{i,p} = [\psi_i'(\gamma_{i,p}^*) - \psi_i'(r_{i,[p]})]X_i'(\widehat{\gamma}_{ext}-\widetilde{b}) \ .$$

Clearly,

$$\begin{aligned}\psi_i(\epsilon_i + X_i'\beta_0 - X_i'\widetilde{b}) &= \psi_i(r_{i,[p]}) + \psi_i'(r_{i,[p]})X_i'(\widehat{\gamma}_{ext}-\widetilde{b}) + \delta_{i,p} \ ,\\ &= \psi_i(r_{i,[p]}) + \psi_i'(r_{i,[p]})[\mathfrak{b}_p-\beta_0(p)]\left[V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p - X_i(p)\right] + \delta_{i,p} \ .\end{aligned}$$

We therefore see that

$$\begin{aligned}[f(\widetilde{b})]_p + \frac{1}{n}\sum_{i=1}^{n}X_i(p)\delta_{i,p} &= -\frac{1}{n}\sum_{i=1}^{n}X_i(p)\left(\psi_i(r_{i,[p]}) + \psi_i'(r_{i,[p]})(\mathfrak{b}_p-\beta_0(p))\left[V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p - X_i(p)\right]\right) + \tau\widetilde{b}_p \ ,\\ &= -\frac{1}{n}\sum_{i=1}^{n}X_i(p)\psi_i(r_{i,[p]}) - (\mathfrak{b}_p-\beta_0(p))u_p'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\\ &\quad + (\mathfrak{b}_p-\beta_0(p))\frac{1}{n}\sum_{i=1}^{n}\psi_i'(r_{i,[p]})X_i^2(p) + \tau\mathfrak{b}_p \ ,\\ &= -\left[\frac{1}{n}\sum_{i=1}^{n}X_i(p)\psi_i(r_{i,[p]}) - \tau\mathfrak{b}_p\right]\\ &\quad + (\mathfrak{b}_p-\beta_0(p))\left(\frac{1}{n}\sum_{i=1}^{n}\psi_i'(r_{i,[p]})X_i^2(p) - u_p'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\right) \ ,\\ &= -\left[\frac{1}{\sqrt{n}}N_p - \tau\mathfrak{b}_p\right] + (\mathfrak{b}_p-\beta_0(p))\xi_n \ ,\\ &= 0 \ .\end{aligned}$$

We conclude that

$$[f(\widetilde{b})]_p = -\frac{1}{n}\sum_{i=1}^{n}X_i(p)\delta_{i,p} = -\frac{1}{n}\sum_{i=1}^{n}\mathsf{d}_{i,p}X_i(p)X_i'(\widehat{\gamma}_{ext}-\widetilde{b}) \ .$$

Representation of $f(\widetilde{b})$

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

$$f(\widetilde{b}) = \left(-\frac{1}{n}\sum_{i=1}^{n} \mathfrak{d}_{i,p} X_i X_i'\right)(\widehat{\gamma}_{ext} - \widetilde{b}) \ ,$$
$$= -(\mathfrak{b}_p - \beta_0(p))\left(\frac{1}{n}\sum_{i=1}^{n} \mathfrak{d}_{i,p} X_i X_i'\right)\begin{bmatrix}(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p \\ -1\end{bmatrix} \ .$$

We conclude immediately that

$$\|f(\widetilde{b})\| \leq |\mathfrak{b}_p - \beta_0(p)| \sup_{1\leq i\leq n} |\mathfrak{d}_{i,p}| \, |||\widehat{\Sigma}|||_2 \sqrt{\|(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p\|^2 + 1} \ . \tag{D-28}$$

This gives Equation (D-26). The rest of the proof follows easily with mild modifications from El Karoui (2013) and we do not repeat it here.

□

D-2 Stochastic aspects

From now on, we assume that $\mathcal{X}(p)$, is independent of $\{\mathcal{V}_i, \epsilon_i\}_{i=1}^n$. This is consistent with Assumption **P1**. (Recall that $X_i = \lambda_i \mathcal{X}_i$ and therefore $V_i = \lambda_i \mathcal{V}_i$.) Note that Assumption **O4** is satisfied for $\mathcal{V}_i$ if it is satisfied for $\mathcal{X}_i$: convex 1-Lipschitz function of $\mathcal{V}_i$ can be trivially made to be convex 1-Lipschitz function of $\mathcal{X}_i$ by simply not acting on the last coordinate of $\mathcal{X}_i$.

Naturally, a large amount of the rest of the proof consists in showing that we can bound $\|f(\widetilde{b})\|$ sufficiently finely for our results to hold true. So we will work on bounding each term in the product appearing in Equation (D-26) in the rest of this section.

The last term is very easy to bound. In fact, using Equation (D-27), we have

$$\|(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p\|^2 \leq \frac{1}{\tau}\frac{1}{n}\sum_{i=1}^{n} \|\psi_i'\|_\infty \lambda_i^2 \mathcal{X}_i^2(p) \ .$$

Under Assumption **O3**, this translates into

$$\|(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p\|^2 \leq \frac{\sup_i \|\psi_i'\|_\infty}{\tau}\frac{1}{n}\sum_{i=1}^{n} \lambda_i^2 \mathcal{X}_i^2(p) \ .$$

Hence, under assumptions **O3-O4** and **O6**, we see that, for any fixed k and at τ fixed,

$$\boxed{\|(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p\|^2 = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)) \ .}$$

(Note that p does not play a particular role here. If we considered the same quantity when we remove the k-th predictor, and took the sup over $1 \leq k \leq p$ of the corresponding random variables, the same inequality would hold, in light of our work in e.g Lemma G.3.)

This guarantees that

$$\left\|\begin{matrix}(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p \\ -1\end{matrix}\right\|^2 \leq (1 + \|(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p\|^2) = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)) \ .$$

We conclude, using Equation (D-26), that

$$\|\widehat{\beta} - \widetilde{b}\| = \mathrm{O}_{L_k}\left(\frac{1}{\tau}\mathrm{polyLog}(n) |\mathfrak{b}_p - \beta_0(p)| \sup_{1\leq i\leq n} |\mathfrak{d}_{i,p}| \, |||\widehat{\Sigma}|||_2\right) \ ,$$

provided the terms appearing inside the O_{L_k} have enough moments to enable us to use Holder's inequality.

Recall that Lemma G.4 gives $|||\widehat{\Sigma}|||_2 = \mathrm{O}_{L_k}(\mathrm{polyLog}(n))$ under our assumptions **O1-O7**. At a high level, we expect $\sup_{1\le i\le n}|\mathsf{d}_{i,p}|$ and $[\mathfrak{b}_p - \beta_0(p)]$ to be small, which "should give us" that

$$\|\widehat{\beta} - \widetilde{b}\| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n) \sup_{1\le i\le n}|\mathsf{d}_{i,p}||\mathfrak{b}_p - \beta_0(p)|) .$$

In fact, we will show in Proposition D.2 that $\mathfrak{b}_p - \beta_0(p) = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)[n^{-1/2} \vee n^{-\mathsf{e}}])$ and in Proposition D.9 that $\sup_{1\le i\le n}|\mathsf{d}_{i,p}| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)[n^{\alpha-1/2} \vee n^{\alpha-\mathsf{e}}])$.

These are the key bounds we will need in showing that $\|\widehat{\beta} - \widetilde{b}\|$ is small.

We now turn our attention to showing these two results.

D-2.1 On $\mathfrak{b}_p - \beta_0(p)$

We recall the notations

$$N_p = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_i(r_{i,[p]})X_i(p) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \lambda_i\psi_i(r_{i,[p]})\mathcal{X}_i(p) ,$$
$$\xi_n = \frac{1}{n}\sum_{i=1}^n \psi_i'(r_{i,[p]})X_i^2(p) - u_p'(\mathfrak{S}_p + \tau\mathrm{Id})^{-1}u_p .$$

Under our assumptions, we have $\mathbf{E}(\mathcal{X}_i) = 0$ and $\mathrm{cov}(\mathcal{X}_i) = \mathrm{Id}_p$ and hence $\mathbf{E}\left(\mathcal{X}_i^2(p)\right) = 1$. Recall that since we assume that $\mathcal{X}(p)$ is independent of $\{\mathcal{V}_i, \epsilon_i\}_{i=1}^n$, $\mathcal{X}(p)$ is independent of $\{r_{i,[p]}\}_{i=1}^n$.

Proposition D.2. *We have*

$$|\mathfrak{b}_p - \beta_0(p)| \le \frac{1}{\sqrt{n}\tau}|N_p| + |\beta_0(p)| .$$

*Furthermore, under assumptions **O1-O7** and **P1**, $N_p = \mathrm{O}_{L_k}(\mathit{polyLog}(n))$ and therefore, when τ is held fixed,*

$$|\mathfrak{b}_p - \beta_0(p)| = \mathrm{O}_{L_k}(\mathit{polyLog}(n)n^{-1/2} + \|\beta_0\|_\infty) .$$

Proof. From the definition of $\mathfrak{b}_p$, we see that, when $\xi_n \neq 0$

$$\mathfrak{b}_p - \beta_0(p) = \frac{1}{\sqrt{n}}\frac{N_p}{\tau + \xi_n} - \frac{\tau\beta_0(p)}{\tau + \xi_n} .$$

We will see later, in Subsubsection D-2.2, that $\xi_n \ge 0$ (there is no circular arguments, it is simply more convenient to postpone the investigation of the properties of ξ_n). It immediately then follows that

$$|\mathfrak{b}_p| \le \frac{1}{\sqrt{n}\tau}|N_p| + |\beta_0(p)| .$$

Using independence of $\mathcal{X}(p)$ and $\{\mathcal{V}_i, \epsilon_i\}_{i=1}^n$, we have for instance

$$\mathbf{E}\left(N_p^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\mathcal{X}_i^2(p)\right)\mathbf{E}\left(\lambda_i^2\psi_i^2(r_{i,[p]})\right) ,$$

whether the right-hand side is finite or not. Using our bounds on $\max\lambda_i^2$ and $\sup_i\|\psi_i\|_\infty$, we therefore have

$$\mathbf{E}\left(N_p^2\right) \le \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\mathcal{X}_i^2(p)\right)\|\psi_i\|_\infty^2\mathbf{E}\left(\lambda_i^2\right) = \mathrm{O}(1) = \mathrm{O}(\mathrm{polyLog}(n)) .$$

Simple computations also show that N_p has as many moments as we need and that for any finite k, under our assumptions,

$$N_p = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)) .$$

We therefore have

$$|\mathfrak{b}_p - \beta_0(p)| \le \frac{1}{\sqrt{n}\tau}\mathrm{O}_{L_k}(\mathrm{polyLog}(n)) + \sup_{1\le k\le p}|\beta_0(k)| .$$

□

D-2.2 On ξ_n

Let us write ξ_n using matrix notations. Let $D_{\psi_i'(r_{.,[p]})}$ be the $n \times n$ diagonal matrix such that

$$D_{\psi_i'(r_{.,[p]})}(i,i) = \psi_i'(r_{i,[p]}) \ .$$

We also denote by $X(p)$ is the last column of the design matrix X. Then we have

$$\xi_n = \frac{1}{n} X(p)' D^{1/2}_{\psi_i'(r_{.,[p]})} M D^{1/2}_{\psi_i'(r_{.,[p]})} X(p) \ , \tag{D-29}$$

where

$$M = \mathrm{Id}_n - \frac{D^{1/2}_{\psi_i'(r_{.,[p]})} V}{\sqrt{n}} \left(\frac{1}{n} V' D_{\psi_i'(r_{.,[p]})} V + \tau \mathrm{Id} \right)^{-1} \frac{V' D^{1/2}_{\psi_i'(r_{.,[p]})}}{\sqrt{n}} \ . \tag{D-30}$$

This simply comes from elementary linear algebra and representing u_p and $\mathfrak{S}_p$ in matrix form. For example, $nu_p' = X(p)' D_{\psi_i'(r_{.,[p]})} V$.

We are now ready to investigate in more detail the properties of ξ_n.

Lemma D.3. *We have*

$$\xi_n \geq 0 \ .$$

*Furthermore, under Assumptions **O1-O7** and **P1**, if D_{λ_i} is the diagonal matrix with i-th entry λ_i,*

$$\left| \xi_n - \frac{1}{n} trace \left(D_{\lambda_i} D^{1/2}_{\psi_i'(r_{.,[p]})} M D^{1/2}_{\psi_i'(r_{.,[p]})} D_{\lambda_i} \right) \right| = \mathrm{O}_{L_k}(\sup_{1 \leq i \leq n} \lambda_i^2 \psi_i'(r_{i,[p]}) / (\sqrt{n \mathsf{c}_n})) \ . \tag{D-31}$$

Proof. Let us first focus on

$$M = \mathrm{Id}_n - \frac{1}{n} D^{1/2}_{\psi_i'(r_{.,[p]})} V (\frac{V' D_{\psi_i'(r_{.,[p]})} V}{n} + \tau \mathrm{Id})^{-1} V' D^{1/2}_{\psi_i'(r_{.,[p]})} \ .$$

The first part of the proof is very similar to the corresponding arguments in El Karoui (2013). 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 $n^{-1/2} D^{1/2}_{\psi_i'(r_{.,[p]})} V$ are denoted by σ_i, the eigenvalues of M are $\tau/(\sigma_i^2 + \tau)$.

Therefore, since $\xi_n = \frac{1}{n} v' M v$ with $v = D^{1/2}_{\psi_i'(r_{.,[p]})} X(p)$, $\xi_n \geq 0$.

Since M is symmetric and has eigenvalues between 0 and 1, we also have, using e.g Lemma V.1.5 in Bhatia (1997),

$$0 \preceq D^{1/2}_{\psi_i'(r_{.,[p]})} M D^{1/2}_{\psi_i'(r_{.,[p]})} \preceq D_{\psi_i'(r_{.,[p]})} \ .$$

The matrix M is independent of $\mathcal{X}(p)$ under Assumption **P1**. $D_{\psi_i'(r_{.,[p]})}$ is also independent of $\mathcal{X}(p)$. Of course, we have $X(p) = D_{\lambda_i} \mathcal{X}(p)$, where D_{λ_i} is the diagonal matrix with i-th entry λ_i.

Since $\mathcal{X}_p$ satisfy the necessary concentration assumptions under Assumption **P1**, we can now appeal to Lemma G.3 to obtain

$$\left| \frac{1}{n} X(p)' D^{1/2}_{\psi_i'(r_{.,[p]})} M D^{1/2}_{\psi_i'(r_{.,[p]})} X(p) - \frac{1}{n} \mathrm{trace} \left(D_{\lambda_i} D^{1/2}_{\psi_i'(r_{.,[p]})} M D^{1/2}_{\psi_i'(r_{.,[p]})} D_{\lambda_i} \right) \right| = \mathrm{O}_{L_k}(\frac{1}{\sqrt{n \mathsf{c}_n}} \sup_{1 \leq i \leq n} \lambda_i^2 \psi_i'(r_{i,[p]})) \ .$$

□

We now take a slight detour from the aim of showing that we have a very good approximation of $\widehat{\beta}$ through $\widetilde{b}$ by working on finer properties of ξ_n and $\mathfrak{b}_p$. These properties will be essential in establishing the validity of the system (4).

To get a finer understanding of ξ_n, we now focus on the properties of

$$\frac{1}{n} \mathrm{trace} \left(D_{\lambda_i} D^{1/2}_{\psi_i'(r_{.,[p]})} M D^{1/2}_{\psi_i'(r_{.,[p]})} D_{\lambda_i} \right) \ .$$

The previous lemma shows clearly why this is natural.

About $\frac{1}{n}\mathbf{trace}\left(D_{\lambda_i}D^{1/2}_{\psi_i'(r_{\cdot,[p]})}MD^{1/2}_{\psi_i'(r_{\cdot,[p]})}D_{\lambda_i}\right)$

Lemma D.4. *Let us call* $\mathfrak{S}_p = \frac{1}{n}\sum_{i=1}^n \psi_i'(r_{i,[p]})V_iV_i'$ *and* $\mathfrak{S}_p(i) = \mathfrak{S}_p - \frac{1}{n}\psi_i'(r_{i,[p]})V_iV_i'$. *Let us also call*

$$\begin{aligned}\mathsf{c}_{\tau,p} &= \frac{1}{n}trace\left((\mathfrak{S}_p+\tau\mathrm{Id})^{-1}\right)\ ,\\ \zeta_i &= \frac{1}{n}V_i'(\mathfrak{S}_p(i)+\tau\mathrm{Id})^{-1}V_i - \lambda_i^2\mathsf{c}_{\tau,p}\ .\end{aligned}$$

Then we have under Assumptions ***O1-O7*** *and* ***P1****, if* M *is the matrix defined in Equation* (D-30),

$$\left|\frac{1}{n}trace\,(\mathrm{Id}_n - M) - \left(\frac{1}{n}trace\left(D_{\lambda_i}D^{1/2}_{\psi_i'(r_{\cdot,[p]})}MD^{1/2}_{\psi_i'(r_{\cdot,[p]})}D_{\lambda_i}\right)\right)\mathsf{c}_{\tau,p}\right| \le \left[\sup_i|\zeta_i|\right]\frac{1}{n}\sum_{i=1}^n\psi_i'(r_{i,[p]})\ . \quad \text{(D-32)}$$

We also have

$$\frac{1}{n}trace\,(\mathrm{Id}_n - M) = \frac{p-1}{n} - \tau\mathsf{c}_{\tau,p}\ .$$

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

$$\begin{aligned}M_{i,i} &= 1 - d_{i,i}V_i'(V'D_{\psi_i'(r_{\cdot,[p]})}V/n + \tau\mathrm{Id})^{-1}V_i\ ,\\ &= 1 - d_{i,i}\frac{V_i'(\mathfrak{S}_p(i)+\tau\mathrm{Id})^{-1}V_i}{1+d_{i,i}V_i'(\mathfrak{S}_p(i)+\tau\mathrm{Id})^{-1}V_i}\ ,\\ &= \frac{1}{1+d_{i,i}V_i'(\mathfrak{S}_p(i)+\tau\mathrm{Id})^{-1}V_i}\ .\end{aligned}$$

Recall that we are interested in $\frac{1}{n}\sum_i\lambda_i^2\psi_i'(r_{i,[p]})M_{i,i} = \frac{1}{n}\mathrm{trace}\left(D_{\lambda_i}D^{1/2}_{\psi_i'(r_{\cdot,[p]})}MD^{1/2}_{\psi_i'(r_{\cdot,[p]})}D_{\lambda_i}\right)$. Note that, since $\mathrm{trace}\,(AB) = \mathrm{trace}\,(BA)$,

$$\mathrm{trace}\,(\mathrm{Id}_n - M) = \mathrm{trace}\left((\mathfrak{S}_p+\tau\mathrm{Id})^{-1}\mathfrak{S}_p\right) = p - 1 - \tau\mathrm{trace}\left((\mathfrak{S}_p+\tau\mathrm{Id})^{-1}\right) = p - 1 - n\tau\mathsf{c}_{\tau,p}\ .$$

This shows the second result of the lemma.

On the other hand,

$$\mathrm{trace}\,(\mathrm{Id}_n - M) = \sum_i(1-M_{i,i}) = \sum_i d_{i,i}\frac{V_i'(\mathfrak{S}_p(i)+\tau\mathrm{Id})^{-1}V_i}{1+d_{i,i}V_i'(\mathfrak{S}_p(i)+\tau\mathrm{Id})^{-1}V_i}\ . \quad \text{(D-33)}$$

With our definitions, we have, since $\lambda_i^2\mathsf{c}_{\tau,p} + \zeta_i = \frac{1}{n}V_i'(\mathfrak{S}_p(i)+\tau\mathrm{Id})^{-1}V_i$,

$$\frac{1}{n}\mathrm{trace}\,(\mathrm{Id}_n - M) = \left(\frac{1}{n}\sum_i\lambda_i^2\psi_i'(r_{i,[p]})M_{i,i}\right)\mathsf{c}_{\tau,p} + \frac{1}{n}\sum_i\psi_i'(r_{i,[p]})\frac{\zeta_i}{1+d_{i,i}V_i'(\mathfrak{S}_p(i)+\tau\mathrm{Id})^{-1}V_i}\ .$$

It immediately follows that

$$\left|\frac{1}{n}\mathrm{trace}\,(\mathrm{Id}_n - M) - \left(\frac{1}{n}\sum_i\lambda_i^2\psi_i'(r_{i,[p]})M_{i,i}\right)\mathsf{c}_{\tau,p}\right| \le \left[\sup_i|\zeta_i|\right]\frac{1}{n}\sum_i\psi_i'(r_{i,[p]})\ ,$$

as announced. □

The previous result will be especially useful as an approximation result if we can show that ζ_i's are small, since assumption **P2** - which we will use later - implies that $\frac{1}{n}\sum_{i=1}^n\|\psi_i'\|_\infty$ cannot be too large. This is what we do in the next few pages.

Controlling ζ_i

The main problem that arises when trying to control ζ_i is the fact that $r_{j,[p]}$ appearing in $\mathfrak{S}_p(i)$ depend on V_i. This prevents us from using concentration of quadratic forms results such as those shown in Lemma G.3. So further approximations arguments are needed. Of course, the idea of using a leave-two-out residuals to approximate $\{r_{j,[p]}\}_{j\neq i}$ immediately comes to mind. Hence our work in Section C will later play a key role in showing that ζ_i's are small.

Lemma D.5. *Suppose we can find $\{\mathsf{r}^{(i)}_{j,[p]}\}_{j\neq i}$ independent of $(\lambda_i, \mathcal{V}_i)$ and K_n such that*

$$\sup_i \sup_{j\neq i} |\psi_j'(\mathsf{r}^{(i)}_{j,[p]}) - \psi_j'(r_{j,[p]})| \leq K_n$$

Then

$$\sup_i |\zeta_i| = \mathrm{O}_{L_k}\left(\left[\frac{1}{\tau^2}K_n |||\widehat{\Sigma}|||_2 + \frac{polyLog(n)}{\tau\sqrt{n\mathsf{c}_n}} + \frac{1}{n\tau}\right] polyLog(n)\right) , \tag{D-34}$$

provided K_n has $3k$ uniformly bounded moments.

Proof. We call

$$AM_{i,p} = \frac{1}{n}\sum_{j\neq i} \psi_j'(\mathsf{r}^{(i)}_{j,[p]}) V_j V_j' .$$

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

$$|||(\mathfrak{S}_p(i) + \tau \mathrm{Id})^{-1} - (AM_{i,p} + \tau\mathrm{Id})^{-1}|||_2 \leq \frac{1}{\tau^2}K_n |||\widehat{\Sigma}|||_2 ,$$

since $|||\frac{1}{n}\sum_i V_i V_i'|||_2 \leq |||\widehat{\Sigma}|||_2$. In particular,

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

However, since $AM_{i,p}$ is independent of $(\lambda_i, \mathcal{V}_i)$, we can use Lemma G.3 and see that, since $V_i = \lambda_i \mathcal{V}_i$,

$$\sup_{1\leq i\leq n}\left|\frac{1}{n}V_i'(AM_{i,p} + \tau\mathrm{Id})^{-1}V_i - \frac{\lambda_i^2}{n}\mathrm{trace}\left((AM_{i,p} + \tau\mathrm{Id})^{-1}\right)\right| = \mathrm{O}_{L_k}(\frac{\mathrm{polyLog}(n)}{\tau\sqrt{n\mathsf{c}_n}}\sup_{1\leq i\leq n}\lambda_i^2) ,$$

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

Using the operator norm bound we gave above, we also have

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

We conclude that

$$\sup_{1\leq i\leq n}\left|\frac{1}{n}V_i'(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}V_i - \frac{\lambda_i^2}{n}\mathrm{trace}\left((\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}\right)\right| = \tag{D-35}$$

$$\mathrm{O}_{L_k}\left(\left[\frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2 \sup_{1\leq i\leq n}\left[\frac{p}{n} + \frac{\|\mathcal{V}_i\|^2}{n}\right] + \frac{\mathrm{polyLog}(n)}{\tau\sqrt{n\mathsf{c}_n}}\right][\sup_{1\leq i\leq n}\lambda_i^2 \vee 1]\right) . \tag{D-36}$$

Now, it is clear that under **O1** and **O4**, $\sup_{1\leq i\leq n}\|\mathcal{V}_i\|^2/n = \mathrm{O}_{L_k}(1)$ and finally

$$\sup_{1\leq i\leq n}\left|\frac{1}{n}V_i'(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}V_i - \frac{\lambda_i^2}{n}\mathrm{trace}\left((\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}\right)\right| = \mathrm{O}_{L_k}\left(\left[\frac{1}{\tau^2}K_n|||\widehat{\Sigma}|||_2 + \frac{\mathrm{polyLog}(n)}{\tau\sqrt{n\mathsf{c}_n}}\right][\sup_{1\leq i\leq n}\lambda_i^2 \vee 1]\right) .$$

Control of $\frac{1}{n}$trace $\left((\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}\right) - \frac{1}{n}$trace $\left((\mathfrak{S}_p + \tau\mathrm{Id})^{-1}\right)$

Using the Sherman-Woodbury-Morrison formula, we have

$$(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1} - (\mathfrak{S}_p + \tau\mathrm{Id})^{-1} = \frac{\psi_i'(r_{i,[p]})}{n}\frac{(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}V_iV_i'(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}}{1 + \frac{\psi_i'(r_{i,[p]})}{n}V_i'(\mathfrak{S}_p(i) + \tau\mathrm{Id})^{-1}V_i} .$$

After taking traces, we see that

$$0 \leq \text{trace}\left((\mathfrak{S}_p(i) + \tau \text{Id})^{-1}\right) - \text{trace}\left((\mathfrak{S}_p + \tau \text{Id})^{-1}\right) \leq \frac{1}{\tau} ,$$

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

Therefore,

$$0 \leq \frac{1}{n}\text{trace}\left((\mathfrak{S}_p(i) + \tau \text{Id})^{-1}\right) - \frac{1}{n}\text{trace}\left((\mathfrak{S}_p + \tau \text{Id})^{-1}\right) \leq \frac{1}{n\tau} .$$

We conclude that

$$\sup_{1 \leq i \leq n} |\zeta_i| = \text{O}_{L_k}\left(\left[\frac{1}{\tau^2} K_n |||\widehat{\Sigma}|||_2 + \frac{\text{polyLog}(n)}{\tau\sqrt{n\mathsf{c}_n}} + \frac{1}{n\tau}\right] [\sup_{1 \leq i \leq n} \lambda_i^2 \vee 1]\right) ,$$

provided we can use Holder's inequality. In effect, this requires K_n to have $3k$ moments. □

D-2.3 Control of K_n

A natural choice for $\mathsf{r}_{j,[p]}^{(i)}$ defined in Lemma D.5 is to use a leave one out estimator of $\widehat{\gamma}$, where the i-th observation (and hence V_i) is ommitted. Hence, all the work done in Theorem C.6 becomes immediately relevant.

Lemma D.6. *Suppose we use for $\{\mathsf{r}_{j,[p]}^{(i)}\}_{j \neq i}$ the residuals we would get by using a leave-one-out estimator of $\widehat{\gamma}$, i.e excluding (V_i, ϵ_i) from problem* (D-21).

*With the notations of Lemma D.5, we have under assumptions **O1-O7** and **P1***

$$K_n = \text{O}_{L_k}\left(n^{2\alpha - 1/2} \textit{polyLog}(n)\right) .$$

In particular, for any fixed τ,

$$\sup_i |\zeta_i| = \text{O}_{L_k}\left(n^{2\alpha - 1/2} \textit{polyLog}(n)\right)$$

Proof. Let us call $\delta_n(i)$ random variables such that

$$\sup_{j \neq i} |\mathsf{r}_{j,[p]}^{(i)} - r_{j,[p]}| \leq \delta_n(i) .$$

Applying Theorem C.6 with $R_j = r_{j,[p]}$ and $\tilde{r}_{j,(i)} = \mathsf{r}_{j,[p]}^{(i)}$, we get

$$\sup_i(\delta_n(i)) = \text{O}_{L_k}\left(\frac{\text{polyLog}(n)}{n^{1/2-\alpha}}\right) .$$

The control of K_n follows immediately by using our assumptions on ψ_i', specifically the fact that it is Cn^{α}-Lipschitz. □

Important remark: the previous remark has important consequences for c_i defined in Equation (C-14). Indeed, we have the following corollary.

Corollary D.7. *Let c_i be defined as in Equation* (C-14) *and c_τ be defined as in Equation* (A-7). *Then, under assumptions **O1-O7** and **P1**, we have*

$$\sup_i |c_i - \lambda_i^2 c_\tau| = \text{O}_{L_k}(n^{2\alpha - 1/2} \textit{polyLog}(n)) . \tag{D-37}$$

The corollary follows from drawing analogy between these quantities and the situation investigated in Lemmas D.4, D.5, and D.6; we now give a detailed proof.

Proof. We have now established that

$$\sup_i |\frac{1}{n} V_i'(\mathfrak{S}_p(i) + \tau \mathrm{Id})^{-1} V_i - \lambda_i^2 \mathsf{c}_{\tau,p}| = \mathrm{O}_{L_k}(\frac{\mathrm{polyLog}(n)}{n^{1/2-2\alpha}}) \ .$$

Recalling the notation

$$c_\tau = \frac{1}{n} \mathrm{trace} \left(\left[\frac{1}{n} \sum_{i=1}^n \psi_i'(R_i) X_i X_i' + \tau \mathrm{Id}_p \right]^{-1} \right) \ ,$$

we see that this quantity is the analog of $\mathsf{c}_{\tau,p}$ when we use all the predictors and not only $(p-1)$ of them.

Indeed, c_i in Equation (C-14) is defined, in the notation of the proof of Lemma D.5 as an analog of $\frac{1}{n} V_i'(AM_{i,p} + \tau \mathrm{Id})^{-1} V_i$, with the role of $\{\mathsf{r}_{j,[p]}^{(i)}\}_{j \neq i}$ being played by the residuals obtained from the leave-one-out estimate of $\widehat{\beta}$, excluding (X_i, ϵ_i) from the problem. Lemma D.5 in connection with Theorem D.10 shows that $\sup_i |\frac{1}{n} V_i'(AM_{i,p} + \tau \mathrm{Id})^{-1} V_i - \lambda_i^2 \mathsf{c}_{\tau,p}| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)/n^{1/2-2\alpha})$ under our assumptions. Passing from the $p-1$ dimensional version of this result, i.e Lemma D.5, to the p-dimensional version gives the approximation stated in the corollary.

We therefore see that

$$\sup_i |c_i - \lambda_i^2 c_\tau| = \mathrm{O}_{L_k}(n^{2\alpha - 1/2} \mathrm{polyLog}(n)) \ .$$

□

D-2.4 Further results on ξ_n and $\mathfrak{b}_p$

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

Proposition D.8. *We have, under Assumptions **O1-O7** and **P1**,*

$$\left| \mathsf{c}_{\tau,p}(\xi_n + \tau) - \frac{p-1}{n} \right| = \mathrm{O}_{L_k} \left(\frac{\mathit{polyLog}(n)}{n^{1/2-2\alpha}} \right) \ . \tag{D-38}$$

*Furthermore, under Assumptions **O1-O7** and **P1-P3**, since $\|\beta_0\|_\infty = \mathrm{O}(n^{-\mathsf{e}})$,*

$$\left(\frac{p}{n}\right)^2 n \mathbf{E} \left([\mathfrak{b}_p - \beta_0(p)]^2 \right) = \frac{1}{n} \sum_{i=1}^n \mathbf{E} \left([\mathsf{c}_{\tau,p} \lambda_i \psi_i(r_{i,[p]})]^2 \right) + n\tau^2 \beta_0^2(p) \mathbf{E} \left(\mathsf{c}_{\tau,p}^2 \right) + \mathrm{o}(1) \ . \tag{D-39}$$

Both equations in this proposition are very important for this paper. The first one gives us a very precise idea of the behavior of ξ_n in terms of $\mathsf{c}_{\tau,p}$, which we will see in Section E is relatively easy to understand. This first equation is also a stepping stone towards the first equation of the System (4).

The second equation, on the other hand, is a stepping stone towards the second equation of System (4) in our main theorem, Theorem 2.1.

Proof. • **First equation**
The proof of Equation (D-38) consists just in aggregating all the previous results and noticing that $\mathsf{c}_{\tau,p} \leq (p-1)/(n\tau)$ and therefore remains bounded. Indeed, we have

$$\frac{p-1}{n} - \tau \mathsf{c}_{\tau,p} = \frac{1}{n} \mathrm{trace} \left(\mathrm{Id} - M \right) \geq 0 \ .$$

This latter quantity was approximated in Lemma D.4 by

$$\left(\frac{1}{n} \mathrm{trace} \left(D^{1/2}_{\psi_i'(r_{\cdot,[p]})} M D^{1/2}_{\psi_i'(r_{\cdot,[p]})} \right) \right) \mathsf{c}_{\tau,p} \ .$$

And in Lemma D.3, we approximated ξ_n by $\left(\frac{1}{n} \mathrm{trace} \left(D^{1/2}_{\psi_i'(r_{\cdot,[p]})} M D^{1/2}_{\psi_i'(r_{\cdot,[p]})} \right) \right)$. This gives the result of Equation (D-38), by simply keeping track of the approximation errors we make at each step.

• **Second equation**

Recall that by definition (see Equations (D-24) and (D-23)),

$$\sqrt{n}\left[(\tau+\xi_n)\mathfrak{b}_p-\xi_n\beta_0(p)\right]=N_p=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\lambda_i\psi_i(r_{i,[p]})\mathcal{X}_i(p)\ .$$

Therefore,

$$\mathsf{c}_{\tau,p}\sqrt{n}\left[(\tau+\xi_n)[\mathfrak{b}_p-\beta_0(p)]+\tau\beta_0(p)\right]=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathsf{c}_{\tau,p}\lambda_i\psi_i(r_{i,[p]})\mathcal{X}_i(p)\ ,$$

or

$$\mathsf{c}_{\tau,p}\sqrt{n}(\tau+\xi_n)[\mathfrak{b}_p-\beta_0(p)]=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathsf{c}_{\tau,p}\lambda_i\psi_i(r_{i,[p]})\mathcal{X}_i(p)-\mathsf{c}_{\tau,p}\sqrt{n}\tau\beta_0(p)\ .$$

We note that $\mathsf{c}_{\tau,p}\lambda_i\psi_i(r_{i,[p]})$, which depends only on $\{\lambda_i,\mathcal{V}_i,\epsilon_i\}_{i=1}^n$, is independent of $\{\mathcal{X}_i(p)\}_{i=1}^n$. (If needed, see the definition of $\mathsf{c}_{\tau,p}$ in Lemma D.4.)

Since $\mathcal{X}_i(p)$'s are independent with mean 0 and variance 1, we conclude that

$$\mathbf{E}\left(\mathsf{c}_{\tau,p}^2 n(\tau+\xi_n)^2\left[\mathfrak{b}_p-\beta_0(p)\right]^2\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left([\mathsf{c}_{\tau,p}\lambda_i\psi_i(r_{i,[p]})]^2\right)+n\tau^2\beta_0^2(p)\mathbf{E}\left(\mathsf{c}_{\tau,p}^2\right)\ .$$

Given the result in Equation (D-38) and our bound on $\sqrt{n}[\mathfrak{b}_p-\beta_0(p)]$ in Proposition D.2, this means that

$$\left(\frac{p}{n}\right)^2 n\mathbf{E}\left([\mathfrak{b}_p-\beta_0(p)]^2\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left([\mathsf{c}_{\tau,p}\lambda_i\psi_i(r_{i,[p]}]^2\right)+n\tau^2\beta_0^2(p)\mathbf{E}\left(\mathsf{c}_{\tau,p}^2\right)+\mathrm{o}(1)\ .$$

In this last equation, we make use of Proposition D.2 and Assumption **P3** since under this assumption $n\|\beta_0\|_\infty^2\mathrm{polyLog}(n)n^{2\alpha-1/2}\to 0$. This is what allows us to replace $\mathsf{c}_{\tau,p}(\tau+\xi_n)$ by p/n without loss of accuracy in going from the second-to-last to the last equation. □

We now need to control $\mathsf{d}_{i,p}$ to show that our approximation of $\widehat{\beta}$ by $\widetilde{b}$ in Proposition D.1 will yield sufficiently good results that they can be used to prove Theorem 2.1.

D-2.5 On $\mathsf{d}_{i,p}$

Recall the definition

$$\mathsf{d}_{i,p}=\left[\psi_i'(\gamma_{i,p}^*)-\psi_i'(r_{i,[p]})\right]\ ,$$

where $\gamma_{i,p}^*\in(r_{i,[p]},r_{i,[p]}+\nu_i)$, with

$$\nu_i=[\mathfrak{b}_p-\beta_0(p)]X_i'\begin{bmatrix}(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p\\-1\end{bmatrix}=[\mathfrak{b}_p-\beta_0(p)]\pi_i\ .$$

(The fact that $\gamma_{i,p}^*\in(r_{i,[p]},r_{i,[p]}+\nu_i)$ follows from writing the definition of $Y_i-X_i'\widetilde{b}$.)

We have the following result.

Proposition D.9. *We have, under Assumptions **O1-O7** and **P1-P3**, at fixed τ,*

$$\sup_i|\mathsf{d}_{i,p}|=\mathrm{O}_{L_k}\left(\frac{\mathit{polyLog}(n)n^\alpha}{n^{1/2}\wedge n^{\mathsf{e}}}\right)\ .$$

Proof. Recall the definition

$$\mathsf{d}_{i,p}=\left[\psi_i'(\gamma_{i,p}^*)-\psi_i'(r_{i,[p]})\right]\ ,$$

where $\gamma_{i,p}^*\in(r_{i,[p]},r_{i,[p]}+\nu_i)$, with ν_i defined above. Therefore,

$$\pi_i=V_i'(\mathfrak{S}_p+\tau\mathrm{Id})^{-1}u_p-X_i(p)\ .$$

Recall that $u_p = \frac{1}{n} V' D_{\psi_i'(r_{\cdot,[p]})} X(p)$. We can also rewrite it as

$$u_p = \frac{1}{n} \mathcal{V}' D_{\lambda_i^2 \psi_i'(r_{\cdot,[p]})} \mathcal{X}(p) \ .$$

Using independence of $\mathcal{X}(p)$ with $\{(\mathcal{V}_i, \epsilon_i)\}_{i=1}^n$, and our concentration assumptions on $\mathcal{X}(p)$ formulated in **P1**, we see that according to Lemma G.2, we have

$$\sup_i |V_i'(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p| = \mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\mathsf{c}_n^{1/2}} \sup_i \| \frac{1}{n} D_{\lambda_i^2 \psi_i'(r_{\cdot,[p]})} \mathcal{V} (\mathfrak{S}_p + \tau \mathrm{Id})^{-1} \mathcal{V}_i \| \right) \ ,$$

where we look at $V_i'(\mathfrak{S}_p + \tau \mathrm{Id})^{-1} u_p$ as a linear form in $\mathcal{X}(p)$. Note that we have absorbed the $\sup_i |\lambda_i|$ in the $\mathrm{polyLog}(n)$ term.

Now,

$$\| \frac{1}{n} D_{\lambda_i^2 \psi_i'(r_{\cdot,[p]})} \mathcal{V} (\mathfrak{S}_p + \tau \mathrm{Id})^{-1} \mathcal{V}_i \|^2 = \frac{1}{n} \mathcal{V}_i' (\mathfrak{S}_p + \tau \mathrm{Id})^{-1} \frac{\mathcal{V}' D^2_{\lambda_i^2 \psi_i'(r_{\cdot,[p]})} \mathcal{V}}{n} (\mathfrak{S}_p + \tau \mathrm{Id})^{-1} \mathcal{V}_i \ .$$

Notice that $\mathfrak{S}_p = \frac{\mathcal{V}' D_{\lambda_i^2 \psi_i'(r_{\cdot,[p]})} \mathcal{V}}{n}$. Hence, $\frac{\mathcal{V}' D^2_{\lambda_i^2 \psi_i'(r_{\cdot,[p]})} \mathcal{V}}{n} \preceq |||D_{\lambda_i^2 \psi_i'(r_{\cdot,[p]})}|||_2 \mathfrak{S}_p$ and we conclude that

$$\frac{1}{n} \mathcal{V}_i' (\mathfrak{S}_p + \tau \mathrm{Id})^{-1} \frac{\mathcal{V}' D^2_{\lambda_i^2 \psi_i'(r_{\cdot,[p]})} \mathcal{V}}{n} (\mathfrak{S}_p + \tau \mathrm{Id})^{-1} \mathcal{V}_i \leq \frac{\|\mathcal{V}_i\|^2}{n\tau} |||D_{\lambda_i^2 \psi_i'(r_{\cdot,[p]})}|||_2 = \frac{\|\mathcal{V}_i\|^2}{n\tau} \sup_i \lambda_i^2 \psi_i'(r_{i,[p]}) \ .$$

We also note that $\sup_i |X_i(p)| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n)/\sqrt{\mathsf{c}_n})$ under **O4**, **O6** and **P1**, using the results of Subsection G. So we conclude that

$$\begin{aligned}
\sup_i |\pi_i| &= \mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\mathsf{c}_n^{1/2}} \left[1 + \sqrt{\sup_i \lambda_i^2 \psi_i'(r_{i,[p]}) \sup_i \frac{\|\mathcal{V}_i\|^2}{n\tau}} \right] \right) \ , \\
&= \mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\mathsf{c}_n^{1/2}} \left[1 + \sqrt{\sup_i \lambda_i^2 \psi_i'(r_{i,[p]})} \right] \right) \ , \\
&= \mathrm{O}_{L_k}\left(\mathrm{polyLog}(n) \right) \ .
\end{aligned}$$

Recalling that $|\mathfrak{b}_p - \beta_0(p)| = \mathrm{O}_{L_k}(n^{-1/2}\mathrm{polyLog}(n) + \|\beta_0\|_\infty)$, we finally see that

$$\sup_i |\nu_i| = \mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n)}{\sqrt{n} \wedge n^{\mathsf{e}}} \right)$$

Under our assumption that ψ_i' is Cn^α-Lipschitz, we see that

$$\sup_i |\mathsf{d}_{i,p}| = \mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n) n^\alpha}{n^{1/2} \wedge n^{\mathsf{e}}} \right) \ .$$

□

D-3 Final conclusions

We can now gather together our approximation results in the following Theorem.

Theorem D.10. *Under Assumptions **O1-O7** and **P1-P3**, we have, for any fixed $\tau > 0$,*

$$\|\widehat{\beta} - \widetilde{b}\| \leq \mathrm{O}_{L_k}\left(\frac{\mathit{polyLog}(n) n^\alpha}{[n^{1/2} \wedge n^{\mathsf{e}}]^2} \right)$$

In particular,

$$\sqrt{n}(\widehat{\beta}_p - \mathfrak{b}_p) = \mathrm{O}_{L_k}\left(\frac{polyLog(n)n^{\alpha+1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right) ,$$
$$\sup_i |X_i'(\widehat{\beta}-\widetilde{b})| = \mathrm{O}_{L_k}\left(\frac{polyLog(n)n^{\alpha+1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right) ,$$
$$\sup_i |R_i - r_{i,[p]}| = \mathrm{O}_{L_k}\left(\left[\frac{polyLog(n)}{\sqrt{n}\wedge n^{\mathsf{e}}}\right]\vee\left[\frac{polyLog(n)n^{\alpha+1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right]\right) .$$

We note that the index p in the previous theorem plays no particular role and similar results holds when p is replaced by $1\le k\le p$.

Proof. The Theorem is just the aggregation of all of our results, using the key bound on $\|\widehat{\beta}-\widetilde{b}\|$ in Proposition D.1.

The last statement is the only one that might need an explanation. With the notations of the proof of Proposition D.9, we have $R_i - r_{i,[p]} = X_i'(\widetilde{b}-\widehat{\beta}) - \nu_i$. The results on $\sup_i|\nu_i|$ in the proof of Proposition D.9 as well as the bound on $\sup_i|X_i'(\widetilde{b}-\widehat{\beta})|$ give us the announced result. □

Combining the results of Equation (D-39) and the previous theorem, we see that under Assumptions **O1-O7** and **P1-P4**,

$$\left(\frac{p}{n}\right)^2 n\mathbf{E}\left((\widehat{\beta}_p-\beta_0(p))^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left([\mathsf{c}_{\tau,p}\lambda_i\psi_i(r_{i,[p]})]^2\right) + n\tau^2\beta_0^2(p)\mathbf{E}\left(\mathsf{c}_{\tau,p}^2\right) + \mathrm{o}(1) .$$

Since p did not play any particular role as compared to any other index in our analysis, the same result holds when p is replaced by k, $1\le k\le p$.

Dividing the previous expression by n on both sides and summing over all the indices $1\le k\le p$, we finally get

$$\left(\frac{p}{n}\right)^2 \mathbf{E}\left(\|\widehat{\beta}-\beta_0\|_2^2\right) = \frac{1}{n}\sum_{k=1}^p\left[\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left([\mathsf{c}_{\tau,k}\lambda_i\psi_i(r_{i,[k]})]^2\right)\right] + \tau^2\sum_{k=1}^p\beta_0^2(k)\mathbf{E}\left(\mathsf{c}_{\tau,k}^2\right) + \mathrm{o}(1) . \tag{D-40}$$

Our aim now is to further simplify the above expression to get the second equation of our system.

D-3.1 On $\mathsf{c}_{\tau,p}$ and c_τ

We now show that $\mathsf{c}_{\tau,k}$'s are all close to the same quantity, which turns out to be c_τ.

Proposition D.11. *We have, under Assumptions **O1-O7** and **P1-P3**,*

$$\sup_{1\le k\le p}|c_\tau - \mathsf{c}_{\tau,k}| = \mathrm{O}_{L_k}\left(\left[\frac{polyLog(n)n^{\alpha}}{\sqrt{n}\wedge n^{\mathsf{e}}}\right]\vee\left[\frac{polyLog(n)n^{2\alpha+1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right]\vee\frac{polyLog(n)}{n}\right) .$$

Of course, we also have $0\le c_\tau\le p/(n\tau)$ and $0\le\mathsf{c}_{\tau,k}\le p/(n\tau)$.

Proof. Let us recall the notation

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

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

$$S = \begin{pmatrix}\Gamma & \mathsf{v}\\ \mathsf{v} & a\end{pmatrix} .$$

According to Lemma H.2, we see that, since $c_\tau = \frac{1}{n}\text{trace}\left((S+\tau \text{Id}_p)^{-1}\right)$,

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

It is clear that under our assumptions, $a = \text{O}_{L_k}(\text{polyLog}(n))$, since

$$a = \frac{1}{n}\sum_{i=1}^n \lambda_i^2 \mathcal{X}_i^2(p)\psi_i'(R_i) \leq polyLog(n)\frac{1}{n}\sum_{i=1}^n \lambda_i^2\mathcal{X}_i^2(p) = \text{O}_{L_k}(\text{polyLog}(n)) ,$$

(using e.g our work in Section G). Since ψ_i' is Cn^α-Lipschitz and

$$\sup_i |R_i - r_{i,[p]}| = \text{O}_{L_k}\left(\left[\frac{\text{polyLog}(n)}{\sqrt{n}\wedge n^{\mathsf{e}}}\right] \vee \left[\frac{\text{polyLog}(n)n^{\alpha+1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right]\right) ,$$

we have

$$\sup_i |\psi_i'(R_i) - \psi_i'(r_{i,[p]})| = \text{O}_{L_k}\left(\left[\frac{\text{polyLog}(n)n^{\alpha}}{\sqrt{n}\wedge n^{\mathsf{e}}}\right] \vee \left[\frac{\text{polyLog}(n)n^{2\alpha+1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right]\right) .$$

Hence, using arguments similar to the ones we have used in the proof of Lemma D.5 (i.e first resolvent identity, etc...), we see that

$$\left|\frac{1}{n}\text{trace}\left((\Gamma+\tau \text{Id})^{-1}\right) - \frac{1}{n}\text{trace}\left((\mathfrak{S}_p+\tau \text{Id})^{-1}\right)\right| = \text{O}_{L_k}\left(\left[\frac{\text{polyLog}(n)n^{\alpha}}{\sqrt{n}\wedge n^{\mathsf{e}}}\right] \vee \left[\frac{\text{polyLog}(n)n^{2\alpha+1/2}}{[n^{1/2}\wedge n^{\mathsf{e}}]^2}\right]\right) .$$

Since $\mathsf{c}_{\tau,p} = \frac{1}{n}\text{trace}\left((\mathfrak{S}_p+\tau \text{Id})^{-1}\right)$, the result we announced follows immediately.

We note that p did not play a particular role here and hence taking the sup over those indices only adds a polyLog(n) term to the approximation. Hence our approximation is valid also for $\sup_{1\leq k\leq n}|c_\tau - \mathsf{c}_{\tau,k}|$. □

We are now ready to prove the last proposition of this section, which will help us get the second equation of our System (4).

Proposition D.12. *Under Assumptions **O1-O7** and **P1-P4**,*

$$\left(\frac{p}{n}\right)^2 \mathbf{E}\left(\|\widehat{\beta}-\beta_0\|_2^2\right) = \frac{p}{n}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left([c_\tau\lambda_i\psi_i(prox(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)}))]^2\right) + \tau^2\|\beta_0\|^2\mathbf{E}\left(c_\tau^2\right) + \text{o}(1) . \tag{D-41}$$

Furthermore, when all λ_i's are non-zero,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left([c_\tau\lambda_i\psi_i(prox(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)}))]^2\right) = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\frac{[\tilde{r}_{i,(i)} - prox(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})]^2}{\lambda_i^2}\right)$$

Proof. In light of the result in Proposition D.11 and Assumption **P3** which guarantees that $\|\beta_0\|_2^2$ is uniformly bounded in p and n, we see that

$$\sum_{k=1}^p \beta_0^2(k)\mathbf{E}\left(\mathsf{c}_{\tau,k}^2\right) = \sum_{k=1}^p \beta_0^2(k)\mathbf{E}\left(c_\tau^2\right) + \text{o}(1) = \|\beta_0\|_2^2\mathbf{E}\left(c_\tau^2\right) + \text{o}(1) .$$

Using Theorems C.6 and D.10, Equation (D-40) implies that

$$\left(\frac{p}{n}\right)^2 \mathbf{E}\left(\|\widehat{\beta}-\beta_0\|_2^2\right) = \frac{p}{n}\frac{1}{p}\sum_{k=1}^p\left[\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left([\mathsf{c}_{\tau,k}\lambda_i\psi_i(r_{i,[k]})]^2\right)\right] + \tau^2\|\beta_0\|_2^2\mathbf{E}\left(c_\tau^2\right) + \text{o}(1) .$$

Using Theorem D.10 and our bound on $\|\psi_i'\|_\infty$ from Assumption **O3**, we see that

$$\frac{1}{p}\sum_{k=1}^p \mathbf{E}\left([\mathsf{c}_{\tau,k}\lambda_i\psi_i(r_{i,[k]})]^2\right) = \frac{1}{p}\sum_{k=1}^p \mathbf{E}\left([\mathsf{c}_{\tau,k}\lambda_i\psi_i(R_i)]^2\right) + \text{o}(1) .$$

Thanks to Proposition D.11 we also have

$$\frac{1}{p}\sum_{k=1}^{p}\mathbf{E}\left([\mathsf{c}_{\tau,k}\lambda_i\psi_i(R_i)]^2\right)=\frac{1}{p}\sum_{k=1}^{p}\mathbf{E}\left([c_\tau\lambda_i\psi_i(R_i)]^2\right)+\mathrm{o}(1)\ .$$

In light of Equation (C-20) and Assumption **O3**, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left([c_\tau\lambda_i\psi_i(R_i)]^2\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left([c_\tau\lambda_i\psi_i(\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)}))]^2\right)+\mathrm{o}(1)$$

Using Lemma F.2 and specifically the computation of the derivative of $\mathrm{prox}(c\rho)(x)$ with respect to c, we see that, by using Corollary D.7, we can re-express the previous equation as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left([c_\tau\lambda_i\psi_i(R_i)]^2\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left([c_\tau\lambda_i\psi_i(\mathrm{prox}(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)}))]^2\right)+\mathrm{o}(1)\ ,$$

since Equation (D-37) in Corollary D.7, gives $\sup_i|c_i-\lambda_i^2c_\tau|=\mathrm{O}_{L_k}(n^{2\alpha-1/2}\mathrm{polyLog}(n))$.

When λ_i's are all different from 0, we can rewrite this equation as

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left([c_\tau\lambda_i\psi_i(R_i)]^2\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\frac{[c_\tau\lambda_i^2\psi_i(\mathrm{prox}(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)}))]^2}{\lambda_i^2}\right)+\mathrm{o}(1)$$

Finally, since almost by definition,

$$\forall x\in\mathbb{R}\ ,x=\mathrm{prox}(c\rho)(x)+c\psi(\mathrm{prox}(c\rho)(x))\ ,$$

we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\frac{[c_\tau\lambda_i^2\psi_i(\mathrm{prox}(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)}))]^2}{\lambda_i^2}\right)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\frac{[\tilde{r}_{i,(i)}-\mathrm{prox}(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})]^2}{\lambda_i^2}\right)\ .$$

□

E Last steps of the proof

We now reach the last steps of the proof and two imporant tasks remain. The first one is understanding the limiting behavior of $\tilde{r}_{i,(i)}$ and showing that it behaves like $\epsilon_i+\lambda_i r_\rho(\kappa)Z_i$ in the limit, where $Z_i\sim\mathcal{N}(0,1)$. With a little bit of further work, the corresponding results will give us in connection with Proposition D.12 the second equation of our main system (4).

The second main task is then to show that c_τ is asymptotically deterministic, i.e it converges towards a non-random number.

E-1 On the asymptotic distribution of $\tilde{r}_{i,(i)}$

We have the following lemma.

Lemma E.1. *Under Assumptions **O1-O7** and **P1-P4**, as n and p tend to infinity, $\tilde{r}_{i,(i)}=Y_i-X_i'\widehat{\beta}_{i,(i)}$ behaves like $\epsilon_i+\lambda_i\sqrt{\mathbf{E}\left(\|\widehat{\beta}-\beta_0\|^2\right)}Z_i$, where $Z_i\sim\mathcal{N}(0,1)$ is independent of ϵ_i and λ_i, in the sense of weak convergence.*

Furthermore, if $i\neq j$, $\tilde{r}_{i,(i)}$ and $\tilde{r}_{j,(j)}$ are asymptotically (pairwise) independent. The same is true for the pairs $(\tilde{r}_{i,(i)},\lambda_i)$ and $(\tilde{r}_{j,(j)},\lambda_j)$

Proof. We recall that $X_i = \lambda_i \mathcal{X}_i$ and hence $\tilde{r}_{i,(i)} = Y_i - X_i' \widehat{\beta}_{(i)} = \epsilon_i - \lambda_i \mathcal{X}_i'(\widehat{\beta}_{(i)} - \beta_0)$.

• **First part** The only problem is of course showing that $(\widehat{\beta}_{(i)} - \beta_0)' \mathcal{X}_i$ is approximately $\mathcal{N}(0, \mathbf{E}\left(\|\widehat{\beta} - \beta_0\|^2\right))$. Recall that $\widehat{\beta}_{(i)}$ is independent of $\mathcal{X}_i$ and that $\mathcal{X}_i$ has mean 0, $\mathrm{cov}\left(\mathcal{X}_i\right) = \mathrm{Id}_p$ and that, for any finite k, the first k absolute moments of its entries are assumed to be bounded uniformly in n.

Recall that we showed in Proposition C.7 that $\mathrm{var}\left(\|\widehat{\beta} - \beta_0\|^2\right) \to 0$. Thanks to Lemma B.3, we also know that $\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|^2\right)$ is uniformly bounded. Furthermore, in the proof of Proposition C.7, we showed that $\mathbf{E}\left(\|\widehat{\beta}\|^2 - \|\widehat{\beta}_{(i)}\|^2\right) \to 0$ and that $\mathbf{E}\left(\|\widehat{\beta} - \beta_0\|^2 - \|\widehat{\beta}_{(i)} - \beta_0\|^2\right) \to 0$.

Let us now show that $(\widehat{\beta}_{(i)} - \beta_0)' \mathcal{X}_i$ behaves like $\mathcal{N}(0, \mathbf{E}\left(\|\widehat{\beta}_{(i)} - \beta_0\|^2\right))$. We employ a similar strategy as was done in El Karoui (2013) but give the argument in details since it requires some new work.

We need a simple generalization of the standard Lindeberg-Feller theorem (see e.g Stroock (1993)). Indeed, if $a_{n,p}(k)$ are random variables with $\sqrt{\sum_{k=1}^p a_{n,p}(k)^2} = A_n$, $\mathbf{E}\left(A_n^2\right)$ remains bounded in n, and $a_{n,p}(k)'s$ are independent of $\mathcal{X}_i$, we see that: a) if $\mathsf{Z} \sim \mathcal{N}(0, \mathrm{Id}_p)$, independent of $a_{n,p}(k)$, then $a_{n,p}'\mathsf{Z} \sim A_n \mathsf{N}$ where $\mathsf{N} \sim \mathcal{N}(0,1)$ and independent of A_n (conditionally and unconditionally on $a_{n,p}$); b) Theorem 2.1.5 and its proof in Stroock (1993) hold provided $\sum_{i=1}^n \mathbf{E}\left(|a_{n,p}(k)|^3\right) = \mathrm{o}(1)$. The proof simply needs to be started conditionally on $a_{n,p}$, and the final moment bounds are then taken unconditionally. This very mild generalization gives, if ϕ is a $\mathcal{C}^3$ function, with bounded 2nd and third derivatives,

$$\forall \epsilon > 0\ , \left|\mathbf{E}\left(\phi(a_{n,p}'\mathcal{X}_i)\right) - \mathbf{E}\left(\phi(A_n \mathsf{N})\right)\right| \leq K\left(\epsilon \|\phi^{(3)}\|_\infty \mathbf{E}\left(\sum_{k=1}^p a_{n,p}(k)^2\right) + \frac{\|\phi^{(2)}\|_\infty}{\epsilon} \sum_{k=1}^p \mathbf{E}\left(|a_{n,p}(k)|^3\right)\right)\ ,$$

where K is a constant that depends on the second and third absolute moments of the entries of $\mathcal{X}_i$. It is therefore independent of n and p under our assumptions on $\mathcal{X}_i$.

To make matters clearer, we allow ourselves to use the notation v_k or $v(k)$ to refer to the k-th coordinate of the vector v.

In our setting, $a_{n,p}(k) = \widehat{\beta}_{(i)}(k) - \beta_0(k)$. Recall that we have shown that

$$\widehat{\beta}(p) - \beta_0(p) = \mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n) n^\alpha}{[n^{1/2} \wedge n^{\mathsf{e}}]^2}\right)\ .$$

The same arguments we used apply also to $\widehat{\beta}_{(i)}(p)$, the p-th coordinate of the leave-one-out estimate $\widehat{\beta}_{(i)}$. So it is clear that

$$\mathbf{E}\left(|\widehat{\beta}_{(i)}(p) - \beta_0(p)|^3\right) = \mathrm{O}\left(\frac{\mathrm{polyLog}(n) n^{3\alpha}}{[n^{1/2} \wedge n^{\mathsf{e}}]^6}\right)\ .$$

We conclude that

$$\mathbf{E}\left(\sum_{k=1}^p |\widehat{\beta}_{(i)}(k) - \beta_0(k)|^3\right) = \mathrm{O}\left(\frac{\mathrm{polyLog}(n) n^{3\alpha+1}}{[n^{1/2} \wedge n^{\mathsf{e}}]^6}\right) = \mathrm{o}(1)\ .$$

This, in connection with Corollary 2.1.9 in Stroock (1993), shows that $(\widehat{\beta}_{(i)} - \beta_0)' \mathcal{X}_i$ behaves asymptotically like $\|\widehat{\beta}_{(i)} - \beta_0\| \mathsf{N}$ in the sense of weak convergence.

Since $\|\widehat{\beta}_{(i)} - \beta_0\| - \mathbf{E}\left(\|\widehat{\beta}_{(i)} - \beta_0\|\right) \to 0$ in probability and $\mathbf{E}\left(\|\widehat{\beta}_{(i)} - \beta_0\|\right)$ remains bounded under our assumptions, Slutsky's lemma guarantees that

$$(\widehat{\beta}_{(i)} - \beta_0)' \mathcal{X}_i \text{ behaves like } \mathbf{E}\left(\|\widehat{\beta}_{(i)} - \beta_0\|\right) \mathsf{N}$$

asymptotically, in the sense of weak convergence - by which we mean that the difference of their characteristic functions goes to 0. Using the fact, which can be shown using results in the proof of Proposition C.7, that $\mathbf{E}\left(\|\widehat{\beta}_{(i)} - \beta_0\|\right) - \mathbf{E}\left(\|\widehat{\beta} - \beta_0\|\right) \to 0$ and Slutsky's lemma, we see that

$$(\widehat{\beta}_{(i)} - \beta_0)' \mathcal{X}_i \text{ behaves like } \mathbf{E}\left(\|\widehat{\beta} - \beta_0\|\right) \mathsf{N}\ ,$$

in the sense of weak convergence.

We note that the same reasoning applies when replacing $a_{n,p}(k) = \widehat{\beta}_{(i)}(k) - \beta_0(k)$ by $\tilde{a}_{n,p}(k) = \lambda_i[\widehat{\beta}_{(i)}(k) - \beta_0(k)]$, provided λ_i has 3 moments. This shows that

$$\lambda_i(\widehat{\beta}_{(i)} - \beta_0)'\mathcal{X}_i = (\widehat{\beta}_{(i)} - \beta_0)'X_i \text{ behaves like } \mathbf{E}\left(\|\widehat{\beta} - \beta_0\|\right)\lambda_i \mathsf{N} \ ,$$

This shows the first part of the lemma.

• **Second part** For the second part, we use a leave-two-out approach, namely we use the approximation $\tilde{r}_{i,(i)} = \epsilon_i + X_i'\beta_0 - \widehat{\beta}'_{(i)}X_i = \epsilon_i + X_i'\beta_0 - \widehat{\beta}'_{(ij)}X_i + \mathrm{o}_{L_k}(1)$ and similarly for $\tilde{r}_{j,(j)}$ (this is clear from Theorem 2.2; $\widehat{\beta}_{(ij)}$ is computed by solving Problem (3) without (X_i, ϵ_i) and (X_j, ϵ_j)). It is clear that $\tilde{r}_{i,(i)}$ and $\tilde{r}_{j,(j)}$ are asymptotically independent conditional on $X_{(ij)}$, i.e all the predictors except X_i and X_j. But because their dependence on $X_{(ij)}$ is only through $\|\widehat{\beta}_{(ij)} - \beta_0\|$, which is asymptotically deterministic by arguments similar to those used in the proof of Proposition C.7, we see that $\tilde{r}_{i,(i)}$ and $\tilde{r}_{j,(j)}$ are asymptotically independent.

After this high-level explanation, let us now give a detailed proof. The arguments we gave above apply to $\widehat{\beta}_{(ij)}$ as they did to $\widehat{\beta}_{(i)}$. In particular, since

$$\mathbf{E}\left(\sum_{k=1}^{p} |\widehat{\beta}_{(ij)}(k) - \beta_0(k)|^3\right) = \mathrm{O}(\frac{\mathrm{polyLog}(n)n^{3\alpha+1}}{[n^{1/2} \wedge n^{\mathsf{e}}]^6}) = \mathrm{o}(1) \ ,$$

we also have

$$\sum_{k=1}^{p} |\widehat{\beta}_{(ij)}(k) - \beta_0(k)|^3 = \mathrm{o}_P(1).$$

Of course, $\widehat{\beta}_{(ij)}$ depends only on $\{X_{(ij)}, \epsilon_{(ij)}\}$. We call $P_{(ij)}$ the joint probability measure $P_{(ij)} = \prod_{k \neq (i,j)} P_{X_k, \epsilon_k}$, i.e probability computed with respect to all our random variables except (X_i, ϵ_i) and (X_j, ϵ_j) (we slightly abuse notation and do not index this probability measure by n for the sake of clarity).

So we have found $E^n_{(ij)}$, depending only on $(X_{(ij)}, \epsilon_{(ij)})$, such that $P_{(ij)}(E^n_{(ij)}) \to 1$ and $\sum_{k=1}^{p} |\widehat{\beta}_{(ij)}(k) - \beta_0(k)|^3 = \mathrm{o}(1)$ when $(X_{(ij)}, \{\epsilon_k\}_{k \neq (i,j)}) \in E^n_{(ij)}$. The arguments we gave above (treating $a_{n,p}$'s as deterministic quantities) then imply that, when $(X_{(ij)}, \epsilon_{(i,j)}) \in E^n_{(ij)}$,

$$(\widehat{\beta}_{(ij)} - \beta_0)'\mathcal{X}_i | (X_{(ij)}, \epsilon_{(ij)}) \text{ behaves like } \|\widehat{\beta}_{(ij)} - \beta_0\| \mathsf{N} \ .$$

Let us now use characteristic function arguments. Let $\alpha_i = (\widehat{\beta}_{(ij)} - \beta_0)'\mathcal{X}_i$ and $\alpha_j = (\widehat{\beta}_{(ij)} - \beta_0)'\mathcal{X}_j$

Let $(w_i, w_j) \in \mathbb{R}^2$ be fixed and

$$\chi(w_i, w_j) = \mathbf{E}\left(\mathrm{e}^{\imath(w_1\alpha_i + w_2\alpha_j)}\right) = \mathbf{E}\left(\mathrm{e}^{\imath(w_1\alpha_i + w_2\alpha_j)}\left[1_{E^n_{(ij)}} + 1_{[E^n_{(ij)}]^c}\right]\right) \ .$$

Since $P([E^n_{(ij)}]^c) = P_{(ij)}([E^n_{(ij)}]^c) \to 0$, we can just focus on $\mathbf{E}\left(\mathrm{e}^{\imath(w_1\alpha_i + w_2\alpha_j)}1_{E^n_{(ij)}}\right)$, since the modulus of the functions we are integrating is bounded by 1.

Now

$$\mathbf{E}\left(\mathrm{e}^{\imath(w_1\alpha_i + w_2\alpha_j)}1_{E^n_{(ij)}}\right) = \mathbf{E}\left(1_{E^n_{(ij)}}\mathbf{E}\left(\mathrm{e}^{\imath(w_1\alpha_i + w_2\alpha_j)} | X_{(ij)}, \epsilon_{(ij)}\right)\right) \ ,$$

since $1_{E^n_{(ij)}}$ is a deterministic function of $(X_{(ij)}, \epsilon_{(ij)})$. Independence of $\mathcal{X}_i$ and $\mathcal{X}_j$ implies that

$$\mathbf{E}\left(\mathrm{e}^{\imath(w_1\alpha_i + w_2\alpha_j)} | X_{(ij)}, \epsilon_{(ij)}\right) = \mathbf{E}\left(\mathrm{e}^{\imath w_1\alpha_i} | X_{(ij)}, \epsilon_{(ij)}\right)\mathbf{E}\left(\mathrm{e}^{\imath w_2\alpha_j} | X_{(ij)}, \epsilon_{(ij)}\right) \ .$$

Also, our conditional asymptotic normality arguments above imply that

$$1_{E^n_{(ij)}}\left[\mathbf{E}\left(\mathrm{e}^{\imath w_1\alpha_i} | X_{(ij)}, \epsilon_{(ij)}\right) - \mathrm{e}^{-w_1^2/2\|\widehat{\beta}_{(ij)} - \beta_0\|^2}\right] \to 0$$

in $P_{(ij)}$-probability. We therefore have

$$1_{E^n_{(ij)}}\left[\mathbf{E}\left(\mathrm{e}^{\imath(w_1\alpha_i + w_2\alpha_j)} | X_{(ij)}, \epsilon_{(ij)}\right) - \mathrm{e}^{-(w_1^2/2 + w_2^2/2)\|\widehat{\beta}_{(ij)} - \beta_0\|^2}\right] \to 0$$

in $P_{(ij)}$-probability.

So we conclude that

$$\mathbf{E}\left(1_{E^n_{(ij)}}\mathrm{e}^{\imath(w_1\alpha_i+w_2\alpha_j)}\right)-\mathbf{E}\left(1_{E^n_{(ij)}}\mathrm{e}^{-(w_1^2/2+w_2^2/2)\|\widehat{\beta}_{(ij)}-\beta_0\|^2}\right)\to 0\,.$$

Since $P(E^n_{(ij)})\to 1$ and $\|\widehat{\beta}_{(ij)}-\beta_0\|^2$ is asymptotically deterministic by arguments similar to those used in the proof of Proposition C.7, we see that

$$\mathbf{E}\left(1_{E^n_{(ij)}}\mathrm{e}^{-(w_1^2/2+w_2^2/2)\|\widehat{\beta}_{(ij)}-\beta_0\|^2}\right)-\mathrm{e}^{-\left[(w_1^2/2+w_2^2/2)\mathbf{E}\left(\|\widehat{\beta}_{(ij)}-\beta_0\|^2\right)\right]}\to 0\,.$$

Therefore,

$$\mathbf{E}\left(\mathrm{e}^{\imath(w_1\alpha_i+w_2\alpha_j)}\right)-\mathbf{E}\left(\mathrm{e}^{\imath w_1\alpha_i}\right)\mathbf{E}\left(\mathrm{e}^{\imath w_2\alpha_j}\right)\to 0\,.$$

This proves that α_i and α_j are asymptotically independent. It easily follows that the same is true for $\tilde{r}_{i,(i)}$ and $\tilde{r}_{j,(j)}$.

The same leave-two-out approach also shows asymptotic pairwise independence of the pairs $(\lambda_i,\tilde{r}_{i,(i)})$ and $(\lambda_j,\tilde{r}_{j,(j)})$, since $\widehat{\beta}_{(ij)}$ is independent of λ_i and λ_j under Assumption **O6**, which guarantees independence of the λ_i's.

The lemma is shown. □

E-2 On the asymptotic behavior of c_τ

We are now in position to show that $c_\tau=\frac{1}{n}\mathrm{trace}\left((S+\tau\mathrm{Id}_p)^{-1}\right)$ is asymptotically deterministic. This result will require several steps.

Lemma E.2. *We work under Assumptions **O1-O7**, **P1-P4** and **F2-F4**.*

Consider the random function

$$g_n(x)=\frac{1}{n}\sum_{i=1}^n\frac{1}{1+x\lambda_i^2\psi_i'(prox(x\lambda_i^2\rho_i)(\tilde{r}_{i,(i)}))}\ ,\ \textit{defined for } x\geq 0.$$

Let $B>0$ be in $\mathbb{R}_+$. We have, for any $(x,y)\in\mathbb{R}_+^2$, and $x\leq B$, $y\leq B$

$$\sup_{(x,y):|x-y|\leq\eta,x\leq B,y\leq B}|g_n(x))-g_n(y)|\leq\eta\frac{1}{n}\sum_{i=1}^n\left(\lambda_i^2\|\psi_i'\|_\infty+B\,\mathsf{L}_i(n)\lambda_i^4\|\psi_i\|_\infty\right)\ .$$

*In particular, under **P2** and **F3-F4** we have, for C a constant independent of n and p,*

$$P^*\left(\sup_{(x,y):|x-y|\leq\eta,x\leq B,y\leq B}|g_n(x)-g_n(y)|>\delta\right)\leq\frac{\eta}{\delta}\mathsf{C}\,. \tag{E-42}$$

Hence, g_n is stochastically equicontinuous on $[0,B]$ for any $B>0$ given.

We used the notation P^* above to denote outer probability and avoid a discussion of potential measure theoretic issues associated with taking a supremum over a non-countable collection of random variables (see e.g van der Vaart (1998), Section 18.2). We refer the reader to e.g Pollard (1984) for more details on stochastic equicontinuity. (While we could probably avoid appealing to abstract concepts like outer measures here, we use this approach because it is a standard tool in empirical process theory and helps us avoid side measurability discussions that would distract us from the main focus of our proof.)

Proof. Let us consider the function, defined for $x\geq 0$,

$$h_u^{(i)}(x)=\frac{1}{1+x\lambda_i^2\psi_i'(\mathrm{prox}(x\lambda_i^2\rho_i)(u))}=\frac{\partial}{\partial u}\mathrm{prox}(x\lambda_i^2\rho)(u)\,.$$

The last equality comes from Lemma F.3.

We have, since ψ_i' is non-negative because ρ_i is convex,

$$\left|h_u^{(i)}(x) - h_u^{(i)}(y)\right| \leq |x\lambda_i^2\psi_i'(\mathrm{prox}(x\lambda_i^2\rho_i)(u)) - y\lambda_i^2\psi_i'(\mathrm{prox}(y\lambda_i^2\rho_i)(u))| \wedge 1 \; .$$

Therefore, since $x, y \geq 0$,

$$\left|h_u^{(i)}(x) - h_u^{(i)}(y)\right| \leq \lambda_i^2|x-y|\psi_i'(\mathrm{prox}(x\lambda_i^2\rho_i)(u)) + \lambda_i^2 y|\psi_i'(\mathrm{prox}(x\lambda_i^2\rho_i)(u)) - \psi_i'(\mathrm{prox}(y\lambda_i^2\rho_i)(u))| \; .$$

In particular, if $|x-y| \leq \eta$, and $x \vee y \leq B$

$$\begin{aligned}\sup_{y:|x-y|\leq\eta; x\vee y\leq B}\left|h_u^{(i)}(x) - h_u^{(i)}(y)\right| &\leq \lambda_i^2\eta\psi_i'(\mathrm{prox}(x\lambda_i^2\rho_i)(u)) \\ &\quad + B\lambda_i^2 \sup_{y:|x-y|\leq\eta, x\vee y\leq B}|\psi_i'(\mathrm{prox}(x\lambda_i^2\rho_i)(u)) - \psi_i'(\mathrm{prox}(y\lambda_i^2\rho_i)(u))| \; .\end{aligned}$$

Under assumption **O3**, ψ_i' is $\mathsf{L}_i(n)$-Lipschitz, with $\mathsf{L}_i(n) \leq Cn^\alpha$. Therefore, for $x_i, y_i \geq 0$,

$$|\psi_i'(\mathrm{prox}(x_i\rho_i)(u)) - \psi_i'(\mathrm{prox}(y_i\rho_i)(u))| \leq \mathsf{L}_i(n)|\mathrm{prox}(x_i\rho_i)(u) - \mathrm{prox}(y_i\rho_i)(u)| \; .$$

We recall that, according to Lemma F.2,

$$\frac{\partial}{\partial x}\mathrm{prox}(x\rho_i)(u) = -\frac{\psi_i(\mathrm{prox}(x\rho_i)(u))}{1 + x\psi_i'(\mathrm{prox}(x\rho_i)(u))} \; .$$

Hence,

$$\sup_x |\frac{\partial}{\partial x}\mathrm{prox}(x\rho_i)(u)| \leq \|\psi_i\|_\infty \; .$$

We finally conclude that

$$|\psi_i'(\mathrm{prox}(x_i\rho)(u)) - \psi_i'(\mathrm{prox}(y_i\rho)(u))| \leq [\mathsf{L}_i(n)\|\psi_i\|_\infty|x_i - y_i|] \wedge 2\|\psi_i'\|_\infty \; .$$

We therefore have, when $x \vee y \leq B$

$$\sup_{y:|x-y|\leq\eta}\left|h_u^{(i)}(x) - h_u^{(i)}(y)\right| \leq \lambda_i^2\eta\psi_i'(\mathrm{prox}(x\lambda_i^2\rho_i)(u)) + B\lambda_i^4\mathsf{L}_i(n)\|\psi_i\|_\infty\eta \; .$$

Therefore,

$$\sup_{(x,y):|x-y|\leq\eta, x\vee y\leq B}\left|h_u^{(i)}(x) - h_u^{(i)}(y)\right| \leq \lambda_i^2\eta\|\psi_i'\|_\infty + \eta B\mathsf{L}_i(n)\lambda_i^4\|\psi_i\|_\infty \; .$$

This analysis shows that for x given, if $|x-y| \leq \eta$ and $x \vee y \leq B$, we have

$$\sup_{(x,y):|x-y|\leq\eta, x\leq B, y\leq B}|g_n(x)) - g_n(y)| \leq \eta\frac{1}{n}\sum_{i=1}^n\left(\lambda_i^2\|\psi_i'\|_\infty + B\mathsf{L}_i(n)\lambda_i^4\|\psi_i\|_\infty\right) \; .$$

Under assumptions **P2** and **F3-F4**, we can now take expectations and get the result in L_1, since all the terms on the right hand side are bounded in L_1 under those assumptions.

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

Lemma E.3. *Let us call $G_n(x) = \mathbf{E}\left(g_n(x)\right)$. Let $B > 0$ be given. For any given $x_0 \leq B$,*

$$g_n(x_0) - G_n(x_0) = \mathrm{o}_{L_2}(1) \; .$$

*Under Assumptions **O1-O7**, **P1-P4** and **F1-F5**, we also have*

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

Proof. Under assumptions **F1** and **F5**, we can divide the index set $\{1,\dots,n\}$ into K subsets $A_1,\dots,A_K$, where K is finite (with n), in which $(X_i,\epsilon_i)_{i\in A_j}$ play a symmetric role. Hence, $\mathrm{var}\left(g_n(x_0)\right)$ can be expressed as a sum of variances and covariances of finitely many functions of finitely many random variables $(\lambda_i,\tilde{r}_{i,(i)})$: for those random variables, we just need to pick a representative in each subset $\{A_j\}_{j=1}^K$.

We note that since ψ_i' is Lipschitz and hence continuous, g_n is an average of bounded continuous functions of the random variables of interest to us.

Asymptotic pairwise independence of $(\lambda_i,\tilde{r}_{i,(i)})$'s, and the fact that ψ_i' can only be one of finitely many functions imply that

$$\mathrm{var}\left(g_n(x_0)\right)\to 0$$

and therefore gives the first result.

Let us now pick $\epsilon>0$. By the stochastic equicontinuity of g_n and our bound in Equation (E-42), we can find $x_1,\dots,x_K$, independent of n, such that for all $x\in[0,B]$, there exists l such that, when n is large enough,

$$\mathbf{E}\left(|g_n(x)-g_n(x_l)|\right)\le\epsilon\ .$$

Note that

$$|g_n(x)-G_n(x)|\le|g_n(x)-g_n(x_l)|+|g_n(x_l)-G_n(x_l)|+|G_n(x_l)-G_n(x)|\ .$$

We immediately get

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

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 $\sup_{1\le l\le K}|g_n(x_l)-G_n(x_l)|\to 0$ in L_2. In particular, if n is sufficiently large,

$$\mathbf{E}\left(\sup_{1\le l\le K}|g_n(x_l)-G_n(x_l)|\right)\le\epsilon\ .$$

The lemma is shown. □

Lemma E.4. *Assume **O1-O7**, **P1-P4** and **F1-F5**. Call $c_\tau=\frac{1}{n}\mathrm{trace}\left((S+\tau\mathrm{Id}_p)^{-1}\right)$. Call as before*

$$g_n(x)=\frac{1}{n}\sum_{i=1}^n\frac{1}{1+x\lambda_i^2\psi_i'(prox(x\lambda_i^2\rho_i)(\tilde{r}_{i,(i)}))}\ .$$

Then c_τ is a near solution of

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

i.e $\frac{p}{n}-\tau c_\tau-1+g_n(c_\tau)=\mathrm{o}_{L_k}(1)$, when $3\alpha-1/2<0$.

Asymptotically, near solutions of

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

are close to solutions of

$$\Delta_n(x)=\frac{p}{n}-\tau x-1+\mathbf{E}\left(g_n(x)\right)=0\ .$$

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

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

Hence c_τ is asymptotically deterministic.

Proof. Note that $g_n(x) \leq 1$.

Let δ_n be the function

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

and $\Delta_n(x) = \mathbf{E}\left(\delta_n(x)\right)$. Call x_n a solution $\delta_n(x_n) = 0$ and $x_{n,0}$ a 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_τ is such that $\delta_n(c_\tau) = \mathrm{o}(1)$**

An important remark is that c_τ is a near solution of $\delta_n(x) = 0$. This follows most clearly for arguments we have developed for $\mathsf{c}_{\tau,p}$ so we start by giving details through arguments for this random variable. Recall that in the notation of Lemma D.4, we had

$$\frac{p-1}{n} - \tau \mathsf{c}_{\tau,p} = \frac{1}{n}\mathrm{trace}\left(\mathrm{Id}_n - M\right).$$

Now, according to Equation (D-33),

$$\frac{1}{n}\mathrm{trace}\left(\mathrm{Id}_n - M\right) = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + \psi_i'(r_{i,[p]})\frac{1}{n}V_i'(\mathfrak{S}_p(i) + \tau \mathrm{Id})^{-1}V_i}.$$

According to Lemmas D.4, D.5 and D.6, we have

$$\sup_i \left| \frac{1}{n} V_i'(\mathfrak{S}_p(i) + \tau \mathrm{Id})^{-1} V_i - \lambda_i^2 \mathsf{c}_{\tau,p} \right| = \mathrm{O}_{L_k}\left(\frac{\mathrm{polyLog}(n)}{n^{1/2-2\alpha}} \right) \ .$$

Of course, when $x \geq 0$ and $y \geq 0$, $|1/(1+x) - 1/(1+y)| \leq |x-y| \wedge 1$. Hence, we see that

$$\left| \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + \psi_i'(r_{i,[p]})\frac{1}{n}V_i'(\mathfrak{S}_p(i) + \tau \mathrm{Id})^{-1}V_i} - \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + \psi_i'(r_{i,[p]})\lambda_i^2 \mathsf{c}_{\tau,p}} \right| \leq$$
$$\sup_{1 \leq i \leq n} \left| \frac{1}{n} V_i'(\mathfrak{S}_p(i) + \tau \mathrm{Id})^{-1} V_i - \lambda_i^2 \mathsf{c}_{\tau,p} \right| \frac{1}{n}\sum_{i=1}^{n} \|\psi_i'\|_\infty \ .$$

We conclude that

$$p/n - \tau \mathsf{c}_{\tau,p} - 1 + \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + \lambda_i^2 \mathsf{c}_{\tau,p}\psi_i'(r_{i,[p]})} = \mathrm{O}_{L_k}(n^{-1/2+2\alpha}\mathrm{polyLog}(n)) \ .$$

Exactly the same computations can be made with c_τ, so we have established that

$$\boxed{p/n - \tau c_\tau - 1 + \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + c_\tau \lambda_i^2 \psi_i'(R_i)} = \mathrm{O}_{L_k}(n^{-1/2+2\alpha}\mathrm{polyLog}(n)) \ .} \tag{E-43}$$

Now we have seen in Theorem 2.2 that

$$\sup_i |R_i - \mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})| = \mathrm{O}_{L_k}(n^{-1/2+\alpha}\mathrm{polyLog}(n)) \ .$$

Through our assumptions on ψ_i', this of course implies that

$$\sup_i |\psi_i'(R_i) - \psi_i'[\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})]| = \mathrm{O}_{L_k}(n^{-1/2+2\alpha}\mathrm{polyLog}(n))$$

We have furthermore noted that $\sup_i |c_i - \lambda_i^2 c_\tau| = \mathrm{O}_{L_k}(n^{-1/2+2\alpha}\mathrm{polyLog}(n))$ in Corollary D.7. Using Lemma F.2, this implies that

$$\left|\mathrm{prox}(c_i\rho_i)(\tilde{r}_{i,(i)}) - \mathrm{prox}(\lambda_i^2 c_\tau)(\tilde{r}_{i,(i)})\right| \leq \|\psi_i\|_\infty |c_i - \lambda_i^2 c_\tau|$$

and hence

$$\left|\psi_i'[\text{prox}(c_i\rho_i)(\tilde{r}_{i,(i)})] - \psi_i'[\text{prox}(\lambda_i^2 c_\tau \rho_i)(\tilde{r}_{i,(i)})]\right| = \text{O}_{L_k}\left(\|\psi_i\|_\infty n^{-1/2+3\alpha}\text{polyLog}(n)\right) .$$

Gathering everything together, we get

$$\left|\psi_i'(R_i) - \psi_i'(\text{prox}(\lambda_i^2 c_\tau \rho_i)(\tilde{r}_{i,(i)}))\right| = \text{O}_{L_k}([\|\psi_i\|_\infty + 1]n^{-1/2+3\alpha}\text{polyLog}(n))$$

So we have established that $\delta_n(c_\tau) = \text{O}_{L_k}(n^{-1/2+3\alpha}\text{polyLog}(n))$.

• **Final details**

Note that for any given x, $\delta_n(x) - \Delta_n(x) = \text{o}_P(1)$ by using Lemma E.3. In our case, with the notation of this lemma, $B = p/(n\tau) + \eta/\tau$, for $\eta > 0$ given.

This implies that, for any given $\epsilon > 0$

$$\sup_{x\in(0,p/(n\tau)+\eta/\tau]} |\delta_n(x) - \Delta_n(x)| < \epsilon ,$$

with high-probability when n is large. Therefore, for any $\epsilon > 0$, if x_n is a solution of $\delta_n(x_n) = 0$,

$$|\Delta_n(x_n)| \leq \epsilon \text{ with high-probability.}$$

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 $G_n = \mathbf{E}(g_n)$ is continuous.

If $T_{n,0}$ were reduced to a single point, we would have established the asymptotically deterministic character of c_τ.

Given our work concerning the limiting behavior of $\tilde{r}_{i,(i)}$ and our assumptions about ϵ_i's, we see that Lemma H.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_τ is asymptotically non-random. (Note that assumption **F1** is stated in terms of the properties of densities of random variables of the form $\epsilon_i + rZ_i$ where Z_i is $\mathcal{N}(0,1)$, independent of ϵ_i and r is arbitrary; Assumption **F1** also gives us guarantees for $\epsilon_i + r\lambda_i Z_i$ at λ_i given by a simple change of variable. The W_i's appearing in Lemma H.1 are of the form $\epsilon_i + |\lambda_i| rZ_i$, so assumption **F1** is all we need for Lemma H.1 to apply.) □

E-3 Proof of Theorem 2.1

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1

As we had noted in El Karoui et al. (2011),

$$\frac{\partial}{\partial t}\text{prox}(c\rho)(t) = \text{prox}(c\rho)'(t) = \frac{1}{1 + c\psi'(\text{prox}(c\rho)(t))} .$$

So Δ_n can be interpreted as

$$\Delta_n(x) = \frac{p}{n} - \tau x - 1 + \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\text{prox}(x\lambda_i^2\rho_i)'(\tilde{r}_{i,(i)})\right) .$$

The fact that c_τ is asymptotically arbitrarily close to the root of $\Delta_n(x) = 0$ gives us the first equation in the system appearing in Theorem 2.1. The second equation of the system comes from Equation (D-41). Theorem 2.1 is shown, with $c_\rho(\kappa)$ being the limit of c_τ.

E-4 About c_i's, ξ_n, N_p, and the limiting distribution of $\widehat{\beta}(p)$

Theorem 2.1 as well as many of our intermediate results have interesting consequences for various quantities we encountered. Let us now state them.

When we use the expression "under our assumptions", we mean assumptions **O1-O7**, **P1-P4** and **F1-F5**.

E-4.1 On c_i's

Recall that in Corollary D.7 we had shown that under our assumptions **O1-O7** and **P1-P4**

$$\sup_i |c_i - \lambda_i^2 c_\tau| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n) n^{2\alpha - 1/2}) \ .$$

Since we have now shown that c_τ has a deterministic limit $c_\rho(\kappa)$, we have the following lemma.

Lemma E.5. *We have under our assumptions **O1-O7** and **P1-P4***

$$\sup_i |R_i - \mathit{prox}(\lambda_i^2 c_\tau \rho_i)(\tilde{r}_{i,(i)})| = \mathrm{O}_{L_k}(n^{2\alpha - 1/2} \mathit{polyLog}(n)) \ .$$

Hence, under (all of) our assumptions we have asymptotically, for any given i

$$|R_i - \mathit{prox}(\lambda_i^2 c_\rho(\kappa) \rho_i)(\tilde{r}_{i,(i)})| = \mathrm{o}_{L_1}(1) \ .$$

If we furthermore assume that λ_i's are uniformly bounded, we have

$$\sup_{1 \le i \le n} |R_i - \mathit{prox}(\lambda_i^2 c_\rho(\kappa) \rho_i)(\tilde{r}_{i,(i)})| = \mathrm{o}_{L_k}(1)$$

Proof. Recall Lemma F.2 implies that

$$\sup_{x \in \mathbb{R}} |\mathrm{prox}(c_1 \rho)[x] - \mathrm{prox}(c_2 \rho)[x]| \le \|\psi\|_\infty |c_1 - c_2| \ .$$

Therefore, Corollary D.7 implies that

$$\sup_{1 \le i \le n} |\mathrm{prox}(\lambda_i^2 c_\tau \rho_i)(\tilde{r}_{i,(i)}) - \mathrm{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})| \le \sup_i \|\psi_i\|_\infty \sup_i |c_i - \lambda_i^2 c_\tau| = \mathrm{O}_{L_k}(\mathrm{polyLog}(n) n^{2\alpha - 1/2}) \ .$$

So in light of Theorem C.6 we conclude that

$$\sup_i |R_i - \mathrm{prox}(\lambda_i^2 c_\tau \rho_i)(\tilde{r}_{i,(i)})| = \mathrm{O}_{L_k}(n^{2\alpha - 1/2} \mathrm{polyLog}(n))$$

Since c_τ is bounded by $p/(n\tau)$ and therefore so is $c_\rho(\kappa)$, we see that converge in probability of c_τ to $c_\rho(\kappa)$ implies convergence in L_k for any k. Using Holder's inequality, we therefore see that

$$\mathbf{E}\left(\sup_{x \in \mathbb{R}} |\mathrm{prox}(\lambda_i^2 c_\tau \rho_i)[x] - \mathrm{prox}(\lambda_i^2 c_\rho(\kappa) \rho_i)[x]| \right) \le \sqrt{\mathbf{E}\left((c_\tau - c_\rho(\kappa))^2 \right) \mathbf{E}\left(\lambda_i^4 \right)} \ .$$

This gives us the second result of the lemma.

The last result is shown by simply remarking that

$$\sup_i \sup_{x \in \mathbb{R}} |\mathrm{prox}(\lambda_i^2 c_\tau \rho_i)[x] - \mathrm{prox}(\lambda_i^2 c_\rho(\kappa) \rho_i)[x]| \le (\sup_i \lambda_i^2) |c_\tau - c_\rho(\kappa)| \ .$$

□

E-4.2 On ξ_n

We have the following lemma.

Lemma E.6. *Under our assumptions, $\xi_n \to \xi$ in probability, where ξ is deterministic.*

Furthermore, ξ_n is bounded in L_1 and hence in probability.

We also have

$$\xi_n = \frac{p-1}{n \mathsf{c}_{\tau,p}} - \tau + \mathrm{o}_P(1) = \frac{p-1}{n c_\tau} - \tau + \mathrm{o}_P(1) \ ,$$

and $\mathsf{c}_{\tau,p}$ as well as c_τ are bounded away from 0 in probability.

We note that using the last result and arguments in the proof below, we see that, with the notations of Theorem 2.1

$$\xi = \frac{p-1}{nc_\rho(\kappa)} - \tau \ .$$

Proof. The proof follows easily from the result of Proposition D.8 which gives us that

$$\mathsf{c}_{\tau,p}(\xi_n + \tau) - \frac{p-1}{n} = \mathrm{O}_{L_k}(n^{-1/2+2\alpha}\mathrm{polyLog}(n)) = \mathrm{o}_P(1) \ .$$

Since we have shown that $\mathsf{c}_{\tau,p}$ convergences to a deterministic constant (recall that $c_\tau - \mathsf{c}_{\tau,p} \to 0$), we see that it is also the case for ξ_n. Note also that $\xi_n \leq \frac{1}{n}\sum_{i=1}^n X_i^2(p)\|\psi_i'\|_\infty$, so $\mathbf{E}\left(\xi_n\right)$ remains bounded under our assumptions.

To get the last result of the lemma, we just need to show that we can divide in the above display by $\mathsf{c}_{\tau,p}$ and still have something that converges to 0. We now show that $\mathsf{c}_{\tau,p}$ is bounded below. Note that

$$\mathsf{c}_{\tau,p} - \frac{p-1}{n(\xi_n+\tau)} = \mathrm{o}_P(1) \ .$$

Since ξ_n is bounded in probability, we see that $\frac{p-1}{n(\xi_n+\tau)}$ is bounded away from 0 in probability, which guarantees that $\mathsf{c}_{\tau,p}$ is bounded away from 0 in probability.

The results involving c_τ immediately follow by appealing to Proposition D.11.

□

E-4.3 On N_p

Recall that by definition, we had

$$N_p = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i(p)\psi_i(r_{i,[p]}) \ .$$

We have the following result.

Lemma E.7. *Under our assumptions, N_p is asymptotically $\mathcal{N}(0, v_n^2)$, with*

$$v_n^2 = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\lambda_i^2\psi_i^2(prox(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)}))\right) \ .$$

Furthermore, there exists v such that $v_n^2 \to v^2$, so that

$$N_p \Longrightarrow \mathcal{N}(0, v^2) \ .$$

As the proof makes clear, we can replace in the asymptotic statements above v_n^2 by

$$\hat{v}_n^2 = \frac{1}{n}\sum_{i=1}^n \lambda_i^2\psi_i^2(\mathrm{prox}(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})) \ .$$

Proof. Note that we can write

$$N_p = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathcal{X}_i(p)\lambda_i\psi_i(r_{i,[p]}) \ .$$

Under our assumptions $\mathcal{X}_i(p)$'s are independent and independent of $\{\lambda_i\psi_i(r_{i,[p]})\}_{i=1}^n$. The mild generalization of the Lindeberg-Feller argument given in the proof of Lemma E.1 now applies, using in the notation of that lemma $a_{n,p}(k) = n^{-1/2}\lambda_i\psi_i(r_{i,[p]})$ and recalling that $|\psi_i(r_{i,[p]})| \leq \|\psi_i\|_\infty$. Since λ_i's have 4 uniformly bounded moments under our assumptions, the fact that

$$N_p \text{ behaves like } A_n\mathcal{N}(0,1)$$

follows immediately, where $A_n^2 = \sum_{k=1}^n a_{n,p}^2(k) = \frac{1}{n}\sum_{i=1}^n \lambda_i^2\psi_i^2(r_{i,[p]})$. We note that $\mathbf{E}\left(A_n^2\right) \leq \frac{1}{n}\sum_{i=1}^n \|\psi_i\|_\infty^2 = \mathrm{O}(1)$ under our assumptions.

Work similar to the one done in the proof of Proposition D.12 shows that under our assumptions

$$A_n^2 - \frac{1}{n}\sum_{i=1}^n \lambda_i^2\psi_i^2[\mathrm{prox}(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})] = \mathrm{o}_{L_2}(1) \ .$$

Asymptotic pairwise independence of $(\lambda_i, \tilde{r}_{i,(i)})$ and $(\lambda_j, \tilde{r}_{j,(j)})$ - see Lemma E.1 - in connection with Assumption **F5** guarantees that

$$\mathrm{var}\left(\frac{1}{n}\sum_{i=1}^n \lambda_i^2\psi_i^2[\mathrm{prox}(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})]\right) \to 0 \ .$$

We conclude that A_n^2 is asymptotically deterministic and so is A_n. By Slutsky's lemma we have

$$N_p \text{ behaves like } \mathcal{N}(0, v_n^2)$$

where

$$v_n^2 = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\lambda_i^2\psi_i^2[\mathrm{prox}(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})]\right) \ ,$$

since $\mathbf{E}\left(A_n^2\right) - v_n^2 \to 0$.

We note that under our assumptions c_τ has limit $c_\rho(\kappa)$ and ψ_i is a bounded continuous function (one of only finitely many possible functions). Also, λ_i's have 4 moments. Therefore, since $\tilde{r}_{i,(i)}$ behaves asymptotically like $\epsilon_i + \lambda_i r_\rho(\kappa) Z_i$, where $Z_i \sim \mathcal{N}(0,1)$ independent of ϵ_i and $r_\rho(\kappa)$ is deterministic, we see that v_n^2 has a limit v^2. Of course,

$$v^2 = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\lambda_i^2\psi_i^2[\mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})]\right) \ .$$

In the notation of Theorem 2.1, we can rewrite v^2 as

$$v^2 = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left(\lambda_i^2\psi_i^2[\mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)(W_i)]\right)$$

• **Minor technical point**: it is true that $\lambda_i^2\psi_i^2[\mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})])$ is not a bounded continuous function of $(\lambda_i, \tilde{r}_{i,(i)})$. However, $[\lambda_i^2 \wedge M]\psi_i^2[\mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})])$ is, for any M. Since λ_i has 4 moments, it is easy to see that

$$\mathbf{E}\left(\left|[\lambda_i^2 \wedge M]\psi_i^2[\mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})]) - \lambda_i^2\psi_i^2[\mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})])\right|\right) \leq \frac{\mathbf{E}\left(\lambda_i^4\right)}{M^2}\|\psi_i\|_\infty^2 \ .$$

This standard approximation/uniform integrability argument shows that

$$\mathbf{E}\left(\lambda_i^2\psi_i^2[\mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})]\right) - \mathbf{E}\left(\lambda_i^2\psi_i^2[\mathrm{prox}(c_\rho(\kappa)\lambda_i^2\rho_i)(W_i)]\right) \to 0,$$

since M can be chosen arbitrarily large. □

E-4.4 Asymptotic normality of $\widehat{\beta}_p$

One of the aims of the previous results was to lead to a fluctuation result for $\widehat{\beta}_p$.

Proposition E.8. *We have, with the notation of the previous lemmas,*

$$\sqrt{n}[(\tau + \xi_n)\widehat{\beta}_p - \beta_0(p)\xi_n] \Longrightarrow \mathcal{N}(0, v^2) \ .$$

Furthermore, provided $\beta_0(p) = \mathrm{O}(n^{-1/2})$, we also have

$$\sqrt{n}[(\tau + \xi)\widehat{\beta}_p - \beta_0(p)\xi] \Longrightarrow \mathcal{N}(0, v^2) \ .$$

The proof of the proposition we give below shows that ξ in the previous display can be replaced by any quantity ω_n such that $\xi_n - \omega_n = \mathrm{o}_P(1)$. This in particular the case if we choose $\omega_n = p/(nc_\tau) - \tau$, according to Lemma E.6.

The main advantage of this ω_n is that it is computable from the data. And we can therefore test the null hypothesis that $\beta_0(p) = 0$, since we can approximate v^2 by $\frac{1}{n}\sum_{i=1}^n \lambda_i^2\psi_i^2[\mathrm{prox}(c_\tau\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})]$ according to the proof of Lemma E.7.

Proof. Recall that we have shown in Theorem D.10 that

$$\sqrt{n}(\widehat{\beta}_p - \mathfrak{b}_p) = \mathrm{o}_P(1) \; .$$

Recall that we showed that $\xi_n = \mathrm{O}_{L_k}(1)$ under our assumptions. It is easy to verify that the same is true for ξ, its limit. We also see that

$$\sqrt{n}(\tau + \xi_n)(\widehat{\beta}_p - \mathfrak{b}_p) = \mathrm{o}_P(1) \; .$$

Recall that by definition,

$$\sqrt{n}[(\tau + \xi_n)\mathfrak{b}_p - \xi_n\beta_0(p)] = N_p \; .$$

So we conclude, using Slutsky's lemma that

$$\sqrt{n}[(\tau + \xi_n)\widehat{\beta}_p - \xi_n\beta_0(p)] \Longrightarrow \mathcal{N}(0, v^2) \; .$$

When $\beta_0(p) = \mathrm{O}(n^{-1/2})$, we see that

$$\sqrt{n}(\xi - \xi_n)(\beta_0(p)) = \mathrm{o}_P(1) \; .$$

Furthermore, in this setting $\sqrt{n}\mathfrak{b}_p = \mathrm{O}_P(1)$ and hence $\sqrt{n}\widehat{\beta}_p = \mathrm{O}_P(1)$. We conclude that then

$$\sqrt{n}\widehat{\beta}_p(\xi_n - \xi) = \mathrm{o}_P(1) \; .$$

Therefore,

$$\sqrt{n}[(\tau + \xi_n)\widehat{\beta}_p - \xi_n\beta_0(p)] = \sqrt{n}[(\tau + \xi)\widehat{\beta}_p - \xi\beta_0(p)] + \mathrm{o}_P(1)$$

and we get the second result of the proposition through Slutsky's lemma. □

F Notes on the proximal mapping

In this section of the Appendix we remind the reader of elementary properties of the proximal mapping. The proofs, when needed, can be found in e.g El Karoui (2013).

Lemma F.1. *Almost by definition, we have*

$$prox(c\rho)(x) + c\psi(prox(c\rho)(x)) = x \; .$$

Let ρ be differentiable and such that ψ changes sign at 0, i.e $sign(\psi(x)) = sign(x)$ for $x \neq 0$. Then,

$$prox(c\rho)(0) = 0 \; .$$

Furthermore,

$$|\psi(prox(c\rho)(x))| \leq |\psi(x)| \; .$$

Also,

$$|\psi(prox(c\rho)(x))| \leq |x|/c \; .$$

We will also need the following simple result.

Lemma F.2. *Suppose x is a real and ρ is twice differentiable and convex. Then, for $c > 0$, we have*

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

and

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

In particular, at x given $c \to \rho(prox(c\rho)(x))$ is decreasing in c.

We also make the following observation, which is useful to obtain a compact representation for the system of equations (4).

Lemma F.3. *We have*

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

Moreover, at c fixed, when ψ' is continuous, $x \to \frac{1}{1+c\psi'(prox(c\rho)(x))}$ is a bounded, continuous function of x.

A proof of the first fact follows immediately from the well-known representation (see Moreau (1965))

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

The second result is also immediate, since $\psi' \geq 0$.

We finally make notice of the following simple fact.

Lemma F.4. *The function $c \to [c\psi(prox(c\rho)(x))]^2$ (defined on $\mathbb{R}_+$) is increasing, for any x.*

Examples : for the sake of concreteness, we now give a couple examples of proximal mappings.

1. if $\rho(x) = x^2/2$, $\text{prox}(c\rho)[x] = \frac{x}{1+c}$.
2. if $\rho(x) = |x|$, $\text{prox}(c\rho)[x] = sgn(x)(|x| - c)_+$, i.e the "soft-thresholding" function.

G On convex Lipschitz functions of random variables

In this section, we provide a brief reminder concerning convex Lipschitz functions of random variables. The proofs can be found in El Karoui (2013)

Lemma G.1. *Suppose that $\{X_i\}_{i=1}^n \in \mathbb{R}^p$ satisfy the following concentration property: $\exists C_n, \mathsf{c}_n$ such that for any G_i, a convex, 1-Lipschitz (with respect to Euclidean norm) function of X_i,*

$$P(|G_i(X_i) - m_i| \geq t) \leq C_n \exp(-\mathsf{c}_n t^2) ,$$

where m_i is deterministic.

Let us now fix $\{F_i\}_{i=1}^n$, n functions which are convex and 1-Lipschitz in X_i. Then if $\mathcal{F}_n = \sup_i |F_i(X_i) - m_i|$, we have, even when the X_i's are dependent:

1. *if $u_n = \sqrt{\log(n)/\mathsf{c}_n}$, $\mathbf{E}(\mathcal{F}_n) \leq u_n + C_n/(2\sqrt{\mathsf{c}_n}\sqrt{\log n}) = \frac{\sqrt{\log n}}{\sqrt{\mathsf{c}_n}}(1 + C_n/(2\log n))$. 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 $\mathcal{F}_n/u_n \leq K$ with overwhelming probability, i.e probability asymptotically smaller than any power of $1/n$.*
3. *m_i can be chosen to be the mean or the median of $F_i(X_i)$.*

In particular,

$$\mathcal{F}_n = \text{O}(polyLog(n)/\sqrt{\mathsf{c}_n})$$

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 $P(|G_i(X_i) - m_i| \geq t) \leq C_n \exp(-\mathsf{c}_n t^\beta)$, with $\beta \neq 2$. Of course, the order of magnitudes of the bounds then change: in particular, wherever $\sqrt{\mathsf{c}_n}$ appears, it would have to be replaced by $\mathsf{c}_n^{1/\beta}$. But since under our assumptions $1/\mathsf{c}$ is at most polyLog(n), this would effectively have no impact on our results.

We now turn our attention to a slightly more complicated setting.

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_{I^c} the remaining random variable.

Lemma G.2. *Suppose X_i's are independent and satisfy the concentration inequalities as above. Consider the situation where $F_{I_k}(\cdot)$ is a convex Lipschitz function of 1 variable; $F_{I_k}(\xi)$ depends on X through X_{I_k} only and we call $\mathcal{L}_{I_k}$ the Lipschitz constant of $F_{I_k}(\cdot)$ (at X_{I_k} given). $\mathcal{L}_{I_k}$ is assumed to be random, since X_{I_k} is. Call $m_{F_{I_k}} = m_{F_i(X_{I_k^c})}|X_{I_k}$, m being the mean or the median. As before, call $\mathcal{F}_n = \sup_{j=1,\ldots,n} |F_{I_j}(X_{I_j^c}) - m_{F_{I_j}}|$ Then $\mathcal{F}_n = \mathrm{O}(\sqrt{\log n/\mathsf{c}_n} \sup_{1\leq j\leq n} \mathcal{L}_{I_j})$ in probability and in $\sqrt{L_{2k}}$, i.e there exists $K > 0$, independent of n, such that*

$$\mathbf{E}\left(\mathcal{F}_n{}^k\right) \leq K(\sqrt{\log n/\mathsf{c}_n})^k \sqrt{\mathbf{E}\left(\sup_{1\leq j\leq n} \mathcal{L}_{I_j}^{2k}\right)} .$$

Hence, $\mathcal{F}_n$ is polyLog$(n)/\mathsf{c}_n^{1/2} \sup_{1\leq j\leq n} \mathcal{L}_{I_j}$ in $\sqrt{L_{2k}}$.

We repeatedly use the following lemma in the proof.

Lemma G.3. *Suppose the assumptions of the previous Lemma are satisfied. Consider $Q_{I_j} = \frac{1}{n} X'_{I_j^c} M_{I_j} X_{I_j^c}$, where M_{I_j} is a random positive-semidefinite matrix depending only on X_{I_j} whose largest eigenvalue is λ_{max,I_j}. Assume that $\mathbf{E}(X_i) = 0$, $\mathrm{cov}(X_i) = \mathrm{Id}_p$ and $n\mathsf{c}_n \to \infty$. Then, we have in L_k,*

$$\sup_{1\leq j\leq n} \left| Q_{I_j} - \frac{1}{n} \mathit{trace}\left(M_{I_j}\right) \right| = \mathrm{O}_{L_k}\left(\frac{\mathit{polyLog}(n)}{\sqrt{n\mathsf{c}_n}} \sup_{1\leq j\leq n} \lambda_{max,I_j}\right) .$$

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

On the spectral norm of covariance matrices

Lemma G.4. *Suppose X_i's are independent random vectors in $\mathbb{R}^p$, satisfying **O4**, and having mean 0 and covariance Id_p. Suppose that λ_i's satisfy **O6**. Let $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$. Then,*

$$|||\widehat{\Sigma}|||_2 = \mathrm{O}_P(\mathit{polyLog}(n)\mathsf{c}_n^{-1}) .$$

The results hold also in L_k.

Proof. The proof is exactly similar to that given in El Karoui (2013), which gives, following a simple adaption of the well-known ϵ-net argument explained e.g in Talagrand (2003), Appendix A.4 that

$$|||\frac{1}{n}\sum_{i=1}^n \mathcal{X}_i \mathcal{X}_i'|||_2 = \mathrm{O}_{L_k}(\mathsf{c}_n^{-1}) .$$

It is clear that

$$|||\widehat{\Sigma}|||_2 \leq (\sup_{1\leq i\leq n} \lambda_i^2)|||\frac{1}{n}\sum_{i=1}^n \mathcal{X}_i \mathcal{X}_i'|||_2 ,$$

and the result follows immediately. □

H Miscellaneous results

H-1 An analytic result

We now study the roots of $F(x) = 0$, where

$$F(x) = \frac{p}{n} - \tau x - 1 + \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left((\mathrm{prox}(x\lambda_i^2\rho_i))'(W_i)\right)$$

where W_i's are random variables and $(\mathrm{prox}(x\rho))'(t) = \frac{\partial}{\partial t}\mathrm{prox}(x\rho)(t) = \frac{1}{1+x\psi'(\mathrm{prox}(x\rho)(t))}$.

We now show that under mild conditions on W_i's this equation has a unique solution. We allow W_i to depend on the random variables λ_i's.

Lemma H.1. *Suppose that W_i's have smooth densities $f_i(t, \lambda_i)$ with $sign(f_i'(x, \lambda_i)) = -sign(x)$. Suppose further that $\lim_{|t|\to\infty} |t| f_i(t, \lambda_i) = 0$ and that $sign(\psi_i(x)) = sign(x)$. Then, if*

$$F_i(x) = \frac{p}{n} - \tau x - 1 + \mathbf{E}\left((prox(x\lambda_i^2\rho))'(W_i)\right) \ ,$$

the function F_i is decreasing, with $F_i'(x) \le -\tau$. Hence, the same applies to F.

In particular, the equation $F(x) = 0$ has a unique solution.

Proof. We call

$$G_i(x) \triangleq \mathbf{E}\left((\mathrm{prox}(x\lambda_i^2\rho_i))'(W_i)\right) \ ,$$

and

$$G_i(x, \lambda_i) \triangleq \mathbf{E}\left((\mathrm{prox}(x\lambda_i^2\rho_i))'(W_i)|\lambda_i\right)$$

Of course,

$$\mathbf{E}\left((\mathrm{prox}(x\lambda_i^2\rho))'(W_i)|\lambda_i\right) = \int (\mathrm{prox}(x\lambda_i^2\rho))'(t) f_i(t, \lambda_i) dt.$$

Using contractivity of the proximal mapping (see Moreau (1965)) we see that $\lim_{|t|\to\infty} \mathrm{prox}(x\lambda_i^2\rho_i)(t) f_i(t, \lambda_i) = 0$ under our assumptions.

Integrating the previous equation by parts, we see that

$$G_i(x, \lambda_i) = -\int (\mathrm{prox}(x\lambda_i^2\rho_i))(t) f_i'(t, \lambda_i) dt \ .$$

To compute $G_i'(x, \lambda_i)$, we differentiate under the integral sign (under our assumptions the conditions of Theorem 9.1 in Durrett (1996) are satisfied) to get

$$G_i'(x, \lambda_i) = \int \frac{\psi_i(\mathrm{prox}(x\lambda_i^2\rho_i)(t)) f_i'(t, \lambda_i)}{1 + x\lambda_i^2\psi_i'(\mathrm{prox}(x\lambda_i^2\rho_i)(t))} dt \ .$$

Under our assumptions, $\mathrm{sign}(\psi_i(\mathrm{prox}(x\lambda_i^2\rho_i)(t))) = \mathrm{sign}(t)$ and $\mathrm{sign}(f_i'(t, \lambda_i)) = -\mathrm{sign}(t)$, so that

$$\forall t \neq 0, \mathrm{sign}(\psi_i(\mathrm{prox}(x\lambda_i^2\rho_i)(t)) f_i'(t, \lambda_i)) = -1 \ .$$

Since the denominator of the function we integrate is positive, we conclude that

$$G_i'(x, \lambda_i) \le 0 \ \text{ and } G_i'(x) \le 0 \ .$$

Since $F_i'(x) = -\tau + G_i'(x)$, we see that $F_i'(x) \le -\tau < 0$. Therefore F_i is a decreasing function on $\mathbb{R}_+$. Of course, $\mathrm{prox}(0\rho)(t) = t$, so that $F_i(0) = p/n$ and $\lim_{x\to\infty} F_i(x) = -\infty$, since, for instance,

$$0 \le \mathrm{prox}(x\rho)'(t) = \frac{1}{1 + x\psi'[\mathrm{prox}(x\rho)(t)]} \le 1 \ .$$

So we conclude that the equation $F_i(x) = 0$ has a unique root. (Since F_i is differentiable, F_i is of course continuous.)

We note that

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x) \ .$$

Therefore, $F(0) = p/n$ and $F'(x) \le -\tau$. So F is decreasing, differentiable and hence has a unique root. □

Remark: the conditions on the density of W are satisfied in many situations. For instance if $W_i = \epsilon + r\lambda_i Z$, where ϵ is symmetric about 0 and log-concave, Z is $\mathcal{N}(0,1)$, independent of λ_i and ϵ, and $r > 0$, it is clear that the density of W satisfies the conditions of our lemma. Similar results hold under weaker assumptions on ϵ of course. For more details, we refer the reader to e.g Karlin (1968), Prékopa (1973), Ibragimov (1956), and Dharmadhikari and Joag-Dev (1988).

In particular, we recall Theorem 1.6 in Dharmadhikari and Joag-Dev (1988) which says that the convolution of two symmetric unimodal distributions on $\mathbb{R}$ is unimodal. Hence, when ϵ has a symmetric and unimodal distribution, so does $W_i = \epsilon + \lambda_i r Z$, for any r. This is for instance the case when ϵ has a Cauchy distribution.

H-2 A linear algebraic remark

We need the following lemma at some point in the proof.

Lemma H.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 τ be a strictly positive real. Call $\Gamma_\tau = \Gamma + \tau \mathrm{Id}_{p-1}$. Then we have

$$trace\left((A + \tau \mathrm{Id}_p)^{-1}\right) = trace\left(\Gamma_\tau^{-1}\right) + \frac{1 + v'\Gamma_\tau^{-2} v}{a + \tau - v'\Gamma_\tau^{-1} v} .$$

In particular,

$$\left| trace\left((A + \tau)^{-1}\right) - trace\left(\Gamma_\tau^{-1}\right) \right| \leq \frac{1 + a/\tau}{\tau} .$$

The proof is simple and we refer the reader to El Karoui (2013) for details if needed.

References

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis.* Wiley.

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., El Karoui, N., Lim, C., and Yu, B. (2012). Penalized robust regression in high dimension. Technical report, Department of Statistics, UC Berkeley. Submitted to AISTATs in October 2011. Not under consideration anymore.

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 Teboulle, 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.

Bhatia, R. (1997). *Matrix analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.

Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, convexity, and applications.* Probability and Mathematical Statistics. Academic Press, Inc., Boston, MA.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815. URL `http://dx.doi.org/10.1214/aos/1176346703`.

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

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. (2010). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation. *Ann. Statist.* **38**, 3487–3566. URL `http://dx.doi.org/10.1214/10-AOS795`.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv:1311.2445* ArXiv:1311.2445.

El Karoui, N., Bean, D., Bickel, P., Lim, C., and Yu, B. (2011). On robust regression with high-dimensional predictors. Technical Report 811, UC, Berkeley, Department of Statistics. Originally submitted as manuscript AoS1111-009. Not under consideration anymore.

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`.

El Karoui, N. and Koesters, H. (2011). Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *Submitted to Bernoulli* Available at arXiv:1105.1404 (68 pages).

El Karoui, N. and Purdom, E. (2015). Can we trust the bootstrap in high-dimension? Technical Report 824, UC Berkeley, Department of Statistics. Submitted to JASA.

Hall, P., Marron, J. S., and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 427–444.

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

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`.

Michael B. Eisen, P. O. B., Paul T. Spellman and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS* **95**, 14863–14868.

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. (1986). Asymptotic behavior of the empiric distribution of M-estimated residuals from a regression model with many parameters. *Ann. Statist.* **14**, 1152–1170. URL http://dx.doi.org/10.1214/aos/1176350056.

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.

Relles, D. (1968). *Robust Regression by Modified Least Squares*. Ph.D. thesis, Yale University.

Ruszczyński, A. (2006). *Nonlinear optimization*. Princeton University Press, Princeton, NJ.

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.

Stroock, D. W. (1993). *Probability theory, an analytic view*. Cambridge University Press, Cambridge.

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.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Yin, Y. Q., Bai, Z. D., and Krishnaiah, P. R. (1988). On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. *Probab. Theory Related Fields* **78**, 509–521.