Electronic Journal of Statistics
Vol. 11 (2017) 2832–2875
ISSN: 1935-7524
DOI: 10.1214/17-EJS1304

Error bounds for the convex loss Lasso in linear models

Mark Hannay

Division of Mathematical Sciences, School of Physical and Mathematical Sciences
Nanyang Technological University, 637371, Singapore
e-mail: smark@ntu.edu.sg

and

Pierre-Yves Deléamont

Research Center for Statistics and Geneva School of Economics and Management
University of Geneva, Blv. du Pont-d'Arve 40, CH-1211 Geneva, Switzerland
e-mail: pierre-yves.deleamont@unige.ch

Abstract: In this paper we investigate error bounds for convex loss functions for the Lasso in linear models, by first establishing a gap in the theory with respect to the existing error bounds. Then, under the compatibility condition, we recover bounds for the absolute value estimation error and the squared prediction error under mild conditions, which appear to be far more appropriate than the existing bounds for the convex loss Lasso. Interestingly, asymptotically the only difference between the new bounds of the convex loss Lasso and the classical Lasso is a term solely depending on a well-known expression in the robust statistics literature appearing multiplicatively in the bounds. We show that this result holds whether or not the scale parameter needs to be estimated jointly with the regression coefficients. Finally, we use the ratio to optimize our bounds in terms of minimaxity.

MSC 2010 subject classifications: Primary 62F35; secondary 62J07.
Keywords and phrases: Robust Lasso, high dimensions, error bounds, joint scale and location estimation.

Received January 2017.

Contents

1 Introduction 2833
2 Literature review and motivation 2834
2.1 Classical Lasso 2835
2.2 Convex loss Lasso 2835
2.3 Robust Lasso 2837
2.4 Problems with existing bounds 2837
3 Error bounds with known scale 2838
3.1 Set-up 2838
3.2 Basic bounds 2839
3.3 Controlling the empirical process 2840

3.4 New error bounds for the convex loss Lasso 2843
3.5 Asymptotic results 2844
4 Error bounds with estimated scale 2845
4.1 Scaled convex loss Lasso estimator 2845
4.2 Basic bounds 2845
4.3 New error bounds for the scaled convex loss Lasso 2846
4.4 Asymptotic results 2847
5 Minimaxity of bounds and relevance of the ratios 2848
5.1 Minimaxity of error bounds for known scale 2848
5.2 Minimaxity of error bounds for estimated scale 2849
5.3 Relevance of the ratios 2850
5.3.1 Estimating equations 2850
5.3.2 Ratio in distribution 2851
5.3.3 Ratio for projections onto true span 2851
6 Discussion 2852
A Proofs and technical arguments: Section 3 2853
A.1 Proof of Lemma 3.1 2853
A.2 Proof of Lemma 3.2 2853
A.3 Proof of Lemma 3.3 2853
A.4 Proof of Theorem 3.1 2854
A.5 Proof of Lemma 3.4 2855
A.6 Proof of Lemma 3.5 2856
A.7 Probability bounds 2858
A.8 Proof of Theorem 3.2 2860
B Proofs and technical arguments: Section 4 2864
B.1 Proof of Lemma 4.1 2864
B.2 Technical arguments on scale error bound 2864
B.3 Technical arguments on bounds between norms 2868
B.4 Proof of Theorem 4.1 2870
Acknowledgements 2874
References 2874

1. Introduction

Among the many techniques that have been proposed to address the estimation of a location parameter in the high-dimensional linear model, the Lasso [12] remains one of the most widely studied. Arguably, this method, which consists of penalizing the sum of squared residuals with the l_1 norm of the vector of coefficients, has many advantages. It leads to accurate predictions while setting some coefficients exactly to zero, thus achieving model selection simultaneously. Additionally, the estimates can be computed in a highly efficient manner. Since the seminal work of [12], the classical Lasso[1] has been generalized in various ways, in terms of the loss and penalty functions under consideration.

[1]We will refer to the Lasso given in [12] as the classical Lasso, as opposed to the convex loss Lasso or the robust Lasso that we study in this paper.

One of the main reasons for considering alternative loss functions is the issue of robustness. Indeed, it is well known that the classical Lasso can be largely affected by contamination of the error distribution.

In this paper, we investigate estimation and prediction error bounds for the Lasso with a general convex loss function. Our motivation comes from the observation that there is a kind of theoretical gap in the literature, in the sense that the bounds developed for the convex loss Lasso are not related in a natural way to the ones given for the classical Lasso. Our main contribution is to show explicitly the presence in our bounds of an additional term compared to the classical case. We demonstrate that this same term appears in the bounds whether or not the scale parameter needs to be estimated. Interestingly, this extra term corresponds to the ratio found by [7] in his minimax problem, which serves as a justification for the use of the famous Huber loss function in the low-dimensional setting. We provide theoretical arguments for the relevance of the ratio in terms of optimality of the bounds. To the best of our knowledge, these findings have not appeared previously in the literature.

The outline of the paper is as follows. In Section 2, we discuss some key results related to the classical Lasso and to a more general convex loss Lasso. We also provide an overview of the literature, focusing on robust versions of the Lasso, and finally we provide an account of what we believe to be problematic with the existing error bounds, thus motivating the present paper. In Section 3, we establish bounds for the estimation and prediction errors in the case of a known scale parameter. In Section 4, we relax that assumption and consider joint estimation of scale and regression parameters, inspired by Huber's Proposal 2 [7]. The main result of the analyses carried out in Sections 3 and 4 is that the bounds on the prediction and estimation errors contain an extra term, in the form of a ratio, relative to the classical case. In Section 5, we give a rationale for the importance of this ratio from a theoretical point of view. In Section 6, we summarize our results and mention opportunities for future research.

2. Literature review and motivation

In order to motivate our paper, we provide a selective overview of the existing literature on the linear Lasso. This mainly encompasses the classical Lasso, the convex loss Lasso and the robust Lasso. We then show that the existing literature is unsatisfactory in some respects. One specific problem is the lack of a link between the choice of the loss function and the resulting prediction or estimation error depending on the error distribution. This is most notably a problem in justifying the choice of the loss function in robust statistics.

Throughout the paper, we consider the linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\epsilon},$$

where $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\beta}^0 \in \mathbb{R}^p$, $\boldsymbol{\epsilon} \in \mathbb{R}^n$, $\sigma > 0$ and where p is (potentially) larger than n.

2.1. Classical Lasso

We now summarize the properties of the classical Lasso derived, among others, by [1], focusing on the aspects which are most relevant for our purposes. In doing so, our aim is to provide a basis for comparison to the bounds that we will develop in this paper.

As briefly mentioned in the Introduction, the classical Lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{Lasso} = \hat{\boldsymbol{\beta}}_{Lasso}(\lambda) \in \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{\sigma^2} \frac{1}{n} \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right),$$

where $\lambda \geq 0$ is a penalty parameter and σ is the scale parameter.

We will consider the case where the design is fixed and the columns of $\boldsymbol{X}$ are normalized, so that $\frac{1}{n}\left(\boldsymbol{X}^T\boldsymbol{X}\right)_{jj} = 1$. Now if we pick $\lambda = \frac{4}{\sigma}\sqrt{\frac{t^2+2\log(p)}{n}}$ and we assume $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix, we have the following inequality with a probability bigger than $1 - 2\exp[-t^2/2]$:

$$\frac{1}{\sigma^2}\frac{1}{n}\left\| \boldsymbol{X}\left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_{Lasso}\right)\right\|_2^2 \leq \frac{3}{2}\lambda \left\|\boldsymbol{\beta}^0\right\|_1.$$

This shows that the Lasso is consistent in terms of prediction when we take a penalty parameter of order $\frac{1}{\sigma}\sqrt{\frac{\log(p)}{n}}$ and when $\left\|\boldsymbol{\beta}^0\right\|_1$ is small enough.

Assuming that the true parameter vector is sufficiently sparse, and imposing additional conditions on the design matrix ("compatibility conditions", see below), we can obtain a result that is sometimes referred to as an oracle inequality. It states that by selecting λ as above, we have the following inequality with probability bigger than $1 - 2\exp[-t^2/2]$:

$$\frac{1}{\sigma^2}\frac{1}{n}\left\| \boldsymbol{X}\left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_{Lasso}\right)\right\|_2^2 + \lambda \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_{Lasso}\right\|_1 \leq 4\lambda^2\sigma^2\frac{s_0}{\phi_0^2}, \tag{2.1}$$

where s_0 is the number of true non-zero coefficients and ϕ_0^2 is a compatibility constant. This inequality includes two interesting results. On the one hand, it gives us a bound for the l_1 estimation error:

$$\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_{Lasso}\right\|_1 \leq 4\lambda\sigma^2\frac{s_0}{\phi_0^2}.$$

On the other hand, we also get a bound for the prediction error:

$$\frac{1}{\sigma^2}\frac{1}{n}\left\| \boldsymbol{X}\left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_{Lasso}\right)\right\|_2^2 \leq 4\lambda^2\sigma^2\frac{s_0}{\phi_0^2}.$$

2.2. Convex loss Lasso

We now provide results in the more general case where we replace the squared loss function with a convex loss function. More specifically, in a linear model

and for a given convex loss function ρ, the corresponding convex loss Lasso is defined as

$$\hat{\boldsymbol{\beta}}_{CLasso} = \hat{\boldsymbol{\beta}}_{CLasso}(\lambda) \in \arg\min_{\boldsymbol{\beta}} \left(\frac{2}{n} \sum_{i=1}^{n} \rho \left(\frac{Y_i - (\boldsymbol{X}\boldsymbol{\beta})_i}{\sigma} \right) + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right), \quad (2.2)$$

where σ is the scale parameter. We note that the constant 2 is only included to ensure that the classical Lasso is recovered when $\rho(x) = \frac{1}{2}x^2$. This problem can be seen as a particular instance of the Lasso for general convex loss which has notably been studied by [15] and [1]. We briefly mention some relevant results derived by these authors. Define

$$\begin{aligned} \mathcal{E}(\boldsymbol{\beta}) &:= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[\rho \left(\frac{Y_i - (\boldsymbol{X}\boldsymbol{\beta})_i}{\sigma} \right) \right] - \mathbb{E} \left[\rho(\epsilon) \right], \\ \nu_n(\boldsymbol{\beta}) &:= \frac{1}{n} \sum_{i=1}^{n} \rho \left(\frac{Y_i - (\boldsymbol{X}\boldsymbol{\beta})_i}{\sigma} \right) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[\rho \left(\frac{Y_i - (\boldsymbol{X}\boldsymbol{\beta})_i}{\sigma} \right) \right], \\ Z_{M^*} &:= \sup_{\left\| \boldsymbol{\beta} - \boldsymbol{\beta}^* \right\|_1 \leq \sigma M^*} \left| \nu_n(\boldsymbol{\beta}) - \nu_n(\boldsymbol{\beta}^*) \right|, \end{aligned}$$

where the ϵ_i are independent and identically distributed replicas of ϵ, which is assumed to have a symmetric distribution.

Assuming that there exists a $K > 0$ such that $\max_{i,j} \left| \boldsymbol{X}_i^{(j)} \right| \leq K$, then for $\left\| \boldsymbol{\beta} - \boldsymbol{\beta}^0 \right\|_1 \leq \sigma M^*$ small enough depending on K, we have by a Taylor expansion the margin condition

$$\mathcal{E}(\boldsymbol{\beta}) \geq c \frac{1}{\sigma^2} \frac{1}{n} \left\| \boldsymbol{X} \left(\boldsymbol{\beta}^0 - \boldsymbol{\beta} \right) \right\|_2^2$$

where $c \approx \mathbb{E}\left[\psi'(\epsilon) \right]$ with $\psi(x) := \rho'(x)$. Under the additional assumptions that $\sup_x |\psi(x)| = L < +\infty$ and that $\frac{1}{n} \left(\boldsymbol{X}^T \boldsymbol{X} \right)_{jj} = 1$, for $\lambda_0 \asymp \frac{L}{\sigma} \sqrt{\frac{\log(p)}{n}}$ and $\mathcal{J} := \{ Z_{M^*} \leq \lambda_0 \sigma M^* \}$, we have that $\mathcal{J}$ holds with high probability for arbitrary symmetric error distributions if ψ is odd.

Regarding convergence, under the above conditions on $\mathcal{J}$, for $\lambda \geq 8\lambda_0$, we have

$$\mathcal{E}\left(\hat{\boldsymbol{\beta}}_{CLasso} \right) + \lambda \left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_{CLasso} \right\|_1 \leq \frac{16}{c} \lambda^2 \sigma^2 \frac{s_0}{\phi_0^2},$$

where s_0 is the number of true non-zero coefficients and ϕ_0^2 is a compatibility constant. This inequality includes two interesting results. On the one hand, it gives us a bound for the l_1 estimation error:

$$\left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_{CLasso} \right\|_1 \leq \frac{16}{c} \lambda \sigma^2 \frac{s_0}{\phi_0^2}.$$

On the other hand, we also get a bound for the prediction error:

$$\frac{1}{\sigma^2} \frac{1}{n} \left\| \boldsymbol{X} \left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}_{CLasso} \right) \right\|_2^2 \leq \frac{16}{c^2} \lambda^2 \sigma^2 \frac{s_0}{\phi_0^2}.$$

[9] has also studied this case and recovers similar results. Although he makes a stronger assumption on ψ by imposing $\inf_i \psi'(\left|\left(\boldsymbol{X}\boldsymbol{\beta}^0\right)_i\right|/\sigma) > 0$, this can be relaxed by using the same ideas as [1].

2.3. Robust Lasso

The robust Lasso, for convex loss functions, can be studied as a special case of the convex loss Lasso. In spite of this, it has also been studied independently. For instance, [17] introduced the LAD-Lasso:

$$\hat{\boldsymbol{\beta}}_{LAD-Lasso}(\lambda) \in \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{n} \left\| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \right\|_1 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right)$$

Using an adaptive version of the LAD-Lasso, they were able to show root-n consistency and asymptotic normality. Yet they did not investigate the growth of p with respect to n, and they assumed a positive definite covariance matrix.

More recently, [2] and [3] investigated the high-dimensional case for the LAD-Lasso. In both papers, under mild conditions, consistency in l_2 was shown. However, since the main goal of these papers was to investigate the properties of the adaptive LAD-Lasso (especially the model selection properties), they did not investigate bounds on l_1 estimation error or on squared prediction error. In addition, the constants appearing in the consistency theorems were in no way specified, contrary to the classical Lasso.

Another method to make the Lasso more robust was used by [10]. They focused on a Huberized adaptive Lasso, which can be defined as

$$\hat{\boldsymbol{\beta}}_{Hub-Lasso}(\lambda) \in \arg\min_{\alpha,\boldsymbol{\beta},\sigma} \left(\frac{2}{n} \sum_{i=1}^{n} \rho_{Hub,L} \left(\frac{Y_i - \alpha - \left(\boldsymbol{X}\boldsymbol{\beta}\right)_i}{\sigma} \right) \sigma + \lambda \sum_{j=1}^{p} w_j \left|\beta_j\right| \right)$$

when $\sigma > 0$, where $\rho_{Hub,L}(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le L \\ L|x| - \frac{1}{2}L^2 & \text{else} \end{cases}$ is the Huber ρ-function.

This method is computationally interesting since it follows from [11] that we only need to deal with partial linear solutions. In their paper, [10] developed the theory for model selection and asymptotic normality, where they estimated $\boldsymbol{\beta}$, α and σ jointly. However, they did not discuss the high-dimensional case (i.e., where $p > n$).

2.4. Problems with existing bounds

The bounds derived in both the general study of the convex loss Lasso and the robust Lasso for specific loss functions depend heavily on $L = \sup_x |\psi(x)| < \infty$. At first glance, this suits well to robust statistics, where one usually bounds the influence of single observations by bounding ψ, implying that any reasonable choice of ψ would automatically satisfy $\sup_x |\psi(x)| < \infty$.

However, the classical Lasso cannot be studied under such assumptions, since in such a case $\psi(x) = x$. This means that there is no unified framework within which both bounded and unbounded ψ-functions can be studied. Unfortunately, this lack of a unified framework can lead to unreasonable results for a fixed n. For instance, under the assumption that the errors are Gaussian, suppose that we want to approximate the classical Lasso by using the convex loss Lasso with the Huber loss function with a large tuning parameter L. For such an L, the corresponding convex loss Lasso is basically nearly always equal to the classical Lasso. Yet the bounds from the convex loss Lasso become useless, despite the fact that we know that the error of the convex loss Lasso is approximately equal to that of the classical Lasso.

An additional cause for concern is that the bounds from the convex loss Lasso only depend on the distribution of ϵ through $\mathbb{E}[\psi'(\epsilon)]$. Therefore, for a given distribution of ϵ, it is impossible, with the existing theory, to improve the error bounds by selecting an appropriate ρ function other than by minimizing $\frac{\sup_x |\psi(x)|}{\mathbb{E}[\psi'(\epsilon)]}$. This does not take into account the second moment and excludes the study of the classical Lasso as a special case of the convex loss Lasso.

Finally, it is imperative to jointly estimate the scale parameter σ with the location parameter $\boldsymbol{\beta}$. This is because, just as the choice of ρ can affect the location estimation, so does the value of the scale. Also, this joint estimation should be studied in the high-dimensional setting, with the same asymptotic assumptions as those for a known scale.

We believe that all these points lead to a gap in the theory which we address in this paper, by providing a unified framework to study error bounds for different types of loss functions satisfying a new moment condition. This framework includes many types of loss functions (most importantly, the classical Lasso and the Huberized Lasso) and leads to error bounds which smoothly depend on ψ through the term $\mathbb{E}\left[\psi(\epsilon)^2\right] / \mathbb{E}[\psi'(\epsilon)]^2$.

3. Error bounds with known scale

Throughout this section, we consider the scale parameter as known. While this assumption is obviously not realistic, it allows for easier derivations which may be more insightful. This assumption will be relaxed in Section 4.

3.1. Set-up

We consider the general convex loss Lasso as defined in Subsection 2.2, where for the remainder of the paper we work with a continuous, odd and monotone increasing ψ. For better readability, the assumptions on ψ' will slightly change throughout the relevant subsections, as we now explain.

The purpose of Subsection 3.2 is to describe the construction of basic bounds. For the sake of brevity, we impose ψ' to be well defined and continuous here. In Subsection 3.3, the study of the empirical process does not require the use of

ψ'. For the new error bounds that we develop in Subsection 3.4, we refine our assumptions on ψ' and require ψ' to be well defined and uniformly continuous but for a finite set not including 0, while bounded everywhere, and satisfying $\mathbb{E}\left[\psi'\left(\epsilon\right)\right] > 0$ (see Assumptions 3.1 and 3.2). No further restriction is needed for the asymptotic bounds in Subsection 3.5.

The conditions on ψ' for the error bounds imply that we require ρ to be locally strictly convex over some intervals, but not necessarily over the entire space. In fact, ρ is allowed to be partially affine outside of a compact set. We note that the assumption that $\psi'(0)$ is well defined excludes the LAD-Lasso but still includes many other relevant instances of the convex loss Lasso such as the Huberized Lasso.

Moreover, we consider a fixed design matrix $\boldsymbol{X}$, and we assume that the ϵ_i are i.i.d. replicas of ϵ, whose distribution is only required to be symmetric and continuous at the points of discontinuity of ψ'. The symmetry assumption is essential to robust statistics in order to avoid inevitable bias in estimation (see [8]), while the continuity assumption makes the points of discontinuity of ψ' asymptotically irrelevant.

3.2. Basic bounds

We start by providing an inequality which can be interpreted as a generalization of the basic inequality for the classical Lasso. For better readability, we denote the prediction error for observation i by

$$a_i := \left(\boldsymbol{X}\left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right)\right)_i.$$

Lemma 3.1. *There exists* $t^\lambda \in [0,1]$ *such that*

$$\frac{1}{n}\sum_{i=1}^{n}\psi'\left(\epsilon_i + t^\lambda \frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2} + \lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1 \leq -\frac{2}{n}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)^T\frac{\boldsymbol{X}}{\sigma}\left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right) + \lambda\left\|\boldsymbol{\beta}^0\right\|_1.$$

The generalized basic bound, just as the basic bound given in [1], contains an empirical process component, namely $\frac{2}{n}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)^T\frac{\boldsymbol{X}}{\sigma}\left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right)$. This term can easily be bounded as follows:

$$\left|\frac{2}{n}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)^T\frac{\boldsymbol{X}}{\sigma}\left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right)\right| \leq \max_{1\leq j\leq p}\left|\frac{2}{n}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)^T\frac{\boldsymbol{X}^{(j)}}{\sigma}\right|\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1.$$

This in turn motivates the following definition:

$$\mathcal{J}_0 := \left\{\max_{1\leq j\leq p}\left|\frac{2}{n}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)^T\frac{\boldsymbol{X}^{(j)}}{\sigma}\right| \leq \lambda_0\right\}. \tag{3.1}$$

For specific choices of λ_0 and λ, we can easily bound a type of prediction error on $\mathcal{J}_0$.

Lemma 3.2. *Let $2\lambda_0 \leq \lambda$. Then there exists $t^\lambda \in [0,1]$, such that on $\mathcal{J}_0$,*

$$\frac{1}{n}\sum_{i=1}^{n}\psi'\left(\epsilon_i + t^\lambda \frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2} \leq \frac{3}{2}\lambda\left\|\boldsymbol{\beta}^0\right\|_1.$$

If ρ is strictly convex, we have $\psi' > 0$ and so we recover a bound on the prediction error. This type of condition on ρ is not very useful in robust statistics though, since we want to work with a bounded ψ-function. We will need to rely on sparsity and on a compatibility condition to recover stronger error bounds, which also apply in the case where $\inf_x \psi'(x) = 0$.

3.3. Controlling the empirical process

Before we investigate how sparsity can be useful in deriving a stronger bound than the one given in Lemma 3.2, we elaborate on the conditions which ensure that the set $\mathcal{J}_0$ defined in Equation (3.1) has sufficiently large probability.

We start by giving an easily derived bound for $\mathbb{P}[\mathcal{J}_0]$ for a bounded ψ-function.

Lemma 3.3. *Let $\boldsymbol{X}$ have normalized columns, i.e. $\frac{1}{n}\left(\boldsymbol{X}^T\boldsymbol{X}\right)_{jj} = 1$, and assume that $\sup_x |\psi(x)| = L$. Then, for $\lambda_0 := 2\frac{L}{\sigma}\sqrt{\frac{t^2+2log(p)}{n}}$, we have $\mathbb{P}[\mathcal{J}_0] \geq 1 - 2\exp\left[-t^2/2\right]$.*

The previous lemma does not allow us to make a connection with the results of the classical Lasso, where ψ is the identity function and is thus obviously unbounded. Therefore, we provide the following theorem, which gives a bound based on $\mathbb{E}\left[\psi(\epsilon)^2\right]$.

Theorem 3.1. *Suppose $\frac{1}{n}\left(\boldsymbol{X}^T\boldsymbol{X}\right)_{jj} = 1$ and $\mathbb{E}\left[\psi(\epsilon)^{2k}\right] \leq \frac{(2k)!}{k!}2^{-k}\mathbb{E}\left[\psi(\epsilon)^2\right]^k$ for all $k \in \mathbb{N}$. Then, for $\lambda_0 := 2\frac{\sqrt{\mathbb{E}\left[\psi(\epsilon)^2\right]}}{\sigma}\sqrt{\frac{t^2+2log(p)}{n}}$, we have $\mathbb{P}[\mathcal{J}_0] \geq 1 - 2\exp\left[-t^2/2\right]$.*

If $\psi(\epsilon)/L$ follows a Rademacher distribution, i.e. if $\mathbb{P}[\psi(\epsilon)/L = -1] = 1/2 = \mathbb{P}[\psi(\epsilon)/L = 1]$, the choice of λ_0 in Lemma 3.3 and in Theorem 3.1 are the same, since in such a case we have $\mathbb{E}\left[\psi(\epsilon)^2\right] = L^2$. For instance, this is the case if $\psi(x) = sign(x)L$ and $\mathbb{P}[\epsilon = 0] = 0$.

Generally the difference between $\sup_x |\psi(x)|^2 = L^2$ and $\mathbb{E}\left[\psi(\epsilon)^2\right]$ can be arbitrarily big. It is therefore interesting to give conditions under which the moment condition in Theorem 3.1, namely $\mathbb{E}\left[\psi(\epsilon)^{2k}\right] \leq \frac{(2k)!}{k!}2^{-k}\mathbb{E}\left[\psi(\epsilon)^2\right]^k$ for all $k \in \mathbb{N}$, is satisfied. For instance, in the classical uncontaminated case, i.e. $\psi(x) := x$ and $\epsilon \sim \mathcal{N}(0,1)$, the moment condition is obviously satisfied since a key property of the Gaussian distribution is that $\mathbb{E}\left[\epsilon^{2k}\right] = \frac{(2k)!}{k!}2^{-k}$. On the

other hand, in that case, we have $\sup_x |\psi(x)|^2 = +\infty$ and $\mathbb{E}\left[\psi(\epsilon)^2\right] = 1$. In the following lemma, we show that the moment condition is satisfied by a wide class of ψ-functions under the assumption that $\epsilon \sim \mathcal{N}\left(0, \tilde{\sigma}^2\right)$.

Lemma 3.4. *Let ψ be monotone increasing, with $\frac{x}{\psi(x)}$ monotone increasing in $|x|$, and assume that $\epsilon \sim \mathcal{N}\left(0, \tilde{\sigma}^2\right)$. Then, for all $k \in \mathbb{N}$, we have*

$$\mathbb{E}\left[\psi(\epsilon)^{2k}\right] \leq \frac{(2k)!}{k!} 2^{-k} \mathbb{E}\left[\psi(\epsilon)^2\right]^k.$$

Accordingly, the Huber ψ-function, i.e. $\psi(x) := \min\{\max\{x, -L\}, L\}$ for a given threshold $L > 0$, satisfies the moment condition for Gaussian errors, since ψ is obviously monotone increasing and $\frac{x}{\psi(x)} = \max\left\{1, \frac{|x|}{L}\right\}$. Lemma 3.4 also applies to unbounded ψ-functions. For instance, let $\psi(x) := x$ if $|x| < L$ and $\psi(x) := ax + sign(x)(1-a)L$ otherwise, for $0 \leq a \leq 1$ and $L > 0$. This ψ-function is obviously monotone increasing, since $a \geq 0$, and we have that $\frac{x}{\psi(x)} = \max\left\{1, \frac{|x|}{a|x|+(1-a)L}\right\}$ is monotone increasing in $|x|$ since $a \leq 1$. Therefore, this particular ψ-function also satisfies the moment condition for Gaussian errors, although it is unbounded.

More generally, in a case where ϵ does not follow a Gaussian distribution or ψ does not satisfy the conditions in Lemma 3.4, we can still check the moment condition, provided there exists $L < +\infty$ with $\sup_x |\psi(x)| = L$. When ψ is bounded, there are only finitely many conditions that one must check. Specifically, for $k_0 = \left\lceil \frac{2L^2}{\mathbb{E}\left[\psi(\epsilon)^2\right]} \right\rceil$ and $k \geq k_0$ in $\mathbb{N}$, we have the following:

$$\mathbb{E}\left[\psi(\epsilon)^{2k}\right]^{\frac{1}{k}} \leq L^2 \leq \frac{k}{2}\mathbb{E}\left[\psi(\epsilon)^2\right] \leq \left(\frac{(2k)!}{k!}2^{-k}\right)^{\frac{1}{k}} \mathbb{E}\left[\psi(\epsilon)^2\right]. \tag{3.2}$$

This means that we only need to check the moments for $2 \leq k < k_0$ in order to ensure that the moment condition is satisfied for all k.

Example 3.1. *To illustrate the use of this inequality, we apply it to errors with a t-distribution and Tukey's biweight ψ-function. Let $\psi(x) = x(1 - \frac{x^2}{c^2})^2$ if $|x| \leq c$ and $\psi(x) = 0$ otherwise, with $c = 4.685$. Moreover, we assume that $\epsilon \sim t_3$ and $\sigma = 1$. These choices imply that $\mathbb{E}\left[\psi(\epsilon)^2\right] \approx 0.638$ and $L = \sup_x |\psi(x)| \approx 1.341$. Therefore, we have $k_0 = 6$ and thus only need to check that $\mathbb{E}\left[\psi(\epsilon)^{2k}\right]\mathbb{E}\left[\psi(\epsilon)^2\right]^{-k} \leq \frac{(2k)!}{k!}2^{-k}$ for $k \in \{2,3,4,5\}$:*

$$\begin{aligned}
\mathbb{E}\left[\psi(\epsilon)^4\right]\mathbb{E}\left[\psi(\epsilon)^2\right]^{-2} &\approx 1.881 \quad \leq \quad 3 = \frac{4!}{2!}2^{-2}, \\
\mathbb{E}\left[\psi(\epsilon)^6\right]\mathbb{E}\left[\psi(\epsilon)^2\right]^{-3} &\approx 4.168 \quad \leq \quad 15 = \frac{6!}{3!}2^{-3}, \\
\mathbb{E}\left[\psi(\epsilon)^8\right]\mathbb{E}\left[\psi(\epsilon)^2\right]^{-4} &\approx 9.927 \quad \leq \quad 105 = \frac{8!}{4!}2^{-4},
\end{aligned}$$

$$\mathbb{E}\left[\psi(\epsilon)^{10}\right]\mathbb{E}\left[\psi(\epsilon)^{2}\right]^{-5} \approx 24.608 \quad \leq \quad 945 = \frac{10!}{5!}2^{-5}.$$

Since $\mathbb{E}\left[\psi(\epsilon)^{2k}\right]\mathbb{E}\left[\psi(\epsilon)^{2}\right]^{-k} \leq \frac{(2k)!}{k!}2^{-k}$ *for all relevant values of* k*, the moment condition is satisfied in this case despite the fact that the* ψ*-function is non-increasing.*

We note that this method can be used for arbitrary bounded odd ψ-functions and symmetric error distributions. It does however require the knowledge of the error distribution, which may reduce its use in some important cases.

In robust statistics, for instance, the distribution of ϵ is only approximately known [7]. In such a situation, we cannot use Lemma 3.4 or even Equation (3.2) to check the moment condition directly. Since this is an important application of our work, we study the moment condition for a fixed k in contamination models. In the following lemma, we provide a method to verify the moment condition in contamination models.

Lemma 3.5. *Suppose* $\epsilon^* \sim G$*, where* G *is a distribution function,* $\sup_x |\psi(x)| = L$*,* $0 < \delta < 1$ *and* $k \in \mathbb{N}$*. Furthermore we assume that*

$$\mathbb{E}\left[\psi(\epsilon^*)^{2k}\right] \leq \frac{(2k)!}{k!}2^{-k}(1-\delta)^{k-1}\mathbb{E}\left[\psi(\epsilon^*)^{2}\right]^{k} \tag{3.3}$$

and

$$(1-\delta)\mathbb{E}\left[\psi(\epsilon^*)^{2k}\right] + \delta L^{2k} \leq \frac{(2k)!}{k!}2^{-k}\left[(1-\delta)\mathbb{E}\left[\psi(\epsilon^*)^{2}\right] + \delta L^{2}\right]^{k}. \tag{3.4}$$

Then for any $\epsilon \sim F$*, where* $F(x) := (1-\delta)G(x) + \delta H(x)$ *and* H *is an arbitrary symmetric distribution, we have:*

$$\mathbb{E}\left[\psi(\epsilon)^{2k}\right] \leq \frac{(2k)!}{k!}2^{-k}\mathbb{E}\left[\psi(\epsilon)^{2}\right]^{k}. \tag{3.5}$$

Lemma 3.5 allows for a verification of the moment condition based only on L and the distribution G. As shown in the Appendix (see Lemmas A.1 and A.2), $H(x) = 1_{0\leq x}$ and $H(x) = \frac{1}{2}1_{-u\leq x} + \frac{1}{2}1_{u\leq x}$ with $u = \arg\max \psi(x)$, are the most challenging types of contaminating distributions. In fact, for $H(x) = 1_{0\leq x}$, Equations (3.3) and (3.5) are equal, while for $H(x) = \frac{1}{2}1_{-u\leq x} + \frac{1}{2}1_{u\leq x}$, Equations (3.4) and (3.5) are equal. This means that the conditions in Lemma 3.5 are necessary, since if either Equation (3.3) or (3.4) fails then there exists a distribution H so that $\mathbb{E}\left[\psi(\epsilon)^{2k}\right] > \frac{(2k)!}{k!}2^{-k}\mathbb{E}\left[\psi(\epsilon)^{2}\right]^{k}$.

Example 3.2. *To showcase the use of Lemma 3.5, we apply it to the contaminated normal case with a Huber* ψ*-function as first studied by [7]. More specifically, for a given* $\delta > 0$*, we assume* $\epsilon \sim F$*, where* G *is the standard normal distribution* $\mathcal{N}(0,1)$*,* H *is an unknown symmetric distribution and* F *is as in Lemma 3.5. We are now interested in the maximal value of* L *so that Equations* (3.3) *and* (3.4) *hold for all* $k \geq 2$*. Obviously they are satisfied for* $L = 0$*,*

and for any value of L below this maximal value. Moreover, for $k = 2$, we have $\frac{(2k)!}{k!}2^{-k}(1-\delta)^{k-1} = 3(1-\delta)$, and since $\mathbb{E}\left[\psi\left(\epsilon^*\right)^{2k}\right] \geq \mathbb{E}\left[\psi\left(\epsilon^*\right)^2\right]^k$ by Jensen's inequality, Equation (3.3) is only met for $L = 0$ when $\delta > \frac{2}{3}$. More generally, using the same methods as in Example 3.1, we can recover this maximal value which is always strictly bigger than 0 for $\delta < \frac{2}{3}$. Table 1 shows the value of L for which the moment condition is satisfied with equality for the various levels of contamination δ considered in [7].

TABLE 1
Maximal values of L with respect to δ.

Level δ	0.0008	0.0023	0.0061	0.0156	0.0376	0.0855	0.1825	0.3599
Maximal L	4.1082	3.8167	3.5209	3.2010	2.8552	2.4652	1.9976	1.3645

3.4. New error bounds for the convex loss Lasso

In this subsection, we derive new prediction and estimation error bounds for the convex loss Lasso. To do so, we start by exploiting the sparsity of $\boldsymbol{\beta}^0$ just as in [1]. Let $S \subset \{1, ..., p\}$ and $\boldsymbol{\beta}_{j,S} = \boldsymbol{\beta}_j 1_{j \in S}$. Throughout, let S_0 be the true set of non zeros of $\boldsymbol{\beta}^0$ and $s_0 = |S_0|$ is the number of non zeros of $\boldsymbol{\beta}^0$.

Following [1], we say that the *compatibility condition* is met for the set S_0 if for some $\phi_0 > 0$ and for all $\boldsymbol{\theta} \in \mathbb{R}^p$ satisfying $\left\|\boldsymbol{\theta}_{S_0^c}\right\|_1 \leq 3\left\|\boldsymbol{\theta}_{S_0}\right\|_1$, it holds that

$$\left\|\boldsymbol{\theta}_{S_0}\right\|_1^2 \leq \|\boldsymbol{\theta}\|_{\hat{\boldsymbol{\Sigma}}}^2 \frac{s_0}{\phi_0^2},$$

where $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i^T\boldsymbol{X}_i$ is the Gram matrix and $\|\boldsymbol{\theta}\|_{\hat{\boldsymbol{\Sigma}}}^2 := \boldsymbol{\theta}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\theta}$.

Assumption 3.1. *There exists a monotone increasing sequence $\{\theta_j\}_{j=0}^{N+1}$ with $\theta_0 = -\infty$ and $\theta_{N+1} = +\infty$, where ψ' restricted to the intervals (θ_{j-1}, θ_j) is uniformly continuous. If $N \geq 1$, then ϵ has a continuous distribution at all the points in $\Theta := \{\theta_j : j \in \{1, ..., N\}\}$.*

Assumption 3.2. $\inf_{x \notin \Theta} \psi'(x) \geq 0$ *(convexity of ρ) and* $0 < \sup_{x \notin \Theta} \psi'(x) \leq 1$ *(bounded ψ').*

Assumption 3.1 is a restriction on the type of ψ-functions that we investigate. We point out that this assumption is fulfilled by the Huber ψ-function and many other ψ-functions. Additionally, the restriction on ϵ still allows for a discrete distribution; the masses just have to be away from the points $\{\theta_j\}_{j=1}^N$.

Assumption 3.2 is an assumption of convexity, while also of bounded ψ'. Bounding ψ' by 1, as opposed to another constant, is by no means restrictive, since if for instance $1 < \sup_x \psi'(x) < \infty$ we can work with $\frac{1}{\sup_x \psi'(x)}\rho$ instead of ρ. The reason to bound ψ' is that we do not want ρ to grow strictly faster than a quadratic function.

We can now provide the following theorem.

Theorem 3.2. *There exists $\alpha^0, c_0 > 0$, such that if the compatibility condition holds for S_0, $12\lambda K \frac{\sigma}{\mathbb{E}[\psi'(\epsilon)]} \frac{s_0}{\phi_0^2} \leq \alpha \leq \alpha^0$, $\tilde{\lambda} \frac{s_0}{\phi_0^2} \leq c_0$ and $2\lambda_0 \leq \lambda$, then, on $\mathcal{J}_0 \cap \mathcal{I}_\alpha$,*

$$\frac{\mathbb{E}\left[\psi'(\epsilon)\right] - \Delta_1(\alpha)}{\sigma^2} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 + [1 - \Delta_2(\tilde{\lambda}\frac{s_0}{\phi_0^2})]\lambda \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq 4\lambda^2 \frac{\sigma^2}{\mathbb{E}\left[\psi'(\epsilon)\right]} \frac{s_0}{\phi_0^2},$$

where $\mathcal{I}_\alpha$ is a set depending on α and $\tilde{\lambda}$ defined in the Appendix, the Δ_i are continuous, monotone increasing and satisfy $\Delta_i(0) = 0$, and $\max_{i,j} \left|\boldsymbol{X}_i^{(j)}\right| \leq K$.

3.5. Asymptotic results

Here we study the asymptotic implications of Theorem 3.2, where we allow both p and n to tend to infinity. In order to recover asymptotic results, we impose a condition on the covariates as the dimensions diverge.

Assumption 3.3. *The covariates are bounded, i.e.* $\max_{i,j} \left|\boldsymbol{X}_i^{(j)}\right| \leq K$.

Assumption 3.3 is rather common in robust statistics, since it limits the influence of single covariates and so the leverage of single covariates is limited. Because of $\left(\boldsymbol{X}^T\boldsymbol{X}\right)_{jj} = n$, we must have $1 \leq K$.

In the Appendix, under the same condition for consistency as for the classical Lasso, namely $\sqrt{\frac{\log(p)}{n}} \frac{s_0}{\phi_0^2} \to 0$ as $n \to \infty$, we show that it is possible to let $\tilde{\lambda}$ depend on n such that $\tilde{\lambda}\frac{s_0}{\phi_0^2}$ tends to 0 with $\mathbb{P}\left[\mathcal{I}_\alpha\right]$ tending to 1.

Therefore, for $\lambda_0 = 2\frac{\sqrt{\mathbb{E}\left[\psi(\epsilon)^2\right]}}{\sigma}\sqrt{\frac{t^2+2log(p)}{n}}$, $\lambda = 2\lambda_0$ and $\alpha = 12\lambda K \frac{\sigma}{\mathbb{E}[\psi'(\epsilon)]} \frac{s_0}{\phi_0^2}$ in Theorem 3.2, and by using Theorem 3.1, we recover

$$\frac{\mathbb{E}\left[\psi'(\epsilon)\right]}{\sigma^2} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 + 2\lambda_0 \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq 16\lambda_0^2 \frac{\sigma^2}{\mathbb{E}\left[\psi'(\epsilon)\right]} \frac{s_0}{\phi_0^2}$$

with probability approximately at least $1 - 2\exp\left[-t^2/2\right]$ (asymptotically in n). Consequently, we recover estimation and prediction error bounds,

$$\begin{aligned}
\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 &\leq 16\sigma\sqrt{\frac{t^2 + 2\log(p)}{n}} \frac{\sqrt{\mathbb{E}\left[\psi^2(\epsilon)\right]}}{\mathbb{E}\left[\psi'(\epsilon)\right]} \frac{s_0}{\phi_0^2}, \\
\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 &\leq 16\sigma^2 \frac{t^2 + 2\log(p)}{n} \frac{\mathbb{E}\left[\psi^2(\epsilon)\right]}{\mathbb{E}\left[\psi'(\epsilon)\right]^2} \frac{s_0}{\phi_0^2},
\end{aligned}$$

with probability approximately at least $1 - 2\exp\left[-t^2/2\right]$ (asymptotically in n). In both cases, for estimation and prediction errors, we recover the same error bound as that of the classical Lasso in [1] but for a term solely depending on $\mathbb{E}\left[\psi^2(\epsilon)\right]/\mathbb{E}\left[\psi'(\epsilon)\right]^2$. We note that this ratio has primarily emerged from the analysis conducted in Subsection 3.3, where our focus was on giving conditions

for controlling the empirical process component. Obviously, the examples we gave there can be considered as well in this asymptotic setting since they fulfil the moment condition.

4. Error bounds with estimated scale

In most applications, σ is unknown and must be estimated. This section addresses joint estimation of the regression and scale parameters. In particular, we show how the results obtained in the previous section carry over to this more realistic setting.

4.1. Scaled convex loss Lasso estimator

Just as in [10], we propose the following estimating equation

$$\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_a\right) \in \underset{\boldsymbol{\beta},\sigma}{\arg\min} \left(\frac{2}{n} \sum_{i=1}^{n} \left(\rho\left(\frac{Y_i - \left(\boldsymbol{X}\boldsymbol{\beta}\right)_i}{\sigma} \right) + a \right) \sigma + \lambda_* \left\| \boldsymbol{\beta} \right\|_1 \right),$$

where $a \in \mathbb{R}$. It follows from [8] (equation 7.110, p. 174) that if ρ is convex then the function in the equation above is convex. This was used by [13] in the classical Lasso case. We extend it here to a general class of ρ-functions. The idea is that, in the case where σ is known and fixed, we recover the same definition as in (2.2) for $\lambda_* = \sigma\lambda$.

The estimation procedure now depends on a new parameter a, which mainly affects the estimation of σ. As we will see, $\hat{\sigma}_a$ converges in probability to σ_a, the solution in $\tilde{\sigma}$ of $a = \mathbb{E}\left[\chi_0\left(\frac{\sigma\epsilon}{\tilde{\sigma}}\right)\right]$, where $\chi_0(x) = \psi(x)\, x - \rho(x)$. Consequently, for $a = \mathbb{E}\left[\chi_0(\epsilon)\right]$, $\hat{\sigma}_a$ converges in probability to σ. We will study the choice of a from a robustness point of view in Section 5.

Just as in Section 3, the assumptions on ψ' will slightly change throughout the following subsections. The purpose of Subsection 4.2 is to describe the construction of basic bounds, where once again, for the sake of brevity, we impose ψ' to be well defined and continuous here. For the new error bounds that we develop in Subsection 4.3, we refine our assumptions on ψ', similarly to what was done in Subsection 3.4 (see Assumptions 4.1 and 4.2). Finally, for the asymptotic results in Subsection 4.4, no further restriction is needed.

4.2. Basic bounds

We now provide a new basic inequality in this case depending on a.

Lemma 4.1. *There exists* $t^{\lambda_*} \in [0,1]$ *such that*

$$\begin{aligned} \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0, \hat{\sigma}_a - \sigma_a \right\|_{\boldsymbol{\Gamma}(t^{\lambda_*})} + \lambda_* \left\| \hat{\boldsymbol{\beta}} \right\|_1 \leq & - \frac{2}{n} \sum_{i=1}^{n} \left(a - \chi_0\left(\frac{\sigma\epsilon_i}{\sigma_a} \right) \right) \left(\hat{\sigma}_a - \sigma_a \right) \\ & - \frac{2}{n} \boldsymbol{\psi}\left(\frac{\sigma\boldsymbol{\epsilon}}{\sigma_a} \right)^T \boldsymbol{X} \left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right) + \lambda_* \left\| \boldsymbol{\beta}^0 \right\|_1, \end{aligned}$$

where $\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0, \hat{\sigma}_a-\sigma_a\right\|_{\boldsymbol{\Gamma}(t^{\lambda_*})}$ *is non negative and is equal to*

$$\frac{1}{n}\sum_{i=1}^{n}\psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{a_i^2}{\tilde{\sigma}_a}+\frac{1}{n}\sum_{i=1}^{n}\chi_0'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\frac{\left(\hat{\sigma}_a-\sigma_a\right)^2}{\tilde{\sigma}_a}$$
$$-\frac{2}{n}\sum_{i=1}^{n}\psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\frac{a_i}{\tilde{\sigma}_a}\left(\hat{\sigma}-\sigma_a\right)$$

with $\tilde{\sigma}_a=\sigma_a+t^{\lambda_*}\left(\hat{\sigma}_a-\sigma_a\right)$ *and* $\tilde{\epsilon}_i=\epsilon_i+t^{\lambda_*}\frac{a_i}{\sigma}$.

It is interesting to see that if $\hat{\sigma}_a=\sigma=\sigma_a$, we recover the same bound as in Lemma 3.1, by setting $\lambda_*=\sigma\lambda$ in the optimization.

As opposed to the case with fixed σ, there are now two empirical processes, namely $\frac{2}{n}\sum_{i=1}^{n}\left(a-\chi_0\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)\right)\left(\hat{\sigma}_a-\sigma_a\right)$ and $\frac{2}{n}\boldsymbol{\psi}\left(\frac{\sigma\boldsymbol{\epsilon}}{\sigma_a}\right)^T\boldsymbol{X}\left(\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right)$. Similarly to Subsection 3.1, this motivates the following definitions:

$$\mathcal{J}_{a;0} := \left\{\max_{1\le j\le p}\left|\frac{2}{n}\boldsymbol{\psi}\left(\frac{\sigma\boldsymbol{\epsilon}}{\sigma_a}\right)^T\frac{\boldsymbol{X}^{(j)}}{\sigma_a}\right|\le\lambda_0\right\},$$
$$\mathcal{J}_{a;1} := \left\{\left|\frac{1}{\sigma_a}\frac{2}{n}\sum_{i=1}^{n}\left(a-\chi_0\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)\right)\right|\le\lambda_1\right\}.$$

In the special case $a=\mathbb{E}\left[\chi_0\left(\epsilon\right)\right]$ and consequently $\sigma_a=\sigma$, $\mathcal{J}_{a;0}=\mathcal{J}_0$. More generally, the same technics used to study $\mathcal{J}_0$ can be used to study $\mathcal{J}_{a;0}$, since $x\mapsto\psi\left(\frac{\sigma x}{\sigma_a}\right)$ can itself be studied as a ψ-function.

By definition, under the mild assumption, that $\mathbb{E}\left[\chi_0\left(\frac{\sigma\epsilon}{\sigma_a}\right)^2\right]<\infty$, we have that $\frac{2}{n}\sum_{i=1}^{n}\left(a-\chi_0\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)\right)=O_{\mathbb{P}}(n^{-\frac{1}{2}})$ and thus for $\lambda_1\sim n^{-\frac{1}{2}}$, $\mathcal{J}_{a;1}$ can hold with arbitrarily high probability.

4.3. New error bounds for the scaled convex loss Lasso

Before stating the theorem on the joint error bounds, we go through the required assumptions for the theorem to hold. We need to alter Assumption 3.1, since the discontinuous points now need to be scaled. This leads to the following assumption.

Assumption 4.1. *There exists a monotone increasing sequence* $\{\theta_j\}_{j=0}^{N+1}$ *with* $\theta_0=-\infty$ *and* $\theta_{N+1}=+\infty$*, where* ψ'' *restricted to the intervals* (θ_{j-1},θ_j) *is uniformly continuous. If* $N\ge1$*, then* ϵ *has a continuous distribution at all the points in* $\Theta_a:=\left\{\frac{\sigma_a}{\sigma}\theta_j:j\in\{1,...,N\}\right\}$.

There are two differences with respect to Assumption 3.1, namely that we are working with ψ'' instead of ψ' and that the set over which ϵ must have a continuous distribution is Θ_a rather than Θ.

Moreover, we alter Assumption 3.2, to impose a new assumption on ψ.

Assumption 4.2. $\inf_{x\notin\Theta}\psi'(x)\geq 0$ *(convexity of* ρ*),* $0<\sup_{x\notin\Theta}\psi'(x)\leq 1$ *(bounded* ψ'*), and* $\sup_{x\notin\Theta}\psi''(x)\,x^2<+\infty$.

The additional restriction that we impose with respect to Assumption 3.1 is that $\sup_{x\notin\Theta}\psi''(x)\,x^2<+\infty$.

While the classical Lasso obviously satisfies both of these assumptions (since in that case $\psi''(x)=1$), we stress that the Huberized Lasso also satisfies the above assumptions if $\mathbb{P}\left[\frac{\sigma_a}{\sigma}|\epsilon|=L\right]=0$, since in that case $\psi_L''(x)=0$ for $|x|\neq L$.

Theorem 4.1. *There exist* $c_0,c_1,c_2,c_3,\alpha_*^0,\delta_*^0,>0$*, such that if the compatibility condition holds for* S_0*,* $\frac{10\lambda_*}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}c_3K\leq\delta_*\leq\delta_*^0$*,* $\frac{10\lambda_*}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}K\leq\alpha_*\leq\alpha_*^0$*,* $\tilde{\lambda}_*\frac{s_0}{\phi_0^2}\leq c_0$*,* $\lambda_1\sigma_a\leq c_1\min\left\{\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_1/\sigma_a,\delta_*^0\right\}$*,* $\lambda_2\sigma_a\leq 1$*,* $\gamma_*\leq c_2$*,* $\frac{2\lambda_0\sigma_a+2\lambda_2^2\sigma_a^2+2\lambda_1\sigma_a}{1-\frac{\tilde{\lambda}_*}{c_0}\frac{s_0}{\phi_0^2}}\leq\lambda_*$*, then, on* $\mathcal{J}_{a;0}\cap\mathcal{J}_{a;1}\cap\mathcal{J}_{a;2}\cap\mathcal{I}_{\alpha_*,\delta_*}\cap\mathcal{G}_a$*,*

$$\frac{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]-\Delta_3(\alpha_*+\delta_*+\gamma_*)}{\sigma_a}\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2+\lambda_*\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1\leq\frac{4\lambda_*^2\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2},$$

where $\mathcal{I}_{\alpha_*,\delta_*}$ *is a set depending on* α_**,* δ_* *and* $\tilde{\lambda}_*$ *defined in the Appendix,* Δ_3 *is continuous, monotone increasing and satisfies* $\Delta_3(0)=0$*,* $\mathcal{J}_{a;2}$ *depends on* λ_2 *and while* $\mathcal{G}_a$ *depends on* γ_* *defined in the Appendix, and* $\max_{i,j}\left|\boldsymbol{X}_i^{(j)}\right|\leq K$.

Moreover, there exist $C_1,C_2,C_3,C_4>0$ *such that,*

$$\frac{|\hat{\sigma}_a-\sigma_a|}{\sigma_a}\leq C_1\lambda_1\sigma_a+C_2\lambda_2\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_1+\frac{C_3}{n}\sum_{i\in J_{\alpha_*,\delta_*}}\frac{|a_i|}{\sigma_a}+\frac{C_4}{n}\sum_{i=1}^n\frac{a_i^2}{\sigma_a^2},$$

where $J_{\alpha_*,\delta_*}=\left\{i\in\{1,...,n\}:\inf_{1\leq j\leq N}\left\{\left|\frac{\sigma\epsilon_i}{\sigma_a}-\theta_j\right|-\frac{\delta_*}{1-\delta_*}\left|\frac{\sigma\epsilon_i}{\sigma_a}\right|\right\}\leq\frac{\alpha_*}{1-\delta_*}\right\}$.

4.4. Asymptotic results

In this subsection we study the implications of Theorem 4.1 in the asymptotic set-up. First of all, we point out that there is an unusual condition in this theorem, namely $\lambda_1\sigma_a\leq c_1\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_1/\sigma_a$. This condition is in no way restrictive, since, as we have we can pick $\lambda_1=o\left(n^{-\frac{1}{2}}\right)$, which is a much faster rate of convergence than the one we show.

Set $\lambda_*=\frac{2\lambda_0\sigma_a+2\lambda_2^2\sigma_a^2+2\lambda_1\sigma_a}{1-\frac{\tilde{\lambda}_*}{c_0}\frac{s_0}{\phi_0^2}}$, $\delta_*=\frac{10\lambda_*}{\mathbb{E}[\psi'(\epsilon)]}\frac{s_0}{\phi_0^2}c_3K$ and $\alpha_*=\frac{10\lambda_*}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}K$. Under the standard asymptotic assumption, namely $\sqrt{\frac{\log p}{n}}\frac{s_0}{\phi_0^2}\to 0$ as $n\to\infty$, it is shown in the Appendix that $\mathbb{P}\left[\mathcal{I}_{\alpha_*,\delta_*}\cap\mathcal{G}_a\cap\mathcal{J}_{a;1}\cap\mathcal{J}_{a;2}\right]\to 1$ with $\frac{\lambda_*}{2\lambda_0\sigma_a}\to 1$ and $\delta_*\to 0$, $\alpha_*\to 0$ and $\gamma_*\to 0$.

Following the ideas in Subsection 3.5, set $\lambda_0 = 2\frac{\sqrt{\mathbb{E}\left[\psi\left(\frac{\sigma\epsilon}{\sigma_a}\right)^2\right]}}{\sigma_a}\sqrt{\frac{t^2+2\log(p)}{n}}$, in which case $\mathbb{P}\left[\mathcal{J}_{a;0}\right] \geq 1-2\exp\left[-t^2/2\right]$, under the moment condition. Therefore, by Theorem 4.1, just as in Section 3, we recover

$$\frac{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{\sigma_a^2}\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 + 2\lambda_0\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1 \leq 16\lambda_0^2\frac{\sigma_a^2}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}$$

with probability approximately at least $1-2\exp\left[-t^2/2\right]$ (asymptotically in n). Consequently, in this case, we recover estimation and prediction error bounds,

$$\begin{aligned}\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1 &\leq 16\sigma\sqrt{\frac{t^2+2\log(p)}{n}}\frac{\sigma_a\sqrt{\mathbb{E}\left[\psi^2\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}}{\sigma\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2},\\ \left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 &\leq 16\sigma^2\frac{t^2+2\log(p)}{n}\frac{\sigma_a^2\mathbb{E}\left[\psi^2\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{\sigma^2\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]^2}\frac{s_0}{\phi_0^2},\end{aligned}$$

with probability approximately at least $1-2\exp\left[-t^2/2\right]$ (asymptotically in n).

Interestingly, in terms of asymptotics for $\hat{\boldsymbol{\beta}}$, the error bounds that we recover depend on the ψ-function and on the distribution of ϵ only through the ratio $\sigma_a^2/\sigma^2\mathbb{E}\left[\psi^2\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]/\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]^2$. The main difference between the case of known σ and this one is the scale parameter σ_a, where we remind the reader that σ_a is the solution to $a = \mathbb{E}\left[\chi_0\left(\frac{\sigma\epsilon}{\tilde{\sigma}}\right)\right]$ in $\tilde{\sigma}$.

5. Minimaxity of bounds and relevance of the ratios

In this section, we first study the derived error bounds in terms of minimaxity, which basically amounts to studying $\mathbb{E}\left[\psi^2(\epsilon)\right]/\mathbb{E}\left[\psi'(\epsilon)\right]^2$ for known σ and $\sigma_a^2/\sigma^2\mathbb{E}\left[\psi^2\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]/\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]^2$ for estimated σ in terms of minimaxity as done by [7]. Then, we show that these ratios appear asymptotically in a couple of other interesting settings.

5.1. Minimaxity of error bounds for known scale

Under the moment condition, the ratio $\mathbb{E}\left[\psi(\epsilon)^2\right]/\mathbb{E}\left[\psi'(\epsilon)\right]^2$ appears in the error bounds derived in Sections 3 and 4. Therefore, to optimize the error bounds with respect to ψ, we need to minimize $\mathbb{E}\left[\psi(\epsilon)^2\right]/\mathbb{E}\left[\psi'(\epsilon)\right]^2$. Then, if the corresponding ψ satisfies the moment condition it is obviously the optimal ψ for the bound. For instance, if $\epsilon \sim F$, where F is the normal distribution $\mathcal{N}(0,1)$, $\psi(x) = x$ is

clearly optimal, since this ψ-function minimizes $\mathbb{E}\left[\psi(\epsilon)^2\right]/\mathbb{E}\left[\psi'(\epsilon)\right]^2$ while also satisfying the moment condition.

More generally, just as in [7], we define $V(\psi, F) = \mathbb{E}_F\left[\psi(\epsilon)^2\right]/\mathbb{E}_F\left[\psi'(\epsilon)\right]^2$. As noted earlier, in robust statistics, it is assumed that we only approximately know the true distribution F, which can be modelled as $F = (1-\delta)G+\delta H$, where δ is the contamination level and H is a contaminating distribution. This then leads to a minimax problem, namely, solving $\min_\psi \max_{F=(1-\delta)G+\delta H} V(\psi, F)$.

Example 5.1. *We continue our study of the case where G is the normal distribution $\mathcal{N}(0,1)$ as considered in Example 3.2. In this case, the solution to the minimax problem was derived by [7] and was shown to be the Huber ψ-function, with a tuning parameter L depending on δ. In fact, we can combine these results with the ones given in Example 3.2 to obtain the following table.*

TABLE 2
Optimal and maximal values with respect to δ.

Level δ	0.0008	0.0023	0.0061	0.0156	0.0376	0.0855	0.1825	0.3599
Maximal L	4.1082	3.8167	3.5209	3.2010	2.8552	2.4652	1.9976	1.3645
Optimal L	2.7000	2.4000	2.1000	1.8000	1.5000	1.2000	0.9000	0.6000

Table 2 reveals that the optimal tuning parameter L is always smaller for the selected contamination levels than the maximal value L for which the moment condition is still satisfied. This implies that for reasonable contamination levels, the Huber-ψ function produces minimax error bounds under the assumption that G is the standard normal distribution.

5.2. Minimaxity of error bounds for estimated scale

To begin, let $V_a(\psi, F) = \sigma_a(\psi, F)^2/\sigma^2\mathbb{E}_F\left[\psi\left(\frac{\sigma\epsilon}{\sigma_a(\psi,F)}\right)^2\right]/\mathbb{E}_F\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a(\psi,F)}\right)\right]^2$, where we make the definition of σ_a explicit through F and ψ. Additionally, let $\tilde{\psi}(x) = \frac{1}{\tilde{\sigma}}\psi(\tilde{\sigma}x)$. Consequently, we recover $\mathbb{E}_F\left[\tilde{\psi}'(\epsilon)\right]^2 = \mathbb{E}_F\left[\psi'(\tilde{\sigma}\epsilon)\right]^2$ and $\mathbb{E}_F\left[\tilde{\psi}(\epsilon)^2\right] = \frac{1}{\tilde{\sigma}^2}\mathbb{E}_F\left[\psi(\tilde{\sigma}\epsilon)^2\right]$. Thus, for $\tilde{\sigma} = \sigma/\sigma_a(\psi, F)$, we have $V_a(\psi, F) = V(\tilde{\psi}, F)$.

With these relations in mind, we now show that finding the ψ-function solving $\min_\psi \max_{F=(1-\delta)G+\delta H} V_a(\psi, F)$ reduces to the previously studied problem of finding the $\tilde{\psi}$-function solving $\min_{\tilde{\psi}} \max_{F=(1-\delta)G+\delta H} V(\tilde{\psi}, F)$. In fact, let $(\tilde{\psi}, F)$ be a solution to the minimax problem involving V. Define $\tilde{\sigma}$ through the equation $a = \tilde{\sigma}^2\mathbb{E}_F\left[\tilde{\psi}(\epsilon)\epsilon - \tilde{\rho}(\epsilon)\right]$, where $\tilde{\rho}$ is the ρ-function corresponding to the ψ-function $\tilde{\psi}$. Now let $\rho(x) = \tilde{\sigma}^2\tilde{\rho}\left(\frac{x}{\tilde{\sigma}}\right)$ and $\psi(x) = \rho'(x) = \tilde{\sigma}\tilde{\psi}\left(\frac{x}{\tilde{\sigma}}\right)$. Correspondingly, we recover $a = \mathbb{E}_F\left[\psi(\tilde{\sigma}\epsilon)\tilde{\sigma}\epsilon - \rho(\tilde{\sigma}\epsilon)\right] = \mathbb{E}_F\left[\chi_0(\tilde{\sigma}\epsilon)\right]$, and therefore we identify $\tilde{\sigma} = \sigma/\sigma_a(\psi, F)$. Thus, (ψ, F) is a solution to the minimax problem involving V_a.

As shown above, the minimizer ψ depends on a only through scaling. Just as in [7], we propose setting $a = \mathbb{E}_G[\chi_0(\epsilon)]$. In such a case, we recover $\sigma_a(\psi, F) = \sigma$ in the non-contaminated model (i.e., $F = G$). This leads to the equation $\mathbb{E}_G\left[\tilde{\sigma}\tilde{\psi}\left(\frac{\epsilon}{\tilde{\sigma}}\right)\epsilon - \tilde{\sigma}^2\tilde{\rho}\left(\frac{\epsilon}{\tilde{\sigma}}\right)\right] = \tilde{\sigma}^2\mathbb{E}_F\left[\tilde{\psi}(\epsilon)\epsilon - \tilde{\rho}(\epsilon)\right]$ defining $\tilde{\sigma}$.

Example 5.2. *Once again we study the case where G is the normal distribution $\mathcal{N}(0,1)$. In this case, the solution to the minimax problem involving V, the Huber ψ-function with parameter L, was studied in Example 5.1. Obviously if $\tilde{\psi}$ is the Huber ψ-function with parameter L, then ψ defined through $\psi(x) = \tilde{\sigma}\tilde{\psi}\left(\frac{x}{\tilde{\sigma}}\right)$ is the Huber ψ-function with parameter $\tilde{\sigma}L$. Thus the optimal ψ-functions in solving the new minimax problem are Huber ψ-functions.*

The worst case contamination, as shown by [7], involves any symmetric distribution with mass outside of $[-L, L]$. This allows us to recover the parameter in the ψ-function in the minimax problem involving V_a from that of the optimal ψ-function in the minimax problem involving V, by computing $\tilde{\sigma}$. Table 3 contains the results for the same levels of contamination as in Example 5.1.

TABLE 3
Optimal L and corresponding $\tilde{\sigma}$ with respect to δ.

Level δ	0.0008	0.0023	0.0061	0.0156	0.0376	0.0855	0.1825
V:Optimal L	2.7000	2.4000	2.1000	1.8000	1.5000	1.2000	0.9000
$\tilde{\sigma}$ for given L	0.9973	0.9938	0.9866	0.9720	0.9439	0.8914	0.7952
V_a:Optimal L	2.6928	2.3850	2.0718	1.7496	1.4159	1.0697	0.7157

5.3. Relevance of the ratios

In this subsection we provide some theory to compare the estimation and prediction errors associated to the convex loss Lasso estimators for different ρ-functions. We focus on the case of known scale, noting that it is straightforward to recover the corresponding results in the case of estimated scale.

5.3.1. Estimating equations

The general Karush-Kuhn-Tucker (KKT) conditions for convex functions will play a key part in the analysis. Indeed, these conditions allow us to characterize all possible solutions to the Lasso problem in a straightforward manner. More specifically, we have that $\boldsymbol{\beta}$ is a solution if and only if

$$\begin{aligned} \frac{2}{n}\boldsymbol{X}^T_{[,S(\boldsymbol{\beta})]}\boldsymbol{\psi}\left(Y - \boldsymbol{X}\boldsymbol{\beta}\right) &= \lambda sign(\boldsymbol{\beta}_{[S(\boldsymbol{\beta})]}) \\ \text{and } \left\|\frac{2}{n}\boldsymbol{X}^T\boldsymbol{\psi}\left(Y - \boldsymbol{X}\boldsymbol{\beta}\right)\right\|_{+\infty} &\leq \lambda, \end{aligned}$$

where $S(\boldsymbol{\beta}) = \{i \in \{1, ..., p\} : \boldsymbol{\beta}_i \neq 0\}$ and, for simplicity, we set $\sigma = 1$.

For given observations $(Y, \boldsymbol{X})$, let $\hat{\boldsymbol{\beta}}$ be the convex loss Lasso corresponding to $\lambda = \sqrt{\mathbb{E}\left[\psi\left(\epsilon\right)^2\right]}\lambda^*$. This leads to the following scores:

$$\begin{aligned}\frac{2}{n}\boldsymbol{X}^T\boldsymbol{\psi}\left(Y - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) &= \frac{2}{n}\boldsymbol{X}^T\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right) + \frac{2}{n}\boldsymbol{X}^T D\boldsymbol{\psi}\left(\tilde{\boldsymbol{\epsilon}}\right)\boldsymbol{X}\left[\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right], \\ &= \frac{2}{n}\boldsymbol{X}^T\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right) + \mathbb{E}\left[\psi'\left(\epsilon\right)\right]\frac{2}{n}\boldsymbol{X}^T\boldsymbol{X}\left[\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right] + \Delta.\end{aligned}$$

We will study the ratio under different sets of hypotheses.

5.3.2. Ratio in distribution

We start by making a strong assumption, namely that there exists a fixed $\boldsymbol{S}^*$ such that $\boldsymbol{S}^* = sign\left(\hat{\boldsymbol{\beta}}\right)$ with high probability. In that case, we can define $S_* = S\left(\hat{\boldsymbol{\beta}}\right)$, which is well defined in the event that $sign\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{S}^*$. Furthermore, we can assume that $S\left(\boldsymbol{\beta}^0\right) = S_0 \subset S_*$ (by assuming that the non zero elements of $\boldsymbol{\beta}^0$ are big enough, this follows directly). Then, by dividing the components in S_* of the scores by $\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}$ we recover:

$$\lambda^*\boldsymbol{S}^*_{[S_*]} = \frac{2}{n}\frac{\boldsymbol{X}_{[,S_*]}{}^T\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}} + \frac{\mathbb{E}\left[\psi'\left(\epsilon\right)\right]}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}}\frac{2}{n}\boldsymbol{X}^T\boldsymbol{X}_{[S_*,]}\left[\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right] + \frac{\Delta_{[S_*]}}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}}.$$

Now since $\left[\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right]_{[j]} = 0$ for all $j \notin S_*$, by the previous equation we recover:

$$\frac{\mathbb{E}\left[\psi'\left(\epsilon\right)\right]}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}}\frac{2}{n}\boldsymbol{X}^T\boldsymbol{X}_{[S_*,S_*]}\left[\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right]_{[S_*]} = \lambda^*\boldsymbol{S}^*_{[S_*]} - \frac{2\boldsymbol{X}_{[,S_*]}{}^T\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right) + n\Delta_{[S_*]}}{n\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}}.$$

Under the extra assumption that $\sqrt{n}\left\|\Delta\right\|_\infty$ is very small, which is true for instance if $\sqrt{\log(p)}\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1$ is small, the term $\frac{\Delta_{[S_*]}}{\sqrt{\mathbb{E}[\psi^2(\epsilon)]}}$ can be ignored, since its variability is overshadowed by that of $\frac{2}{n}\frac{\boldsymbol{X}_{[,S_*]}{}^T\boldsymbol{\psi}(\boldsymbol{\epsilon})}{\sqrt{\mathbb{E}[\psi^2(\epsilon)]}}$. Under mild conditions, the distribution of $\frac{2}{\sqrt{n}}\frac{\boldsymbol{X}_{[,S_*]}{}^T\boldsymbol{\psi}(\boldsymbol{\epsilon})}{\sqrt{\mathbb{E}[\psi^2(\epsilon)]}}$ is close to a normal distribution by the central limit theorem and thus does not depend on ψ asymptotically. Therefore, under all these assumptions, the efficiency of $\left[\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right]_{[S_*]}$ depends only on the ratio.

In spite of the fact that the main assumption is rather strong, we do not assume that $S_* = S_0$. Consistency in model selection is however a special case, i.e. if $S_* = S_0$, in which case the ratio directly reflects the loss of efficiency.

5.3.3. Ratio for projections onto true span

We now relax some of the assumptions made in the last part, namely we only assume that $\boldsymbol{S}^0_{[S_0]} = sign\left(\hat{\boldsymbol{\beta}}\right)_{[S_0]}$, where once again $\boldsymbol{S}^0 = sign\left(\boldsymbol{\beta}^0\right)$ and $S_0 =$

$S\left(\boldsymbol{\beta}^0\right)$. The above assumption is rather weak, since for instance it is satisfied if the smallest non zero coefficient of $\boldsymbol{\beta}^0$ is bigger in absolute value than a certain threshold. In that case, by dividing the components in S_0 of the scores by $\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}$ we recover:

$$\lambda^* \boldsymbol{S}^0_{[S_0]} = \frac{2}{n}\frac{\boldsymbol{X}^T_{[,S_0]}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}} + \frac{\mathbb{E}\left[\psi'\left(\epsilon\right)\right]}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}}\frac{2}{n}\boldsymbol{X}^T_{[,S_0]}\boldsymbol{X}\left[\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right] + \frac{\Delta_{[S_0]}}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}}.$$

This leads to the following equations:

$$\frac{\mathbb{E}\left[\psi'\left(\epsilon\right)\right]}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}}\frac{2}{n}\boldsymbol{X}^T_{[,S_0]}\boldsymbol{X}\left[\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right] = \lambda^* \boldsymbol{S}^0_{[S_0]} - \frac{\Delta_{[S_0]}}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}} - \frac{2}{n}\frac{\boldsymbol{X}^T_{[,S_0]}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)}{\sqrt{\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]}}.$$

Relying once again on the assumption that $\sqrt{n}\left\|\Delta\right\|_\infty$ is small, the term $\frac{\Delta_{[S_0]}}{\sqrt{\mathbb{E}[\psi^2(\epsilon)]}}$ can be ignored and so the distribution of any projection of $\boldsymbol{X}\left[\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right]$ onto a vector in the span of $\boldsymbol{X}_{[,S_0]}$ only depends on ψ through the ratio asymptotically.

6. Discussion

In this paper, we have given explicit bounds for the estimation and prediction errors for the Lasso with a general convex loss function. We have shown that both of these bounds are a natural extension of the well-known bounds in the classical setting (i.e., with a squared error loss function), with an additional term given by $\mathbb{E}\left[\psi^2\left(\epsilon\right)\right]/\mathbb{E}\left[\psi'\left(\epsilon\right)\right]^2$. Interestingly, this term is exactly the same as the one found by [7] in the low-dimensional setting, underlying the minimax property of the Huber loss function. We have provided theoretical arguments supporting the importance of this ratio in the high-dimensional setting. Our work establishes a clear and explicit link between the bounds for the prediction and estimation errors on the one hand and the choice of the loss function motivated by robustness considerations on the other hand. To the best of our knowledge, such findings have never appeared in this form in the literature.

An interesting direction for future work would be to further assess the sharpness of the bounds that we have obtained. This could notably give an indication of how relevant the ratio from Huber's minimax problem could be in other contexts. It would also be useful to consider the impact of having an intercept in the model, even though we believe that our techniques can be easily adapted to handle that case. Finally, we have excluded from our analysis the possibility of outliers in the design matrix. It is clear that contamination of the covariates is highly plausible in applications, perhaps even more so in the high-dimensional setting. It would thus be of interest to examine the impact of such contaminations on the bounds.

Appendix A: Proofs and technical arguments: Section 3

A.1. Proof of Lemma 3.1

It follows directly from the following inequality

$$\frac{2}{n}\sum_{i=1}^{n}\rho\left(\epsilon_i+\frac{a_i}{\sigma}\right)+\lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1\leq\frac{2}{n}\sum_{i=1}^{n}\rho\left(\epsilon_i\right)+\lambda\left\|\boldsymbol{\beta}^0\right\|_1$$

and a Taylor expansion on $\frac{2}{n}\sum_{i=1}^{n}\rho\left(\epsilon_i+\frac{a_i}{\sigma}\right)$, which implies $\exists t^\lambda\in[0,1]$ such that

$$\frac{2}{n}\sum_{i=1}^{n}\rho\left(\epsilon_i+\frac{a_i}{\sigma}\right)=\frac{2}{n}\sum_{i=1}^{n}\rho\left(\epsilon_i\right)+\frac{2}{n}\sum_{i=1}^{n}\psi\left(\frac{\epsilon_i}{\sigma}\right)\frac{a_i}{\sigma}+\frac{1}{n}\sum_{i=1}^{n}\psi'\left(\epsilon_i+t^\lambda\frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2}.$$

A.2. Proof of Lemma 3.2

Because of $2\lambda_0\leq\lambda$ and the definition of $\mathcal{J}_0$, we have

$$\left|\frac{2}{n}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)^T\frac{\boldsymbol{X}}{\sigma}\left(\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right)\right|\leq\frac{\lambda}{2}\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1.$$

By using Lemma 3.1, this in turn implies that on $\mathcal{J}_0$ we have

$$\frac{1}{n}\sum_{i=1}^{n}\psi'\left(\epsilon_i+t^\lambda\frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2}\leq\frac{\lambda}{2}\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1+\lambda\left\|\boldsymbol{\beta}^0\right\|_1-\lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1\leq\frac{3}{2}\lambda\left\|\boldsymbol{\beta}^0\right\|_1.$$

A.3. Proof of Lemma 3.3

Define $V^{(j)}=\frac{1}{L}\frac{1}{\sqrt{n}}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)^T\boldsymbol{X}^{(j)}=\frac{1}{L}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi\left(\epsilon_i\right)\boldsymbol{X}_i^{(j)}$. Since $\sup_x\left|\psi\left(x\right)\right|\leq L$, we have $\frac{\psi(\epsilon_i)}{L}\boldsymbol{X}_i^{(j)}\in[-\boldsymbol{X}_i^{(j)},\boldsymbol{X}_i^{(j)}]$ and so by Hoeffding's inequality as in [6]

$$\begin{aligned}\mathbb{P}\left[\left|V^{(j)}\right|\geq t\right] &\leq 2\exp\left[-\frac{2nt^2}{\sum_{i=1}^{n}\left[2\boldsymbol{X}_i^{(j)}\right]^2}\right]\\ &= 2\exp\left[-\frac{2nt^2}{2^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)_{jj}}\right]\\ &= 2\exp\left[-\frac{t^2}{2}\right].\end{aligned}$$

We can now bound $\max_{1\leq j\leq p}\left|V^{(j)}\right|$ in probability by using the union bound

$$\mathbb{P}\left[\max_{1\leq j\leq p}\left|V^{(j)}\right|\geq\sqrt{t^2+2\log(p)}\right]\leq p\mathbb{P}\left[\left|V^{(j)}\right|\geq\sqrt{t^2+2\log(p)}\right]$$

$$\begin{aligned} &\leq 2p\exp\left[-\frac{t^2+2\log(p)}{2}\right] \\ &= 2\exp\left[-\frac{t^2}{2}\right]. \end{aligned}$$

The lemma follows directly from $\mathcal{J}_0 := \left\{\max_{1\leq j\leq p}\left|V^{(j)}\right| 2\frac{L}{\sigma\sqrt{n}} \leq \lambda_0\right\}$.

A.4. Proof of Theorem 3.1

Let $U_i = \frac{\psi(\epsilon_i)}{\sqrt{\mathbb{E}[\psi(\epsilon)^2]}}$ and $W_i \sim \mathcal{N}(0,1)$ iid. For fixed j, we define $V^{(j)}$, respectively $Z^{(j)}$, as the linear combinations of U_i, respectively W_i, with $\boldsymbol{X}_i^{(j)}$:

$$\begin{aligned} V^{(j)} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} V_i^{(j)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i \boldsymbol{X}_i^{(j)} \\ Z^{(j)} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i^{(j)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} W_i \boldsymbol{X}_i^{(j)}. \end{aligned}$$

Because of the normalized columns, we have $Z^{(j)} \sim \mathcal{N}(0,1)$, from which we know that $\mathbb{E}\left[\left(Z^{(j)}\right)^{2k}\right] = (2k)!/(2^k(k)!)$. Thus we have the following equation:

$$\mathbb{E}\left[\exp\left[tZ^{(j)}\right]\right] = \sum_{k=0}^{+\infty}\frac{\mathbb{E}\left[\left(Z^{(j)}\right)^{2k}\right]t^{2k}}{(2k)!} = \sum_{k=0}^{+\infty}\frac{t^{2k}}{2^k(k)!} = \exp\left[t^2/2\right].$$

If $\mathbb{E}\left[\left(V^{(j)}\right)^{2k}\right] \leq \mathbb{E}\left[\left(Z^{(j)}\right)^{2k}\right]\ \forall k \geq 0$, then $\mathbb{E}\left[\exp\left[tV^{(j)}\right]\right] \leq \mathbb{E}\left[\exp\left[tZ^{(j)}\right]\right]$, since the uneven moments are all 0 in both cases (this is because ϵ_i has a symmetric distribution and ψ is odd). With the help of enumerative combinatorics we recover

$$\mathbb{E}\left[\left(V^{(j)}\right)^{2k}\right] = n^{-k}\sum_{1\leq l_1<\ldots<l_m\leq n} c_{2k,a_1,\ldots,a_m}\mathbb{E}\left[\left(V_{l_1}^{(j)}\right)^{a_1}\right]\cdot\ldots\cdot\mathbb{E}\left[\left(V_{l_m}^{(j)}\right)^{a_m}\right]$$

where $1 \leq m \leq 2k$, $\sum_{l=1}^{m} a_l = 2k$, $a_l > 0$ and $c_{2k,a_1,\ldots,a_m} = \frac{(2k)!}{a_1!\cdot\ldots\cdot a_m!}$.

Now since we have $\mathbb{E}\left[U_i^{2k}\right] \leq \mathbb{E}\left[W_i^{2k}\right]$, it follows that $\mathbb{E}\left[\left(V_i^{(j)}\right)^{2k}\right] \leq \mathbb{E}\left[\left(Z_i^{(j)}\right)^{2k}\right]$ and thus that $\mathbb{E}\left[\left(V^{(j)}\right)^{2k}\right] \leq \mathbb{E}\left[\left(Z^{(j)}\right)^{2k}\right]$. This in turn implies for $t \geq 0$:

$$\mathbb{E}\left[\exp\left[tV^{(j)}\right]\right] \leq \mathbb{E}\left[\exp\left[tZ^{(j)}\right]\right] = \exp\left[t^2/2\right].$$

We now have the following by Markov's inequality for $t \geq 0$:

$$\mathbb{P}\left[V^{(j)} \geq t\right] \leq \frac{\mathbb{E}\left[\exp\left[tV^{(j)}\right]\right]}{\exp\left[t^2\right]} \leq \frac{\exp\left[t^2/2\right]}{\exp\left[t^2\right]} = \exp\left[-t^2/2\right].$$

Consequently by the union bound we have

$$\mathbb{P}\left[\max_{1\leq j\leq p}\left|V^{(j)}\right|\geq\sqrt{t^2+2\log(p)}\right]\leq 2p\exp\left[-\frac{t^2+2\log(p)}{2}\right]=2\exp\left[-t^2/2\right].$$

The theorem follows directly from the definition of $V^{(j)}$, since we have

$$\mathcal{J}_0:=\left\{\max_{1\leq j\leq p}\left|V^{(j)}\right|2\frac{\sqrt{\mathbb{E}\left[\psi\left(\epsilon\right)^2\right]}}{\sigma\sqrt{n}}\leq\lambda_0\right\}.$$

A.5. Proof of Lemma 3.4

Without loss of generality we set $\sigma=1$. In the case $\psi(x)=x$ the inequality is obviously satisfied. More generally, let $\psi_t(x)=(1-t)\psi(x)+tx$, where ψ respects the conditions of the lemma. If we show that $\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k}\right]\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]^{-k}$ is monotone increasing in $t\in[0,1]$, then by the above remark the lemma follows directly.

Let $g:[0,1]\to\mathbb{R}$ where $g(t)=\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k}\right]\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]^{-k}$. We then have

$$\begin{aligned}\frac{\partial g(t)}{\partial t} &= 2k\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k-1}\left[\epsilon-\psi\left(\epsilon\right)\right]\right]\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]^{-k}\\ &\quad -2k\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k}\right]\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]^{-k-1}\mathbb{E}\left[\psi_t\left(\epsilon\right)\left[\epsilon-\psi\left(\epsilon\right)\right]\right]\\ &= 2k\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]^{-k}\\ &\quad\times\left[\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k-1}\left[\epsilon-\psi\left(\epsilon\right)\right]\right]-\frac{\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k}\right]}{\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]}\mathbb{E}\left[\psi_t\left(\epsilon\right)\left[\epsilon-\psi\left(\epsilon\right)\right]\right]\right]\\ &= \frac{2k}{1-t}\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]^{-k}\\ &\quad\times\left[\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k-1}\left[\epsilon-\psi_t\left(\epsilon\right)\right]\right]-\frac{\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k}\right]}{\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]}\mathbb{E}\left[\psi_t\left(\epsilon\right)\left[\epsilon-\psi_t\left(\epsilon\right)\right]\right]\right].\end{aligned}$$

In order to show that $\frac{\partial g(t)}{\partial t}\geq 0$, we study the expression

$$\mathbb{E}\left[\left[\psi_t\left(\epsilon\right)^{2k-1}-\frac{\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k}\right]}{\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]}\psi_t\left(\epsilon\right)\right]\left[\epsilon-\psi_t\left(\epsilon\right)\right]\right],$$

which is equal to

$$\mathbb{E}\left[\psi_t\left(\epsilon\right)^{2k-2}\psi_t\left(\epsilon\right)^2\left[\frac{\epsilon}{\psi_t\left(\epsilon\right)}-\frac{\mathbb{E}\left[\psi_t\left(\epsilon\right)\epsilon\right]}{\mathbb{E}\left[\psi_t\left(\epsilon\right)^2\right]}\right]\right].$$

For $k = 1$, the above expression is obviously 0. Since $x/\psi(x)$ is monotone increasing in $|x|$ and even, there exists $\tilde{x} > 0$ such that for all $x \in [-\tilde{x}, \tilde{x}]$: $\frac{x}{\psi_t(x)} \leq \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}$, while for all $x \in [-\tilde{x}, \tilde{x}]^c$: $\frac{x}{\psi_t(x)} \geq \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}$.

Let $k \geq 1$, $F_0(\epsilon) = 1_{\epsilon \in [-\tilde{x}, \tilde{x}]}$ and $F_1(\epsilon) = 1 - F_0(\epsilon)$. We then have

$$\mathbb{E}\left[\psi_t(\epsilon)^{2k}\left[\frac{\epsilon}{\psi_t(\epsilon)} - \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}\right]\right] =$$
$$\mathbb{E}\left[F_0(\epsilon)\psi_t(\epsilon)^{2k}\left[\frac{\epsilon}{\psi_t(\epsilon)} - \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}\right]\right] + \mathbb{E}\left[F_1(\epsilon)\psi_t(\epsilon)^{2k}\left[\frac{\epsilon}{\psi_t(\epsilon)} - \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}\right]\right].$$

By definition of $\tilde{x}$ and the fact that $\psi_t(x)^{2k-2}$ is monotone increasing in $|x|$, we have

$$\mathbb{E}\left[F_0(\epsilon)\psi_t(\epsilon)^{2k}\left[\frac{\epsilon}{\psi_t(\epsilon)} - \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}\right]\right] \geq$$
$$\psi_t(\tilde{x})^{2k-2}\,\mathbb{E}\left[F_0(\epsilon)\psi_t(\epsilon)^2\left[\frac{\epsilon}{\psi_t(\epsilon)} - \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}\right]\right],$$

$$\mathbb{E}\left[F_1(\epsilon)\psi_t(\epsilon)^{2k}\left[\frac{\epsilon}{\psi_t(\epsilon)} - \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}\right]\right] \geq$$
$$\psi_t(\tilde{x})^{2k-2}\,\mathbb{E}\left[F_1(\epsilon)\psi_t(\epsilon)^2\left[\frac{\epsilon}{\psi_t(\epsilon)} - \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}\right]\right].$$

By adding the terms, we recover $\mathbb{E}\left[\psi_t(\epsilon)^{2k}\left[\frac{\epsilon}{\psi_t(\epsilon)} - \frac{\mathbb{E}[\psi_t(\epsilon)\epsilon]}{\mathbb{E}\left[\psi_t(\epsilon)^2\right]}\right]\right] \geq 0$. Thus $\frac{\partial g(t)}{\partial t} \geq 0$ and so the lemma follows.

A.6. Proof of Lemma 3.5

Lemma A.1. *Suppose $0 < \delta < 1$, $k \in \mathbb{N}$, $k \geq 2$ and let $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ defined by:*

$$f(a) = (1-\delta)\mathbb{E}\left[\psi(\epsilon)^{2k}\right] + \delta a^{2k} - \frac{(2k)!}{k!}2^{-k}\mathbb{E}\left[(1-\delta)\psi(\epsilon)^2 + \delta a^2\right]^k.$$

Then, if $\frac{(2k)!}{k!}2^{-k}\delta^{k-1} \leq 1$, f is monotone decreasing on $[0, \infty)$. On the other hand, if $\frac{(2k)!}{k!}2^{-k}\delta^{k-1} > 1$, there exists $a^* > 0$ such that f is monotone decreasing on $[0, a^*]$ and monotone increasing on $(a^*, +\infty)$.

Proof. We study the derivative of f:

$$\begin{aligned}\frac{d}{da}f(a) &= 2k\delta a^{2k-1} - \frac{(2k)!}{k!}2^{-k}\mathbb{E}\left[(1-\delta)\psi(\epsilon)^2 + \delta a^2\right]^{k-1} 2k\delta a \\ &= 2k\delta a\left(a^{2k-2} - \frac{(2k)!}{k!}2^{-k}\mathbb{E}\left[(1-\delta)\psi(\epsilon)^2 + \delta a^2\right]^{k-1}\right).\end{aligned}$$

For a fixed $0<\delta<1$, we can define $P(a) = \frac{(2k)!}{k!}2^{-k}\mathbb{E}\left[(1-\delta)\psi(\epsilon)^2 + \delta a^2\right]^{k-1}$, where $P(a) = \sum_{j=1}^{k} c_j a^{2j-2}$ and $c_j > 0$. By inspecting the polynomial coefficients, we recover that $c_k = \frac{(2k)!}{k!}2^{-k}\delta^{k-1}$. Furthermore, we recognize that $\frac{d}{da}f(a) = 2k\delta a\left(a^{2k-2} - P(a)\right)$ and so the sign of $\frac{d}{da}f(a)$ for $a>0$ is equal to the sign of $a^{2k-2} - P(a)$.

Since $P(0)>0$, $a^{2k-2} - P(a)$ must be locally negative near $a=0$. For $a>0$ we have:

$$\begin{aligned}sign\left(a^{2k-2} - P(a)\right) &= sign\left(1 - a^{-2k+2}P(a)\right) = sign\left(1 - \sum_{j=1}^{k} c_j a^{2(j-k)}\right) \\ &= sign\left(1 - c_k - \sum_{j=1}^{k-1} c_j a^{-2(k-j)}\right).\end{aligned}$$

Therefore, if $c_k = \frac{(2k)!}{k!}2^{-k}\delta^{k-1} \geq 1$, then $sign\left(a^{2k-2} - P(a)\right) = -1$ for all $a>0$. On the other hand, if $c_k = \frac{(2k)!}{k!}2^{-k}\delta^{k-1} < 1$, by using the fact that $\sum_{j=1}^{k-1} c_j a^{-2(k-j)}$ is positive, strictly monotone decreasing in $a>0$ and tending to 0 as a tends to $+\infty$, as well as the fact that $1-c_k>0$, there exists a unique $a^*>0$ such that $sign\left(a^{2k-2} - P(a)\right) = -1$ for $0<a<a^*$ and $sign\left(a^{2k-2} - P(a)\right) = 1$ for $a^*<a$. The lemma now follows directly from these observations. □

Lemma A.2. *Suppose $0<\delta<1$, $k\in\mathbb{N}$, $k\geq 2$ and let $F:\mathbb{R}^n_{\geq 0}\to\mathbb{R}$, where $F(a_1,\ldots,a_n)$ is defined by*

$$(1-\delta)\mathbb{E}\left[\psi(\epsilon)^{2k}\right] + \delta\sum_{i=1}^{n} a_i^{2k}p_i - \frac{(2k)!}{k!}2^{-k}\left((1-\delta)\mathbb{E}\left[\psi(\epsilon)^2\right] + \delta\sum_{i=1}^{n} a_i^2 p_i\right)^k,$$

where $p_i>0$ and $\sum_{i=1}^{n} p_i = 1$.

Then maximizing F under the constraint that $\sum_{i=1}^{n} a_i p_i \leq L$ for a given $L>0$ leads to either $(a_1,\ldots,a_n) = (L,\ldots,L)$ or $(a_1,\ldots,a_n) = (0,\ldots,0)$.

Proof. We study the Lagrangian of F:

$$L(a_1,\ldots,a_n,\lambda) = F(a_1,\ldots,a_n) + \lambda\sum_{i=1}^{n} a_i p_i.$$

We now study the partial derivative of L:

$$\frac{\partial L}{\partial a_i} = 2k\delta a_i^{2k-1} p_i - \frac{(2k)!}{k!} 2^{-k} \mathbb{E}\left[(1-\delta)\psi(\epsilon)^2 + \delta \sum_{j=1}^{n} a_j^2 p_j\right]^{k-1} 2k\delta a_i p_i + \lambda p_i$$

$$= p_i \left[2k\delta a_i \left[a_i^{2k-2} - \frac{(2k)!}{k!} 2^{-k} \mathbb{E}\left[(1-\delta)\psi(\epsilon)^2 + \delta \sum_{j=1}^{n} a_j^2 p_j\right]^{k-1}\right] + \lambda\right].$$

Now $\frac{\partial}{\partial a_i} L = 0$ implies that all the a_i must be equal. Therefore, the problem of optimizing $F(a_1, ..., a_n)$ under this constraint is equivalent to optimizing $f(a)$ in Lemma A.1 under the same constraint, where we fix $(a_1, \ldots, a_n) = (a, \ldots, a)$ and the corresponding constraint is $a \leq L$. The lemma now follows directly from Lemma A.1. □

It is well known that any continuous distribution can be approximated by a discrete one, by assigning sufficiently small probability mass to a sufficiently large number of points. The lemma follows directly from this observation along with Lemma A.2, which shows that the worst possible contaminating distribution (the most challenging for the moment condition) is either a point mass at 0 or at the maximizing value of ψ.

A.7. Probability bounds

Let $\hat{\boldsymbol{\Sigma}}_{\psi'} := \frac{1}{n} \sum_{i=1}^{n} \frac{\psi'(\epsilon_i)}{\mathbb{E}[\psi'(\epsilon)]} \boldsymbol{X}_i^T \boldsymbol{X}_i$, which is well defined by Assumption 3.1.

Lemma A.3. *Let* $\left\|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\psi'}\right\|_{+\infty} \leq \tilde{\lambda}$ *and* $\left\|\boldsymbol{\theta}_{S_0^c}\right\|_1 \leq 3 \left\|\boldsymbol{\theta}_{S_0}\right\|_1$. *Suppose the compatibility condition holds for* S_0, *then,*

$$\left|\boldsymbol{\theta}^T \left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\psi'}\right) \boldsymbol{\theta}\right| \leq 16\tilde{\lambda} \left\|\boldsymbol{\theta}_{S_0}\right\|_1^2.$$

Proof. We start by making the following observations:

$$\begin{aligned} \left|\boldsymbol{\theta}^T \left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\psi'}\right) \boldsymbol{\theta}\right| &\leq \left\|\left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\psi'}\right) \boldsymbol{\theta}\right\|_{+\infty} \|\boldsymbol{\theta}\|_1 \\ \left|\left(\left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\psi'}\right) \boldsymbol{\theta}\right)_i\right| &= \left|\sum_{j=1}^{p} \left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\psi'}\right)_{ij} \boldsymbol{\theta}_j\right| \leq \tilde{\lambda} \|\boldsymbol{\theta}\|_1. \end{aligned}$$

Combining both inequalities we have $\left|\boldsymbol{\theta}^T \left(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\psi'}\right) \boldsymbol{\theta}\right| \leq \tilde{\lambda} \|\boldsymbol{\theta}\|_1^2$, while on the other hand

$$\begin{aligned} \|\boldsymbol{\theta}\|_1^2 &= \|\boldsymbol{\theta}_{S_0}\|_1^2 + 2 \|\boldsymbol{\theta}_{S_0}\|_1 \left\|\boldsymbol{\theta}_{S_0^c}\right\|_1 + \left\|\boldsymbol{\theta}_{S_0^c}\right\|_1^2 \\ &\leq \|\boldsymbol{\theta}_{S_0}\|_1^2 + 6 \|\boldsymbol{\theta}_{S_0}\|_1^2 + 9 \|\boldsymbol{\theta}_{S_0}\|_1^2 = 16 \|\boldsymbol{\theta}_{S_0}\|_1^2. \end{aligned}$$

The lemma follows directly from these observations. □

Clearly, Lemma A.3 implies that under the assumptions that the compatibility condition holds, $\left\|\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\psi'}\right\|_{+\infty} \leq \tilde{\lambda}$ and $\left\|\boldsymbol{\theta}_{S_0^c}\right\|_1 \leq 3\left\|\boldsymbol{\theta}_{S_0}\right\|_1$, we have

$$\left\|\boldsymbol{\theta}_{S_0}\right\|_1^2 \leq \frac{1}{1-16\tilde{\lambda}s_0/\phi_0^2}\left\|\boldsymbol{\theta}\right\|_{\hat{\boldsymbol{\Sigma}}_{\psi'}}^2 \frac{s_0}{\phi_0^2} = \left\|\boldsymbol{\theta}\right\|_{\hat{\boldsymbol{\Sigma}}_{\psi'}}^2 \frac{s_0}{\phi_{\tilde{\lambda}}^2},$$

with $\phi_{\tilde{\lambda}}^2 = \phi_0^2\left(1-16\tilde{\lambda}\frac{s_0}{\phi_0^2}\right)$. This means that for $\tilde{\lambda}\frac{s_0}{\phi_0^2}$ small enough we approximately recover the same compatibility constant as that of $\hat{\boldsymbol{\Sigma}}$. By definition, even for fixed $\boldsymbol{X}$, $\hat{\boldsymbol{\Sigma}}_{\psi'}$ is non degenerate with high probability and so we now introduce the set

$$\mathcal{A} := \left\{\left\|\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\psi'}\right\|_{+\infty} \leq \tilde{\lambda}\right\}.$$

The following lemma shows that $\mathcal{A}$ is met with high probability for small $\tilde{\lambda}$, assuming bounded covariates and that Assumptions 3.1 and 3.2 hold. For readability, we set $\Theta = \{\theta_j : j \in \{1,...,N\}\}$.

Lemma A.4. *Let* $\tilde{\lambda} = L_{\psi'}\sqrt{\frac{\tilde{t}^2+4\log(p)}{n}}\sqrt{\max_{1\leq j\leq p}\frac{1}{n}\left\|\boldsymbol{X}^{(j)}\right\|_4^4}$, *then we have* $\mathbb{P}\left[\mathcal{A}\right] \geq 1-2\exp\left[-\tilde{t}^2/2\right]$, *where* $L_{\psi'} := \sup_{x\neq\Theta}\left|\frac{\psi'(x)}{\mathbb{E}[\psi'(\epsilon)]}-1\right|$.

Proof. We have $\left(\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\psi'}\right)_{kl} = \frac{1}{n}\sum_{i=1}^n\left(1-\frac{\psi'(\epsilon_i)}{\mathbb{E}[\psi'(\epsilon)]}\right)\boldsymbol{X}_i^{(k)}\boldsymbol{X}_i^{(l)}$ and so it follows from Hoeffding's inequality as in [6] that for $t \geq 0$:

$$\mathbb{P}\left[\left|\left(\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\psi'}\right)_{kl}\right| \geq \tilde{t}\right] \leq 2\exp\left[-\frac{\tilde{t}^2 n}{2(L_{\psi'})^2\max_j\frac{1}{n}\left\|\boldsymbol{X}^{(j)}\right\|_4^4}\right].$$

Indeed,

$$\sum_{i=1}^n\left(\boldsymbol{X}_i^{(k)}\right)^2\left(\boldsymbol{X}_i^{(l)}\right)^2 \leq \sqrt{\sum_{i=1}^n\left(\boldsymbol{X}_i^{(k)}\right)^4}\sqrt{\sum_{i=1}^n\left(\boldsymbol{X}_i^{(l)}\right)^4} \leq \max_j\left\|\boldsymbol{X}^{(j)}\right\|_4^4,$$

and so $\mathbb{P}\left[\left|\left(\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\psi'}\right)_{kl}\right| \geq \tilde{t}\right] \leq 2\exp\left[-\frac{\tilde{t}^2 n}{2(L_{\psi'})^2\max_j\frac{1}{n}\left\|\boldsymbol{X}^{(j)}\right\|_4^4}\right]$. We can now show by the union bound that

$$\begin{aligned}\mathbb{P}\left[\left\|\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\psi'}\right\|_{+\infty} \geq \tilde{\lambda}\right] &\leq \sum_{k,l\in\{1,...,p\}}\mathbb{P}\left[\left|\left(\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\psi'}\right)_{kl}\right| \geq \tilde{\lambda}\right] \\ &\leq p^2 2\exp\left[-\frac{\tilde{\lambda}^2 n}{2(L_{\psi'})^2\max_j\frac{1}{n}\left\|\boldsymbol{X}^{(j)}\right\|_4^4}\right].\end{aligned}$$

The lemma follows directly from

$$2\exp\left[2\log(p)-\frac{\tilde{\lambda}^2 n}{2(L_{\psi'})^2\max_j\frac{1}{n}\left\|\boldsymbol{X}^{(j)}\right\|_4^4}\right]=2\exp\left[-\tilde{t}^2/2\right]. \qquad \square$$

As one would expect, we recover that in the case of the classical Lasso we can take $\tilde{\lambda}=0$ with probability 1, since in that case $L_{\psi'}=0$.

A.8. Proof of Theorem 3.2

In this subsection, we relax the continuity assumption on ψ', by only imposing Assumptions 3.1 and 3.2 on ψ'.

Let $\psi'_k(x)=\frac{x-(\theta_j-k)}{2k}\left(\psi'(\theta_j+k)-\psi'(\theta_j-k)\right)+\psi'(\theta_j-k)$ if $\exists j\in\{1,...,N\}$ s.t. $|x-\theta_j|<k$, and $\psi'_k(x)=\psi'(x)$ otherwise. For $0<k<\inf_{j_1\neq j_2}|\theta_{j_1}-\theta_{j_2}|$, ψ'_k is well defined, continuous, non negative and bounded by 1. Set ρ_k and ψ_k as the ρ and ψ functions corresponding to ψ'_k. Consequently, ρ_k converges pointwise to ρ and therefore there exists $c_k\geq 0$ such that $\lim_{k\to 0}c_k=0$ satisfying:

$$\frac{1}{n}\sum_{i=1}^n\rho_k\left(\epsilon_i+\frac{a_i}{\sigma}\right)+\lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1\leq\frac{1}{n}\sum_{i=1}^n\rho_k\left(\epsilon_i\right)+\lambda\left\|\boldsymbol{\beta}^0\right\|_1+c_k.$$

Accordingly, just as in Lemma 3.1, there exists $t^\lambda\in[0,1]$ such that:

$$\frac{1}{n}\sum_{i=1}^n\psi'_k\left(\epsilon_i+t^\lambda\frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2}+\lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1\leq-\frac{2}{n}\boldsymbol{\psi}_k\left(\boldsymbol{\epsilon}\right)^T\frac{\boldsymbol{X}}{\sigma}\left(\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right)+\lambda\left\|\boldsymbol{\beta}^0\right\|_1+c_k.$$

By continuity, we have $\inf_{t\in[0,1]}\psi'\left(\epsilon_i+t\frac{a_i}{\sigma}\right)=\lim_{k\to 0}\inf_t\psi'_k\left(\epsilon_i+t\frac{a_i}{\sigma}\right)$, where we abuse the notation somewhat by only taking the infinimum over Θ in the first part and restrict $t\in[0,1]$. Since ψ_k converges pointwise to ψ, we recover:

$$\frac{1}{n}\sum_{i=1}^n\inf_t\psi'\left(\epsilon_i+t\frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2}+\lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1\leq-\frac{2}{n}\boldsymbol{\psi}\left(\boldsymbol{\epsilon}\right)^T\frac{\boldsymbol{X}}{\sigma}\left(\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right)+\lambda\left\|\boldsymbol{\beta}^0\right\|_1. \tag{A.1}$$

Lemma A.5. *Let $2\lambda_0\leq\lambda$. Then on $\mathcal{J}_0$, we have*

$$\frac{2}{n}\sum_{i=1}^n\inf_t\psi'\left(\epsilon_i+t\frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2}+\lambda\left\|\hat{\boldsymbol{\beta}}_{S_0^c}\right\|_1\leq 3\lambda\left\|\hat{\boldsymbol{\beta}}_{S_0}-\boldsymbol{\beta}^0_{S_0}\right\|_1.$$

Proof. The proof is the same as in [1] (Lemma 6.3., proof on page 105). We do provide it again for a self-sufficient reading. By using Inequality (A.1) on $\mathcal{J}_0$ we have

$$\frac{2}{n}\sum_{i=1}^n\inf_t\psi'\left(\epsilon_i+t\frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2}+\lambda\left\|\hat{\boldsymbol{\beta}}_{S_0^c}\right\|_1\leq 2\lambda\left\|\boldsymbol{\beta}^0\right\|_1-\lambda\left\|\hat{\boldsymbol{\beta}}_{S_0^c}\right\|_1-2\lambda\left\|\hat{\boldsymbol{\beta}}_{S_0}\right\|_1$$
$$+\lambda\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1,$$

where we used $\left\|\hat{\boldsymbol{\beta}}\right\|_1 = \left\|\hat{\boldsymbol{\beta}}_{S_0}\right\|_1 + \left\|\hat{\boldsymbol{\beta}}_{S_0^c}\right\|_1$. On the other hand, we also have $\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 = \left\|\boldsymbol{\beta}^0_{S_0} - \hat{\boldsymbol{\beta}}_{S_0}\right\|_1 + \left\|\hat{\boldsymbol{\beta}}_{S_0^c}\right\|_1$ and $\left\|\boldsymbol{\beta}^0\right\|_1 = \left\|\boldsymbol{\beta}^0_{S_0}\right\|_1$. This together implies that the right side of the above inequality is equal to $\lambda\left\|\boldsymbol{\beta}^0_{S_0} - \hat{\boldsymbol{\beta}}_{S_0}\right\|_1 + 2\lambda\left(\left\|\boldsymbol{\beta}^0_{S_0}\right\|_1 - \left\|\hat{\boldsymbol{\beta}}_{S_0}\right\|_1\right)$. The lemma follows directly from the triangle inequality. □

We now turn our attention to producing a joint estimation and prediction error bound on the event $\mathcal{J}_0 \cap \mathcal{A}$.

Lemma A.6. *Suppose the compatibility condition holds for S_0. Then on $\mathcal{J}_0 \cap \mathcal{A}$, we have for $2\lambda_0 \leq \lambda$:*

$$\frac{2}{n}\sum_{i=1}^{n} \inf_t \psi'\left(\epsilon_i + t\frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2} - \frac{1}{n}\sum_{i=1}^{n}\psi'(\epsilon_i)\frac{a_i^2}{\sigma^2} + \lambda\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1 \leq 4\lambda^2 \frac{\sigma^2}{\mathbb{E}\left[\psi'(\epsilon)\right]}\frac{s_0}{\phi^2_{\tilde{\lambda}}}.$$

Proof. The proof is basically the same as the one from [1] (Theorem 6.1, proof on page 107). We provide it again for a self-sufficient reading. With the help of Lemma A.5 and Lemma A.3, we have that $\frac{2}{n}\sum_{i=1}^{n}\inf_t \psi'\left(\epsilon_i + t\frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2} + \lambda\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1$ satisfies the following:

$$\begin{aligned}
& \frac{2}{n}\sum_{i=1}^{n}\inf_t \psi'\left(\epsilon_i + t\frac{a_i}{\sigma}\right)\frac{a_i^2}{\sigma^2} + \lambda\left\|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^0_{S_0}\right\|_1 + \lambda\left\|\hat{\boldsymbol{\beta}}_{S_0^c}\right\|_1 \\
\leq\ & 4\lambda\left\|\hat{\boldsymbol{\beta}}_{S_0} - \boldsymbol{\beta}^0_{S_0}\right\|_1 \\
\leq\ & 4\lambda\sqrt{\frac{s_0}{\phi^2_{\tilde{\lambda}}}\frac{1}{\mathbb{E}\left[\psi'(\epsilon)\right]}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)^T\frac{\sum_{i=1}^{n}\psi'(\epsilon_i)\boldsymbol{X}_i\boldsymbol{X}_i^T}{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)} \\
=\ & 4\sqrt{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)^T\frac{\sum_{i=1}^{n}\psi'(\epsilon_i)\boldsymbol{X}_i\boldsymbol{X}_i^T}{\sigma^2 n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right)\lambda^2\frac{\sigma^2}{\mathbb{E}\left[\psi'(\epsilon)\right]}\frac{s_0}{\phi^2_{\tilde{\lambda}}}} \\
\leq\ & \frac{1}{n}\sum_{i=1}^{n}\psi'(\epsilon_i)\frac{a_i^2}{\sigma^2} + 4\lambda^2\frac{\sigma^2}{\mathbb{E}\left[\psi'(\epsilon)\right]}\frac{s_0}{\phi^2_{\tilde{\lambda}}},
\end{aligned}$$

where we use the inequality $4\sqrt{uv} \leq u + 4v$, for $u, v \geq 0$. □

It is once again interesting to point out that we recover the classical bounds, as given in Equation (2.1), for $\rho(x) = \frac{1}{2}x^2$, since in that case, for $\tilde{\lambda} = 0$, $\mathbb{P}[\mathcal{A}] = 1$ and $\psi' = 1$.

In the case where $N \geq 1$ in Assumption 3.1, we need to introduce some new notation to deal with the points of discontinuity of ψ'. For $\alpha \geq 0$, we define $\psi'_\alpha(\epsilon_i) = -\psi'(\epsilon_i)$ if $\inf_{1\leq j\leq N}\{|\epsilon_i - \theta_j|\} \leq \alpha$ and $\psi'_\alpha(\epsilon_i) = \psi'(\epsilon_i)$ otherwise. Clearly, it follows from Assumption 3.2 that $\lim_{\alpha\to 0}\mathbb{E}\left[\psi'_\alpha(\epsilon)\right] = \mathbb{E}\left[\psi'(\epsilon)\right]$. Let $\hat{\boldsymbol{\Sigma}}_{\psi'_\alpha} := \frac{1}{n}\sum_{i=1}^{n}\frac{\psi'_\alpha(\epsilon_i)}{\mathbb{E}[\psi'_\alpha(\epsilon)]}\boldsymbol{X}_i^T\boldsymbol{X}_i$, and so for $\alpha = 0$, we recover the definition

of $\hat{\mathbf{\Sigma}}_{\psi'}$. On the other hand, in the case where $N = 0$, we set $\hat{\mathbf{\Sigma}}_{\psi'_\alpha} = \hat{\mathbf{\Sigma}}_{\psi'}$. Furthermore we set

$$\mathcal{I}_\alpha := \left\{ \left\| \hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}_{\psi'_\alpha} \right\|_{+\infty} \leq \tilde{\lambda} \right\}.$$

We point out that $\mathbb{P}[\mathcal{I}_\alpha]$ can be studied in exactly the same way as $\mathbb{P}[\mathcal{A}]$.

Finally let $h : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, where

$$h(z) = \sup_{j \in \{1,\dots,N\}} \sup_{(x,x+y) \in I_j^2, |y| \leq z} 2 \left| \psi'(x+y) - \psi'(x) \right|.$$

From Assumption 3.1, it follows directly that this function is well defined, continuous, non negative, bounded and monotone increasing in z with $h(0) = 0$.

Lemma A.7. *Suppose the compatibility condition holds for S_0 and $\frac{\sup_i |a_i|}{\sigma} \leq \alpha$. Then on $\mathcal{J}_0 \cap \mathcal{A} \cap \mathcal{I}_\alpha$, we have for $2\lambda_0 \leq \lambda$:*

$$\frac{\mathbb{E}[\psi'_\alpha(\epsilon)] - h(\alpha)}{\sigma^2} \left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right\|_{\hat{\Sigma}}^2 + (1 - \gamma_1)\lambda \left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right\|_1 \leq 4\lambda^2 \frac{\sigma^2}{\mathbb{E}[\psi'(\epsilon)]} \frac{s_0}{\phi_{\tilde{\lambda}}^2},$$

where $\gamma_1 = 16 \frac{\tilde{\lambda}}{\lambda} \frac{\mathbb{E}[\psi'_\alpha(\epsilon)]}{\sigma} \frac{\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\|_1}{\sigma}$.

Proof. We begin by defining $Z_i^\alpha = 1$ if $\inf_{1 \leq j \leq N} \{|\epsilon_i - \theta_j|\} \leq \alpha$ and $Z_i^\alpha = 0$ otherwise. Thus for $\frac{\sup_i |a_i|}{\sigma} \leq \alpha$, we have by definition that $Z_i^\alpha = 0$ implies that if $\epsilon_i \in I_j$, then $\epsilon_i + \frac{a_i}{\sigma} \in I_j$.

This and the definition of h implies that we have

$$\frac{2}{n} \sum_{i=1}^n \left[\inf_t \psi' \left(\epsilon_i + t \frac{a_i}{\sigma} \right) - \psi'(\epsilon_i) \right] \frac{a_i^2}{\sigma^2} \geq - \frac{2}{n} \sum_{i=1}^n \left[\psi'(\epsilon_i) Z_i^\alpha + \frac{1 - Z_i^\alpha}{2} h\left(\frac{|a_i|}{\sigma} \right) \right] \frac{a_i^2}{\sigma^2}.$$

This in turn implies that $\frac{1}{n} \sum_{i=1}^n \left[2 \inf_t \psi' \left(\epsilon_i + t \frac{a_i}{\sigma} \right) - \psi'(\epsilon_i) \right] \frac{a_i^2}{\sigma^2}$ is bounded from below by

$$\frac{1}{n} \sum_{i=1}^n \psi'(\epsilon_i)(1 - 2Z_i^\alpha) \frac{a_i^2}{\sigma^2} - \frac{1}{\sigma^2} h\left(\frac{\sup_j |a_j|}{\sigma} \right) \left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right\|_{\hat{\Sigma}}^2.$$

In turn, by using Lemma A.3 and by definition of ψ'_α, we can bound the first of these two terms from below by

$$\frac{\mathbb{E}[\psi'_\alpha(\epsilon)]}{\sigma^2} \left[\left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right\|_{\hat{\Sigma}}^2 - 16\tilde{\lambda} \left\| \boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}} \right\|_1^2 \right].$$

Now since $h\left(\frac{\sup_j |a_j|}{\sigma} \right) \leq h(\alpha)$ by the assumption $\frac{\sup_j |a_j|}{\sigma} \leq \alpha$, the lemma follows directly from combining these inequalities with Lemma A.6. ∎

Finally, we set the values $0 < \alpha^0$ and $0 < \lambda^0$. Firstly, let $0 < \alpha^0$ such that for all $\alpha \leq \alpha^0 : \mathbb{E}\left[\psi'_\alpha(\epsilon)\right] \geq \frac{3}{4}\mathbb{E}\left[\psi'(\epsilon)\right]$. On the other hand, because of the monotonicity of $\mathbb{E}\left[\psi'_\alpha(\epsilon)\right]$ with respect to α, we have $\mathbb{E}\left[\psi'_\alpha(\epsilon)\right] \leq \mathbb{E}\left[\psi'(\epsilon)\right]$. Furthermore let $\alpha^0 > 0$ small enough such that $h\left(\alpha^0\right) \leq \frac{1}{4}\mathbb{E}\left[\psi'(\epsilon)\right]$.

Lemma A.8. *Suppose the compatibility condition holds for* S_0, $\frac{10\lambda K\sigma}{\mathbb{E}[\psi'(\epsilon)]}\frac{s_0}{\phi^2_{\tilde{\lambda}}} \leq \alpha \leq \alpha^0$, $640\tilde{\lambda}\frac{s_0}{\phi^2_{\tilde{\lambda}}} \leq \frac{3}{5}$, *and* $\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq 10\lambda\frac{\sigma^2}{\mathbb{E}[\psi'(\epsilon)]}\frac{s_0}{\phi^2_{\tilde{\lambda}}}$. *Then on* $\mathcal{J}_0 \cap \mathcal{A} \cap \mathcal{I}_\alpha$, *for* $2\lambda_0 \leq \lambda$,

$$\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq 5\lambda\frac{\sigma^2}{\mathbb{E}\left[\psi'(\epsilon)\right]}\frac{s_0}{\phi^2_{\tilde{\lambda}}}.$$

Proof. Under the conditions we recover

$$\begin{aligned}
\mathbb{E}\left[\psi'_\alpha(\epsilon)\right] - h(\alpha) &\geq \frac{3}{4}\mathbb{E}\left[\psi'(\epsilon)\right] - h\left(\alpha^0\right) \geq \frac{1}{2}\mathbb{E}\left[\psi'(\epsilon)\right] > 0, \\
1 - 16\frac{\tilde{\lambda}}{\lambda}\frac{\mathbb{E}\left[\psi'_\alpha(\epsilon)\right]}{\sigma^2}\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 &\geq 1 - 160\tilde{\lambda}\frac{\mathbb{E}\left[\psi'_\alpha(\epsilon)\right]}{\mathbb{E}\left[\psi'(\epsilon)\right]}\frac{s_0}{\phi^2_{\tilde{\lambda}}} \geq 1 - \frac{640}{3}\tilde{\lambda}\frac{s_0}{\phi^2_{\tilde{\lambda}}} \geq \frac{4}{5}.
\end{aligned}$$

Therefore the lemma follows from Lemma A.7. □

We now combine these results to prove Theorem 3.2. Here the main idea is to use the margin condition on convex loss functions as in [1]. Let $\boldsymbol{\beta}^t = (1-t)\boldsymbol{\beta}^0 + t\hat{\boldsymbol{\beta}}$ for $t \in [0,1]$ and $a_i^t = \left(\boldsymbol{X}\left(\boldsymbol{\beta}^0 - \boldsymbol{\beta}^t\right)\right)_i$. By convexity, $\frac{2}{n}\sum_{i=1}^n \rho\left(\epsilon_i + \frac{a_i^t}{\sigma}\right) + \lambda\left\|\boldsymbol{\beta}^t\right\|_1$ can be bounded by

$$(1-t)\left[\frac{2}{n}\sum_{i=1}^n \rho(\epsilon_i) + \lambda\left\|\boldsymbol{\beta}^0\right\|_1\right] + t\left[\frac{2}{n}\sum_{i=1}^n \rho\left(\epsilon_i + \frac{a_i}{\sigma}\right) + \lambda\left\|\hat{\boldsymbol{\beta}}\right\|_1\right],$$

which itself is bounded by $\frac{2}{n}\sum_{i=1}^n \rho(\epsilon_i) + \lambda\left\|\boldsymbol{\beta}^0\right\|_1$. As a consequence,

$$\frac{1}{n}\sum_{i=1}^n \inf_{\tilde{t}\in[0,1]} \psi'\left(\epsilon + \tilde{t}\frac{a_i^t}{\sigma}\right)\frac{(a_i^t)^2}{\sigma^2} + \lambda\left\|\boldsymbol{\beta}^t\right\|_1 \leq -\frac{2}{n}\boldsymbol{\psi}(\boldsymbol{\epsilon})^T\frac{\boldsymbol{X}}{\sigma}\left(\boldsymbol{\beta}^0 - \boldsymbol{\beta}^t\right) + \lambda\left\|\boldsymbol{\beta}^0\right\|_1.$$

Therefore, assuming $\left\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^t\right\|_1 \leq 10\lambda\frac{\sigma^2}{\mathbb{E}[\psi'(\epsilon)]}\frac{s_0}{\phi^2_{\tilde{\lambda}}}$, $10\lambda L\frac{\sigma^2}{\mathbb{E}[\psi'(\epsilon)]}\frac{s_0}{\phi^2_{\tilde{\lambda}}} \leq \alpha \leq \alpha^0$, $640\tilde{\lambda}\frac{s_0}{\phi^2_{\tilde{\lambda}}} \leq \frac{3}{5}$ and that the compatibility condition holds for S_0, we have on $\mathcal{J}_0 \cap \mathcal{A} \cap \mathcal{I}_\alpha$ for $2\lambda_0 \leq \lambda \leq \lambda^0 : \left\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^t\right\|_1 \leq 5\lambda\frac{\sigma^2}{\mathbb{E}[\psi'(\epsilon)]}\frac{s_0}{\phi^2_{\tilde{\lambda}}}$. This is because Lemma A.8 follows from Inequality (A.1) and the compatibility condition.

For $H^* = 10\lambda\frac{\sigma^2}{\mathbb{E}[\psi'(\epsilon)]}\frac{s_0}{\phi^2_{\tilde{\lambda}}}$ and $t = \frac{H^*}{H^* + \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1}$, it follows that $\left\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^t\right\|_1 \leq H^*$. So by the above observation it follows that on $\mathcal{J}_0 \cap \mathcal{A} \cap \mathcal{I}_\alpha$ we have $\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^0\right\|_1 \leq 5\lambda\frac{\sigma^2}{\mathbb{E}[\psi'(\epsilon)]}\frac{s_0}{\phi^2_{\tilde{\lambda}}} = \frac{H^*}{2}$. On the other hand we have

$$\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 = \frac{1}{t}\left\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^t\right\|_1 \leq \frac{H^*}{2t} = \frac{1}{2}\left[H^* + \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1\right].$$

This in turn implies that $\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq H^*$ and so by Lemma A.8 we have $\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq \frac{H^*}{2} = 5\lambda \frac{\sigma^2}{\mathbb{E}[\psi'(\epsilon)]} \frac{s_0}{\phi^2_{\tilde{\lambda}}}$ and $\sup_i |a_i| \leq L \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq 5L\lambda \frac{\sigma^2}{\mathbb{E}[\psi'(\epsilon)]} \frac{s_0}{\phi^2_{\tilde{\lambda}}}$.

We can plug this into the inequality in Lemma A.7 and it then follows for $2\lambda_0 \leq \lambda$:

$$\frac{\mathbb{E}\left[\psi'_\alpha(\epsilon)\right] - h(\alpha)}{\sigma^2} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|^2_{\hat{\Sigma}} + (1-\gamma_1)\lambda \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq 4\lambda^2 \frac{\sigma^2}{\mathbb{E}\left[\psi'(\epsilon)\right]} \frac{s_0}{\phi^2_{\tilde{\lambda}}},$$

where $\gamma_1 = 16\frac{\tilde{\lambda}}{\lambda}\frac{\mathbb{E}\left[\psi'_\alpha(\epsilon)\right]}{\sigma^2}\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1$. The theorem now follows directly from:

$$\gamma_1 = 16\frac{\tilde{\lambda}}{\lambda}\frac{\mathbb{E}\left[\psi'_\alpha(\epsilon)\right]}{\sigma^2}\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq 16\frac{\tilde{\lambda}}{\lambda}\frac{\mathbb{E}\left[\psi'_\alpha(\epsilon)\right]}{\sigma^2} 5\lambda \frac{\sigma^2}{\mathbb{E}\left[\psi'(\epsilon)\right]}\frac{s_0}{\phi^2_{\tilde{\lambda}}} \leq 80\tilde{\lambda}\frac{s_0}{\phi^2_{\tilde{\lambda}}}.$$

Appendix B: Proofs and technical arguments: Section 4

B.1. Proof of Lemma 4.1

Let $Q(\boldsymbol{\beta}, \sigma) = \frac{2}{n}\sum_{i=1}^n \left(\rho\left(\frac{Y_i - (\boldsymbol{X}\boldsymbol{\beta})_i}{\sigma}\right) + a\right)\sigma + \lambda_* \|\boldsymbol{\beta}\|_1$. From the definition of $\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_a\right)$ we have $Q\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_a\right) \leq Q\left(\boldsymbol{\beta}^0, \sigma\right)$. By a Taylor expansion there exists $t^{\lambda_*} \in [0,1]$ such that $\frac{2}{n}\sum_{i=1}^n \left(\rho\left(\frac{Y_i - (\boldsymbol{X}\hat{\boldsymbol{\beta}})_i}{\hat{\sigma}_a}\right) + a\right)\hat{\sigma}_a$ is equal to

$$\frac{2}{n}\sum_{i=1}^n \left(\rho(\epsilon_i) + a\right)\sigma + \frac{2}{n}\sum_{i=1}^n \left(a - \chi_0(\epsilon_i)\right)(\hat{\sigma}_a - \sigma) + \frac{2}{n}\sum_{i=1}^n \psi(\epsilon_i)\boldsymbol{X}_i^T\left(\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right) + \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0, \hat{\sigma}_a - \sigma\right\|_{\boldsymbol{\Gamma}(t^{\lambda_*})}.$$

Here we point out that, since the function is convex, $\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0, \hat{\sigma}_a - \sigma\right\|_{\tilde{\boldsymbol{\Gamma}}} \geq 0$.

The lemma follows directly from combining both these equations.

B.2. Technical arguments on scale error bound

We begin with a technical lemma.

Lemma B.1. *Let $\epsilon, a, \sigma, \gamma \in \mathbb{R}$ such that $f : [0,1] \to \mathbb{R}$ with $f(t) = \frac{\sigma\epsilon + ta}{\sigma_a + t\gamma}$ and $h : [0,1] \to \mathbb{R}$ with $h(t) = \psi'(f(t))$ are well defined. Suppose h is differentiable, then $\exists \tilde{t} \in [0,1]$ s.t.*

$$h(1) - h(0) = \psi''\left(f(\tilde{t})\right)\frac{a}{\sigma_a + \tilde{t}\gamma} - \psi''\left(f(\tilde{t})\right)f(\tilde{t})\frac{\gamma}{\sigma_a + \tilde{t}\gamma}.$$

Proof. This is a direct consequence of the mean value theorem. □

Here we provide an extremely useful result to bound $|\hat{\sigma}_a - \sigma_a|$ locally, where we assume ψ' to be well defined and continuous.

Lemma B.2. *There exists* $t^{\lambda_*} \in [0,1]$ *such that*

$$\hat{\sigma}_a - \sigma_a = \frac{\tilde{\sigma}_a \frac{1}{n}\sum_{i=1}^n \left(\chi_0\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) - a\right) + \frac{1}{n}\sum_{i=1}^n \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}a_i}{\frac{1}{n}\sum_{i=1}^n \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma^2}{\tilde{\sigma}_a^2}\tilde{\epsilon}_i^2},$$

with $\tilde{\sigma}_a = \sigma_a + t^{\lambda_*}(\hat{\sigma}_a - \sigma_a)$ *and* $\tilde{\epsilon}_i = \epsilon_i + t^{\lambda_*}\frac{a_i}{\sigma}$.

Proof. By a Taylor expansion on $t \mapsto \frac{1}{n}\sum_{i=1}^n \chi_0\left(\frac{\sigma\epsilon_i + t a_i}{\sigma_a + t(\hat{\sigma}_a - \sigma_a)}\right)$, there exists $t^{\lambda_*} \in [0,1]$ such that $\frac{1}{n}\sum_{i=1}^n \chi_0\left(\frac{\sigma\epsilon_i + a_i}{\tilde{\sigma}_a}\right)$ equals

$$\frac{1}{n}\sum_{i=1}^n \chi_0\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) + \frac{1}{n}\sum_{i=1}^n \chi_0'\left(\frac{\sigma\epsilon_i + t^{\lambda_*}a_i}{\sigma_a + t^{\lambda_*}(\hat{\sigma}_a - \sigma_a)}\right)\frac{a_i}{\sigma_a + t^{\lambda_*}(\hat{\sigma}_a - \sigma_a)}$$
$$- \frac{1}{n}\sum_{i=1}^n \chi_0'\left(\frac{\sigma\epsilon_i + t^{\lambda_*}a_i}{\sigma_a + t^{\lambda_*}(\hat{\sigma}_a - \sigma_a)}\right)\frac{\sigma\epsilon_i + t^{\lambda_*}a_i}{(\sigma_a + t^{\lambda_*}(\hat{\sigma}_a - \sigma_a))^2}(\hat{\sigma}_a - \sigma_a).$$

The lemma follows from the fact that $\frac{1}{n}\sum_{i=1}^n \chi_0\left(\frac{\epsilon_i + a_i}{\tilde{\sigma}_a}\right) = a$. □

We now relax the assumptions on ψ', by only imposing Assumptions 4.1 and 4.2. By using the same technique as in Subsection A.8, we recover that $|\hat{\sigma}_a - \sigma_a|$ is bounded from above by

$$\frac{\sup_t \tilde{\sigma}_a \left|\frac{1}{n}\sum_{i=1}^n \left(\chi_0\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) - a\right)\right| + \frac{1}{n}\sum_{i=1}^n \sup_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}a_i}{\frac{1}{n}\sum_{i=1}^n \inf_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{(\sigma\epsilon_i + t a_i)^2}{(\sigma_a + t(\hat{\sigma}_a - \sigma_a))^2}}, \tag{B.1}$$

where $\sigma\tilde{\epsilon}_i = \sigma\epsilon_i + t a_i$, $\tilde{\sigma}_a = \sigma_a + t(\hat{\sigma}_a - \sigma_a)$, and the supremum and infimum are taken over $[0,1]$, under the constraint that $\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}$ is in Θ^c so that everything is well defined.

To recover a local error bound we study the three sums in the upper bound in Equation (B.1) separately.

For the first term, we have $\sup_t \tilde{\sigma}_a \left|\frac{1}{n}\sum_{i=1}^n \left(\chi_0\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) - a\right)\right| \leq \sup_t \tilde{\sigma}_a \sigma_a \lambda_1$ on the set $\mathcal{J}_{1;a}$ by definition.

The second term is $\left|\frac{1}{n}\sum_{i=1}^n \sup_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}a_i\right|$. If $1 \leq N$, let $\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) = 0$ if $\inf_{1\leq j\leq N}\left\{\left|\frac{\sigma\epsilon_i}{\sigma_a} - \theta_j\right| - \frac{\delta_*}{1-\delta_*}\left|\frac{\sigma\epsilon_i}{\sigma_a}\right|\right\} \leq \frac{\alpha_*}{1-\delta_*}$ and $\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) = \psi'\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)$ otherwise. On the other hand, if $N = 0$, we set $\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) = \psi'\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)$. Furthermore, we define

$$\mathcal{J}_{2;a} = \left\{\max_{1\leq j\leq p}\frac{1}{n}\left|\sum_{i=1}^n \psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)\frac{\sigma\epsilon_i}{\sigma_a}\frac{\boldsymbol{X}_i^{(j)}}{\sigma_a}\right| \leq \lambda_2\right\}.$$

In the case where $N \geq 1$, we define $D_1 = 2\theta_N + 1$, whereas for $N = 0$ we define $D_1 = 0$. Moreover, let $D_2 = \sup_{x \notin \Theta} \psi''(x)$, $D_3 = \sup_{x \notin \Theta} \psi''(x)x$ and $D_4 = \sup_{x \notin \Theta} \psi''(x)x^2$. By Assumption 4.2, we have that for all $i \in \{1, ..., 4\}$: $D_i < +\infty$.

Lemma B.3. *Under the assumptions that* $\frac{|\hat{\sigma}_a - \sigma_a|}{\sigma_a} \leq \delta_* \leq \frac{1}{4}$ *and* $\frac{\sup_i |a_i|}{\sigma_a} \leq \alpha_* \leq \frac{1}{4}$, *we have on* $\mathcal{J}_{2:a}$, $\frac{1}{n}\sum_{i=1}^n \sup_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}a_i$ *is bounded from above by*

$$\frac{\lambda_2\sigma_a}{1-\delta_*}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1 + \frac{D_2}{(1-\delta_*)^2}\frac{1}{n}\sum_{i=1}^n \frac{|a_i|^3}{\sigma_a^2} + \frac{D_1}{n}\sum_{i \in J_{\alpha_*,\delta_*}} |a_i| +$$
$$\frac{|\hat{\sigma}_a - \sigma_a|}{\sigma_a}\frac{D_4}{1-\delta_*}\frac{1}{n}\sum_{i=1}^n |a_i| + \left[1 + \frac{D_3}{1-\delta_*} + \frac{D_3\delta_*}{(1-\delta_*)^2}\right]\frac{1}{n}\sum_{i=1}^n \frac{a_i^2}{\sigma_a},$$

where $J_{\alpha_*,\delta_*} = \left\{i \in \{1, ..., n\} : \inf_{1 \leq j \leq N}\left\{\left|\frac{\sigma\epsilon_i}{\sigma_a} - \theta_j\right| - \frac{\delta_*}{1-\delta_*}\left|\frac{\sigma\epsilon_i}{\sigma_a}\right| \leq \frac{\alpha_*}{1-\delta_*}\right\}\right\}$.

Proof. Let $i \in J_{\alpha_*,\delta_*}$, then by definition $\exists j \in \{1, ..., N\}$ with $\left|\frac{\sigma\epsilon_i}{\sigma_a} - \theta_j\right| - \frac{\delta_*}{1-\delta_*}\left|\frac{\sigma\epsilon_i}{\sigma_a}\right| \leq \frac{\alpha_*}{1-\delta_*}$. By using $\delta_*, \alpha_* \leq \frac{1}{4}$, we have $\left|\frac{\sigma\epsilon_i}{\sigma_a} - \theta_j\right| - \frac{1}{3}\left|\frac{\sigma\epsilon_i}{\sigma_a}\right| \leq \frac{1}{3}$ and so $\left|\frac{\sigma\epsilon_i}{\sigma_a}\right| \leq \frac{3}{2}\theta_N + \frac{1}{2}$. This implies that $\sup_t\left|\frac{\sigma\epsilon_i + ta_i}{\sigma_a + t(\hat{\sigma}_a - \sigma_a)}\right| \leq \frac{1}{1-\delta_*}\left[\frac{|\sigma\epsilon_i|}{\sigma_a} + \frac{|a_i|}{\sigma_a}\right] \leq \frac{4}{3}\left[\frac{3}{2}\theta_N + \frac{1}{2}\right] + \frac{4}{3}\frac{1}{4} = 2\theta_N + 1 = D_1$. Therefore, by using $\sup_{x \notin \Theta}\psi'(x) \leq 1$, we have $\sup_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}a_i \leq D_1|a_i|$.

Let $i \in J^c_{\alpha_*,\delta_*}$, then by the choices of δ^* and α^*, $t \mapsto \psi''\left(\frac{\sigma\epsilon_i + ta_i}{\sigma_a + t(\hat{\sigma}_a - \sigma_a)}\right)$ is well defined and continuous over $[0, 1]$. Consequently, by Lemma B.1, for any $t \in [0, 1]$, there exists $t^* \in [0, 1]$ such that $\psi'\left(\frac{\sigma\epsilon_i + ta_i}{\sigma_a + t(\hat{\sigma}_a - \sigma_a)}\right)\frac{\sigma\epsilon_i + ta_i}{\sigma_a + t(\hat{\sigma}_a - \sigma_a)}a_i$ equals

$$\begin{aligned}
&\psi'\left(\frac{\sigma\epsilon_i + ta_i}{\sigma_a + t(\hat{\sigma}_a - \sigma_a)}\right)\frac{ta_i^2}{\sigma_a + t(\hat{\sigma}_a - \sigma_a)} \\
&\qquad + \psi'\left(\frac{\sigma\epsilon_i + ta_i}{\sigma_a + t(\hat{\sigma}_a - \sigma_a)}\right)\frac{\sigma\epsilon_i}{\sigma_a + t(\hat{\sigma}_a - \sigma_a)}a_i \\
= \;& \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{ta_i^2}{\tilde{\sigma}_a} + \psi'\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)\frac{\sigma\epsilon_i}{\tilde{\sigma}_a}a_i \\
&\qquad + \psi''\left(\frac{\sigma\tilde{\tilde{\epsilon}}_i}{\tilde{\tilde{\sigma}}_a}\right)\left[\frac{ta_i}{\tilde{\tilde{\sigma}}_a} - \frac{\sigma\tilde{\tilde{\epsilon}}_i t(\hat{\sigma}_a - \sigma_a)}{(\tilde{\tilde{\sigma}}_a)^2}\right]\frac{\sigma\tilde{\tilde{\epsilon}}_i - tt^*a_i}{\tilde{\sigma}_a}a_i \\
= \;& \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{ta_i^2}{\tilde{\sigma}_a} + \psi'\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)\frac{\sigma\epsilon_i}{\tilde{\sigma}_a}a_i - \psi''\left(\frac{\sigma\tilde{\tilde{\epsilon}}_i}{\tilde{\tilde{\sigma}}_a}\right)\frac{t^2t^*a_i^3}{\tilde{\tilde{\sigma}}_a(\tilde{\sigma}_a)^2} \\
&+ \psi''\left(\frac{\sigma\tilde{\tilde{\epsilon}}_i}{\tilde{\tilde{\sigma}}_a}\right)\frac{\sigma\tilde{\tilde{\epsilon}}_i}{\tilde{\tilde{\sigma}}_a}\left[\frac{ta_i^2}{\tilde{\sigma}_a} + \frac{t^2t^*(\hat{\sigma}_a - \sigma_a)a_i^2}{\tilde{\sigma}_a\tilde{\tilde{\sigma}}_a}\right] - \psi''\left(\frac{\sigma\tilde{\tilde{\epsilon}}_i}{\tilde{\tilde{\sigma}}_a}\right)\frac{(\sigma\tilde{\tilde{\epsilon}}_i)^2}{(\tilde{\tilde{\sigma}}_a)^2}\frac{t(\hat{\sigma}_a - \sigma_a)a_i}{\tilde{\sigma}_a},
\end{aligned}$$

where $\sigma\tilde{\epsilon}_i = \sigma\epsilon_i + ta_i$, $\sigma\tilde{\tilde{\epsilon}}_i = \sigma\epsilon_i + tt^*a_i$, $\tilde{\sigma}_a = \sigma_a + t(\hat{\sigma}_a - \sigma_a)$ and $\tilde{\tilde{\sigma}}_a = \sigma_a + tt^*(\hat{\sigma}_a - \sigma_a)$. Taking the sum over all $i \in J^c_{\alpha_*,\delta_*}$ divided by n, the first

term is bounded from above by $\frac{1}{n}\sum_{i=1}^n \frac{a_i^2}{(1-\delta^*)\sigma_a}$, by using $\frac{1}{\tilde\sigma_a} \leq \frac{1}{1-\delta^*}\frac{1}{\sigma_a}$ and $\sup_{x\notin\Theta}\psi'(x)\leq 1$. Bounding the second term follows from the definition of $\mathcal{J}_{1;a}$, which implies $\left|\frac{1}{n}\sum_{i\in J^c_{\alpha_*,\delta_*}}\psi'\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)\frac{\sigma\epsilon_i}{\tilde\sigma_a}a_i\right| \leq \frac{\lambda_2\sigma_a}{1-\delta_*}\left\|\hat{\boldsymbol\beta}-\boldsymbol\beta^0\right\|_1$.

Combining all these remarks with the definitions of D_2, D_3 and D_4, we have that $\frac{1}{n}\left|\sum_{i=1}^n\psi'\left(\frac{\sigma\tilde\epsilon_i}{\tilde\sigma_a}\right)\frac{\sigma\tilde\epsilon_i}{\tilde\sigma_a}a_i\right| - \frac{D_1}{n}\sum_{i\in J_{\alpha_*,\delta_*}}|a_i| - \frac{\lambda_2\sigma_a}{1-\delta_*}\left\|\hat{\boldsymbol\beta}-\boldsymbol\beta^0\right\|_1$ is bounded from above by

$$\frac{|\hat\sigma_a-\sigma_a|}{\sigma_a}\frac{D_4}{1-\delta_*}\frac{1}{n}\sum_{i=1}^n|a_i| + \left[1+\frac{D_3}{1-\delta_*}+\frac{D_3\delta_*}{(1-\delta_*)^2}\right]\frac{1}{n}\sum_{i=1}^n\frac{a_i^2}{\sigma_a}+\frac{D_2}{(1-\delta_*)^2}\frac{1}{n}\sum_{i=1}^n\frac{|a_i|^3}{\sigma_a^2}.$$

The lemma follows directly from these observations. □

Here we introduce notation to bound $\frac{1}{n}\sum_{i=1}^n \inf_t \psi'\left(\frac{\sigma\tilde\epsilon_i}{\tilde\sigma_a}\right)\frac{(\sigma\tilde\epsilon_i)^2}{(\tilde\sigma_a)^2}$ from below. Now, by the assumption that ψ' is locally continuous and the assumption $\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]>0$, for $\alpha_*,\delta_*>0$ small enough there exists b_1, b_2 such that $c_2 := \inf_{b_1\leq x\leq b_2,|\delta|\leq\delta_*}\psi'\left(\frac{x}{1+\delta}\right)$ and $c_1 := \mathbb{P}\left[b_1+\frac{\alpha^*}{1-\delta^*}\leq\frac{\sigma\epsilon_i}{\sigma_a}\leq b_2-\frac{\alpha^*}{1-\delta^*}\right]$ are strictly positive. We set $D=\mathbb{E}\left[\frac{c_2}{2}\frac{\sigma^2\epsilon^2}{\sigma_a^2}1_{b_1\leq\epsilon\leq b_2}\right]$,

$$\begin{aligned} G_{\alpha_*,\delta_*} &:= \left\{i\in\{1,...,n\}: b_1+\frac{\alpha^*}{1-\delta^*}\leq\frac{\sigma\epsilon_i}{\sigma_a}\leq b_2-\frac{\alpha^*}{1-\delta^*}\right\} \text{ and} \\ \mathcal{G}_{1;a} &:= \left\{\frac{c_2}{(1+\delta_*)^2}\frac{1}{n}\sum_{i\in G_{\alpha_*,\delta_*}}\frac{\sigma^2\epsilon_i^2}{\sigma_a^2}-2(|b_1|+|b_2|)\alpha_*\frac{(1+\delta_*)^2}{(1-\delta_*)^2}\geq D\right\}. \end{aligned}$$

Finally, let $0<\alpha_1\leq\min\left\{\frac{1}{4},\frac{D}{4D_4}\right\}$ and $0<\delta_1\leq\frac{1}{5}$ such that for all $\delta_*\leq\delta_1$ and all $\alpha_*\leq\alpha_1$ we have $\lim_{n\to\infty}\mathbb{P}[\mathcal{G}_{1;a}]=1$.

By using the above definitions and remarks, we can bound $\frac{|\hat\sigma_a-\sigma_a|}{\sigma_a}$ locally, under the assumptions that $\frac{|\hat\sigma_a-\sigma_a|}{\sigma_a}$ and $\frac{\sup_i|a_i|}{\sigma_a}$ are small enough.

Lemma B.4. *There exist $\delta_1,\alpha_1,C_1,C_2,C_3,C_4>0$, such that if $\frac{|\hat\sigma_a-\sigma_a|}{\sigma_a}\leq\delta_*\leq\delta_1$ and $\frac{\sup_i|a_i|}{\sigma_a}\leq\alpha_*\leq\alpha_1$, then, on $\mathcal{G}_{1;a}\cap\mathcal{J}_{1:a}\cap\mathcal{J}_{2:a}$,*

$$\frac{|\hat\sigma_a-\sigma_a|}{\sigma_a}\leq C_1\lambda_1\sigma_a+C_2\lambda_2\left\|\hat{\boldsymbol\beta}-\boldsymbol\beta^0\right\|_1+\frac{C_3}{n}\sum_{i\in J_{\alpha_*,\delta_*}}\frac{|a_i|}{\sigma_a}+\frac{C_4}{n}\sum_{i=1}^n\frac{a_i^2}{\sigma_a^2}.$$

Proof. By using Lemma B.2 and Assumption 4.2:

$$\psi'\left(\frac{\sigma\tilde\epsilon_i}{\tilde\sigma_a}\right)\frac{(\sigma\tilde\epsilon_i)^2}{(\tilde\sigma_a)^2} \geq t^2\psi'\left(\frac{\sigma\tilde\epsilon_i}{\tilde\sigma_a}\right)\frac{a_i^2}{(\tilde\sigma_a)^2}+2t\psi'\left(\frac{\sigma\tilde\epsilon_i}{\tilde\sigma_a}\right)\frac{\sigma\epsilon_i a_i}{(\tilde\sigma_a)^2}+\psi'\left(\frac{\sigma\tilde\epsilon_i}{\tilde\sigma_a}\right)\frac{\sigma^2\epsilon_i^2}{(\tilde\sigma_a)^2}$$

$$\begin{aligned} &\geq \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma^2\epsilon_i^2}{(\tilde{\sigma}_a)^2} - 2\frac{|\sigma\epsilon_i a_i|}{\tilde{\sigma}_a)^2} \\ &\geq \frac{c_2}{(1+\delta_*)^2}\frac{\sigma^2\epsilon_i^2}{\sigma_a^2} - 2(|a|+|b|)\alpha_*\frac{(1+\delta_*)^2}{(1-\delta_*)^2}, \end{aligned}$$

for $i \in G_{\alpha_*,\delta_*}$ and $t \in [0,1]$. Consequently, on $\mathcal{G}_{1;a}$, we have that $0 < D \leq \frac{1}{n}\sum_{i=1}^n \inf_t \psi'\left(\frac{\sigma\epsilon_i + t a_i}{\sigma_a + t(\tilde{\sigma}_a - \sigma_a)}\right)\frac{(\sigma\epsilon_i + t a_i)^2}{(\sigma_a + t(\tilde{\sigma}_a - \sigma_a))^2}$. It now follows from Lemmas B.3 and B.2 that

$$\begin{aligned} |\hat{\sigma}_a - \sigma_a| \quad \leq \quad & \frac{1}{D}\left[(1+\delta_*)\sigma_a^2\frac{\lambda_1}{2} + \frac{\lambda_2\sigma_a}{1-\delta_*}\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1\right. \\ & + \frac{D_1}{n}\sum_{i\in J_{\alpha_*,\delta_*}}|a_i| + \frac{|\hat{\sigma}_a - \sigma_a|}{\sigma_a}\frac{D_4}{1-\delta_*}\frac{1}{n}\sum_{i=1}^n|a_i| \\ & \left. + \left[1 + \frac{D_3}{1-\delta_*} + \frac{D_3\delta_*}{(1-\delta_*)^2}\right]\frac{1}{n}\sum_{i=1}^n\frac{a_i^2}{\sigma_a} + \frac{D_2}{(1-\delta_*)^2}\frac{1}{n}\sum_{i=1}^n\frac{|a_i|^3}{\sigma^2}\right]. \end{aligned}$$

By plugging in $\delta_* = \frac{1}{4}$ and $\frac{|a_i|^3}{\sigma_a^2} \leq \alpha_*\frac{|a_i|^2}{\sigma_a} \leq \frac{1}{4}\frac{|a_i|^2}{\sigma_a}$ in the above inequality and using $1 - 2\frac{D_4}{D}\alpha_* \geq \frac{1}{2}$, we recover the constants C_j for the lemma. □

B.3. Technical arguments on bounds between norms

Here we provide a bound for the off diagonal elements of $\|\cdot\|_{\boldsymbol{\Gamma}}$.

Lemma B.5. *Let* $\frac{|\hat{\sigma}_a - \sigma_a|}{\sigma_a} \leq \delta_* \leq \delta_1$ *and* $\frac{\sup_i|a_i|}{\sigma_a} \leq \alpha_* \leq \alpha_1$. *Then there exist* $Q_1 > 0$ *and* $Q_2 > 0$ *such that on* $\mathcal{G}_{1;a} \cap \mathcal{J}_{1;a} \cap \mathcal{J}_{2;a}$, $\lambda_1\sigma_a|\hat{\sigma}_a - \sigma_a| + \frac{2}{n}\sum_{i=1}^n \sup_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\frac{a_i}{\tilde{\sigma}_a}(\hat{\sigma}_a - \sigma_a)$ *is bounded from above by*

$$\frac{Q_1}{\sigma_a}\left[\lambda_2^2\sigma_a^2\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1^2 + \frac{|J_{\alpha_*,\delta_*}|}{n}\frac{1}{n}\sum_{i=1}^n a_i^2 + \left[\frac{1}{n}\sum_{i=1}^n\frac{a_i^2}{\sigma_a}\right]^2\right] + Q_2\lambda_1^2\sigma_a^3,$$

where Q_1 *and* Q_2 *are constants depending only on* D_1, D_2, D_3 *and* D_4.

Proof. By Lemma B.3, $\delta_* \leq \frac{1}{2}$ and $\sup_i|a_i| \leq \sigma_a\alpha_*$ we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \sup_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}a_i \leq [2 + D_1 + 1 + 4D_3 + 4D_2]\left[\lambda_2\sigma_a\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1\right. \\ \left. + \frac{1}{n}\sum_{i\in J_{\alpha_*,\delta_*}}|a_i| + \frac{1}{n}\sum_{i=1}^n\frac{a_i^2}{\sigma_a} + \alpha_*\frac{1}{n}\sum_{i=1}^n\frac{a_i^2}{\sigma_a}\right] \\ + 2D_4\sigma\alpha_*\frac{|\hat{\sigma}_a - \sigma_a|}{\sigma_a} \end{aligned}$$

$$\leq E_1 \left[\lambda_2 \sigma_a \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1 + \frac{1}{n} \sum_{i \in J_{\alpha_*, \delta_*}} |a_i| + \frac{1}{n} \sum_{i=1}^n \frac{a_i^2}{\sigma_a} \right] + E_2 \alpha_* \left| \hat{\sigma}_a - \sigma_a \right|,$$

where $E_1 = 8 + 2D_1 + 8D_2 + 8D_3$, $E_2 = 2D_4$. Thus $\frac{1}{n} \sum_{i=1}^n \sup_t \psi' \left(\frac{\sigma \tilde{\epsilon}_i}{\tilde{\sigma}_a} \right) \frac{\sigma \tilde{\epsilon}_i}{\tilde{\sigma}_a} \frac{a_i}{\sqrt{\sigma_a}}$ is bounded from above by

$$\frac{E_1}{\sqrt{\sigma_a}} \left[\lambda_2 \sigma_a \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1 + \frac{1}{n} \sum_{i \in J_{\alpha_*, \delta_*}} |a_i| + \frac{1}{n} \sum_{i=1}^n \frac{a_i^2}{\sigma_a} \right] + E_2 \alpha_* \frac{|\hat{\sigma}_a - \sigma_a|}{\sqrt{\sigma_a}}.$$

Similarly by Lemma B.4, we have

$$\begin{aligned} \frac{|\hat{\sigma}_a - \sigma_a|}{\sqrt{\sigma_a}} &\leq C_1 \lambda_1 \sqrt{\sigma_a} \sigma_a \\ &\quad + \frac{C_2 + C_3 + C_4}{\sqrt{\sigma_a}} \left[\lambda_2 \sigma_a \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1 + \frac{1}{n} \sum_{i \in J_{\alpha_*, \delta_*}} |a_i| + \frac{1}{n} \sum_{i=1}^n \frac{a_i^2}{\sigma_a} \right] \\ &\leq E_3 \lambda_1 \sqrt{\sigma_a} \sigma_a \\ &\quad + \frac{E_4}{\sqrt{\sigma_a}} \left[\lambda_2 \sigma_a \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1 + \frac{1}{n} \sum_{i \in J_{\alpha_*, \delta_*}} |a_i| + \frac{1}{n} \sum_{i=1}^n \frac{a_i^2}{\sigma_a} \right], \end{aligned}$$

where $E_3 = 2C_1$ and $E_4 = C_2 + C_3 + C_4$.

$\left| \sum_{i=1}^n \sup_t \psi' \left(\frac{\tilde{\epsilon}_i}{\tilde{\sigma}_a} \right) \frac{\tilde{\epsilon}_i}{\tilde{\sigma}_a} \frac{a_i}{\tilde{\sigma}_a} (\hat{\sigma}_a - \sigma_a) \right| \leq 2 \left| \sum_{i=1}^n \sup_t \psi' \left(\frac{\tilde{\epsilon}_i}{\tilde{\sigma}_a} \right) \frac{\tilde{\epsilon}_i}{\tilde{\sigma}_a} \frac{a_i}{\sqrt{\sigma_a}} \right| \frac{|\hat{\sigma}_a - \sigma_a|}{\sqrt{\sigma_a}}$, a repeated use of $2uv \leq u^2 + v^2$ and $\left[\frac{1}{n} \sum_{i \in J_{\alpha_*, \delta_*}} |a_i| \right]^2 \leq \frac{|J_{\alpha_*, \delta_*}|}{n} \frac{1}{n} \sum_{i \in J_{\alpha_*, \delta_*}} a_i^2$ imply the lemma. □

Building upon Lemma B.5, we provide an upper bound in terms of $\|\cdot\|_1$ in the location space. Let $\mathcal{J}_a = \mathcal{J}_{0;a} \cap \mathcal{J}_{1;a} \cap \mathcal{J}_{2;a}$.

Lemma B.6. *Let* $\frac{|\hat{\sigma}_a - \sigma_a|}{\sigma_a} \leq \delta_* \leq \delta_1$, $\frac{\sup_i |a_i|}{\sigma_a} \leq \alpha_* \leq \alpha_1$ *and* $\frac{\left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1}{\sigma_a} \leq \frac{1}{Q_1}$. *Then* $\frac{1}{1+\delta_*} \frac{1}{n} \sum_{i=1}^n \inf_t \psi' \left(\frac{\sigma \tilde{\epsilon}_i}{\tilde{\sigma}_a} \right) \frac{a_i^2}{\sigma_a} + \lambda_* \left\| \hat{\boldsymbol{\beta}} \right\|_1 - \frac{Q_1}{n} \sum_{i=1}^n \frac{a_i^2}{\sigma_a} \left[\frac{|J_{\alpha_*, \delta_*}|}{n} + \alpha_*^2 \right] - Q_2 \lambda_1^2 \sigma_a^3$ *on* $\mathcal{G}_{1;a} \cap \mathcal{J}_a$ *is bounded from above by*

$$(\lambda_0 \sigma_a + \lambda_2^2 \sigma_a^2) \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1 + \lambda_* \left\| \boldsymbol{\beta}^0 \right\|_1,$$

where Q_1 *and* Q_2 *are the same as in Lemma B.5.*

Proof. We follow the same arguments as in Subsection A.8. Consequently we work with ρ_k instead of ρ in Lemma 4.1 and with $c_k \geq 0$. Furthermore by Lemma 4.1 and the definitions of λ_0 and λ_1, there exists $t^{\lambda_*} \in [0, 1]$ such that

$$\begin{aligned} \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0, \hat{\sigma}_a - \sigma_a \right\|_{\boldsymbol{\Gamma}(t^{\lambda_*})} - \sigma_a \lambda_1 \left| \hat{\sigma}_a - \sigma_a \right| + \lambda_* \left\| \hat{\boldsymbol{\beta}} \right\|_1 &\leq \sigma_a \lambda_0 \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \right\|_1 \\ &\quad + \lambda_* \left\| \boldsymbol{\beta}^0 \right\|_1 + c_k. \end{aligned}$$

On the other hand, we have that $\limsup_{k\to 0} \left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0, \hat{\sigma}_a - \sigma_a\right\|_{\boldsymbol{\Gamma}(t^{\lambda_*})}$ is bounded from below by

$$\frac{1}{1+\delta_*}\frac{1}{n}\sum_{i=1}^{n} \inf_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{a_i^2}{\sigma_a} - \frac{2}{n}\sum_{i=1}^{n}\sup_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\frac{a_i}{\tilde{\sigma}_a}\left(\hat{\sigma}_a - \sigma_a\right).$$

Thus the lemma follows directly from Lemma B.5 and $\frac{1}{n}\sum_{i=1}^{n}\frac{a_i^2}{\sigma_a^2} \leq \alpha^2$. □

B.4. Proof of Theorem 4.1

Here we extend on the definition of $\mathcal{I}_\alpha$ Subsection 3.4, by defining:

$$\mathcal{I}_{\alpha_*,\delta_*} := \left\{\left\|\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\psi'_{\alpha_*,\delta_*}}\right\|_{+\infty} \leq \tilde{\lambda}_*\right\},$$

where $\hat{\boldsymbol{\Sigma}}_{\psi'_{\alpha_*,\delta_*}}$ is defined analogously to $\hat{\boldsymbol{\Sigma}}_{\psi'_\alpha}$. Once again, just as for $\mathcal{I}_\alpha$, we point out that $\mathbb{P}\left[\mathcal{I}_{\alpha_*,\delta_*}\right]$ can be studied in exactly the same way as $\mathbb{P}\left[\mathcal{A}\right]$.

To begin we find a lower bound for $\frac{1}{n}\sum_{i=1}^{n}\inf_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{a_i^2}{\sigma_a}$.

Lemma B.7. *Let* $\frac{|\hat{\sigma}_a - \sigma_a|}{\sigma_a} \leq \delta_* \leq \frac{1}{4}$ *and* $\frac{\sup_i|a_i|}{\sigma_a} \leq \alpha_* \leq \frac{1}{4}$. *Then on* $\mathcal{I}_{\alpha_*,\delta_*}$, $\frac{1}{n}\sum_{i=1}^{n}\inf_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right)\frac{a_i^2}{\sigma_a}$ *is bounded from below by*

$$\left[\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right] - \frac{D_2\alpha_*}{1-\delta_*} - \frac{D_3\delta_*}{1-\delta_*}\right]\frac{\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2}{\sigma_a} - 16\tilde{\lambda}_*\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]\frac{\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1^2}{\sigma_a}.$$

Proof. We first start by bounding $\inf_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}a_i\right)$ from below.

We define $Z_i^{\alpha_*,\delta_*} = 1$ if $i \in J_{\alpha_*,\delta_*}$ and $Z_i^{\alpha_*,\delta_*} = 0$ otherwise. This implies that if $Z_i^{\alpha_*,\delta_*} = 1$ we have the obvious bound: $\inf_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}a_i\right) \geq 0$. On the other hand, if $Z_i^{\alpha_*,\delta_*} = 0$, by Lemma B.1 there exits a $\tilde{t} \in [0,1]$ such that $\psi'\left(\frac{\sigma\epsilon_i + t^{\lambda_*}a_i}{\sigma_a + t^{\lambda_*}(\hat{\sigma}_a - \sigma_a)}\right) - \psi'\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) = \psi''\left(\frac{\sigma\tilde{\tilde{\epsilon}}_i}{\tilde{\tilde{\sigma}}_a}\right)\left[\frac{t^{\lambda_*}a_i}{\tilde{\tilde{\sigma}}_a} - \frac{\sigma\tilde{\tilde{\epsilon}}_i}{\tilde{\tilde{\sigma}}_a}\frac{t^{\lambda_*}(\hat{\sigma}_a - \sigma_a)}{\tilde{\tilde{\sigma}}_a}\right]$, where $\tilde{\tilde{\epsilon}}_i = \epsilon_i + \tilde{t}t^{\lambda_*}\frac{a_i}{\sigma}$ and $\tilde{\tilde{\sigma}}_a = \sigma_a + \tilde{t}t^{\lambda_*}(\hat{\sigma}_a - \sigma_a)$. This implies, in the case where $Z_i^{\alpha_*,\delta_*} = 0$, that $\left|\psi'\left(\frac{\sigma\epsilon_i + t^{\lambda_*}a_i}{\sigma_a + t^{\lambda_*}(\hat{\sigma}_a - \sigma_a)}\right) - \psi'\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)\right| \leq D_2\frac{\alpha_*}{1-\delta_*} + D_3\frac{\delta_*}{1-\delta_*}$.

Therefore, generally we have the following bound:

$$\begin{aligned}\int_t \psi'\left(\frac{\sigma\tilde{\epsilon}_i}{\tilde{\sigma}_a}\right) &\geq \left[1 - Z_i^{\alpha_*,\delta_*}\right]\psi'\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) - \left[D_2\frac{\alpha_*}{1-\delta_*} + D_3\frac{\delta_*}{1-\delta_*}\right] \\ &= \psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon_i}{\sigma_a}\right) - \left[D_2\frac{\alpha_*}{1-\delta_*} + D_3\frac{\delta_*}{1-\delta_*}\right].\end{aligned}$$

Now by using Lemma A.3, we can bound $\frac{1}{n}\sum_{i=1}^{n}\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon_i}{\sigma_a}\right)\frac{a_i^2}{\sigma_a}$ from below by

$$\frac{\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{\sigma}\left[\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2-16\tilde{\lambda}_*\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1^2\right].$$

The lemma follows directly from this. □

Lemma B.8. *Let* $\frac{|\hat{\sigma}_a-\sigma_a|}{\sigma_a}\le\delta_*\le\delta_1$, $\frac{\sup_i|a_i|}{\sigma_a}\le\alpha_*\le\alpha_1$ *and* $\frac{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_1}{\sigma_a}\le\frac{1}{Q_1}$. *Then on* $\mathcal{G}_{1;a}\cap\mathcal{J}_a\cap\mathcal{I}_{\alpha_*,\delta_*}$ *we have*

$$\frac{\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]-\gamma_2}{(1+\delta_*)\sigma_a}\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2+\lambda_*\left\|\hat{\boldsymbol{\beta}}\right\|_1\le\gamma_3\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1+\lambda_*\left\|\boldsymbol{\beta}_0\right\|_1,$$

where $\gamma_2=D_2\frac{\alpha_*}{1-\delta_*}+D_3\frac{\delta_*}{1-\delta_*}+Q_1(1+\delta_*)\left[\frac{|J_{\alpha_*,\delta_*}|}{n}+\alpha_*^2\right]$ *and* $\gamma_3=\lambda_0\sigma_a+\lambda_2^2\sigma_a^2+16\tilde{\lambda}_*\frac{\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{\sigma_a}\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1+\frac{Q_2\sigma_a^3\lambda_1^2}{\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\|_1}$.

Proof. This follows from the combination of Lemmas B.6 and B.7. □

Let $0<\delta_*^0<\delta_1$ and $0<\alpha_*^0<\alpha_1$ satisfy the following inequality:

$$\mathbb{E}\left[\psi'_{\alpha_*^0,\delta_*^0}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]>\max\left\{\frac{9}{10}\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right],\frac{9D_2\alpha_*^0}{1-\delta_*^0}+\frac{9D_3\delta_*^0}{1-\delta_*^0}+9(1+\delta_*^0)Q_1(\alpha_*^0)^2\right\},$$

where Q_1 is a combination of the D_i, defined in Lemma B.6 in the Appendix. Furthermore, let $0<\lambda_*^0$ satisfy the following:

$$10\lambda_*^0\frac{1}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}\le\min\left\{\frac{\delta_*^0}{4K^*},\frac{\alpha_*^0}{K},\frac{1}{Q_1}\right\},$$

where $K^*=(C_2+C_3+C_4)K$, with C_2, C_3 and C_4 as in Lemma B.4 is used to bound the error in scale estimation.

Lemma B.9. *Suppose the compatibility condition holds for* S_0, $\frac{|\hat{\sigma}_a-\sigma_a|}{\sigma_a}\le\delta_*\le\delta^0$, $\frac{\sup_i|a_i|}{\sigma_a}\le\alpha_*\le\alpha^0$, $\left\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\right\|_1\le\min\left\{\frac{10\lambda_*\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2},\frac{\sigma_a}{Q_1}\right\}$, $\frac{320\tilde{\lambda}_*s_0}{\phi_0^2}\le\frac{1}{2}$, $\sigma_a\lambda_1Q_2\le\frac{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^0\|_1}{\sigma_a}$ *and* $\frac{|J_{\alpha_*,\delta_*}|}{n}\le\frac{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{12Q_1}$. *Then for* $\frac{2\lambda_0\sigma_a+2\lambda_2^2\sigma_a^2+2\lambda_1\sigma_a}{1-320\tilde{\lambda}_*\frac{s_0}{\phi_0^2}}\le\lambda_*$ *on* $\mathcal{G}_{1;a}\cap\mathcal{J}_a\cap\mathcal{I}_{\alpha_*,\delta_*}$ *we have that* $\frac{4\lambda_*^2\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}$ *is bounded from below by*

$$\frac{\frac{2}{1+\delta_*}\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]-\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]-\gamma_4}{\sigma_a}\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2+\lambda_*\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1,$$

where $\gamma_4=2\frac{D_2\alpha_*}{1-\delta_*^2}+2\frac{D_3\delta_*}{1-\delta_*^2}+2Q_1\frac{|J_{\alpha_*,\delta_*}|}{n}+2Q_1\alpha_*^2$.

Proof. As $320\tilde{\lambda}_* \frac{s_0}{\phi_0^2} \leq \frac{1}{2}$, the assumption on λ_* implies $0 \leq 2\lambda_0\sigma_a + 2\lambda_2^2\sigma_a^2 + 2\lambda_1\sigma_a \leq \lambda_*$. Furthermore the same assumption, i.e. $\frac{2\lambda_0\sigma_a+2\lambda_2^2\sigma_a^2+2\lambda_1\sigma_a}{1-320\tilde{\lambda}_*\frac{s_0}{\phi_0^2}} \leq \lambda_*$, implies that

$$2\lambda_0\sigma_a + 2\lambda_2^2\sigma_a^2 + 2\lambda_1\sigma_a + 32\tilde{\lambda}_* \frac{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{\sigma_a} \frac{10\lambda_*\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]} \frac{s_0}{\phi_0^2} \leq \lambda_*.$$

This in turn implies that $\lambda_0\sigma_a + \lambda_2^2\sigma_a^2 + \frac{Q_2\sigma_a^3\lambda_1^2}{\left\|\boldsymbol{\beta}^0-\hat{\boldsymbol{\beta}}\right\|_1} + 16\tilde{\lambda}_* \frac{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{\sigma_a} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq \frac{1}{2}\lambda_*$, by the assumptions on λ_1 and $\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1$. We can then plug this in the equation of Lemma B.8 and we recover

$$\frac{\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right] - \gamma_2}{(1+\delta_*)\sigma_a} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 + \lambda_* \left\|\hat{\boldsymbol{\beta}}\right\|_1 \leq \frac{1}{2}\lambda_* \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 + \lambda_* \left\|\boldsymbol{\beta}_0\right\|_1.$$

By the same techniques as in Lemma A.5, we then have:

$$2\frac{\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right] - \gamma_2}{(1+\delta_*)\sigma_a} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 + \lambda_* \left\|\hat{\boldsymbol{\beta}}_{S_0^c}\right\|_1 \leq 3\lambda_* \left\|\boldsymbol{\beta}^0_{S_0} - \hat{\boldsymbol{\beta}}_{S_0}\right\|_1.$$

By the choices of δ^0 and α^0 and the assumption on $\frac{|J_{\alpha_*,\delta_*}|}{n}$, we have $\gamma_2 \leq \mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]$. Therefore we can use the compatibility condition as in Lemma A.6. In fact by using the same techniques as in Lemma A.6, we have that $2\frac{\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]-\gamma_2}{(1+\delta_*)\sigma_a} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 + \lambda_* \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1$ satisfies the following:

$$\begin{aligned}
& 2\frac{\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right] - \gamma_2}{(1+\delta_*)\sigma_a} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 + \lambda_* \left\|\boldsymbol{\beta}^0_{S_0} - \hat{\boldsymbol{\beta}}_{S_0}\right\|_1 + \lambda_* \left\|\hat{\boldsymbol{\beta}}_{S_0^c}\right\|_1 \\
\leq\ & 4\lambda_* \left\|\boldsymbol{\beta}^0_{S_0} - \hat{\boldsymbol{\beta}}_{S_0}\right\|_1 \\
\leq\ & 4\lambda_* \sqrt{\frac{s_0}{\phi_0^2} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2} = 4\sqrt{\frac{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{\sigma_a} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 \lambda_*^2 \frac{\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]} \frac{s_0}{\phi_0^2}} \\
\leq\ & \frac{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{\sigma_a} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 + 4\lambda_*^2 \frac{\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]} \frac{s_0}{\phi_0^2}.
\end{aligned}$$

Consequently $\frac{4\lambda_*^2\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]} \frac{s_0}{\phi_0^2}$ is bounded from below by

$$\frac{\frac{2}{1+\delta_*}\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right] - \frac{2}{1+\delta_*}\gamma_2 - \mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{\sigma_a} \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_{\hat{\Sigma}}^2 + \lambda_* \left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1.$$

The lemma follows directly from $\frac{2}{1+\delta_*}\gamma_2 = \gamma_4$. ∎

Before providing the proof of Theorem 4.1, we introduce new notation. Similarly to the definition of h in Subsection 3.4, we define:

$$h_*(\alpha_*, \delta_*, \gamma_*) := 2\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right] - 2\frac{\mathbb{E}\left[\psi'_{\alpha_*,\delta_*}\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{1+\delta_*} + 2\frac{\alpha_* D_2 + \delta_* D_3}{1-\delta_*^2} + 2Q_1(\alpha_*^2 + \gamma_*).$$

We obviously have $h_* \geq 0$. Moreover, by the choices of α_*^0 and δ_*^0, we have for all $\alpha_* \leq \alpha_*^0$ and $\delta_* \leq \delta_*^0$ that $h_*(\alpha_*, \delta_*, \gamma_*) \leq \frac{4}{6}\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right] + 2Q_1\gamma_*$ and so if $\gamma_* \leq \frac{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}{12Q_1}$ we would have $h_*(\alpha_*, \delta_*, \gamma_*) \leq \frac{5}{6}\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]$.

We now introduce a set, needed specifically for the convergence of the scale parameter. Let σ_a^* be the solution of $\frac{1}{n}\sum_{i=1}^n \chi_0\left(\frac{\sigma\epsilon_i}{\cdot}\right) = a$. This would be the estimate of σ_a in case the true location parameter were known. We can define:

$$\mathcal{G}_{a;2} := \left\{\frac{|\sigma_a^* - \sigma_a|}{\sigma_a} \leq \frac{\delta_*^0}{2}\right\}.$$

Since $\frac{1}{n}\sum_{i=1}^n \chi_0\left(\frac{\epsilon_i}{\cdot}\right)$ is monotone decreasing, by the law of large numbers we obviously have $\lim_{n\to\infty}\mathbb{P}\left[\mathcal{G}_{a;2}\right] = 1$ for any fixed $\delta_*^0 > 0$.

Finally, we define the set of observations close to non differentiable points:

$$\mathcal{G}_{a;3} := \left\{\frac{|J_{\alpha_*,\delta_*}|}{n} \leq \gamma_*\right\}.$$

As an observation, for a differentiable ψ (e.g. the classical case), we have $0 = \frac{|J_{\alpha_*,\delta_*}|}{n}$. Otherwise, for a more general ψ, we have $\lim_{\max\{\alpha_*,\delta_*\}\to 0}\mathbb{E}\left[\frac{|J_{\alpha_*,\delta_*}|}{n}\right] = 0$ and therefore for α_* and δ_* small enough $\mathbb{P}\left[\mathcal{G}_{a;3}\right]$ can be assumed big for relatively small γ_*. To simplify the notation a bit, we define the sets $\mathcal{G}_a = \mathcal{G}_{a;1} \cap \mathcal{G}_{a;2} \cap \mathcal{G}_{a;3}$. The proof is very similar to that of Theorem 3.2 with a few subtleties. Here again we use the convexity of the objective function in order to use previous lemmas on the minimizers.

Defining $\boldsymbol{\beta}^t = \boldsymbol{\beta}^0 + t\left[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right]$, we have $\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^0\right\|_1 = t\left\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0\right\|_1$. Moreover, we define $\sigma_a^*(t)$ as the minimizer of the objective function for a fixed location $\boldsymbol{\beta}^t$. Let $t^* \in [0,1]$ such that $\left\|\boldsymbol{\beta}^{t^*} - \boldsymbol{\beta}^0\right\|_1 = \lambda_1\sigma_a^2 Q_2$. Accordingly we must have $\left\|\boldsymbol{\beta}^{t^*} - \boldsymbol{\beta}^0\right\|_1 \leq \frac{10\lambda_*^0\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}$.

We now show that $\frac{|\sigma_a^*(t) - \sigma_a|}{\sigma_a} \leq \delta_*^0$ for $t \in [0, t^*]$. This is indeed true for $t = 0$, since we are on $\mathcal{G}_{a;2}$. Furthermore if for any $t \in [0, t^*]$ we have $\frac{|\sigma_a^*(t) - \sigma_a|}{\sigma_a} \leq \delta_*^0$, then by Lemma B.4 and the choice of the constant λ_*^0, we must have $\frac{|\sigma_a^*(t) - \sigma_a|}{\sigma_a} \leq \frac{\delta_*^0}{2}$, since:

$$\frac{|\sigma_a^*(t) - \sigma_a|}{\sigma_a} \leq C_1\lambda_1\sigma_a + C_2\lambda_2\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^0\right\|_1 + \frac{C_3}{n}\sum_{i\in J_{\alpha_*,\delta_*}}\frac{|a_i^t|}{\sigma_a} + \frac{C_4}{n}\sum_{i=1}^n\frac{(a_i^t)^2}{\sigma_a^2}$$

$$\begin{aligned} &\leq C_1\lambda_1\sigma_a + \frac{1}{\sigma_a}\left\|\boldsymbol{\beta}^t - \boldsymbol{\beta}^0\right\|_1 \left[C_2\lambda_2\sigma_a + C_3K + C_4K\alpha_*^0\right] \\ &\leq C_1\frac{\delta_0^*}{4C_1} + 10\lambda_*^0\frac{1}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}K^* \leq \frac{\delta_0^*}{4} + \frac{\delta_0^*}{4} = \frac{\delta_0^*}{2}. \end{aligned}$$

The statement then follows from the intermediate value theorem since $\sigma_a^*(t)$ is continuous on $[0,1]$. One may thus apply Lemma B.9, and because of the choice of the constants we recover

$$\left\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^{t^*}\right\|_1 \leq 4\lambda_*\frac{\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}.$$

For the same reasons as before, we have $\left\|\boldsymbol{\beta}^0 - \hat{\boldsymbol{\beta}}\right\|_1 \leq 4\lambda_*\frac{\sigma_a}{\mathbb{E}\left[\psi'\left(\frac{\sigma\epsilon}{\sigma_a}\right)\right]}\frac{s_0}{\phi_0^2}$. By plugging these results in Lemma B.9, the proof is complete.

Acknowledgements

We would like to thank the Associate Editor and a referee for their careful reading of the manuscript and helpful comments which improved the clarity of the paper. Furthermore we are grateful to Elvezio Ronchetti and to Sara van de Geer for helpful discussions and suggestions.

References

[1] Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications.* Heidelberg: Springer. MR2807761

[2] Fan, J., Fan, Y. and Barut, E. (2014). Adaptive robust variable selection. *Ann. Statist.* **42** 324–351. MR3189488

[3] Gao, X. and Huang, J. (2010). Asymptotic analysis of high-dimensional LAD regression with Lasso. *Statist. Sinica* **20** 1485–1506. MR2777333

[4] Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.* **34** 2367–2386. MR2291503

[5] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

[6] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[7] Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35** 73–101. MR0161415

[8] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics.* New York: Wiley. MR2488795

[9] Koltchinskii, V. (2011). *Oracle inequalities in empirical risk minimization and sparse recovery problems.* Heidelberg: Springer. MR2829871

[10] Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive lasso penalty. *Electron. J. Stat.* **5** 1015–1053. MR2836768

[11] Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35** 1012–1030. MR2341696

[12] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[13] Sun, T. and Zhang, C. H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898. MR2999166

[14] van de Geer, S. (2007). The deterministic Lasso. In *JSM Proceedings*, 2007 140. American Statistical Association.

[15] van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *Ann. Statist.* **36** 614–645. MR2396809

[16] van de Geer, S. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316

[17] Wang, H., Li, G. and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. Bus. Econom. Statist.* **25** 347–355. MR2380753