Electronic Journal of Statistics
Vol. 10 (2016) 1296–1340
ISSN: 1935-7524
DOI: 10.1214/16-EJS1136

Minimum Distance Lasso for robust high-dimensional regression

Aurélie C. Lozano

IBM T.J. Watson Research Center
1101 Kitchawan Rd, Yorktown Heights, NY 10598, USA
e-mail: aclozano@us.ibm.com

Nicolai Meinshausen

Seminar für Statistik, ETH Zürich
Raemistrasse 101, 8092 Zürich, Switzerland
e-mail: meinshausen@stat.math.ethz.ch

and

Eunho Yang

IBM T.J. Watson Research Center
1101 Kitchawan Rd Yorktown Heights, NY 10598, USA
e-mail: eunhyang@us.ibm.com

Abstract: We propose a minimum distance estimation method for robust regression in sparse high-dimensional settings. Likelihood-based estimators lack resilience against outliers and model misspecification, a critical issue when dealing with high-dimensional noisy data. Our method, Minimum Distance Lasso (MD-Lasso), combines minimum distance functionals customarily used in nonparametric estimation for robustness, with ℓ_1-regularization. MD-Lasso is governed by a scaling parameter capping the influence of outliers: the loss is locally convex and close to quadratic for small squared residuals, and flattens for squared residuals larger than the scaling parameter. As the parameter approaches infinity the estimator becomes equivalent to least-squares Lasso. MD-Lasso is able to maintain the robustness of minimum distance functionals in sparse high-dimensional regression. The estimator achieves maximum breakdown point and enjoys consistency with fast convergence rates under mild conditions on the model error distribution. These hold for any solution in a convexity region around the true parameter and in certain cases for every solution. We provide an alternative set of results that do not require the solutions to lie within the convexity region but where the ℓ_2-norm of the feasible solutions is constrained within a safety radius. Thanks to this constraint, a first-order optimization method is able to produce local optima that are consistent. A connection is established with re-weighted least-squares that intuitively explains MD-Lasso robustness. The merits of our method are demonstrated through simulation and eQTL analysis.

Keywords and phrases: Lasso, robust estimation, high-dimensional variable selection, sparse learning.

Received August 2014.

1. Introduction

We address the problem of robust sparse estimation in high-dimensional regression. Sparse linear models allow for simultaneous model estimation and variable selection. They have become very popular tools to analyze the high-dimensional data that is prevalent in many domains such as genomics [45], neuroimaging [21, 42], and economics [16]. A widely used approach to sparse learning is via sparsity-inducing regularization. A well known example is the Lasso [38], which employs ℓ_1-penalized least-squares to identify a parsimonious subset of predictors. Beyond Lasso, various structured penalties have been proposed that reflect the underlying structural information among the predictors. For instance the Group Lasso [46] enforces group sparsity via the ℓ_1/ℓ_q norm ($q > 1$), the Path Coding Penalties [26] and Graph Lasso [20] deal with applications where the variables reside in a graph, the Fused Lasso [39] enforces sparsity in both the coefficients and their successive differences for settings where variables are ordered in some meaningful way and a locally constant coefficient profile is desirable. Much attention has been devoted recently to the study of these structured norms and their theoretical properties [29, 30], and to devising efficient algorithms for large scale problems [6].

The issue of robustness, however, has been largely overlooked in the sparse learning literature, while this aspect is critical when dealing with high dimensional noisy data. Traditional likelihood-based estimators (including Lasso and variants) are known to lack resilience to outliers and model misspecification. Despite this fact, there has been limited focus on robust sparse learning methods in high-dimensional modeling. Relevant penalized regression methods include the "extended" Lasso formulation [32] which employs the traditional squared error but incorporates an additional sparse error vector into the model so as to account for corrupted observations, and the LAD-Lasso [43], which uses the least absolute deviation combined with an ℓ_1 penalty. Note that the least absolute deviations estimate also arises as a maximum likelihood estimate if the errors have a Laplace distribution. Hence the aforementioned approaches can still be viewed as likelihood-based, and they share the deficiencies of maximum-likelihood estimators for sparse estimation in high-dimensional regression. In particular their performance drops significantly if the model is mis-specified or outliers are present: a single outlier can make their estimates entirely unreliable [1].

Departing from likelihood-based methods, we propose a penalized minimum distance criterion for robust and consistent estimation of sparse high dimensional regression. Our approach is motivated by minimum distance estimators [44], which are popular in nonparametric methods and have been shown to exhibit excellent robustness and efficiency properties [10, 15]. Their use for parametric estimation has been discussed in Basu et al. [8], Scott [36] and investigated by Chi and Scott [13] for sparse logistic regression. However, the robustness properties of minimum distance estimators have not been formally established in the high-dimensional regression setting. We propose the Minimum Distance Lasso (MD-Lasso) estimator, which is derived from the integrated squared error

distance between the model and the "true" distribution, and imposes sparse model structure via ℓ_1 penalty.

The MD-Lasso loss, taken as a function of a single observation, acts similarly to the squared-loss if the residual squared-error of that observation is small, while the loss becomes flat as the squared-error becomes large. This ensures that the contributions of large outliers to the overall loss are capped. Overall the MD-Lasso loss is invex[1] and locally convex. The extent of the local convexity region and the capping of outliers are both governed by a scaling parameter of our estimator, against which the residual squared error of each observation is being compared. In the extreme case where the scaling parameter goes to zero, only the most "trusted" observation is taken into account, while as the scaling parameter goes to infinity, the estimator becomes equivalent to the traditional Lasso estimator with the same amount of regularization on the ℓ_1 penalty. Our analysis shows that the tradeoff between convexity and robustness, as controlled by the scaling parameter, is, understandably, essential in securing both robustness and consistency of the estimator.

Our results demonstrate that the MD-Lasso enjoys fast convergence rates in high dimensional settings under mild conditions on the model error distribution in relation to the scaling parameter. These conditions are much less restrictive than the traditional sub-gaussian assumption, and cover a broad class of heavy-tailed distributions. We present two sets of consistency results. One holds for any of the solutions in the local convexity region around the true parameter (and in certain cases these are the only existing solutions globally). One does not necessitate that the solutions lie within the local convexity region but requires adding a constraint to the MD-Lasso problem to bound the ℓ_2-norm of the parameters β considered, and assuming that the true parameter vector β^* is feasible. The latter set of results has practical consequences as they allow us to show that a simple incremental algorithm yields consistent estimates.

We also show that MD-Lasso achieves a maximum breakdown point [19] for any finite value of the scaling parameter c, namely MD-Lasso is able to tolerate the maximum percentage of arbitrarily corrupted observations achievable by any method. In contrast the least squares Lasso and LAD-Lasso have a vanishing breakdown point, namely a single corruption can already drastically affect these estimators. We shed further light onto the robustness of MD-Lasso by establishing its connection with a form of iteratively re-weighted ℓ_1-penalized least-squares regression (namely the traditional Lasso) where the weights assigned to the observations can be interpreted in terms of their likelihood.

The performance of our estimator is demonstrated on simulation data under various error distributions, in comparison to the traditional Lasso, LAD-Lasso, and Extended Lasso. This study also confirms that outliers and/or heavy-tailed noise can severely influence the variable selection accuracy of existing sparse

[1] A function f is invex if it is differentiable and there exists a vector-valued function g such that $|f(\boldsymbol{v}) - f(\boldsymbol{u})| \leq \langle \nabla f(\boldsymbol{u}), g(\boldsymbol{v}, \boldsymbol{u}) \rangle$, for all $\boldsymbol{u}, \boldsymbol{v}$. A function is invex if and only if every stationary point is a global minimum [9].

learning methods. Experiments on real eQTL data further illustrate the usefulness of our approach.

The manuscript is organized as follows. Section 2 is devoted to the MD-Lasso estimator, its derivation from a minimum distance criterion, its geometry, and the analysis of its breakdown point. The statistical consistency results and convergence rates are shown in Section 3. The incremental method for efficient and scalable optimization is presented in Section 4. Empirical results are described in Sections 5. All proofs are collected in the Appendix.

2. The Minimum Distance Lasso estimator

2.1. Problem formulation and notation

Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ denote the predictor matrix, whose rows are p-dimensional variable vectors observed for n training examples. Denote by $\boldsymbol{X}_i \in \mathbb{R}^p$ the vector formed by i^{th} observation across all variables.

Denote by $\boldsymbol{X}^j \in \mathbb{R}^n$ the vector formed by the observations for the j^{th} variable. Denote by $X_i^j \in \mathbb{R}$ the entry in matrix $\boldsymbol{X}$ corresponding to the i^{th} observation for the j^{th} variable. Similarly let $\boldsymbol{Y} \in \mathbb{R}^n$ denote the response vector, and Y_i it's i-th observation.

Consider the general regression model:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{\eta}, \tag{1}$$

where $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ is the coefficient vector one wishes to estimate, and $\boldsymbol{\eta} \in \mathbb{R}^n$ is the error term, and for simplicity we assume that the data have been standardized so that we need not consider intercept terms.

We address the sparse estimation of coefficient vector $\boldsymbol{\beta}^\star$ via ℓ_1-penalized loss minimization. Specifically, we consider estimators of the form

$$\hat{\boldsymbol{\beta}}_{\lambda_n} = \arg\min_{\boldsymbol{\beta}} \; L(\boldsymbol{\beta}) + \lambda_n \|\boldsymbol{\beta}\|_1, \tag{2}$$

where the loss function L measures the goodness-of-fit on the response and λ_n is the regularization parameter for the ℓ_1 penalty. However our framework is readily applicable to other sparsity-inducing penalties such as the group or fused Lasso.

Using likelihood-based loss functions such as squared loss is a common practice in estimating and exploring the sparsity structure of the unknown parameters for model (1), whereby $L(\boldsymbol{\beta})$ is derived from a product of probability density functions (p.d.f.). However, the likelihood-based estimators are known to lack resilience to outliers and model misspecification. In contrast, the minimum distance estimators [44] often used in nonparametric function estimation show excellent robustness properties [10, 15]. This motivates our proposed MD-Lasso estimator.

We begin this section by presenting how the MD-Lasso objective can be rigorously derived from a minimum distance criterion.

2.2. Minimum distance estimation

Here we treat response Y and predictors $\boldsymbol{X}$ as random variables, where $Y \in \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^p$.

We first apply the Integrated Squared Error to the conditional distribution of response Y given the predictors $\boldsymbol{X}$. This leads to an ℓ_2 distance between the true conditional distribution $f(Y|\boldsymbol{X})$ and the parametric distribution function $f(Y|\boldsymbol{X};\beta)$ as follows

$$\begin{aligned} d(\beta) &= \int \left[f(Y|\boldsymbol{X};\boldsymbol{\beta}) - f(Y|\boldsymbol{X})\right]^2 dY \\ &= \int f^2(Y|\boldsymbol{X};\boldsymbol{\beta})dY - 2\int f(Y|\boldsymbol{X};\boldsymbol{\beta})f(Y|\boldsymbol{X})dY + \int f^2(Y|\boldsymbol{X})dY \\ &= \int f^2(Y|\boldsymbol{X};\beta)dY - 2\mathbb{E}[f(Y|\boldsymbol{X};\boldsymbol{\beta})] + \text{constant}. \end{aligned} \tag{3}$$

We remark that minimum distance estimators originally involved distances between cumulative distribution functions [44], but the notion was subsequently broadened to encompass distances between probability density functions [10, 15, 36]. We consider the latter, which is easier to work with, and is also more natural in the context of linear regression.

Note that we assume a parametric family for the model while using a non-parametric criterion (the Integrated Squared Error) to measure goodness of fit. From the perspective of the loss function, the Integrated Squared Error is a more robust measure of the goodness-of-fit compared to likelihood-based loss functions. It can match the model with the largest portion of the data because the integration in (3) accounts for the whole range of the squared loss function.

To derive our estimator, we assume that $f(Y|\boldsymbol{X};\boldsymbol{\beta})$ is the p.d.f. of multivariate normal $\mathcal{N}(\boldsymbol{X}'\boldsymbol{\beta},\sigma^2)$. However it is important to note that our methodology and theoretical results go well beyond the normal assumption for the error. $f(Y|\boldsymbol{X};\boldsymbol{\beta}) \equiv f(Y - \boldsymbol{X}'\boldsymbol{\beta})$ because of the conditional distribution assumption, and it holds that $\int f^2(Y|\boldsymbol{X};\boldsymbol{\beta})dY = 1/(2\pi^{1/2}\sigma)$. Since $\eta_i = Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}, i = 1,\ldots,n$ are independently and identically distributed, one can consider estimating $\mathbb{E}[f(Y|\boldsymbol{X};\boldsymbol{\beta})]$ by the empirical mean $n^{-1}\sum_{i=1}^n f(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})$. Such an approximation technique has also been used for Gaussian mixture density estimation [36]. Disregarding the terms that are independent of $\boldsymbol{\beta}$ we can write the resulting empirical criterion as

$$\begin{aligned} d_n(\beta) &= -\frac{2}{n}\sum_{i=1}^n f(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta}) \\ &= -\frac{2}{n}\sum_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2\sigma^2}(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2). \end{aligned} \tag{4}$$

Rather than directly minimizing d_n we will aim to minimize, equivalently,

$$-\log(-d_n(\boldsymbol{\beta})) = -\log\left(\frac{2}{n\sqrt{2\pi\sigma^2}}\sum_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2\right)\right)$$

$$= -\log\left(\sum_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2}(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2\right)\right) + C,$$

where C is a function of σ and n but is independent of $\boldsymbol{\beta}$. As σ is unknown we consider instead

$$L(\boldsymbol{\beta}) = -c\log\left(\sum_{i=1}^{n}\exp\left(-\frac{1}{2c}(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2\right)\right)$$

where c is a scaling parameter. Plugging the resulting loss in (2) yields the MD-Lasso problem:

$$\hat{\boldsymbol{\beta}}_{\lambda_n} = \underset{\beta}{\operatorname{argmin}}\left(-c\log\left[\sum_{i=1}^{n}\exp(-\frac{1}{2c}(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2)\right] + \lambda_n\|\boldsymbol{\beta}\|_1\right). \tag{5}$$

Remarks. We can already gain some intuition on the robustness of MD-Lasso by considering the ratio between data and model probability density functions: $f(Y|\boldsymbol{X})/f(Y|\boldsymbol{X};\boldsymbol{\beta})$. An outlier in the data may drive this ratio to infinity, in which case the log-likelihood becomes infinite as well. In contrast, the difference between $f(Y|\boldsymbol{X}) - f(Y|\boldsymbol{X};\boldsymbol{\beta})$ as in (3) is always bounded. This property makes the ℓ_2-distance a favourable choice when dealing with outliers. A similar argument can be applied to the problem of density estimation and explains why the ℓ_2-distance is also very well suited for this problem (e.g. see Sugiyama et al. [37]). A more heuristical intuition comes from noting that in (5) the logarithm is applied to a *sum* of probability density functions, in contrast to the likelihood-based estimators which involve a *product*: the sum should be more robust to noise and outliers, as often encountered in high-dimensional data.

2.3. The geometry of the MD-Lasso estimator

The geometry of MD-Lasso is worth examining, as it provides some key insights on the estimator's robustness and the theoretical conditions required for fast convergence rates. The MD-Lasso loss, taken as a function of the residual error for a *single* observation with the contributions from the other observations fixed, is depicted in Figure 1, for various values of the scaling parameter c, along with the squared loss and the absolute loss. In the figure, the MD-Loss has been translated w.r.t. the y-axis for ease of comparison. From Figure 1 we can see that the MD-Lasso loss acts similarly to the squared-loss if the residual squared-error of that observation is small, while the loss becomes flat as the squared-error becomes large. This insures that the contributions of large outliers to the overall loss are capped. The range of the similarity to the squared loss is governed by the scaling parameter c of the MD-Lasso estimator, against which the residual squared error of each observation is being compared. Intuitively, the scaling parameter can thus be interpreted as a cut-off on what is an acceptable range for the error.

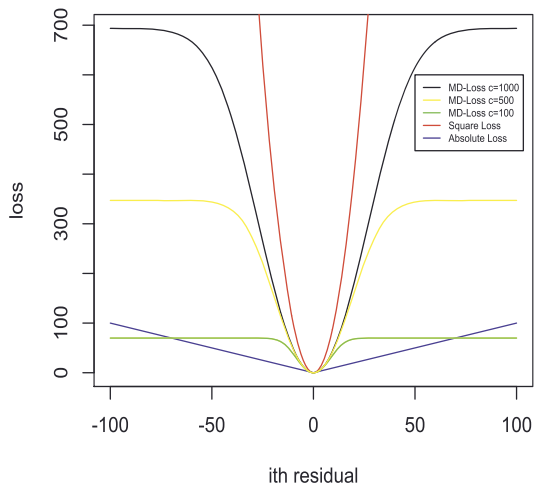


FIG 1. *The MD-Lasso loss, squared loss and absolute loss, as a function of the residual error of a single observation.*

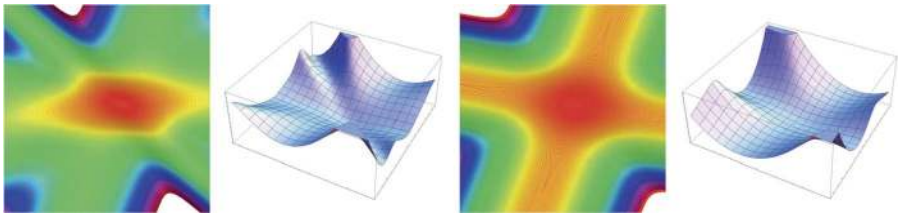


FIG 2. *Contour plot and graph of the MD-Lasso loss for illustrative examples where the dimensionality* $p = 2$, *and sample size* $n > p$ *(two plots on the left), and* $n = p$ *(two plots on the right). The x-axis and the y-axis in the contour plots and graphs correspond to the coordinates of the parameter vector* β. *In the graphs, the z-axis corresponds to the loss.*

The MD-Lasso loss over all observations is depicted in Figure 2 as a function of the regression parameter vector $\boldsymbol{\beta}$, for an illustrative examples with dimensionality $p = 2$ and just a single relevant predictor, namely $\boldsymbol{\beta}^{\star} = (\beta_1^{\star}, 0)$. As shall be formally discussed in Section 3, the MD-Lasso loss is invex and locally convex, yet it is globally non-convex. The extent of the local convexity region is controlled by the scaling parameter c. As the parameter increases, the convexity region becomes larger, and so does the proportion of observations whose squared-error are below the scaling parameter. The robustness, however, becomes weaker, as instances with larger error are allowed to significantly contribute to the overall loss. If the scaling parameter becomes too small, the proportion of observations with squared error below the scaling parameter becomes too small and compromises the convexity of the *overall* loss with respect to the model coefficients, a property that is needed to yield fast convergence rates.

Limit cases. As the parameter c goes to infinity the MD-Lasso estimator becomes equivalent to the traditional Lasso estimator with the same amount of regularisation λ_n. Indeed as $x \to 0$, we have $\exp(x) \sim 1 + x$ and $\log(1 + x) \sim x$. Therefore as $c \to \infty$, the MD-Lasso estimator is identical the minimizer of

$$\frac{1}{2n} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i' \boldsymbol{\beta})^2 + \lambda_n \|\boldsymbol{\beta}\|_1.$$

On the other extreme, for $c \to 0$, the MD-Lasso is equivalent to the minimizer of

$$\min_{i=1,\ldots,n} \frac{1}{2} (Y_i - \boldsymbol{X}_i' \boldsymbol{\beta})^2 + \lambda_n \|\boldsymbol{\beta}\|_1.$$

In the latter case, only the observation with smallest residual error is taken into account, while the other observations are being discarded. This setting can thus be viewed as an extreme case of trimmed Lasso regression, where all but one observation are trimmed out.

Non-convexity and robustness. To illustrate the limitations of convex loss functions and the appropriateness of non-convex loss functions with respect to robustness, it is worthwhile to recall the notion of *influence function* from robust statistics [18]. Consider the loss L as a function of the residual r_i of a single sample. The influence function represents the rate of change in L upon a small amount of contamination on r_i, and thus measures the effect of the size of a residual on the loss. Specifically, for loss functions induced by log-concave densities (e.g. the squared and absolute losses are induced by the Gaussian and Laplace distributions respectively) the influence function is identical to the derivative of the loss with respect to the residual. For squared error loss, the influence function is given for example by $I(r_i) = r_i$ and for the absolute loss it is $I(r_i) = \text{sign}(r_i)$. In our case the influence function can be written as

$$I(r_i) = \frac{r_i}{1 + d \exp(r_i^2/(2c))},$$

where $d > 0$ is a constant due to the contribution from other observations. We can see that in contrast to the case of log-concave densities, the influence function of the MD-Lasso loss is *redescending* as r_i becomes large, which signifies that large residuals are basically ignored. The so-called *redescending* behavior is a desirable property for robustness, which clearly cannot be achieved by convex loss functions. We refer the reader to Hampel et al. [18] for a review of influence-function approaches to robust statistics, including redescending influence functions.

Note that the negative log-likelihood functions of heavy tailed distributions (e.g. Student's t and Cauchy) are non-convex. For instance for a Student's t error model with ν degrees of freedom, the loss becomes $L(r_i) = \log(1 + r_i^2/\nu)$. It thus appears that non-convexity assists in accommodating large outliers or significantly noisy data. See Aravkin et al. [3] for additional pertinent points elucidating the need for non-convex loss functions to achieve robustness.

2.4. Breakdown point of the MD-Lasso estimator

Intuitively, the breakdown point [27] is the proportion of arbitrarily corrupted observations an estimator can tolerate before giving an arbitrarily large result. Thus a high breakdown point reflects a good resistance to corruptions and hence can be considered as a measure of robustness.
We recall the definition of replacement finite-sample breakdown point [27].

Definition 1. *Consider any sample of n points $(\boldsymbol{X}, \boldsymbol{Y})$ and let $\hat{\boldsymbol{\beta}}$ be a regression estimator. Then consider all possible corrupted samples $(\boldsymbol{X}', \boldsymbol{Y}')$ that are obtained by replacing m of the original points by arbitrary values. The breakdown point of the estimator $\hat{\boldsymbol{\beta}}$ at the sample $(\boldsymbol{X}, \boldsymbol{Y})$ is defined as*

$$\epsilon_n^*(\hat{\boldsymbol{\beta}}; \boldsymbol{X}, \boldsymbol{Y}) = \min\left\{\frac{m}{n} : \sup_{(\boldsymbol{X}', \boldsymbol{Y}')} \|\hat{\boldsymbol{\beta}}(\boldsymbol{X}', \boldsymbol{Y}')\|_2 = \infty\right\}.$$

In order to compare different estimators, one usually considers the asymptotic behaviour of $\epsilon_n^*(\hat{\boldsymbol{\beta}}; \boldsymbol{X}, \boldsymbol{Y}) : \epsilon^*(\hat{\boldsymbol{\beta}}; \boldsymbol{X}, \boldsymbol{Y}) = \lim_{n\to\infty} \epsilon_n^*(\hat{\boldsymbol{\beta}}, \boldsymbol{X}, \boldsymbol{Y})$.
The following theorem shows that the MD-Lasso estimator achieves the maximum breakdown point.

Theorem 1. *Let $Q_c(\boldsymbol{\beta})$ denote the objective MD-Lasso seeks to minimize:*

$$Q_c(\boldsymbol{\beta}) = -c \log\left(\sum_{i=1}^{n} \exp\left(-\frac{1}{2c}(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2\right)\right) + \lambda\|\beta\|_1$$

For any finite choice of the scaling parameter c, consider the non-empty set

$$\mathcal{B}_c = \{\boldsymbol{\beta} : \boldsymbol{\beta} \text{ is a local optimum for the MD-Lasso problem and } Q_c(\boldsymbol{\beta}) \leq Q_c(0).\}$$

For every $\alpha \in (0, 1)$,

the breakdown point of any solution in $\mathcal{B}_c$ is at least α. Namely the MD-Lasso can tolerate at least αn arbitrarily corrupted observations and still produce estimates with bounded ℓ_2-norm.

Theorem 1 indicates that even if a majority of observations are corrupted, the estimated regression coefficients will remain bounded. Naturally, if more than 50% of observations are arbitrarily corrupted, it makes no sense to trust *any* model and thus the breakdown point is typically capped at 50%. We can not make a statement about the breakdown point of all local solutions but can show a high breakdown point for the best local solutions, in the sense that the local solutions in a specific levelset of the objective function have a high breakdown point. Recall that as $c \to 0$ the MD-Lasso yields a special case of sparse trimmed least squares regression, where all but the most trusted observation are disregarded. Our results for this limit case are consistent with those on sparse trimmed regression [1]. As $c \to \infty$, MD-Lasso is equivalent to the Lasso estimator. If one sets $c \to \infty$ in the proof of theorem 1, the reasoning guaranteeing high breakdown for MD-Lasso is no longer valid. This is to be

expected: for the traditional Lasso estimator it was shown in Alfons et al. [1] that only one outlier can already send the estimates to infinity and the breakdown point is $1/n$. Finally, we note that the LAD Lasso estimator also shares the poor breakdown point property of Lasso; the absolute loss does not help in improving the breakdown point.

The results of Theorem 1 pertain to *arbitrary* corruptions in *both* the data matrix $\boldsymbol{X}$ and the response vector $\boldsymbol{Y}$. In the next section we further characterize the robustness of MD-Lasso from a different standpoint. Specifically, we consider errors in $\boldsymbol{Y}$ that are incurred via the error term $\boldsymbol{\eta}$, and show that under mild assumptions on the distribution of the error term $\boldsymbol{\eta}$, MD-Lasso achieves consistency with fast convergence rates.

3. Main results

In this section we establish the conditions for consistency and fast convergence rates of the MD-Lasso estimator under the high-dimensional setting ($p \gg n$). The proofs of our results are all relegated to the Appendix. Consistency and fast convergence rates can be secured thanks to two key properties: (i) the *restricted strong convexity* of the loss L in the neighborhood of the true model parameter vector and (ii) the *gradient boundedness* at the true model parameter vector. The importance of these two properties was first identified in [30]. Before defining and establishing them, we introduce some notation and state the assumptions required by our analysis.

Notation. Define t_γ to be the cumulative distribution function of $|\eta_i|$ such that for all $\gamma \geq 0$,

$$t_\gamma := P(|\eta_i| \geq \gamma).$$

Let S denote the set of indices corresponding to the support of the true coefficient vector $\boldsymbol{\beta}^\star$. Writing $\boldsymbol{\Delta}_S$ for the projection of a vector $\boldsymbol{\Delta} \in \mathbb{R}^p$ onto indices S, and $\boldsymbol{\Delta}_{S^C}$ for the projection onto the complement of S, define the cone

$$C(S) := \{\boldsymbol{\Delta} \in \mathbb{R}^p \mid \|\boldsymbol{\Delta}_{S^C}\|_1 \leq 3\|\boldsymbol{\Delta}_S\|_1\}.$$

Assumptions. We make the following assumptions throughout.
[*A*1] Bounded predictors: there exists $M < \infty$ such that $|X_i^j| \leq M$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$.
[*A*2] The error terms $(\eta_i)_{i=1}^n$ form a sequence of independent and identically distributed random variables, with zero-mean or, if the mean is undefined, a probability density function symmetric around zero.
[*A*3] The design matrix $\boldsymbol{X}$ satisfies the following Restricted Eigenvalue condition

$$\frac{\|\boldsymbol{X}\boldsymbol{\Delta}\|_2^2}{n} \geq \kappa_{RE}\|\boldsymbol{\Delta}\|_2^2, \text{for all } \boldsymbol{\Delta} \in C(S) \tag{6}$$

with constant $\kappa_{RE} > 0$.
[*A*4] For every $X_{i=1,\ldots,n}$, the variable $\langle v, X_i \rangle$ is sub-Gaussian with parameter at most $\kappa_u^2\|v\|_2^2$.

Assumptions [A1] and [A4] could be relaxed in some ways but we are mainly interested here in robustness with respect to outliers in the target. The second assumption [A2] is weak since it allows arbitrarily heavy tails in the error distribution, while the last assumption [A3] is standard, see, for example, Bickel et al. [12].

3.1. Gradient boundedness property

The following results provide upper-bounds on the ℓ_∞-norm of the gradient of the MD-Lasso loss evaluated at $\boldsymbol{\beta}^\star$. These bounds are the most important part when establishing rates of convergence.

Lemma 1. *Under Assumptions* [A1] *and* [A2] *for any* $\gamma \leq \sqrt{c}$ *let* $\xi_{c,\gamma} \geq 0$ *be given by*

$$\xi_{c,\gamma}^2 = M^2 t_1^{-2}\big[(1-2t_\gamma)\gamma^2 \exp(-\gamma^2/c) + 2ct_\gamma/e\big]e^{1/c},$$

Then, for some positive constants α_1, α_2,

$$P\left(\|\nabla L(\boldsymbol{\beta}^\star)\|_\infty \leq \xi_{c,\gamma}\sqrt{\frac{\log p}{n}}\right) \geq 1 - \alpha_1 \exp(-\alpha_2 \xi_{c,\gamma}^2 \log p).$$

A proof is given in the Appendix.

Lemma 2. *Under Assumptions* [A1] *and* [A2] *let* $\zeta_c \geq 0$ *be given by*

$$\zeta_c^2 = 4M^2 t_1^{-2} E[\eta_i^2 e^{-\eta_i^2/c}]e^{1/c}.$$

Then, for some positive constants $\alpha_1', \alpha_2', \alpha_3'$,

$$P\left(\|\nabla L(\boldsymbol{\beta}^\star)\|_\infty \leq \zeta_c\sqrt{\frac{\log p}{n}}\right) \geq 1 - \alpha_1' \exp\Big(-\alpha_2'\zeta_c^2 \log p \frac{1-\alpha_3'\sqrt{c\log p/n}}{1+\alpha_3'\sqrt{c\log p/n}}\Big).$$

A proof is given in the Appendix.

Remarks. We note that Lemma 1 rests on establishing the bounded differences property of the gradient coordinates, based on whether or not the amplitude of the error η_i exceeds $\sqrt{c}$. Lemma 2 employs Bernstein's inequality [23], noting that the variance of $\eta_i \exp(-\eta_i^2/(2c))$ is always well-defined regardless of whether or not the variance of η_i exists.

As $c \to \infty$ the bound in Lemma 1 becomes vacuous: it essentially scales with $\sqrt{c}$. For heavy-tailed distributions, this is an accurate indication that large values of c are not an option, as this would essentially mean giving up on the robustness property of the estimator. For lighter-tailed distributions for which the variance of η_i exists (and is finite), Lemma 2 is preferred. Since $\zeta_c \to 4M^2 t_1^{-2} E[\eta_i^2]$ for $c \to \infty$ (by the monotone convergence theorem), Lemma 2 yields finite upper-bounds if $c \to \infty$ but with a rate depending on n such that $c \log p/n \to 0$. We present below some specific examples illustrating this interesting fact. If the variance of η_i is undefined Lemma 1 yields tighter bounds for large values of c.

Examples. GAUSSIAN ERRORS. If the error terms η_i, $i = 1, \ldots, n$ follow a Gaussian distribution $N(0, \sigma^2)$ and c is finite, then Lemma 2 implies that with high probability

$$\|\nabla L(\boldsymbol{\beta}^\star)\|_\infty \leq 2M\sigma \frac{\left(\frac{c}{2\sigma^2+c}\right)^{3/4} e^{1/2c}}{t_1} \sqrt{\frac{\log p}{n}}.$$

If $c \to \infty$ while $c \log p/n \to 0$, we recover the condition for the traditional Lasso (up to a constant factor) namely:

$$\|\nabla L(\boldsymbol{\beta}^\star)\|_\infty \leq 2M\sigma \sqrt{\frac{\log p}{n}}$$

This is consistent with the fact that the MD-Lasso estimator yields the traditional Lasso estimator as $c \to \infty$.

LAPLACE-DISTRIBUTED ERRORS. If the error terms η_i, $i = 1, \ldots, n$ follow a Laplace distribution with scale parameter b, and c is finite, then Lemma 2 implies that with high probability

$$\|\nabla L(\boldsymbol{\beta}^\star)\|_\infty \leq$$
$$\frac{2Me^{1/2c}}{t_1} \sqrt{-\frac{c^2}{4b^2} + \frac{\sqrt{2\pi}}{b} \left(\frac{c}{2}\right)^{3/2} e^{\frac{c}{4b^2}} \left(1 + \frac{c}{2b^2}\right) \bar{F}\left(\frac{1}{b}\sqrt{\frac{c}{2}}\right)} \sqrt{\frac{\log p}{n}},$$

where $\bar{F}(\cdot)$ denotes the tail probability function of the standard normal distribution.

If $c \to \infty$ while $(c \log p)/n \to 0$, Lemma 2 together with the monotone convergence theorem yields the condition

$$\|\nabla L(\boldsymbol{\beta}^\star)\|_\infty \leq \sqrt{8} \frac{Mb}{t_1} \sqrt{\frac{\log p}{n}}.$$

We will use these gradient bounds to show fast convergence rates of the MD-loss under potentially heavy-tailed distributions.

3.2. Restricted strong convexity

The following lemma states conditions that guarantee the restricted strong convexity of the MD-Lasso loss in a restricted neighborhood of the true model coefficients $\boldsymbol{\beta}^\star$.

Lemma 3. *Under Assumptions* [A1], [A3], [A4] *for any* $\mu < \sqrt{c}/(4\kappa_u\sqrt{\log n})$, *consider the set* $K(S, \mu) = \{\boldsymbol{\Delta} \in C(S) : \|\boldsymbol{\Delta}\|_2 = \mu\}$. *Let* $\lambda_\mu \in (0, (\sqrt{c} - 4\mu\kappa_u\sqrt{\log n})/2]$. *If the model error distribution satisfies the tail condition:*

$$t_{\lambda_\mu} < \left(1 + \frac{64}{21} e^{-\frac{3}{2}}\right)^{-1}$$

then for all $\Delta \in K(S,\mu)$ *it holds that*

$$L(\boldsymbol{\beta}^\star + \boldsymbol{\Delta}) - L(\boldsymbol{\beta}^\star) - \langle \nabla L(\boldsymbol{\beta}^\star), \boldsymbol{\Delta} \rangle \;\geq\; \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \sqrt{\frac{\log p}{n}} \|\boldsymbol{\Delta}\|_1 \|\boldsymbol{\Delta}\|_2 \quad (7)$$

with probability at least $1 - \alpha_1 \exp(-\alpha_2 n)$, *for some* $\alpha_1, \alpha_2 > 0$, *where* $\kappa_1 = \frac{1}{4}\kappa_{RE}(C(1 - t_{\lambda_\mu}) - 2e^{-\frac{3}{2}})$ *and* $\kappa_2 = \frac{49}{2} C \kappa_u^2 \sqrt{\log n}$, *with* $C = (21/32) + 2e^{-3/2}$.

A proof is given in the Appendix.

Remarks. Noting that for any $\boldsymbol{\Delta} \in C(S)$, $\|\boldsymbol{\Delta}\|_1 \leq 4\|\boldsymbol{\Delta}_S\|_1 \leq 4\sqrt{s}\|\boldsymbol{\Delta}\|_2$, the bound (7) implies that

$$L(\boldsymbol{\beta}^\star + \boldsymbol{\Delta}) - L(\boldsymbol{\beta}^\star) - \langle \nabla L(\boldsymbol{\beta}^\star), \boldsymbol{\Delta} \rangle \geq \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2 \quad (8)$$

as long as $n > 64(\kappa_2/\kappa_1)^2 s \log p$.

Lemma 3 indicates that the region of restricted strong convexity is controlled by parameter c via the condition $\|\boldsymbol{\Delta}\|_2 \leq \mu$ where $\mu < \sqrt{c}/(4\kappa_u\sqrt{\log n})$. For fixed c, we remark that when n is increasing the region of restricted strong convexity (or μ) will shrink due to the $\log n$ dependency. However, as will be clarified in the next section, this does not compromise convergence rates and consistency of the estimator: when n is increasing, the region within which restricted strong convexity is required to hold is also shrinking with a rate that is faster.

The convexity of the loss rests on a condition related to the tail of the error distribution, which is required so that $\kappa_1 > 0$. We consider a specific example to illustrate that the requirements on the tail of the error distribution are very mild.

Example. Let $\mu = c^{1/4}/(2\sqrt{\log n}\kappa_u)$, $\lambda_\mu = \sqrt{c}/2 - c^{1/4}$. Hence c must be chosen so that $P(|\eta_i| > \sqrt{c}/2 - c^{1/4}) \leq 0.59$. This translates into the conditions

$c > 2.42$ for the Laplace(0,1) distribution,
$c > 2.45$ for the Normal(0,1) distribution,
$c > 2.47$ for the Student's t distribution with 4 degrees of freedom, and
$c > 2.58$ for the Cauchy(0,1) distribution.

These conditions are quite similar. Nevertheless, except for the Laplace distribution, the heavier the tail, the larger the lower bound on c needs to be in order to secure restricted strong convexity, which makes sense as the number of large outliers is expected to increase. It is important to note, however, that while a large value of c extends the convexity region, it reduces the resilience to outliers (via the gradient bound). Thus the choice of c is key in guaranteeing both fast convergence rates and robustness, as shall be made explicit in the next section.

We conclude this section by noting that under similar tail conditions on the error, the MD-Lasso loss function is (simply) convex with asymptotic probability 1 in the set $\mathcal{H}_c = \{\boldsymbol{\beta}^\star + \boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 < \sqrt{c}/(12\kappa_u\sqrt{\log n})\}$.

The proof is similar to that of Lemma 3 and is thus omitted.

3.3. Consistency results

We now state the results on the consistency and convergence rates for the MD-Lasso estimator. These results leverage the restricted strong convexity and gradient boundedness properties.

3.3.1. Consistency in the local convexity region

In this section we provide error bounds for the solutions of MD-Lasso which reside within the local convexity region. The following theorem builds on the gradient bound of Lemma 1, and is thus preferred if the variance of the error is undefined. In the Appendix, we also provide a second theorem (Theorem 4) that uses the gradient bound of Lemma 2 and is thus preferred for errors with finite variance.

Theorem 2. *Consider the linear regression model* (1) *and assume that the support of the true model coefficients* $\boldsymbol{\beta}^\star$ *has cardinality* s. *Let* $\mathcal{H}_c = \{\boldsymbol{\beta}^\star + \boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_2 < \sqrt{c}/(12\kappa_u\sqrt{\log n})\}$

Under Assumptions $[A1 - A4]$, *for any* $\gamma \leq \sqrt{c}$, *with* c *such that* $t_{\sqrt{c}/2} < (1 + (64/21)e^{-3/2})^{-1} < 0.6$ *given the MD-Lasso estimator* (5) *with scaling parameter* c *and regularization parameter* $\lambda_n = 2\xi_{c,\gamma}\sqrt{\log p/n}$, *where* $\xi^2_{c,\gamma} = M^2 t_1^{-2}\big[(1 - 2t_\gamma)\gamma^2 \exp(-\gamma^2/c) + 2ct_\gamma/e\big]e^{1/c}$, *any of the solutions in* $\mathcal{H}_c$ *(there is at least one such solution) satisfies*

$$\|\hat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^\star\|_2 \;\leq\; \frac{32\xi_{c,\gamma}}{\big(C(1 - t_{\sqrt{c}/2}) - 2e^{-\frac{3}{2}}\big)\kappa_{RE}}\sqrt{\frac{s\log p}{n}} \tag{9}$$

with probability at least $1 - \alpha_1\exp(-\alpha_2 n\lambda_n^2)$, *for any* $n > \max\big\{\tilde{\xi}_{c,\gamma}\frac{s}{c}\log p\log n, 64(\kappa_2/\kappa_1)^2 s\log p\big\}$, *where* $C = 21/32 + 2e^{-3/2} \approx 1.1$, $\tilde{\xi}_{c,\gamma} = \alpha_3\frac{\kappa_u^2}{\kappa_1^2}\xi^2_{c,\gamma}$, *and positive constants* $\alpha_1, \alpha_2, \alpha_3$.

A proof is given in the Appendix. Both Theorem 2 and Theorem 4 demonstrate that MD-Lasso is robust with respect to errors in $\boldsymbol{Y}$. The results rest on mild assumptions on the quantiles of the error distribution η.

The bounds on $\|\hat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^\star\|_2$ in Theorem 2 and Theorem 4 scale inversely with the restricted strong convexity constant of Lemma 3. This makes sense as the constant reflects the curvature of the loss function L in a restricted set of directions around the true solution $\boldsymbol{\beta}^\star$: the higher the curvature the faster the convergence. On the other hand, the convergence rates and the regularization parameter λ_n are proportional to the gradient bound of Lemmas 1 and 2. While restricted strong convexity favors large values of c, the gradient bound favors small values, hence the "tension" between the two that we now elaborate upon.

Figure 3 depicts the impact of c on the scaling factor in the convergence rates of (9) and (48) for various error distributions. The values of the y-axis do not reflect the multiplicative factors that do not depend on c and the error distribution. Note that to generate the figure, γ was varied, and t_γ determined

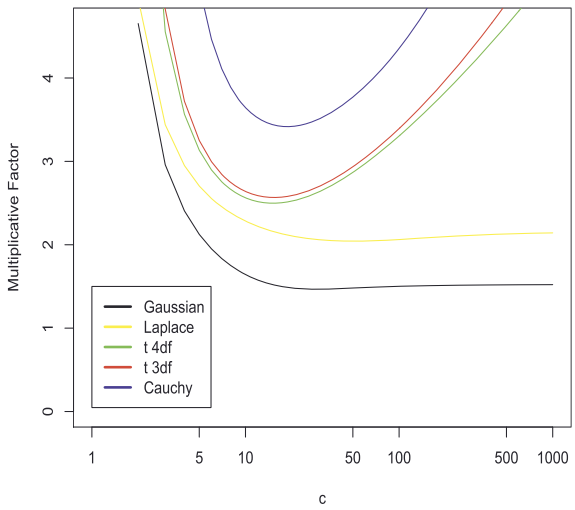


FIG 3. *Impact of the error distribution on the scaling of the convergence rate for* $\|\hat{\beta} - \beta^\star\|_2$ *as a function of* c*. The values of the* y*-axis correspond to the contributions from the error term; multiplicative factors which are independent of* c *are disregarded.*

numerically given the error distribution, then the value with the 'best' γ was selected. The convergence rates along with Figure 3 suggest the following points:

- Regardless of the error distribution, one should not set c to values of much below a value of, say, 5. This makes sense intuitively, since a small c means that we are not using many observations and operating on too small a sample size. Recall that as $c \to 0$ the estimator acts as a trimmed ℓ_1- penalized Least Squares estimator where all but one observations are trimmed out.
- As c grows beyond an "optimum value"[2], the heavier the tail, the faster the multiplicative factor grows. This is aligned with the intuition that one should be more conservative the heavier the tail, and thus not set c too large. Recalling that the MD-Lasso estimator is equivalent to the traditional Lasso estimator as $c \to \infty$, our results also corroborate the fact that the performance of the traditional Lasso estimator degrades dramatically in the presence of heavy-tailed noise.
- Interestingly, for lighter-tailed distribution (e.g. Laplace and Gaussian) the multiplicative factor flattens out and converges to a finite value as $c \to \infty$, provided that c grows in a sample size dependent fashion so that $c \log p/n \to 0$. In particular, for sub-gaussian tails one recovers the results of the traditional Lasso estimator (up to a constant factor) as $c \to \infty$.
- Robustness does not cause significant loss of estimation efficiency in the absence of outliers. This will be confirmed in the simulations of Section 5,

[2]Here we use the term "optimum" in a loose sense as our bounds may not be tight around the actual optimum.

e.g. under the Gaussian setting.

We also want to make a few remarks about global and local solutions. Recall that the MD-Lasso estimator is invex and locally convex but not globally convex. If the constraint region induced by the ℓ_1-penalty resides within the local convexity region, the invexity of the loss is not compromised, every local minimum lies within the local convexity region and is also a global minimum. The results of Theorem 2 and Theorem 4 therefore apply to *any* solution of the MD-Lasso estimator. A sufficient condition for this case to hold is that $\|\boldsymbol{\beta}^\star\|_2 + \bar{\lambda} < \sqrt{c}/(12\kappa_u\sqrt{\log n})$ where $\bar{\lambda}$ is the parameter corresponding to parameter λ in the constrained version of the MD-Lasso problem[3]: minimize $L(\beta)$ s.t. $\|\beta\|_1 \leq \bar{\lambda}$.

If the constraint region induced by the ℓ_1-penalty merely intersects (or even resides outside of the local convexity region), local minima may exist outside of the convexity region.

3.3.2. Consistency within a safety radius

The next set of results does not require the solutions to lie within the convexity region. Here the original MD-Lasso estimator is slightly modified to introduce a "safety" radius for β. Namely we consider

$$\hat{\boldsymbol{\beta}}_{\lambda_n} = \arg \min_{\|\boldsymbol{\beta}\|_2 \leq b_0} \left(L(\boldsymbol{\beta}) + \lambda_n \|\boldsymbol{\beta}\|_1\right), \tag{10}$$

where the safety radius b_0 is chosen such that $\boldsymbol{\beta}^\star$ is feasible. A key benefit of the following results is their practical impact: in Section 4 we will see that a simple incremental algorithm yields consistent estimates.

Since the solutions need not belong to the local convexity region of MD-Lasso, we do not make use of Assumption [A3]. Instead we introduce the following assumption: [$A3'$] The rows of the design matrix $\boldsymbol{X}$ are independently drawn from $N(0, \Sigma)$ where Σ has the minimum eigenvalue $\lambda_{\min}(\Sigma)$.

The following theorem guarantees that any of the local solutions are consistent. The theorem is obtained by adapting the work of [24], noting that global convexity is actually not required for ℓ_2 consistency (see the Appendix for more details).

Theorem 3. *Consider the linear regression model* (1) *and assume that the support of the true model coefficients* $\boldsymbol{\beta}^\star$ *has cardinality* s. *Let* $\widetilde{\boldsymbol{\beta}}_{\lambda_n}$ *be any local optima of* (10) *under the assumptions of Theorem 2. Suppose that the scale parameter* c *and the radius parameter* b_0 *of* (5) *are selected so that (i) the model error distribution can satisfy the tail condition:* $t_{\sqrt{c}/2} < (1+((64/21)e^{-3/2})^{-1} < 0.6$, *and (ii)* $\|\boldsymbol{\beta}^\star\|_2 \leq b_0 \leq \sqrt{c}/(8\kappa_u\sqrt{\log n})$. *Also suppose that* λ_n *is set as* $2\max\left\{\xi_{c,\gamma}\,,\, 2\kappa_2 b_0\right\}\sqrt{\log p/n}$ *where* $\kappa_1 = \frac{1}{32}\big(\lambda_{\min}(\Sigma)\big)^2(C(1-t_{\lambda_\nu}) - 2e^{-2/3})$,

[3]Note that a mapping between this problem and the original penalized MD-Lasso problem is possible due to the invexity of the loss and properties of the ℓ_1-norm [2].

$\kappa_2 = 49C\kappa_u^2\sqrt{\log n} + \frac{9}{4}\lambda_{\min}(\Sigma)\sqrt{\max \Sigma_{jj}}$, *and* $\xi_{c,\gamma}^2$, C *are as defined in Theorem 2. Then, with probability at least* $1 - \alpha_1 \exp(-\alpha_2 n\lambda_n^2)$, *for some universal positive constants* α_1 *and* α_2, *the local optimal error is guaranteed to be consistent:*

$$\|\widetilde{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^\star\|_2 \leq \frac{3}{\kappa_1} \max\left\{\xi_{c,\gamma}\,,\, 2\kappa_2 b_0\right\}\sqrt{\frac{s\log p}{n}}.$$

A proof is given in the appendix. We note that in order to have a valid selection for b_0 from the condition (ii) in the theorem, $\sqrt{c}$ should at least scale with $\|\boldsymbol{\beta}^\star\|_2\sqrt{\log n}$. This is the cost of the non-convexity. However, the cost is mitigated when s and $\|\boldsymbol{\beta}^\star\|_\infty$ are bounded since $\|\boldsymbol{\beta}^\star\|_2 \leq \sqrt{s}\|\boldsymbol{\beta}^\star\|_\infty$.

4. Optimization and parameter tuning

We first describe an incremental method. A connection with re-weighted least squares follows thereafter. We conclude this section by describing a procedure for parameter tuning.

4.1. Incremental algorithm

We first show an incremental method. The need for solving very large problems has led to a recent resurgence of interest in first-order optimization methods, such as the composite gradient method of Nesterov [31] and the incremental methods of Bertsekas [11] (adopted e.g. in [35, 28]). We focus on the MD-Lasso objective with "safety" radius of (10).

The algorithm proceeds with the following updates

$$\boldsymbol{\beta}^{(t+1)} = \Pi_{\{\|\boldsymbol{\beta}\|_2 \leq b_0\}}\Big(\arg\min_{\boldsymbol{\beta}}\Big\{L(\boldsymbol{\beta}^{(t)}) + \langle \nabla L(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\rangle + \frac{\rho}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\Big\}\Big), \tag{11}$$

where $\Pi_{\{\|\boldsymbol{\beta}\|_2 \leq b_0\}}$ denotes the projection onto the ℓ_2 ball of radius b_0. It is important to note that the algorithm is guaranteed to converge to a local minimum even if the loss L is not convex [11] and regardless of the initialization. Hence, the algorithm can be readily instantiated for MD-Lasso.

Instantiation of the incremental algorithm for MD-Lasso. Denote by S the *soft-thresholding operator* defined as

$$S_\lambda(\boldsymbol{u}) = \text{sign}(\boldsymbol{u})\max(|\boldsymbol{u}| - \lambda, \mathbf{0}), \tag{12}$$

where all operations are applied element-wise on a vector u. Each incremental algorithm update can be computed by (at most) two simple operations. The

first operation consists in computing $p_\rho(\boldsymbol{\beta}^{(t)}) = S_{\lambda_n/\rho}(\boldsymbol{\beta}^{(t)} - \frac{1}{\rho}\nabla L(\boldsymbol{\beta}^{(t)}))$. Let $r_i = y_i - \boldsymbol{X}_i\boldsymbol{\beta}^{(t)}$ denote the residual for the i^{th} sample, and define

$$w_i = \frac{\exp\left(-\frac{1}{2c}r_i^2\right)}{\sum_{j=1}^n \exp\left(-\frac{1}{2c}r_j^2\right)}. \tag{13}$$

Let $\tilde{\boldsymbol{R}} = (\tilde{r}_1, \ldots, \tilde{r}_n)'$ where $\tilde{r}_i = r_i w_i$. Then $\nabla L(\boldsymbol{\beta}^{(t)}) = -\boldsymbol{X}'\tilde{\boldsymbol{R}}$, and $\tilde{\boldsymbol{R}}$ can be interpreted as a generalized residual. The thresholding operation boils down to the following simple step:

$$p_\rho(\beta^{(t)}) = S_{\lambda_n/\rho}(\boldsymbol{\beta}^{(t)} + \frac{1}{\rho}\boldsymbol{X}'\tilde{\boldsymbol{R}}). \tag{14}$$

If the ℓ_2-norm of the projection exceeds the safety radius, a second operation has to be carried out to project onto the the ℓ_2 ball of radius b_0. The overall procedure is summarized by Algorithm 1.

Algorithm 1 Incremental Algorithm for MD-Lasso

Initialize $\beta^{(0)}$
repeat
  Given $\beta^{(t-1)}$, compute $w^{(t)}$ from (13) and the corresponding generalized residuals $\tilde{\boldsymbol{R}}^{(t)}$
  Update. $\beta^{(t)} \leftarrow p_\rho(\beta^{(t-1)})$ as in (14)
until stopping criterion is satisfied.

Consistency of the local optima found by the incremental algorithm. The incremental algorithm applied to the objective of (10) is guaranteed to converge to a local optimum [11]. Hence theorem 3 shows that *any* local optimum obtained by the incremental algorithm is consistent, regardless of the initialization.

4.2. Re-weighted penalized least squares

Some interesting insights on the robustness of MD-Lasso can be gained by examining the descent direction in the incremental procedure, as explicated in (14). Given an initial solution $\boldsymbol{\beta}$, under the traditional squared loss one would get $\nabla_j L(\boldsymbol{\beta}) = -\frac{1}{n}\sum_{i=1}^n X_i^j(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})$. For the MD-Lasso we have $\nabla_j L(\boldsymbol{\beta}) = -\sum_{i=1}^n w_i X_i^j(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})$, where the weights w_i are given in (13). Hence the descent direction for MD-Lasso can be seen as a "weighted version" of the direction for usual squared loss, where the weights w_i can be interpreted as being proportional to the likelihood functions of individual data points, i.e., $w_i = \frac{\mathcal{L}(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})}{\sum_{i=1}^n \mathcal{L}(Y_i|\boldsymbol{X}_i;\beta)}$, where $\mathcal{L}$ denotes the likelihood function under Gaussian assumption. Thus data with high likelihood values are given more weights in the computation of the descent direction. Conversely, data with low likelihood values, which are more likely to be outliers, contribute less. The connection between the likelihood functions and weights provides an intuitive insight on the resilience of the original MD-Lasso to outliers.

We remark that a similar conclusion can be obtained by considering a first order approximation of the "log-sum-exp" term in (5) around an initial solution. This yields an approximate iterative procedure where given initial estimates, data are first re-weighted by w_i in (13) and then passed to a traditional Lasso solver to provide new estimates, and the procedure is repeated until convergence. The following algorithm summarizes the procedure.

Algorithm 2 Approximate MD-Lasso as Iteratively Reweighted Lasso

Step1: Given initial estimate $\boldsymbol{\beta}^{(0)}$ for the regression coefficients (e.g. using ridge regression) compute weights $w_i = \frac{\exp(-\frac{1}{2c} r_i^2(\boldsymbol{\beta}^{(0)}))}{\sum_{i=1}^{n} \exp(-\frac{1}{2c} r_i^2(\boldsymbol{\beta}^{(0)}))}$, where $r_i(\boldsymbol{\beta}^{(0)}) = (Y_i - \boldsymbol{X}_i'\boldsymbol{\beta}^{(0)})$.

Step 2: Estimate $\boldsymbol{\beta}$ by minimizing

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} w_i (Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_1.$$

Step 3: If $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}\|^2 \leq \epsilon$ stop. Else, set $\boldsymbol{\beta}^{(\mathbf{0})} = \hat{\boldsymbol{\beta}}$ and go back to Step 1.

While the weighted least squares formulation illustrates most intuitively the robustness of the MD-Lasso loss, it requires running several individual Lasso problems. Even though the procedure can benefit from a warm start in each iteration, it is is computationally more intensive than running the incremental approach. We therefore do not adopt it in practice and instead prefer the incremental approach. Intuitively, the incremental approach can be interpreted as a "lazy update" version of the iteratively reweighted least squares.

4.3. Parameter tuning

In practice, the scaling parameter c is unknown and rather than guessing its value, one might wish to automatically select it, along with the regularization parameter λ. The challenge is that MD-Lasso (unpenalized) losses for different values of c, denoted by L_c, are not directly comparable. This is because $c_1 < c_2$ implies $L_{c_1}(\boldsymbol{\beta}) > L_{c_2}(\boldsymbol{\beta})$. Hence the selection criterion should penalize large values of c somehow. A rigorous criterion can be found by inspecting (3) and (4), and reintroducing the term $\int f^2(Y|\boldsymbol{X};\boldsymbol{\beta})dY = 1/(2\pi^{1/2}\sigma)$ to obtain

$$d_n(\beta) + \int f^2(Y|\boldsymbol{X};\boldsymbol{\beta})dY = \frac{1}{2\sqrt{\pi}\sigma} - \frac{2}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2).$$

Based on this consideration, we propose as selection criterion the minimization of the following loss on evaluation data

$$\tilde{L}_{\text{eval}}(\boldsymbol{\beta}) = \frac{1}{\sqrt{c}} \left(\frac{1}{2^{3/2}} n_{\text{eval}} - \sum_{i=1}^{n_{\text{eval}}} \exp\left(-\frac{(Y_i - \boldsymbol{X}_i'\boldsymbol{\beta})^2}{2c} \right) \right),$$

where n_{eval} is the evaluation sample size.

The overall tuning procedure can then be summarized as follows. For each c in a grid of candidate scaling parameters and each λ in a grid of candidate tuning parameters, solve the MD-Lasso problem on training data and obtain the estimate $\hat{\boldsymbol{\beta}}_{c,\lambda}$. Compute $\tilde{L}_{\text{eval}}(\hat{\boldsymbol{\beta}}_{c,\lambda})$. Pick the pair (c, λ) with smallest loss $\tilde{L}_{\text{eval}}$. The procedure can be naturally extended to cross-validation.

5. Numerical results

We compare the proposed MD-Lasso estimator with the LAD-Lasso [43], the Extended Lasso [32] and the traditional Lasso [38]. We also present a strategy to automatically select the MD-Lasso scaling parameter c.

5.1. Simulation results

Model setup. We simulated data from the linear regression model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}^{\star} + \boldsymbol{\eta}.$$

For the predictors, we consider two data generation models:

(P1) Toeplitz design: The $n \times p$ predictor matrices $\boldsymbol{X}$ have rows sampled independently from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_X)$ where $(\boldsymbol{\Sigma}_X)_i^j = 0.5^{|i-j|}$.
(P2) Factor model with two factors: let ϕ^1 and ϕ^2 be two latent variables following i.i.d. standard normal distributions. Each predictor variable X^k, for $k = 1, , p$, is generated as $X^k = f^{k,1}\phi^1 + f^{k,2}\phi^2 + \epsilon^k$, where f_k^1, f_k^2 and ϵ^k have i.i.d. standard normal distributions for all $k = 1, \ldots, p$.

For the error term distribution, we consider five cases:

(E1) Normal: $\eta \sim \mathcal{N}(0, 1)$.
(E2) Laplace: $\eta \sim \text{Laplace}(0, 1)$.
(E3) Mixture of Gaussians: $\eta \sim \frac{h\mathcal{N}(0,1)+(1-h)\mathcal{N}(0,\sqrt{225})}{\sqrt{0.9*1+0.1*225}}$ where $h \sim \text{Bernoulli}(0.9)$.
(E4) Student's t with degrees of freedom 4: $\eta \sim \text{Student}(0, 4)$.
(E5) Cauchy: $\eta \sim \text{Cauchy}(0, 1)$.

In each simulation study, we consider both $n = 200$ and $n = 1000$ observations, and $p = 1000$ predictors. The entries of true model coefficient vector $\boldsymbol{\beta}^{\star}$ are set to be 0 everywhere, except for a randomly chosen subset of s coefficients, which are chosen independently and uniformly in $(1, 3)$. The size s of the set of non-zero coefficients is randomly set between 3 and 10.

Parameter tuning. We consider holdout-validated estimates, which are obtained by selecting the tuning parameter λ that minimizes the average loss on a validation set. In a first set of experiments, we hold the parameter c fixed at various values so as to examine the impact of c for various error distributions. In a second set of experiments, c is selected automatically along with λ as described in Section 4.3.

Performance metrics. To measure the estimation accuracy, we report the model error defined as

$$ME(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^\star) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star)' \boldsymbol{\Sigma}_X (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star).$$

To measure variable selection accuracy, we use the F_1 score [40] defined by $F_1 = 2PR/(P+R)$, where P is precision (fraction of correctly selected variables among selected variables) and R is recall (fraction of correctly selected variables among true relevant variables).

Results. For each setting, we present the average of the performance measure based on 100 simulations. Figure 4 and Figure 5 provide boxplots for the Model Error and the variable selection accuracy, respectively of MD-Lasso. In the figures MD-x denotes MD-Lasso with $c = x$ and $x = 1, 2, 5, 10, 25, 50, 100$, Lasso denotes the Least Squares Lasso, which is the limiting case of MD-Lasso for $c \to \infty$.

From the figures we can see that the simulations results are in agreement with the theoretical results of Section 3. Specifically if the scaling parameter is too small, the performance of the MD-Lasso method degrades, as the restricted strong convexity property is violated. As expected, the performance of MD-Lasso gets closer to that of Lasso as c becomes large. For light tail distributions (e.g. Gaussian and Laplace) we see that as long as c is larger or equal to the minimum value required for restricted strong convexity, the performance of MD-Lasso is quite insensitive to the choice of c, while the sensitivity increases for heavy tailed distributions (e.g. Student's t and Cauchy). It is intriguing to note that in many cases the variable selection accuracy of MD-Lasso decreases monotonically with the scaling parameter, suggesting that the restricted strong convexity of the loss might be more influential on the model error than on the variable selection accuracy.

Figure 6 and Figure 7 provide boxplots for the Model Error and the variable selection accuracy, respectively of MD-Lasso with automatic selection of the scaling parameter c (denoted by MD in the figures), and comparison methods. Lasso denotes the Least Squares Lasso, LAD denotes the Least Absolute Deviation Lasso, and ExLasso denotes the Extended Lasso.

For Laplace distributed errors, LAD-Lasso performs the best in terms of model error. This can be explained by the fact that the noise distribution matches the LAD-Lasso loss exactly. However, MD-Lasso achieves higher variable selection accuracy. Even for Gaussian errors, MD-Lasso is often able to outperform the Least Squares Lasso in terms of variable selection accuracy. This might be partly attributable to a better parameter selection by MD-Lasso. Indeed we noticed that MD-Lasso tends to select fewer variables than standard Lasso. For instance, under Gaussian errors, Toeplitz design, with $n = 100$, and $p = 1000$, MD-Lasso selects on average 8 variables while Lasso selected over 11 variables.

The results for Cauchy distributed errors underscore the need for a non-convex loss function as offered by MD-Lasso, and the limited ability of convex loss functions (including LAD-Lasso) in dealing with very large outliers.

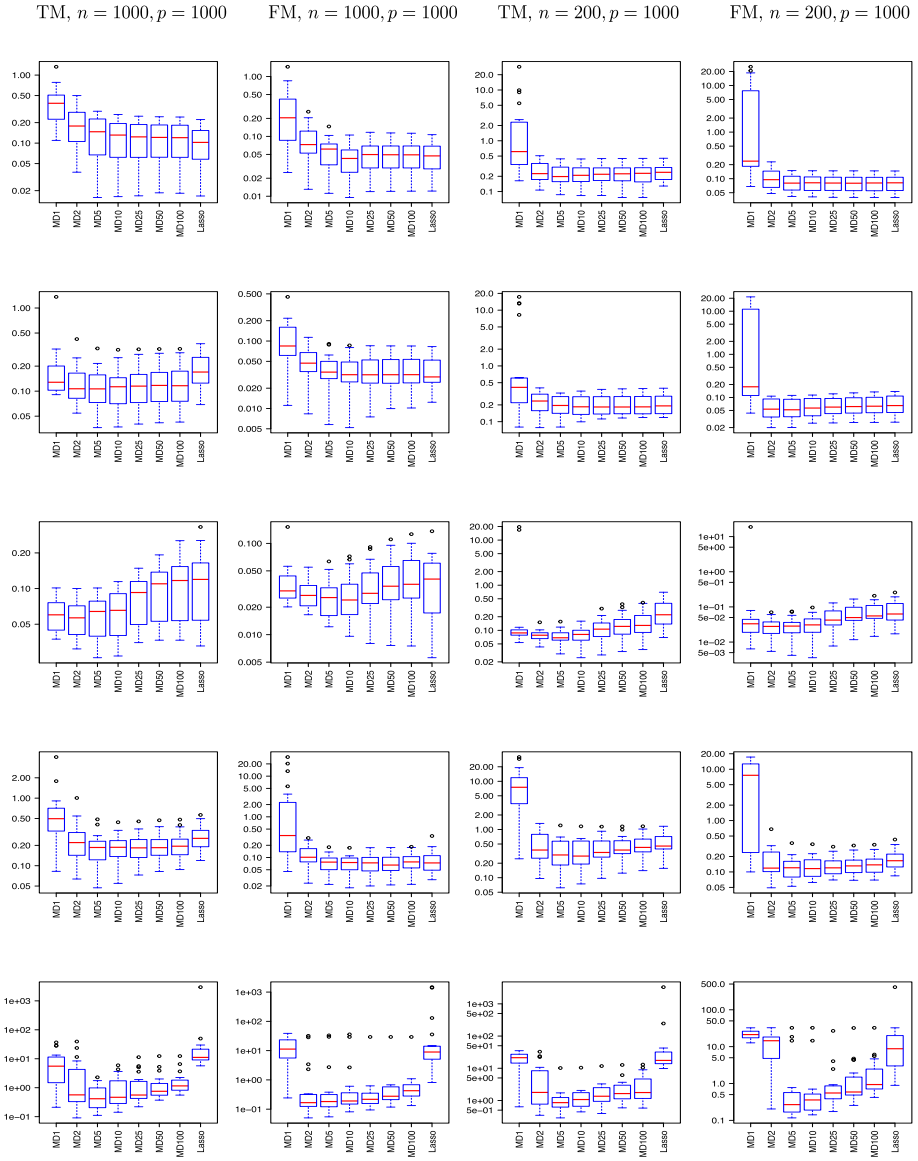


FIG 4. *Influence of the scaling parameter c on the performance of MD-Lasso (Lasso corresponds to $c \to \infty$). Model error (the lower the better), from top to bottom row, errors with a Gaussian, Laplace, Gaussian Mixture, Student t (4 df) and Cauchy distribution. TM=Toeplitz Model, FM=Factor Model.*

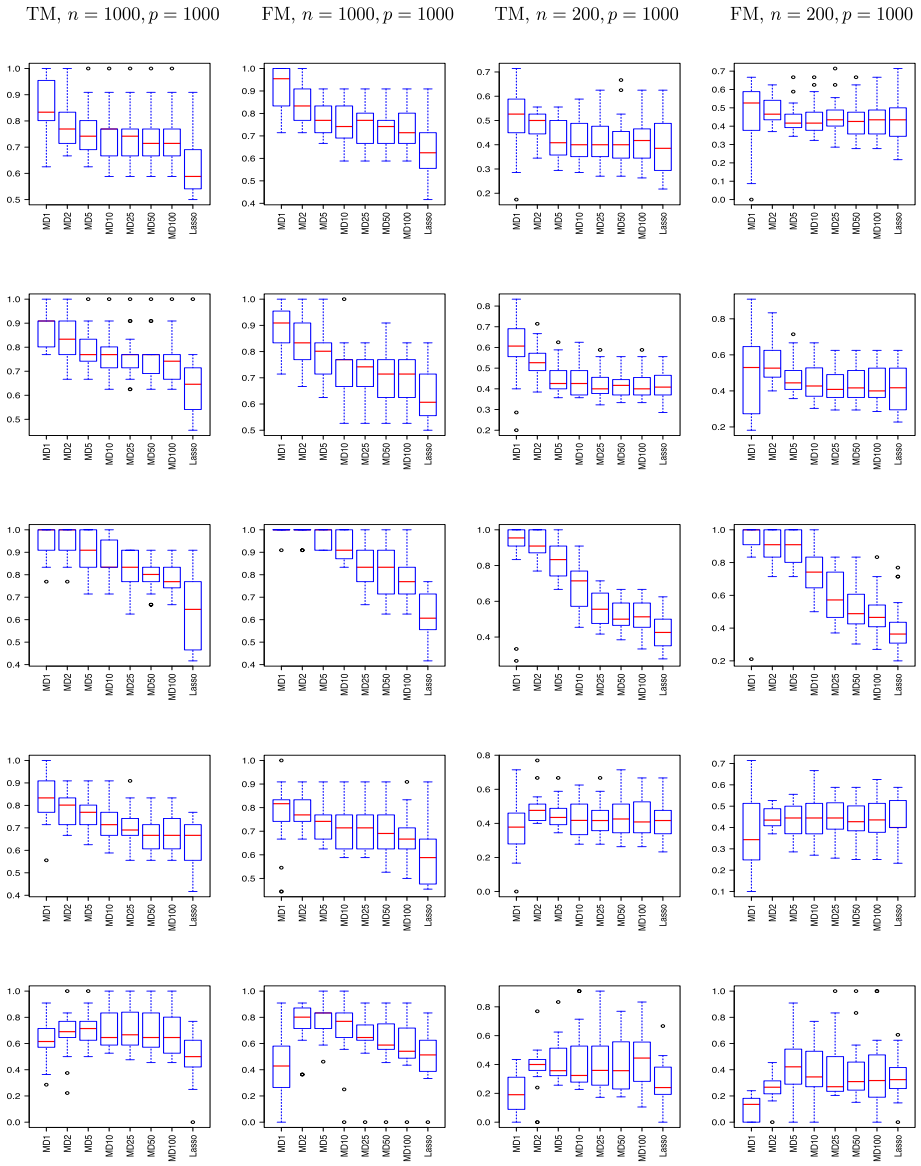


FIG 5. *Influence of the scaling parameter c on the performance of MD-Lasso (Lasso corresponds to $c \to \infty$). Variable selection accuracy (F1 score, the higher the better), from top to bottom row, errors with a Gaussian, Laplace, Gaussian Mixture, Student t (4 df) and Cauchy distribution. TM=Toeplitz Model, FM=Factor Model.*

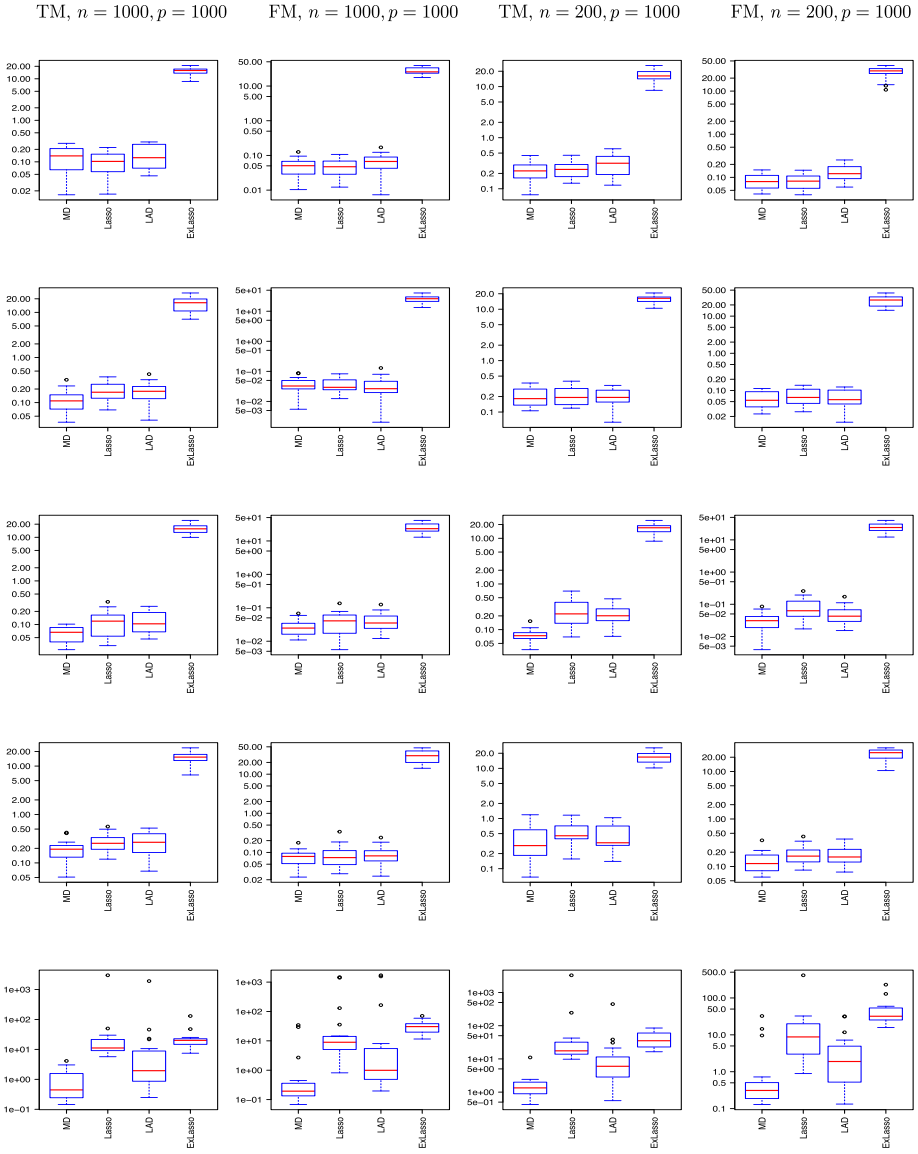


FIG 6. *Model error (the lower the better) for the comparison methods and, from top to bottom row, errors with a Gaussian, Laplace, Gaussian Mixture, Student t (4 df) and Cauchy distribution. TM=Toeplitz Model, FM=Factor Model.*

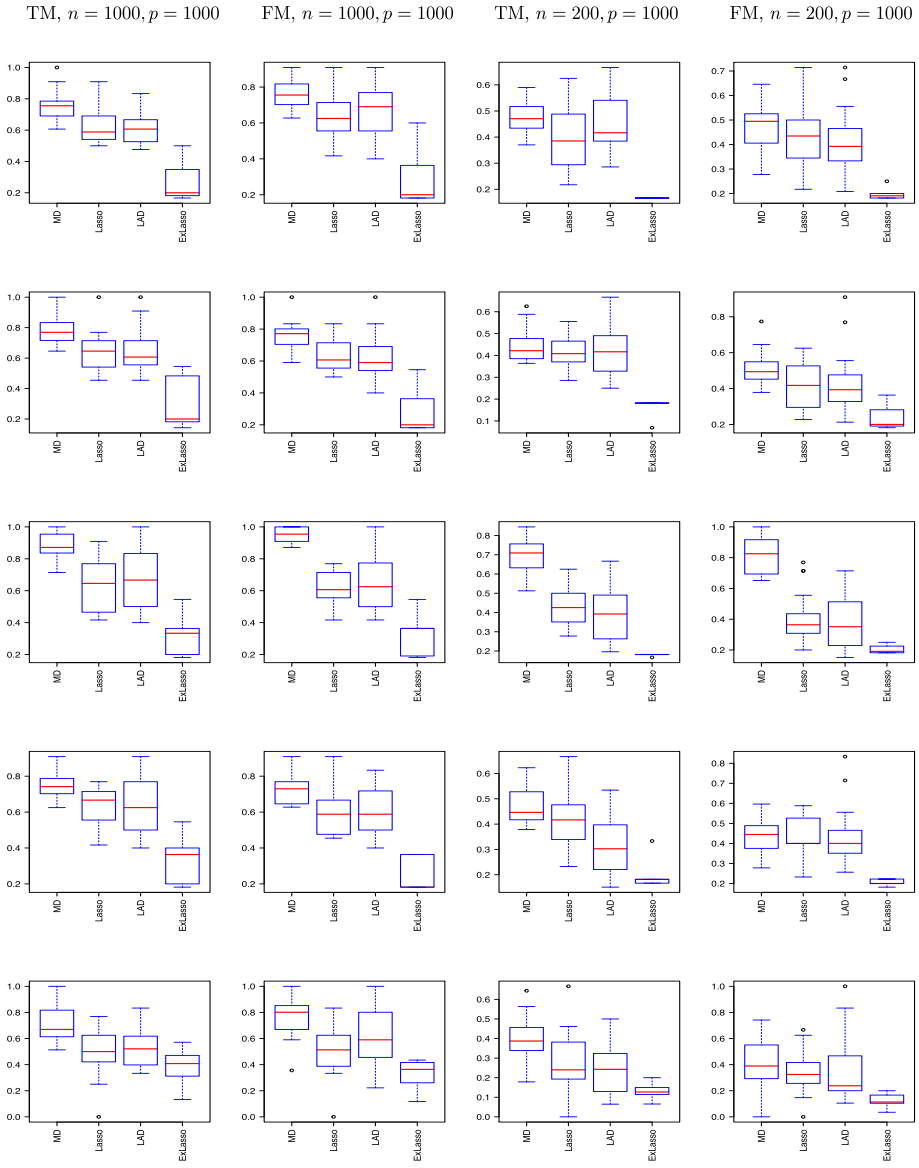
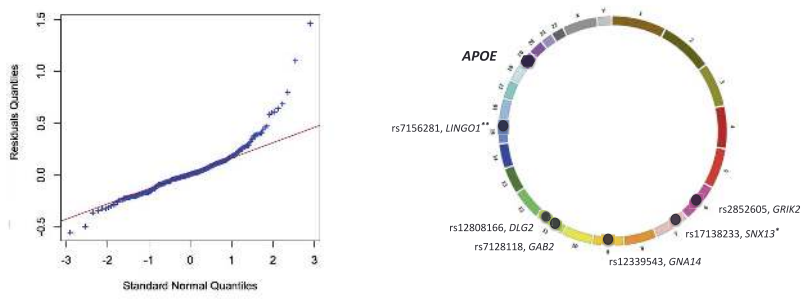


FIG 7. *Variable selection accuracy (F1 score, the higher the better) for the comparison methods and, from top to bottom row, errors with a Gaussian, Laplace, Gaussian Mixture, Student t (4 df) and Cauchy distribution. TM=Toeplitz Model, FM=Factor Model.*



(a) (b)

FIG 8. *(a) Normal QQ-plots of residuals for MD-Lasso in the eQTL study. (b) "Circle" graph of the chromosomes highlighting the location of the target* APOE *gene, a subset of SNPs selected by MD-Lasso only (unmarked), by both MD-Lasso and LAD-Lasso (marked as *), and by Lasso only (marked as **) for the trans-eQTL study.*

5.2. Application to eQTL mapping

We apply MD-Lasso and other methods for comparison to the task of expression quantitative trait locus (eQTL) mapping. The main goal of eQTL studies is to identify the genetic variants (SNPs) that are associated with gene expression traits. In our analysis we use data on Alzheimer's disease (AD) generated by Harvard Brain Tissue Resource Center and Merck Research Laboratories[4]. The dataset concerns $n = 206$ AD cases with SNPs and expression levels in the visual cortex. We study the associations between $p = 18137$ candidate SNPs and the expression levels of *APOE* gene, which is a key Alzheimer's gene [41]. Specifically, persons having an *APOE* $\epsilon 4$ allele have an increased chance of developing the disease; those who inherit two copies of the allele are at even greater risk.

The tuning parameters for all methods were chosen using a five-fold cross validation. To start, we investigated the Normal QQ-plots of the residuals from different regression methods and saw that the residuals from the fitted regressions have very heavy right tails. As an example the plot of the MD-Lasso is shown in Figure 8(a); the plots for the competing methods look similar. This suggests that for this eQTL data analysis it might not be judicious to use methods that lack robustness to noise and model misspecification.

For ease of comparison, we first focus on a *cis*-eQTL analysis, namely we look into the subsets of SNPs within chromosome 19 (where gene *APOE* is located).

To get a measure of confidence in the associations identified, we apply the bootstrap procedure (see Davison and Hinkley [14] for a review) as follows. Given the original data, we randomly draw 100 datasets by sampling with replacement the rows of the original data, so that each dataset has the same number of rows as the original data. We then apply the comparison methods to each of the 100 bootstrap datasets. For each SNP selected using the original dataset, we count

[4] http://sage.fhcrc.org/downloads/downloads.php

TABLE 1

Top 20 selected SNPs on chromosome 19 by comparison methods for the cis-eQTL study and their "confidence score" (the number of times the SNP is selected in 100 bootstrap samples).

MD-Lasso		**LAD-Lasso**		**Extended Lasso**		**Lasso**	
rs5021327	86	rs1120559	81	rs2285751	100	rs280519	87
rs280519	77	rs1654322	75	rs11882861	52	rs2162296	71
rs16964772	76	rs11882490	74	rs1433078	45	rs10408465	69
rs2112460	76	rs3745297	74	rs17314711	43	rs12459372	67
rs7249518	76	rs2116877	70	rs10409463	41	rs16980543	63
rs1599860	74	rs3746006	68	rs12460915	37	rs12327600	42
rs1673130	72	rs353989	67	rs2419549	37	rs13730	40
rs1120559	71	rs10404242	66	rs11665711	32	rs1357879	39
rs11672071	71	rs11878850	66	rs16973403	32	rs1402325	35
rs4805590	71	rs16964772	65	rs1402325	30	rs17314711	28
rs2395891	70	rs3108549	63	rs1422259	29	rs1549951	17
rs352826	69	rs16964420	61	rs12975977	28	rs2304184	16
rs7246997	69	rs184239	59	rs276731	28	rs2395891	16
rs10404242	67	rs2292033	59	rs1013414	27	rs11084566	15
rs3745297	67	rs420703	59	rs12459372	27	rs16964772	13
rs2301742	66	rs1549951	58	rs1560730	27	rs1120559	12
rs2304184	64	rs16973403	58	rs10414066	26	rs11882490	12
rs11084566	63	rs2301742	58	rs12609039	24	rs10412301	11
rs3108549	63	rs280519	58	rs2195948	22	rs276725	11
NA	NA	rs11881644	57	rs12976494	21	rs2304185	10

the number of times it appears in the bootstrap datasets. There is a sharp contrast among methods with respect to the number of selected SNPs. MD-Lasso and LAD-Lasso select fewer SNPs than Extended Lasso and Lasso (19 selected coefficients for MD-Lasso, 20 for LAD-Lasso, and over a 100 for Lasso and Extended Lasso). For each method, the top 20 selected SNPs according to the amplitude of their regression coefficients are listed in Table 1 along with their "confidence score".

From Table 1 we can see that MD-Lasso and LAD-Lasso share 9 common SNPs. In contrast Extended Lasso share no common SNPs with MD-Lasso or LAD-Lasso, and Lasso shares at most 2 common SNPs with other methods. The SNPs identified by MD-Lasso are selected in average 71% of the time in the bootstrap datasets, those by LAD-Lasso 65% of the time, those by Extended Lasso 35% of the time, and those by Lasso 37% of the time. While a low variability in the selection process is not a guarantee that the selection includes the SNPs of interest, a high variability in the selection results (and a corresponding low confidence score like for Lasso and Extended Lasso) makes a consistent selection of interesting SNPs less plausible. We thus hypothesize that non-robust methods may select too many spurious associations due to their inability to cope with outliers and heavy-tailed errors.

We now focus on the results obtained by the comparison methods on the full set of chromosomes. As the genetics of Alzheimer's disease are not yet fully understood, the variable selection results can only lead to qualitative statements about the performance of each method. To provide a more quantitative assessment, we evaluate the predictive accuracy of the various methods by randomly partitioning the data into training and test sets, using 150 observations for

TABLE 2
Average test absolute error (MAE) and square error (MSE) with standard deviation for the models output by MD-Lasso and representative comparison methods on the eQTL dataset. (Smaller values indicate higher predictive accuracy).

	MD-Lasso	**LAD-Lasso**	**Extended Lasso**	**Lasso**	**Trim. Lasso**
MAE	11.90 ± 1.52	20.00 ± 3.04	175.75 ± 14.25	38.62 ± 5.45	35.71 ± 4.46
MSE	6.11 ± 1.39	12.18 ± 2.99	858.90 ± 160.85	20.13 ± 9.87	18.22 ± 9.15

training and the remainder for testing. To get a sense of how a robust criterion performs, we also tested trimmed Lasso regression, removing the worst 10% observations according to the absolute residuals. We computed both the absolute prediction error and squared prediction error for the testing set for the model estimated using the training set. We repeated this process 20 times (using 20 random partitions). The results are presented in Table 2. Overall the predictive performance of MD-Lasso is superior to the other methods. We can also see that trimming is not as beneficial as using MD-Lasso or LAD-Lasso.

To conclude the eQTL analysis, we discuss some biologically interesting SNPs selected by various methods. These are depicted in Figure 8(b), which shows the chromosomes, highlights the position of the target *APOE* gene and the selected SNPs along with their closest gene. To facilitate the following discussion, we refer to the genes close to the corresponding SNPs. We first describe results pertaining to MD-Lasso, which are *not* found by other methods. Gene *DLG2* is a memory-associated protein known to be associated to Schizophrenia. However, a recent study showed that conservation of *DLG2* functions could potentially reduce the symptoms of Alzheimer's [22]. It has been shown that the inactivation of the gene *GRIK2* can cause severe learning disabilities. Gene *GNA14* has been identified by several studies as linked to Alzheimer's disease progression (see, e.g. Arefin et al. [4]). It has been indicated that SNPs in gene *GAB2* can modify the risk of late-onset Alzheimer's disease in *APOE* $\epsilon 4$ carriers and plays an important role in Alzheimer's pathogenesis (see, e.g. Reiman et al. [34]). Remarkably, MD-Lasso was the only method to select SNPs in the coding region of *GAB2*.

We also checked for interesting SNPs selected by other methods. Most of them were also selected by MD-Lasso. For instance, SNP *rs17138233*, located within gene *SNX13*, was selected by *both* MD-Lasso and LAD-Lasso. The carboxyl terminal fragment of *SNX13* was reported to associate with activated *H-Ras* [17], which has been implicated in the process of neurodegeneration in Alzheimer's disease [5]. Our finding is quite intriguing as the functional consequence of the interaction between *SNX13* and *H-Ras* is not fully understood. An example among the few interesting SNPs discarded by MD-Lasso is *rs7156281*, located near gene *LINGO-1* , which was identified by Lasso only. *LINGO-1* is known to be involved in neurodegenerative processes including Alzheimer's disease [25].

Overall, our results suggest that the MD-Lasso method achieves greater predictive accuracy and stability than other methods, and is successful in identifying plausible and relevant SNPs in eQTL mapping.

6. Concluding remarks

We have shown that by combining minimum distance estimation with ℓ_1 penalization the robustness of minimum distance estimation can be preserved in the sparse high-dimensional regression setting. Our theoretical results indicate that the proposed MD-Lasso estimator can achieve optimal convergence rates even under heavy-tailed error distributions. These results hinge on the selection of a scaling parameter of MD-Lasso. If the scaling parameter is very large, MD-Lasso is identical to standard least-squares Lasso. Combining robustness with fast convergence rates requires non-convexity of the loss function, and the objective function can have multiple minima as a consequence. One set of results holds for all local minima within a local convexity region around the desired solution (and we have provided reasonable conditions under which these are the only existing local minima of the objective function). Another set holds beyond the local convexity region but requires constraining the ℓ_2-norm of the feasible solutions within a safety radius. This guarantees the convergence of a simple first-order optimization method to consistent solutions regardless of the initialization. These desirable properties were confirmed by numerical examples. The MD-Lasso framework should prove equally useful in other statistical models such as generalized linear models, which will be investigated in a future study. Another pertinent direction for future work is to consider minimum-distance loss functions beyond those stemming from likelihood-based models.

Appendix A: Proofs of lemmas and theorems

A.1. Breakdown point analysis: Proof of Theorem 1

The proof technique is similar to Alfons et al. [1]. Assume that $m = n-l$ observations are corrupted, and l observations are kept intact. For convenience assume that the uncorrupted observations are placed at the beginning of the sample, so one actually observes $\{(X_1, Y_1), \ldots, (X_l, Y_l), (X'_{l+1}, Y'_{l+1}), \ldots, (X'_n, Y'_n), \}$ where (X'_i, Y'_i) denote the corrupted observations. Let $K = \max_{i=1,\ldots,l} |Y_i|$, where Y_i are the uncorrupted responses. We have

$$\begin{aligned} Q_c(0) &= -c \log \left(\sum_{i=1}^{l} \exp\left(-\frac{Y_i^2}{2c}\right) + \sum_{i=l+1}^{n} \exp\left(-\frac{{Y'_i}^2}{2c}\right)\right) \\ &\leq -c \log(l \exp(-\frac{K^2}{2c}) + 0)) \\ &= -c \log l + \frac{K^2}{2}. \end{aligned} \tag{15}$$

For any β we have

$$Q_c(\beta) \geq -c \log n + \lambda \|\beta\|_1 \geq -c \log n + \lambda \|\beta\|_2. \tag{16}$$

Let $K' = \lambda^{-1}(c\log n - c\log l + K^2/2 + 1)$. If β is such that $\|\beta\|_2 \geq K'$ then (16) and (15) imply

$$Q_c(\beta) \geq -c\log l + \frac{K^2}{2} + 1 > Q(0).$$

By contraposition we thus obtain that

$$Q_c(\beta) \leq Q(0) \Rightarrow \|\beta\|_2 < K'. \tag{17}$$

Recall that the MD-Lasso loss is non-convex, hence multiple local minima may exist. However, there is at least one minimizer, $\hat{\beta}$, such that $Q_c(\hat{\beta}) \leq Q(0)$. Using (17), we get that for such $\hat{\beta}$ we have $\|\hat{\beta}\|_2 \leq K'$. Here $K' = \lambda^{-1}(c\log(n/(n-m)) + K^2/2 + 1)$ is independent of the corrupted sample. Hence for any finite c we can tolerate at least $m = \alpha n$ corruptions where α is arbitrarily close to 1, as in such cases $K' < \infty$ even as $n \to \infty$ and regardless of the nature of the corruptions.

A.2. Gradient bounds: Proof of Lemma 1 and Lemma 2

Define p_i for $i = 1, \ldots, n$ as

$$p_i = \frac{\exp(-\frac{\eta_i^2}{2c})}{\sum_{j=1}^n \exp(-\frac{\eta_j^2}{2c})}$$

and $f^j(\eta_1, \ldots, \eta_n) = \sum_{i=1}^n \eta_i p_i X_i^j$. Note that X_i^j are constants with $|X_i^j| \leq M$. We have $\nabla_j L(\beta^\star) = f^j(\eta_1, \ldots, \eta_n)$.

Notation and lemma. Let the event $E^-_{\gamma,\lambda}$ for $\lambda \geq 0$ and $\gamma < 1$ be defined as $\sum_{i=1}^n 1\{|\eta_i| < \lambda\} < n\gamma(1 - t_\lambda)$. With Hoeffding's inequality we have for $\gamma < 1$

$$P(E^-_{\lambda,\gamma}) \;\leq\; \exp\big(-2n(1-\gamma)^2(1-t_\lambda)^2\big). \tag{18}$$

Likewise let $E^+_{\gamma,\lambda}$ be defined as $\sum_{i=1}^n 1\{|\eta_i| \geq \lambda\} > n\gamma t_\lambda$. With Hoeffding's inequality we have for $\gamma > 1$ again

$$P(E^+_{\lambda,\gamma}) \;\leq\; \exp(-2n(\gamma-1)^2 t_\lambda^2). \tag{19}$$

Showing that $E[f^j] = 0$. If $E[\eta]$ is well defined this is straightforward (e.g. for Gaussian, Laplace, Student's, Gaussian mixture etc). The following deals with the case where $E[\eta]$ is undefined (e.g. Cauchy). Assume that the error term has a symmetrical and bounded probability density function μ. Namely for all $x \in \mathbb{R}$, $\mu(x) = \mu(-x)$ and there exists an $A > 0$ such that $0 \leq \mu(x) \leq A < \infty$.

Since the predictors are bounded, we have by Hölder's inequality

$$|E[f^j]| = |\sum_{i=1}^n E[\eta_i p_i] X_i^j| \leq \|X^j\|_\infty \sum_{i=1}^n |E[\eta_i p_i]| = M \sum_{i=1}^n |E[\eta_i p_i]|. \tag{20}$$

Now

$$|E[\eta_i p_i|\eta_j \text{ for all } j \neq i]| \leq \frac{1}{b}|E[\eta_i \exp\big(-\frac{\eta_i^2}{2c}\big)]|, \tag{21}$$

where $b = \sum_{j\neq i} \exp(-\frac{\eta_j^2}{2c})$. We have

$$\begin{aligned} E[\eta_i \exp\big(-\frac{\eta_i^2}{2c}\big)] &= \int_{-\infty}^{\infty} \mu(x) x \exp\left(-\frac{x^2}{2c}\right) dx \\ &= -\int_0^{\infty} \mu(x) x \exp\left(-\frac{x^2}{2c}\right) dx + \int_0^{\infty} \mu(x) x \exp\left(-\frac{x^2}{2c}\right) dx, \end{aligned}$$

where the last equation comes from the fact that μ is symmetrical.

Now we have

$$\begin{aligned} 0 \leq \int_0^{\infty} \mu(x) x \exp\left(-\frac{x^2}{2c}\right) dx &\leq A \int_0^{\infty} x \exp\left(-\frac{x^2}{2c}\right) dx \\ &= Ac \end{aligned}$$

Hence $\int_0^{\infty} \mu(x) x \exp\big(-x^2/(2c)\big) dx$ is finite and thus $E[\eta_i \exp\big(-\eta_i^2/(2c)\big)]| = 0$. Together with (20) and (21) we conclude that $E[f^j] = 0$.

McDiarmid's inequality. We have to find bounds δ_i such that for all $\eta_1, \ldots, \eta_n$, $\tilde{\eta}_i$ and all i,

$$|f^j(\eta_1, \ldots, \eta_i, \ldots, \eta_n) - f^j(\eta_1, \ldots, \tilde{\eta}_i, \ldots, \eta_n)| \leq \delta_i. \tag{22}$$

Once we have these, since $E[f] = 0$, by McDiarmid's inequality we have for all $t > 0$

$$P(|f| > t) \leq 2 \exp\big(-\frac{2t^2}{\sum_i \delta_i^2}\big) \tag{23}$$

If the bounds δ_i in (22) hold only with probability $1 - \alpha$, we have

$$P(|f| > t) \leq 2 \exp\big(-\frac{2t^2}{\sum_i \delta_i^2}\big) + \alpha \tag{24}$$

Bounds. Now, for a given $\eta_1, \ldots, \eta_{i-1}, \eta_{i+1}, \ldots, \eta_n$, we have, if η_i can take any real value,

$$\begin{aligned} \max_x |f^j(\eta_1, \ldots, \eta_{i-1}, x, \eta_{i+1}, \ldots, \eta_n)| &\leq M \max_{x\geq 0} \left|\frac{x \exp(-\frac{x^2}{2c})}{b + \exp(-\frac{x^2}{2c})}\right| + \text{const} \\ &\leq M \max_{x\geq 0} \left|\frac{x \exp(-\frac{x^2}{2c})}{b}\right| + \text{const}, \end{aligned}$$

where $b = \sum_{j\neq i} \exp(-\frac{\eta_j^2}{c})$. Now, $\max_{x\geq 0} x \exp(-\frac{x^2}{2c})$ is attained at $x = \sqrt{c}$ and the maximal value is $\sqrt{c/e}$. Hence the lhs in (22) is bounded by

$$\max_{\eta,\tilde{\eta}} |f^j(\eta_1, \ldots, \eta_i, \ldots, \eta_n) - f^j(\eta_1, \ldots, \tilde{\eta}_i, \ldots, \eta_n)|$$

$$\begin{aligned} &\leq 2M \max_x |f^j(\eta_1, \ldots, \eta_{i-1}, x, \eta_{i+1}, \ldots, \eta_n)| \\ &\leq 2M \frac{\sqrt{c/e}}{b}. \end{aligned}$$

On the other hand, if $|\eta|$ is constrained to be within λ, and $\lambda \leq \sqrt{c}$ then the equivalent bound becomes

$$2 \max_{x:|x|\leq\lambda} |f^j(\eta_1, \ldots, \eta_{i-1}, x, \eta_{i+1}, \ldots, \eta_n)| \;\leq\; \frac{2\lambda \exp(-\frac{\lambda^2}{2c})}{b}. \tag{25}$$

Fixing a particular value of λ, we have by (19) that with probability at most $\exp(-2nt_\lambda^2)$ a proportion of at least $2t_\lambda$ of all samples have value larger in absolute value than λ. Hence the sum of all δ_i^2 is bounded with probability $1 - \exp(-2nt_\lambda^2)$ by

$$n^{-1} \sum_{i=1}^n \delta_i^2 \leq M(1-2t_\lambda)\Big(2\frac{\lambda \exp(-\frac{\lambda^2}{2c})}{b}\Big)^2 + M(2t_\lambda)\Big(2\frac{\sqrt{c/e}}{b}\Big)^2.$$

Bounding *b*. Let $b = \sum_{j\neq i} \exp(-\eta_j^2/(2c))$. Let $\mathcal{E}_{b,\lambda',\gamma}$ be the event

$$\sum_{j\neq i} 1\{|\eta_j| \geq \lambda'\} > n\gamma t_{\lambda'}.$$

We have $E[\sum_{j\neq i} 1\{|\eta_j| \geq \lambda'\}] = (n-1)t_{\lambda'}$. By Hoeffding's inequality,

$$P\Big(\sum_{j\neq i} 1\{|\eta_j| \geq \lambda'\} \;>\; E[\sum_{j\neq i} 1\{|\eta_j| \geq \lambda'\}] + t\Big) \;\leq\; \exp\Big(\frac{-2t^2}{\sum_{j\neq i}(1-0)^2}\Big),$$

since $1\{|\eta_j| \geq \lambda'\} \in [0,1]$. Setting $t = (n(\gamma-1)+1)t_{\lambda'}$ we obtain

$$P(\mathcal{E}_{b,\lambda',\gamma}) \leq \exp\Big(\frac{-2((n(\gamma-1)+1)t_{\lambda'})^2}{n-1}\Big)$$

and hence, since $\exp(-x)$ is decreasing,

$$P\Big(b < n\gamma t_{\lambda'} \exp\big(-\frac{\lambda'^2}{2c}\big)\Big) \;\leq\; \exp\Big(\frac{-2\,(n(\gamma-1)+1)^2\, t_{\lambda'}^2}{n-1}\Big).$$

For $\gamma = 2$ and $\lambda' = 1$ it follows that

$$P\big(b \geq 2nt_1 \exp(-\frac{1}{2c})\big) \;>\; 1 - \exp\big(\frac{-2\,(n+1)^2\, t_1^2}{n-1}\big).$$

Hence, with probability at least $1 - \exp(-2nt_{\sqrt{1}}^2) - \exp(-2nt_\lambda^2)$,

$$\sum_{i=1}^n \delta_i^2 \;\leq\; \frac{M^2 e^{1/c}}{nt_1^2}\Big[(1-2t_\lambda)\lambda^2 e^{-\lambda^2/c} + \frac{2ct_\lambda}{e}\Big].$$

In summary, we have from (24) that

$$P(|f^j| > t) \;\leq\; 2\exp\big(-2nb_{\lambda,c}t^2\big) + \exp(-2nt_1^2) + \exp(-2nt_\lambda^2), \tag{26}$$

where $b_{\lambda,c} = \frac{t_1^2}{M^2 e^{1/c}}\Big[(1-2t_\lambda)\lambda^2 + \frac{2ct_\lambda}{e}\Big]^{-1}$ with $\lambda \leq \sqrt{c}$. By a union bound over the predictors we obtain

$$\begin{aligned} P(\max_{j=1,\ldots,p} |f_j(\eta_1,\ldots,\eta_n)| > t) \quad &\leq \quad 2\exp\big(-2nb_{\lambda,c}t^2 + \log p\big) \\ &\quad + \exp(-nt_1^2/2 + \log p) + \exp(-2nt_\lambda^2 + \log p), \end{aligned}$$

We can set $t^2 \geq b_{\lambda,c}^{-1}\log p/n$ and $\lambda = \sqrt{c}$. We remark that it seems safe to assume that $\log p/n \to 0$. However, if we choose c as a function of n we still need to make sure that $t_\lambda = \omega(1/\sqrt{n})$, where $\lambda < \sqrt{c}$ and ω denotes the "Small Omega".

Bernstein's inequality. The random variable $z_i = \eta_i \exp(-\eta_i^2/(2c))$ is a mean-zero random variable. Furthermore it is bounded in absolute value by $\sqrt{c/e}$ and hence its variance is guaranteed to exist and is at most c/e (regardless of whether or not the variance of η_i is well-defined). We can thus apply Bernstein's inequality as follows.

$$P\{\frac{\sum_{i=1}^n z_i}{n} > t\} \;\leq\; \exp\Big(-\frac{n^2t^2}{2(\sum_{i=1}^n E[z_i^2] + \sqrt{\frac{c}{e}}\frac{nt}{3})}\Big).$$

We now specialize the above bounds in cases where the variance of z_i is computable in closed form.

Gaussian errors. If (η_i) is a zero-mean gaussian sequence with variance σ^2, let $d_{\sigma,c} = (c/(2\sigma^2 + c))^{3/2}\sigma^2$. Then $E[z_i^2] = d_{\sigma,c}$. By a union bound over the predictors we obtain

$$\begin{aligned} & P(\max_{j=1,\ldots,p} |f_j(\eta_1,\ldots,\eta_n)| > t) \\ \leq \quad & \exp\big(-\frac{n}{2d_{\sigma,c} + 2\sqrt{c/e}(t/3)}t^2 + \log p\big) \\ & + \exp\big(-\frac{n}{2d_{\sigma,c}}t^2 + \log p\big) + \exp(-nt_1^2/2 + \log p). \end{aligned}$$

If c is finite, we can set $t^2 = 4\log(p)d_{\sigma,c}/n$. (This is enough because the term $2\sqrt{c/e}(t/3)$ becomes negligible compared to $d_{\sigma,c}$ as $n \to \infty$ as long as $\log(p)/n \to 0$.) If $c \to \infty$ while $c\log(p)/n \to 0$, we recover the condition for the traditional Lasso, namely:

$$t^2 = 4\sigma^2\frac{\log p}{n}.$$

Note that if $c < \sigma^2$ we obtain $t^2 \sim 4\frac{\log p}{n}c$ which is similar (with respect to the dependence in c) to what we got with McDiarmid's inequality.

Laplacian errors. If (η_i) is a sequence of zero-mean Laplace$(0, 2b)$ random variables then

$$E[z_i^2] = -\frac{c^2}{4b^2} + \frac{\sqrt{2\pi}}{b}\left(\frac{c}{2}\right)^{3/2} e^{\frac{c}{4b^2}}\left(1 + \frac{c}{2b^2}\right)\bar{F}\left(\frac{1}{b}\sqrt{\frac{c}{2}}\right),$$

where $\bar{F}(\cdot)$ denotes the tail probability function of the standard normal distribution. Note that as $c \to \infty$, $E[z_i^2] \to 2b^2$. By a union bound over the predictors,

$$\begin{aligned} & P(\max_{j=1,\ldots,p} |f_j(\eta_1, \ldots, \eta_n)| > t) \\ \leq\ & \exp\big(-\frac{n}{2d_{b,c} + 2\sqrt{c/e}(t/3)}t^2 + \log p\big) + \exp\big(-\frac{n}{2d_{b,c}}t^2 + \log p\big) \\ & + \exp(-nt_1^2/2 + \log p), \end{aligned}$$

where $d_{b,c} = E[z_i^2]$. If c is finite, we can set $t^2 = 4\log(p)d_{b,c}/n$. If $c \to \infty$ while $c\log(p)/n \to 0$, we obtain the condition

$$t^2 = 8b^2\log(p)/n.$$

Error distributions with undefined variances (e.g. Cauchy). If the error distribution has an undefined variance, there is no hope of getting a gradient condition which would guarantee that the gradient is still finite as $c \to \infty$. While Bernstein's inequality is still applicable, as we know that $E[z_i^2] \leq \frac{c}{e}$, McDiarmid's inequality yields tighter bounds in this case.

A.3. Restricted strong convexity: Proof of Lemma 3

Notation. Let $r_i, i = 1, \ldots, n$ be the residual for $\boldsymbol{\beta} = \boldsymbol{\beta}^\star + t\boldsymbol{\Delta}$, . Then $r_i = (\eta_i - \langle \boldsymbol{X}_i, t\boldsymbol{\Delta}\rangle)$. Define $\tilde{p}_i$ for $i = 1, \ldots, n$ as

$$\tilde{p}_i = \frac{\exp(-\frac{r_i^2}{2c})}{\sum_{j=1}^n \exp(-\frac{r_j^2}{2c})}. \tag{27}$$

For any $\lambda \geq 0$ let $t_\lambda = P(|\eta_i| > \lambda)$.

Preliminary lemma. The following technical lemma establishes the conditions required on the Hessian to guarantee restricted strong convexity of the loss L in a neighborhood of the true parameter $\boldsymbol{\beta}^\star$.

Lemma 4. *Let $\boldsymbol{A} \subset \mathbb{R}^p$ be star-shaped with respect to $\boldsymbol{\beta}^\star \in \mathbb{R}^p$, namely for any $\boldsymbol{\beta} \in \boldsymbol{A}$ and $t \in [0, 1]$ it holds that $t\boldsymbol{\beta}^\star + (1-t)\boldsymbol{\beta} \in \boldsymbol{A}$. Let $f : \boldsymbol{A} \to \mathbb{R}$ be a twice-differentiable function. Let the second derivative of f satisfy $\nabla^2 f(\boldsymbol{\beta}^\star + t\boldsymbol{\Delta})(\Delta, \Delta) > 2\kappa_1\|\boldsymbol{\Delta}\|_2(\|\boldsymbol{\Delta}\|_2 - \kappa_2\|\boldsymbol{\Delta}\|_1)$ for all $\boldsymbol{\Delta}$ such that $\boldsymbol{\beta}^\star + \boldsymbol{\Delta} \in \boldsymbol{A}' \subset \boldsymbol{A}$ and t in $(0, 1]$. Then for all $\boldsymbol{\Delta}$ such that $\boldsymbol{\beta}^\star + \boldsymbol{\Delta} \in \boldsymbol{A}' \subset \boldsymbol{A}$ we have $f(\boldsymbol{\beta}^\star + \boldsymbol{\Delta}) - f(\boldsymbol{\beta}^\star) - \langle \nabla f(\boldsymbol{\beta}^\star), \Delta\rangle > \kappa_1\|\boldsymbol{\Delta}\|_2(\|\boldsymbol{\Delta}\|_2 - \kappa_2\|\boldsymbol{\Delta}\|_1)$.*

Proof of Lemma 4: Consider the function $g(t) := f(\boldsymbol{\beta}^\star + t\boldsymbol{\Delta}) - f(\boldsymbol{\beta}^\star) - \langle \nabla f(\boldsymbol{\beta}^\star), t\boldsymbol{\Delta}\rangle - t^2(\kappa_1\|\boldsymbol{\Delta}\|_2(\|\boldsymbol{\Delta}\|_2 - \kappa_2\|\boldsymbol{\Delta}\|_1))$. We need to show that $g(1) > 0$. It holds that $g(0) = g'(0) = 0$. Moreover $g''(t) = \nabla^2 f(\boldsymbol{\beta}^\star + t\boldsymbol{\Delta})(\boldsymbol{\Delta}, \boldsymbol{\Delta}) - 2(\kappa_1\|\boldsymbol{\Delta}\|_2(\|\boldsymbol{\Delta}\|_2 - \kappa_2\|\boldsymbol{\Delta}\|_1)) > 0$ on $(0, 1]$. This implies that g' is positive on $(0, 1]$ and $g(1) > 0$. □

Lemma 4 indicates that it suffices to focus on the Hessian of L and establish that a lower bound of the form $\nabla^2 L(\boldsymbol{\beta}^\star + t\boldsymbol{\Delta})(\boldsymbol{\Delta}, \boldsymbol{\Delta}) \geq 2\kappa_1\|\boldsymbol{\Delta}\|_2(\|\boldsymbol{\Delta}\|_2 - \kappa_2\|\boldsymbol{\Delta}\|_1)$ holds for all $\boldsymbol{\Delta}$ in $K(S, \mu)$ and $t \in (0, 1]$.

Condition on the Hessian of L. We first provide the expression for $\nabla^2 L(\boldsymbol{\beta}^\star + t\boldsymbol{\Delta})(\boldsymbol{\Delta}, \boldsymbol{\Delta})$, for which we then provide a lower-bound. Let $\boldsymbol{\Delta} \in \mathbb{R}^p$. The gradient of L evaluated at a vector $\boldsymbol{\beta} = \boldsymbol{\beta}^\star + t\boldsymbol{\Delta}$, $0 \leq t \leq 1$ can be expressed as

$$\nabla_j(L(\boldsymbol{\beta})) = -\sum_{i=1}^n \tilde{p}_i r_i X_i^j, j = 1, \ldots, p, \tag{28}$$

Differentiating a second time we obtain

$$\nabla^2 L(\boldsymbol{\beta})(\boldsymbol{\Delta}, \boldsymbol{\Delta}) = \sum_{i=1}^n \Big(\tilde{p}_i\big(1 - \frac{1}{c}r_i^2\big)s_i^2\Big) + \frac{1}{c}\big(\sum_{i=1}^n \tilde{p}_i r_i s_i\big)^2, \tag{29}$$

where $s_i = \boldsymbol{X}_i'\boldsymbol{\Delta}$.

Let $z_i = r_i^2/c$. We wish to lower-bound $f(z_i) = e^{-z_i/2}\,(1 - z_i)$. Let $a^2 < 1$. If $z_i \leq a^2$, it holds that $f(z_i) \geq (1-a^2)(1-\frac{a^2}{2})$, noting that $\exp(-z_i/2) \geq (1-\frac{z_i}{2})$. On the other hand if $z_i > a^2$ then $f(z_i) \geq -2e^{-\frac{3}{2}}$. Let

$$\psi(z) = \begin{cases} (1-a^2)(1-\frac{a^2}{2}) & \text{if } z \leq a^2 \\ -2e^{\frac{-3}{2}} & \text{if } z > a^2 \end{cases} \tag{30}$$

Then (29) can be lower-bounded as follows

$$\nabla^2 L(\boldsymbol{\beta}^\star + t\boldsymbol{\Delta})(\boldsymbol{\Delta}, \boldsymbol{\Delta}) \geq \frac{1}{\sum_{i=1}^n \exp\big(-r_i^2/(2c)\big)} \sum_{i=1}^n \psi(z_i)s_i^2.$$

Goal. In what follows we shall consider $K_\nu(S, \mu) = \{\boldsymbol{\Delta} \in C(S) : \|\boldsymbol{\Delta}\|_1 = \nu, \|\boldsymbol{\Delta}\|_2 = \mu\}$ and show that the probability of the event

$$\mathcal{E}(\nu) = \{\frac{1}{n}\sum_{i=1}^n \psi(z_i)s_i^2 < g(\nu, \mu), \text{ for some } \boldsymbol{\Delta} \in K_\nu(S, \mu)\}$$

is very small, where $g(\nu, \mu)$ shall be specified. Then we shall appeal to a peeling argument (see Raskutti et al. [33]) to prove that the event over all $\boldsymbol{\Delta} \in K(S, \mu) = \{\boldsymbol{\Delta} \in C(S) : \|\boldsymbol{\Delta}\|_2 = \mu\}$ is also very small.

We begin by proving a tail bound for

$$\tilde{Z}(\nu) = \sup_{\boldsymbol{\Delta} \in K_\nu(S,\mu)} \left| \frac{1}{n}\sum_{i=1}^n (\psi(z_i) - E(\psi(z_i)))s_i^2 \right|.$$

Step 1: Lower-bounding $\frac{1}{n}\sum_{i=1}^{n} E\psi(z_i)s_i^2$.
Let $\epsilon_n \in \mathbb{R}$ be such that $2\kappa_u\sqrt{\log n}\mu/\sqrt{c} \le \epsilon_n < a$ and let $\lambda_\mu = \sqrt{c}(a-\epsilon_n)$. We first show that

$$E[1\{z_i \ge a^2\}] = P(z_i \ge a^2) \le t_{\lambda_\mu},$$

where $t_{\lambda_\mu} = P(|\eta_i| > \lambda_\mu)$.

Indeed, $P(z_i \ge a^2) = P(|\eta_i - tX_i'\boldsymbol{\Delta}| > a\sqrt{c}) = P(\eta_i > a\sqrt{c} + tX_i'\boldsymbol{\Delta}) + P(\eta_i < -a\sqrt{c} + tX_i'\boldsymbol{\Delta})$. By Assumption [A4] we have

$$P\big(|X_i'\boldsymbol{\Delta}| \ge \delta\big) \le 2\exp\Big(-\frac{\delta^2}{2\kappa_u^2\|\boldsymbol{\Delta}\|_2^2}\Big) \quad \text{for all } \delta > 0, \tag{31}$$

hence it follows $\max_i |X_i'\boldsymbol{\Delta}| \le 2\|\boldsymbol{\Delta}\|_2\kappa_u\sqrt{\log n} \le \sqrt{c}\epsilon_n$ with probability at least $1 - 2/n$.

Now $P(\eta_i > a\sqrt{c} + t\boldsymbol{X}_i\boldsymbol{\Delta}) \le P(\eta_i > a\sqrt{c} - \sqrt{c}\epsilon_n)$ and $P(\eta_i < -a\sqrt{c} + t\boldsymbol{X}_i\boldsymbol{\Delta}) \le P(\eta_i < -a\sqrt{c} + \sqrt{c}\epsilon_n)$. Thus

$$P(|\eta_i - t\boldsymbol{X}_i'\boldsymbol{\Delta}| > a\sqrt{c}) \le P(|\eta_i| > \sqrt{c}(a-\epsilon_n)) = P(|\eta_i| > \lambda_\mu) = t_{\lambda_\mu},$$

as desired.

We also have $E[1\{z_i < a^2\}] = 1 - P(z_i \ge a^2) \ge 1 - t_{\lambda_\mu}$. Thus we obtain

$$\begin{aligned} E[\psi(z_i)] &= (1-a^2)(1-\frac{a^2}{2})P(z_i \le a^2) - 2e^{-\frac{3}{2}}P(z \ge a^2) && (32)\\ &\ge (1-a^2)(1-\frac{a^2}{2}) - t_{\lambda_\mu}\left((1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{3}{2}}\right). && (33)\end{aligned}$$

Hence if

$$t_{\lambda_\mu} \le \frac{(1-a^2)(1-\frac{a^2}{2})}{(1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{3}{2}}},$$

we have $E[\psi(z_i)] \ge 0$. Then, noting that $\sum_{i=1}^{n}\exp\big(-r_i^2/(2c)\big) \le n$, we obtain

$$\begin{aligned} E[\nabla^2 L(\beta^\star + t\boldsymbol{\Delta})(\boldsymbol{\Delta},\boldsymbol{\Delta})] &\ge \frac{1}{n}\sum_{i=1}^{n} E\psi(z_i)s_i^2 && (34)\\ &\ge \frac{1}{n}\Big(C_a(1-t_{\lambda_\mu}) - 2e^{-\frac{3}{2}}\Big)\sum_{i=1}^{n} s_i^2 \\ &\ge \kappa_{RE}\mu^2(C_a(1-t_{\lambda_\mu}) - 2e^{-\frac{3}{2}}), && (35)\end{aligned}$$

where $C_a = (1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{2}{3}}$. Here the second inequality comes from (33) and the last inequality is due to the Restricted Eigenvalue condition (6).

Step 2: Upper-bounding $\tilde{\boldsymbol{Z}}(\boldsymbol{\nu})$.
We first show that

$$|\psi(z_i) - E\psi(z_i)| \le \left[(1-a)^2(1-\frac{a^2}{2}) + 2e^{-\frac{3}{2}}\right].$$

Indeed if $\psi(z_i) \geq E\psi(z_i)$ we have

$$\begin{aligned} |\psi(z_i) - E\psi(z_i)| &= \psi(z_i) - E\psi(z_i) \\ &\leq (1-a^2)(1-\frac{a^2}{2}) - (1-a^2)(1-\frac{a^2}{2}) \\ &\quad + t_{\lambda_\mu}\left((1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{3}{2}}\right) \\ &= t_{\lambda_\mu}\left((1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{3}{2}}\right). \end{aligned}$$

If $\psi(z_i) < E\psi(z_i)$ we have

$$|\psi(z_i) - E\psi(z_i)| = E\psi(z_i) - \psi(z_i) \leq (1-a^2)(1-\frac{a^2}{2}) + 2^{-\frac{3}{2}}.$$

Hence

$$\frac{1}{n}|\psi(z_i) - E\psi(z_i)|s_i^2 \leq \frac{1}{n} C_a 4\kappa_u^2 \log n \|\mathbf{\Delta}\|_2^2, \tag{36}$$

and

$$\tilde{Z}(\nu) \leq 4C_a\kappa_u^2 \log n \mu^2, \tag{37}$$

with $C_a = \left[(1-a)^2(1-\frac{a^2}{2}) + 2e^{-\frac{3}{2}}\right]$.

Step 3: Upper-bounding $E[\tilde{Z}(\nu)]$.
Let $(\xi_i)_{i=1}^n$ be a sequence of i.i.d. Rademacher variables. It holds that

$$\begin{aligned} & E\tilde{Z}(\nu) \\ = \; & E \sup_{\mathbf{\Delta}\in K_\nu(S,\mu)} \left|\frac{1}{n}\sum_{i=1}^n (\psi(z_i) - E\psi(z_i))\, s_i^2\right| && (38) \\ \leq \; & 2E_{\eta,\xi} \sup_{\mathbf{\Delta}\in K_\nu(S,\mu)} \left|\frac{1}{n}\sum_{i=1}^n \xi_i \psi(Z_i) s_i^2\right| && (39) \\ \leq \; & 2(1-a^2)(1-\frac{a^2}{2}) E_{\eta,\xi} \sup_{\mathbf{\Delta}\in K_\nu(S,\mu)} \left|\frac{1}{n}\sum_{i=1}^n \xi_i 1\{z_i \leq a^2\} s_i^2\right| \\ & + 4e^{-\frac{2}{3}} E_{\eta,\xi} \sup_{\mathbf{\Delta}\in K_\nu(S,\mu)} \left|\frac{1}{n}\sum_{i=1}^n \xi_i 1\{z_i > a^2\} s_i^2\right| && (40) \\ \leq \; & 2((1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{2}{3}}) E_\xi \sup_{\mathbf{\Delta}\in K_\nu(S,\mu)} \left|\frac{1}{n}\sum_{i=1}^n \xi_i (s_i(\mathbf{\Delta}))^2\right| && (41) \\ \leq \; & 8\kappa_u\sqrt{\log n}\mu((1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{2}{3}}) E_\xi \sup_{\mathbf{\Delta}\in K_\nu(S,\mu)} \left|\frac{1}{n}\sum_{i=1}^n \xi_i s_i(\mathbf{\Delta})\right| && (42) \\ \leq \; & 8\kappa_u\sqrt{\log n}\mu((1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{2}{3}}) E_\xi \sup_{\mathbf{\Delta}\in K_\nu(S,\mu)} \|\mathbf{\Delta}\|_1 \|\frac{1}{n}\sum_{i=1}^n \xi_i X_i\|_\infty && (43) \\ \leq \; & 8\kappa_u\sqrt{\log n}\mu\nu((1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{2}{3}}) E_\xi \|\frac{1}{n}\sum_{i=1}^n \xi_i X_i\|_\infty. && (44) \end{aligned}$$

Here (39) follows by a standard symmetrization argument, (40) follows from simple structural results on Rademacher complexity (see e.g. Bartlett and Mendelson [7]), (41) follows by the contraction principle (see Ledoux and Talagrand [23]), (42) is obtained by applying Talagrand's comparison theorem (see Theorem 4.12 in Ledoux and Talagrand [23]) noting that for any $\mathbf{\Delta}, \mathbf{\Delta}' :\in K_\nu(S,\mu)$ we have

$$\left|\langle X_i, \mathbf{\Delta}\rangle^2 - \langle X_i, \mathbf{\Delta}'\rangle^2\right| \le (4\kappa_u\sqrt{\log n}\mu)\left|\langle X_i, \mathbf{\Delta}\rangle - \langle X_i, \mathbf{\Delta}'\rangle\right|,$$

and (43) follows by Hölder's inequality.

Applying an existing bound on the expectation of sub-Gaussian maxima (e.g., see Ledoux and Talagrand [23]) we get

$$E_\xi\|\frac{1}{n}\sum_{i=1}^n \xi_i X_i\|_\infty \le 6\kappa_u\sqrt{\frac{\log p}{n}}.$$

Hence we conclude

$$E\tilde{Z}(\nu) \le 48C_a\kappa_u^2\sqrt{\log n}\mu\nu\sqrt{\frac{\log p}{n}}, \tag{45}$$

where $C_a = (1-a^2)(1-\frac{a^2}{2}) + 2e^{-\frac{2}{3}}$.

Step 4: Tail bound on $\tilde{Z}(\nu)$. In view of (36), McDiarmid's inequality implies that for any $t > 0$ we have

$$P(\tilde{Z}(\nu) - E\tilde{Z}(\nu) \ge t) \le \exp\left(-\frac{2nt^2}{(4C_a\kappa_u^2\log n)^2}\right),$$

where $C_a = (1-a)^2(1-\frac{a^2}{2}) + 2e^{-\frac{3}{2}}$. Let

$$t = \frac{1}{2}\kappa_{RE}\mu^2(C_a(1-t_{\lambda_\mu}) - \ 2e^{-\frac{3}{2}}) + C_a\kappa_u^2\sqrt{\log n}\mu\nu\sqrt{\frac{\log p}{n}}.$$

Together with (45) we obtain

$$P\Big(\tilde{Z}(\nu) \ge \frac{1}{2}\kappa_{RE}\mu^2(C_a(1-t_{\lambda_\mu}) - \ 2e^{-\frac{3}{2}}) + 49C_a\kappa_u^2\sqrt{\log n}\mu\nu\sqrt{\log(p)/n}\Big)$$
$$\le \exp\Big(-\frac{2n\big(\frac{1}{2}\kappa_{RE}\mu^2(C_a(1-t_{\lambda_\mu}) - \ 2e^{-\frac{3}{2}}) + C_a\kappa_u^2\sqrt{\log n}\mu\nu\sqrt{\frac{\log p}{n}}\big)^2}{(4C_a\kappa_u^2\log n)^2}\Big),$$

Hence we obtain that the event that for any $\mathbf{\Delta} \in K_\nu(S,\mu)$

$$\frac{1}{n}\sum_{i=1}^n \psi(z_i)s_i^2 \ \le \ \frac{1}{2}\kappa_{RE}\mu^2(C_a(1-t_{\lambda_\mu}) - \ 2e^{-\frac{3}{2}}) - 49C_a\kappa_u^2\sqrt{\log n}\nu\mu\sqrt{\log(p)/n} \tag{46}$$

holds with probability at most

$$\exp\left(-n\frac{\kappa_{RE}^2(C_a(1-t_{\lambda_\mu})-\ 2e^{-\frac{3}{2}})^2}{2(4C_a\kappa_u^2\log n)^2}-\frac{1}{8}\frac{\nu}{\mu\log n}\log p\right), \tag{47}$$

noting that for $a\geq 0, b\geq 0$, it holds that $\exp(-(a+b)^2)\leq\exp(-a^2-b^2)$ and that $\nu/\mu>1$, since for any $\boldsymbol{u}$, $\|\boldsymbol{u}\|_2\leq\|\boldsymbol{u}\|_1$. By a peeling argument (see Raskutti et al. [33] for details) this yields the claim of Lemma 3. □

A.4. Consistency results

We first present a theorem which is the counterpart of Theorem 2, and then prove both theorems. As noted in section 3.3 the following theorem leverages Lemma 2 and is thus better suited for errors with well-defined variance.

Theorem 4. *Consider the linear regression model (1) and assume that the support of the true model coefficients $\boldsymbol{\beta}^\star$ has cardinality s. Let $\mathcal{H}_c=\{\boldsymbol{\beta}^\star+\boldsymbol{\Delta}:\|\boldsymbol{\Delta}\|_2<\sqrt{c}/(12\kappa_u\sqrt{\log n})\}$. Under Assumptions $[A1-A4]$, given the MD Lasso estimator (5) with scaling parameter c such that $t_{\sqrt{c}/2}<(1+(64/21)e^{-3/2})^{-1}<0.6$ and regularization parameter $\lambda_n=2\zeta_c\sqrt{\log p/n}$, where ζ_c is given by $\zeta_c^2=4M^2t_1^{-2}E[\eta_i^2e^{-\eta_i^2/c}]e^{1/c}$, any of the solutions in $\mathcal{H}_c$ (there is at least one such solution) satisfies*

$$\|\hat{\boldsymbol{\beta}}_{\lambda_n}-\boldsymbol{\beta}^\star\|_2\ \leq\ \frac{32\zeta_c}{(C(1-t_{\sqrt{c}/2})-2e^{-\frac{3}{2}})\kappa_{RE}}\sqrt{\frac{s\log p}{n}} \tag{48}$$

with probability at least

$$1-\alpha_1\exp\Big(-\alpha_2 n\lambda_n^2\frac{1-\alpha_3\sqrt{c\log p/n}}{1+\alpha_3\sqrt{c\log p/n}}\Big),$$

for $n\geq\tilde{\zeta}_{c,\gamma}\frac{s}{c}\log p\log n$ where $C=21/32+2e^{-3/2}\approx 1.1$, $\tilde{\zeta}_{c,\gamma}=(96\kappa_u/\kappa_1\zeta_{c,\gamma})^2$, and $\alpha_1,\alpha_2,\alpha_3>0$ constants.

A.5. Proof of Theorem 2 and Theorem 4

The proof follows the same arguments as the proof of Theorem 1 in Negahban et al. [30]. We include details for completeness.

Denote by δ the error tolerances on $\|\hat{\beta}-\beta^\star\|_2$ in the theorem statements. Define the function $F:\mathbb{R}^p\to\mathbb{R}$ as

$$F(\boldsymbol{\Delta})=L(\boldsymbol{\beta}^\star+\boldsymbol{\Delta})-L(\boldsymbol{\beta}^\star)+\lambda_n(\|\boldsymbol{\beta}^\star+\boldsymbol{\Delta}\|_1-\|\boldsymbol{\beta}^\star\|_1).$$

Let $K(\delta,S)=\{\boldsymbol{\Delta}\in C(S):\|\boldsymbol{\Delta}\|_2=\delta\}$, where $\delta\leq\frac{\sqrt{c}}{12\kappa_u\sqrt{\log n}}$. Let $\hat{\boldsymbol{\beta}}$ denote a minimizer of $L(\boldsymbol{\beta})+\lambda_n\|\boldsymbol{\beta}\|_1$ in the local convexity region $\mathcal{H}_c$. Let $\hat{\boldsymbol{\Delta}}=\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star$.

First, Lemma 1 of Negahban et al. [30] can be seamlessly adapted: since λ_n is chosen in the theorem statement such that $\lambda_n \geq 2\|\nabla L(\boldsymbol{\beta}^*)\|_\infty$ we have $\hat{\boldsymbol{\Delta}} \in C(S)$. Next we note that if $F(\boldsymbol{\Delta}) > 0$ for all $\boldsymbol{\Delta} \in K(\delta, S)$ then $\|\hat{\boldsymbol{\Delta}}\|_2 \leq \delta$. This is shown by contraposition. If $\|\hat{\boldsymbol{\Delta}}\|_2 > \delta$, the line joining $\hat{\boldsymbol{\Delta}}$ to 0 intersects $K(\delta, S)$ at some point $t\hat{\boldsymbol{\Delta}}$ with $t \in (0, 1)$. Since L is locally convex on the line joining $(\boldsymbol{\beta}^\star + \hat{\boldsymbol{\Delta}})$ and $\boldsymbol{\beta}^\star$, by Jensen's inequality we have $F(t\hat{\boldsymbol{\Delta}}) = F(t\hat{\boldsymbol{\Delta}} + (1-t)\mathbf{0}) \leq tF(\hat{\boldsymbol{\Delta}})$. Since $F(\hat{\boldsymbol{\Delta}}) \leq 0$ then $F(t\hat{\boldsymbol{\Delta}}) \leq 0$ as well, thus showing the contrapositive statement.

To prove Theorem 2 and Theorem 4 is thus suffices to establish a lower-bound on $F(\boldsymbol{\Delta})$ over $K(\delta, S)$ for the specific values of δ in the theorem statements. For any $\boldsymbol{\Delta} \in K(\delta, S)$ due to restricted strong convexity and decomposability of the ℓ_1-norm with respect to the set S, we have $F(\boldsymbol{\Delta}) \geq \langle \nabla L(\boldsymbol{\beta}^\star), \boldsymbol{\Delta} \rangle + \bar{\kappa}_1 \|\boldsymbol{\Delta}\|^2 + \lambda_n(\|\boldsymbol{\Delta}_{S^c}\|_1 - \|\boldsymbol{\Delta}_S\|_1)$, where $\bar{\kappa}_1 := \kappa_1/2$ comes from the restricted eigenvalue condition in (8) and Lemma 3.

By Cauchy-Schwartz we have $|\nabla L(\boldsymbol{\beta}^\star), \boldsymbol{\Delta}\rangle| \leq \|\nabla L(\boldsymbol{\beta}^\star)\|_\infty \|\boldsymbol{\Delta}\|_1$. Notice that λ_n in the theorems is chosen such that $\lambda_n \geq 2\|\nabla L(\boldsymbol{\beta}^\star)\|_\infty$ based on the gradient bounds of Lemmas 1 and 2. In addition $\|\boldsymbol{\Delta}\|_1 = \|\boldsymbol{\Delta}\|_S + \|\boldsymbol{\Delta}_{S^c}\|_1$. Hence we get $F(\boldsymbol{\Delta}) \geq \bar{\kappa}_1 \|\boldsymbol{\Delta}\|^2 + \lambda_n(\frac{1}{2}\|\boldsymbol{\Delta}_{S^c}\|_1 - \frac{3}{2}\|\boldsymbol{\Delta}_S\|_1) \geq \bar{\kappa}_1 \|\boldsymbol{\Delta}\|^2 - \frac{\lambda_n}{2}(3\|\boldsymbol{\Delta}_S\|_1)$. Since $\|\boldsymbol{\Delta}\|_1 \leq \sqrt{s}\|\boldsymbol{\Delta}\|_2$ and $\|\boldsymbol{\Delta}_S\|_2 \leq \|\boldsymbol{\Delta}\|_2$ we have $F(\boldsymbol{\Delta}) \geq \bar{\kappa}_1 \|\boldsymbol{\Delta}\|^2 - \frac{\lambda_n}{2}(3\sqrt{s}\|\boldsymbol{\Delta}\|_2)$, which is strictly positive as long as $\|\boldsymbol{\Delta}\|_2 \geq \frac{1}{\bar{\kappa}_1}(2\lambda_n\sqrt{s})$. This is possible as long as $\frac{1}{\bar{\kappa}_1}(2\lambda_n\sqrt{s}) \leq \frac{\sqrt{c}}{12\kappa_u\sqrt{\log n}}$. For Theorem 2, this will hold as long as $n \geq \tilde{\xi}_{c,\gamma}\frac{s}{c}\log p \log n$ where $\tilde{\xi}_{c,\gamma} = (96)^2\frac{\kappa_u^2}{\bar{\kappa}_1^2}\xi_{c,\gamma}^2$. For Theorem 4, this will also hold as long as $n \geq \tilde{\zeta}_{c,\gamma}\frac{s}{c}\log p \log n$ where $\tilde{\zeta}_{c,\gamma} = (96)^2\frac{\kappa_u^2}{\bar{\kappa}_1^2}\zeta_{c,\gamma}^2$. The theorem statements then follow. □

A.6. Convergence results of local optima: Proof of Theorem 3

Let $\widetilde{\boldsymbol{\beta}}_{\lambda_n}$ be *any* local optimum found by the incremental algorithm. Throughout the proof, we use the shorthand for local optimal error vector: $\widetilde{\boldsymbol{\Delta}} := \widetilde{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^\star$ for any local optimum $\widetilde{\boldsymbol{\beta}}_{\lambda_n}$. We have two key ingredients in this proof to derive the consistency of arbitrary local optimum.

The first ingredient is the restricted strong convexity condition, which is also required in Theorem 2. Suppose that $b_0 \leq \sqrt{c}/(8\kappa_u\sqrt{\log n})$. Then, for any local optimum $\widetilde{\boldsymbol{\beta}}_{\lambda_n}$, $\|\widetilde{\boldsymbol{\Delta}}\|_2 \leq \|\widetilde{\boldsymbol{\beta}}_{\lambda_n}\|_2 + \|\boldsymbol{\beta}^\star\|_2 \leq \sqrt{c}/(4\kappa_u\sqrt{\log n})$, provided that $\boldsymbol{\beta}^\star$ is also feasible for (10). Now, noting that the proof of Lemma 3 is based on [A3] and the fact that the error vector lies in the structure $C(S)$, we here make the alternate assumption [A3'], and utilize a well known inequality conditioned on [A3'] [33]:

$$\frac{\|X\Delta\|_2}{\sqrt{n}} \geq c_1\|\Delta\|_2 - c_2\sqrt{\frac{\log p}{n}}\|\Delta\|_1 \quad \text{for all } \Delta \in \mathbb{R}^p \tag{49}$$

with probability at least $1 - c_3\exp(-c_4 n)$ for $c_1 = \frac{1}{4}\lambda_{\min}(\Sigma)$, $c_2 = 9\sqrt{\max \Sigma_{jj}}$ and some positive constants c_3 and c_4. With a careful modification of the proof,

we now have for all $\widetilde{\boldsymbol{\Delta}} \in \mathbb{R}^p$ such that $\|\widetilde{\boldsymbol{\Delta}}\|_2 \leq \sqrt{c}/(4\kappa_u\sqrt{\log n})$,

$$L(\boldsymbol{\beta}^\star + \widetilde{\boldsymbol{\Delta}}) - L(\boldsymbol{\beta}^\star) - \langle \nabla L(\boldsymbol{\beta}^\star), \widetilde{\boldsymbol{\Delta}} \rangle \;\geq\; \kappa_1 \|\widetilde{\boldsymbol{\Delta}}\|_2^2 - \kappa_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\boldsymbol{\Delta}}\|_1 \|\widetilde{\boldsymbol{\Delta}}\|_2 \quad (50)$$

similarly as in (7) but now $\kappa_1 = \frac{1}{64}\big(\lambda_{\min}(\Sigma)\big)^2 (C(1-t_{\lambda_\nu}) - 2e^{-2/3})$ and $\kappa_2 = \frac{49}{2} C\kappa_u^2 \sqrt{\log n} + \frac{9}{8}\lambda_{\min}(\Sigma)\sqrt{\max \Sigma_{jj}}$. Furthermore, the proof of Lemma 3 can be seamlessly modified for slightly different version of RSC condition, as follows: as in Lemma 4, we define $g(t) := \big\langle \nabla L(\beta^\star + t\widetilde{\boldsymbol{\Delta}}) - \nabla L(\beta^\star),\, \widetilde{\boldsymbol{\Delta}} \big\rangle - t(\kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \|\boldsymbol{\Delta}\|_1 \|\boldsymbol{\Delta}\|_2))$ for $t \in [0,1]$. As long as $\nabla^2 L(\boldsymbol{\beta}^\star + t\boldsymbol{\Delta})(\boldsymbol{\Delta}, \boldsymbol{\Delta}) \geq \kappa_1 \|\boldsymbol{\Delta}\|_2^2 - \kappa_2 \|\boldsymbol{\Delta}\|_1 \|\boldsymbol{\Delta}\|_2$, we have $g(0) = 0$ and $g'(t) \geq 0$ for $t \in (0,1]$ and hence $g(1) \geq 0$: for all $\widetilde{\boldsymbol{\Delta}} \in \mathbb{R}^p$ such that $\|\widetilde{\boldsymbol{\Delta}}\|_2 \leq \sqrt{c}/(4\kappa_u\sqrt{\log n})$, we have with probability specified in Lemma (3),

$$\Big\langle \nabla L(\beta^\star + \widetilde{\boldsymbol{\Delta}}) - \nabla L(\beta^\star),\, \widetilde{\boldsymbol{\Delta}} \Big\rangle \geq \kappa_1 \|\widetilde{\boldsymbol{\Delta}}\|_2^2 - \kappa_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\boldsymbol{\Delta}}\|_1 \|\widetilde{\boldsymbol{\Delta}}\|_2 \quad (51)$$

where $\kappa_1 = \frac{1}{32}\big(\lambda_{\min}(\Sigma)\big)^2 (C(1-t_{\lambda_\nu}) - 2e^{-2/3})$ and $\kappa_2 = 49C\kappa_u^2\sqrt{\log n} + \frac{9}{4}\lambda_{\min}(\Sigma)\sqrt{\max \Sigma_{jj}}$.

Now, we are ready to show the upper bound of $\widetilde{\boldsymbol{\Delta}}$ as stated. From the constraint of (10), $\|\widetilde{\boldsymbol{\Delta}}\|_2 \leq 2b_0$, hence the RSC inequality can be represented as

$$\begin{aligned}\Big\langle \nabla L(\boldsymbol{\beta}^\star + \widetilde{\boldsymbol{\Delta}}) - \nabla L(\boldsymbol{\beta}^\star),\, \widetilde{\boldsymbol{\Delta}} \Big\rangle &\geq \kappa_1 \|\widetilde{\boldsymbol{\Delta}}\|_2^2 - \kappa_2 \sqrt{\frac{\log p}{n}} \|\widetilde{\boldsymbol{\Delta}}\|_1 \|\widetilde{\boldsymbol{\Delta}}\|_2 \\ &\geq \kappa_1 \|\widetilde{\boldsymbol{\Delta}}\|_2^2 - 2\kappa_2 b_0 \sqrt{\frac{\log p}{n}} \|\widetilde{\boldsymbol{\Delta}}\|_1 \,. \end{aligned} \quad (52)$$

The second ingredient handling the local optima is the first-order necessary condition to be a local optimum:

$$\big\langle \nabla L(\boldsymbol{\beta}^\star + \widetilde{\boldsymbol{\Delta}}),\, \widetilde{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta} \big\rangle \leq -\big\langle \partial \lambda_n \|\boldsymbol{\beta}^\star\|_1,\, \widetilde{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta} \big\rangle \quad \text{for any feasible } \boldsymbol{\beta}. \quad (53)$$

Note that this condition reduces to the usual zero sub-gradient condition when $\widetilde{\boldsymbol{\beta}}_{\lambda_n}$ lies in the interior of the constraint set, but is more general one for the local optimum. (see [24] for further discussion on the local minima and condition (53)).

Therefore, if we take $\boldsymbol{\beta} = \boldsymbol{\beta}^\star$ in (53), we have

$$\begin{aligned}&\big\langle \nabla L(\boldsymbol{\beta}^\star + \widetilde{\boldsymbol{\Delta}}),\, \widetilde{\boldsymbol{\Delta}} \big\rangle \leq -\big\langle \partial \lambda_n \|\boldsymbol{\beta}^\star + \widetilde{\boldsymbol{\Delta}}\|_1,\, \widetilde{\boldsymbol{\Delta}} \big\rangle \overset{(i)}{\leq} \lambda_n \Big(\|\boldsymbol{\beta}^\star\|_1 - \|\widetilde{\boldsymbol{\beta}}_{\lambda_n}\|_1 \Big) \\ &\leq \lambda_n \Big(\|\boldsymbol{\beta}^\star\|_1 + \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 - \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 - \|\widetilde{\boldsymbol{\beta}}_{\lambda_n}\|_1 \Big) \\ &= \lambda_n \Big(\|\boldsymbol{\beta}^\star + \widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 - \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 - \|\widetilde{\boldsymbol{\beta}}_{\lambda_n}\|_1 \Big) \\ &\overset{(ii)}{\leq} \lambda_n \Big(\|\boldsymbol{\beta}^\star + \widetilde{\boldsymbol{\Delta}}_{S^c} + \widetilde{\boldsymbol{\Delta}}_S\|_1 + \|\widetilde{\boldsymbol{\Delta}}_S\|_1 - \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 - \|\widetilde{\boldsymbol{\beta}}_{\lambda_n}\|_1 \Big) \\ &= \lambda_n \Big(\|\widetilde{\boldsymbol{\Delta}}_S\|_1 - \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 \Big) , \end{aligned} \quad (54)$$

where S is true support set of $\boldsymbol{\beta}^\star$ as defined earlier, the inequalities (i) and (ii) hold by respectively the convexity and the triangular inequality of ℓ_1 norm.

Now, combining two ingredients in (52) and (54), we obtain

$$\kappa_1\|\widetilde{\boldsymbol{\Delta}}\|_2^2 - 2\kappa_2 b_0\sqrt{\frac{\log p}{n}}\|\widetilde{\boldsymbol{\Delta}}\|_1 \leq -\langle\nabla L(\boldsymbol{\beta}^\star),\, \widetilde{\boldsymbol{\Delta}}\rangle + \lambda_n\left(\|\widetilde{\boldsymbol{\Delta}}_S\|_1 - \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1\right)$$
$$\leq \left\|\nabla L(\boldsymbol{\beta}^\star)\right\|_\infty\|\widetilde{\boldsymbol{\Delta}}\|_1 + \lambda_n\left(\|\widetilde{\boldsymbol{\Delta}}_S\|_1 - \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1\right).$$

Since the theorem assumes $\max\left\{\|\nabla L(\boldsymbol{\beta}^\star)\|_\infty\,,\, 2\kappa_2 b_0\sqrt{\frac{\log p}{n}}\right\} \leq \frac{\lambda_n}{4}$, we can conclude that

$$\kappa_1\|\widetilde{\boldsymbol{\Delta}}\|_2^2 \leq \left\|\nabla L(\boldsymbol{\beta}^\star)\right\|_\infty\|\widetilde{\boldsymbol{\Delta}}\|_1 + \lambda_n\left(\|\widetilde{\boldsymbol{\Delta}}_S\|_1 - \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1\right) + \left(2\kappa_2 b_0\sqrt{\frac{\log p}{n}}\right)\|\widetilde{\boldsymbol{\Delta}}\|_1$$
$$\leq \frac{3\lambda_n}{2}\|\widetilde{\boldsymbol{\Delta}}_S\|_1 - \frac{\lambda_n}{2}\|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 \tag{55}$$

where we have already shown how the term $\|\nabla L(\boldsymbol{\beta}^\star)\|_\infty$ can be upper bounded in Theorem 2. As a result, we can finally have an ℓ_2 error bound as follows:

$$\kappa_1\|\widetilde{\boldsymbol{\Delta}}\|_2^2 \leq \frac{3\lambda_n}{2}\|\widetilde{\boldsymbol{\Delta}}_S\|_1 \leq \frac{3\lambda_n\sqrt{s}}{2}\|\widetilde{\boldsymbol{\Delta}}_S\|_2 \leq \frac{3\lambda_n\sqrt{s}}{2}\|\widetilde{\boldsymbol{\Delta}}\|_2$$

implying that

$$\|\widetilde{\boldsymbol{\Delta}}\|_2 \leq \frac{3\sqrt{s}\lambda_n}{2\kappa_1}.$$

At the same time we can also derive ℓ_1 error bound using the inequality by (55):

$$\|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 \leq 3\|\widetilde{\boldsymbol{\Delta}}_S\|_1.$$

Hence, combining with ℓ_2 error bound, we obtain

$$\|\widetilde{\boldsymbol{\Delta}}\|_1 \leq \|\widetilde{\boldsymbol{\Delta}}_S\|_1 + \|\widetilde{\boldsymbol{\Delta}}_{S^c}\|_1 \leq 4\|\widetilde{\boldsymbol{\Delta}}_S\|_1 \leq 4\sqrt{s}\|\widetilde{\boldsymbol{\Delta}}_S\|_2 \leq \frac{6s\lambda_n}{\kappa_1}$$

which completes the proof. □

Acknowledgements

The authors would like to thank an anonymous reviewer for comments and suggestions that greatly improved the manuscript and Alekh Agarwal for helpful discussion on the optimization.

References

[1] Alfons, A., Croux, C., and Gelper, S. (2013), "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *Ann. Appl. Stat.*, 7, 226–248. MR3086417

[2] Antczak, T. (2013), "The Exact l1 Penalty Function Method for Constrained Nonsmooth Invex Optimization Problems," in *System Modeling and Optimization*, Springer Berlin Heidelberg, vol. 391 of *IFIP Advances in Information and Communication Technology*, pp. 461–470.

[3] Aravkin, A., Friedlander, M., Herrmann, F. J., and van Leeuwen, T. (2012), "Robust inversion, dimensionality reduction, and randomized sampling," *Mathematical Programming*, 134, 101–125. MR2947554

[4] Arefin, A., Mathieson, L., Johnstone, D., Berretta, R., and Moscato, P. (2012), "Unveiling clusters of RNA transcript pairs associated with markers of Alzheimer's disease progression," *PLoS ONE*, 7 (9), e45535.

[5] Arendt, T., Holzer, M., Stöbe, A., Gärtner, U., Lüth, H. J., Brückner, M. K., and Ueberham, U. (2000), "Activated mitogenic signaling induces a process of dedifferentiation in Alzheimer's disease that eventually results in cell death," *Annals of the New York Academy of Science*, 920–249.

[6] Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012), "Optimization with sparsity-inducing penalties," *Foundations and Trends in Machine Learning*, 4, 1–106.

[7] Bartlett, P. L. and Mendelson, S. (2003), "Rademacher and gaussian complexities: risk bounds and structural results," *Journal of Machine Learning Research*, 3, 463–482.

[8] Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998), "Robust and efficient estimation by minimising a density power divergence," *Biometrika*, 85. MR1665873

[9] Ben-Israel, A. and Mond, B. (1986), "What is invexity," *Journal of the Australian Mathematical Society Series B*, 28, 1–9. MR0846778

[10] Beran, R. (1977), "Robust location estimates," *Annals of Statistics*, 5, 431–444. MR0448699

[11] Bertsekas, D. (2011), "Incremental gradient, subgradient, and proximal methods for convex optimization: a survey," *Optimization for Machine Learning, MIT Press*.

[12] Bickel, P., Ritov, Y., and Tsybakov, A. (2009), "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, 37, 1705–1732. MR2533469

[13] Chi, E. C. and Scott, D. W. (2014), "Robust parametric classification and variable selection by a minimum distance criterion," *Journal of Computational and Graphical Statistics*, 23, 111–128. MR3173763

[14] Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap Methods and Their Applications*, Cambridge: Cambridge University Press, iSBN 0-521-57391-2. MR1478673

[15] Donoho, D. L. and Liu, R. C. (1994), "The "Automatic" robustness of minimum distance functional," *Annals of Statistics*, 16, 552–586. MR0947562

[16] Fan, J., Lv, J., and Qi, L. (2011), "Sparse high dimensional models in economics," *Annual Review of Economics*, 3, 291.

[17] Ghai, R., Mobli, M., Norwood, S. J., Bugarcic, A., Teasdale, R. D., et al. (2011), "Phox homology band 4.1/ezrin/radixin/moesin-like proteins function as molecular scaffolds that interact with cargo receptors and Ras GT-

Pases," *Proceedings of the National Academy of Science USA*, 108, 7763–7768.

[18] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics:The Approach Based on Influence Functions*, Wiley Series in Probability and Statistics. MR0829458

[19] Huber, P. J. (1981), *Robust Statistics*, Wiley New York. MR0606374

[20] Jacob, L., Obozinski, G., and Vert, J.-P. (2009), "Group lasso with overlap and graph lasso," in *Proc. of the 26th Annual International Conference on Machine Learning*, New York, NY, USA: ACM, pp. 433–440.

[21] Jenatton, R., Gramfort, A., Michel, V., Obozinski, G., Eger, E., Bach, F., and Thirion, B. (2012), "Multi-scale mining of fMRI data with hierarchical structured sparsity," *SIAM Journal on Imaging Sciences*, 5, 835–856. MR3022180

[22] Jiang, X., Jia, L. W., Li, X. H., Cheng, X., Xie, J. Z., Ma, Z. W., Xu, W. J., Liu, Y., Yao, Y., Du, L. L., and Zhou, X. W. (2013), "Capsaicin ameliorates stress-induced Alzheimer's disease-like pathological and cognitive impairments in rats," *Journal of Alzheimer's Disease*, 35 (1), 91–105.

[23] Ledoux, M. and Talagrand, M. (1991), *Probability in Banach Spaces: Isoperimetry and Processes*, Ergebnisse der Mathematik und Ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics Series, Springer. MR1102015

[24] Loh, P.-L. and Wainwright, M. J. (2013), "Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima," http://arxiv.org/abs/1305.2436. MR3335800

[25] Lööv, C., Fernqvist, M., Walmsley, A., Marklund, N., and Erlandsson, A. (2012), "Neutralization of LINGO-1 during in vitro differentiation of neural stem cells results in proliferation of immature neurons," *PLoS ONE*.

[26] Mairal, J. and Yu, B. (2013), "Supervised feature selection in graphs with path coding penalties and network flows," http://arxiv.org/abs/1204.4539. MR3111369

[27] Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006), *Robust Statistics: Theory and Methods*, Chichester: Wiley. MR2238141

[28] Martins, A., Figueiredo, M. A. T., Aguiar, P., Smith, N. A., and Xing, E. P. (2011), "Online learning of structured predictors with multiple kernels," in *International Conf. on Artificial Intelligence and Statistics - AISTATS*.

[29] Meinshausen, N. and Bühlmann, P. (2006), "High-dimensional graphs and variable selection with the Lasso," *Annals of Statistics*, 34, 1436–1462. MR2278363

[30] Negahban, S., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012), "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," *Statististical Science*, 27, 538–557. MR3025133

[31] Nesterov, Y. E. (2007), "Gradient methods for minimizing composite objective function," *Technical Report 76, Center of Operations Research and Econometrics, Catholic University of Louvain*.

[32] Nguyen, N. H., Nasrabadi, N. M., and Tran, T. D. (2011), "Robust Lasso

with missing and grossly corrupted observations," *Advances in Neural Information Processing Systems 24*, 1881–1889.

[33] Raskutti, G., Wainwright, M. J., and Yu, B. (2010), "Restricted Eigenvalue Properties for Correlated Gaussian Designs," *Journal of Machine Learning Research*, 11, 2241–2259.

[34] Reiman, E., Webster, J., Myers, A., Hardy, J., Dunckley, T., Zismann, V. L., Joshipura, K. D., Pearson, J. V., Hu-Lince, D., Huentelman, M. J., Craig, D. W., Coon, K. D., et al. (2007), "GAB2 alleles modify Alzheimer's risk in APOE epsilon4 carriers," *Neuron*, 54, 713–720.

[35] Richard, E., Savalle, P., and Vayatis, N. (2012), "Estimation of simultaneously sparse and low rank matrices," in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, New York, NY, USA, pp. 1351–1358.

[36] Scott, D. (2001), "Parametric statistical modeling by minimum integrated square error," *Technometrics*, 43, 274–285. MR1943184

[37] Sugiyama, M., Suzuki, T., Kanamori, T., Du Plessis, M. C., Liu, S., and Takeuchi, I. (2012), "Density-difference estimation," *Advances in Neural Information Processing Systems*, 25, 692–700.

[38] Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. MR1379242

[39] Tibshirani, R., Saunders, M., Rosset, R., Zhu, J., and Knight, K. (2005), "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society Series B*, 91–108. MR2136641

[40] van Rijsbergen, C. J. (1979), *Information Retrieval*, Butterworth.

[41] Vollbach, H., Heun, R., Morris, C. M., Edwardson, J. A., McKeith, I. G., Jessen, F., Schulz, A., Maier, W., and Kölsch, H. (2005), "APOA1 polymorphism influences risk for early-onset nonfamiliar AD," *Annals of Neurology*, 58, 436–441.

[42] Vu, V. Q., Ravikumar, P., Naselaris, T., Kay, K. N., Gallant, J. L., and Yu, B. (2011), "Encoding and decoding V1 FMRI responses to natural images with sparse nonparametric models," *Annals of Applied Statistics*, 5, 1159–1182. MR2849770

[43] Wang, H., Li, G., and Jiang, G. (2007), "Robust regression shrinkage and consistent variable selection through the LAD-lasso," *Journal of Business and Economics Statistics*, 25, 347–355. MR2380753

[44] Wolfowitz, J. (1957), "The minimum distance method," *Annals of Mathematical Statistics*, 28, 75–88. MR0088126

[45] Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. M., and Lange, K. (2009), "Genome-wide association analysis by lasso penalized logistic regression," *Bioinformatics*, 25, 714–721.

[46] Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, 68, 49–67. MR2212574