

# Gaining Outlier Resistance with Progressive Quantiles: Fast Algorithms and Theoretical Studies

Yiyuan She, Zhifeng Wang, Jiahui Shen  
Department of Statistics, Florida State University

## Abstract

Outliers widely occur in big-data applications and may severely affect statistical estimation and inference. In this paper, a framework of outlier-resistant estimation is introduced to robustify an arbitrarily given loss function. It has a close connection to the method of trimming and includes explicit outlyingness parameters for all samples, which in turn facilitates computation, theory, and parameter tuning. To tackle the issues of nonconvexity and nonsmoothness, we develop scalable algorithms with implementation ease and guaranteed fast convergence. In particular, a new technique is proposed to alleviate the requirement on the starting point such that on regular datasets, the number of data resamplings can be substantially reduced. Based on combined statistical and computational treatments, we are able to perform nonasymptotic analysis beyond M-estimation. The obtained resistant estimators, though not necessarily globally or even locally optimal, enjoy minimax rate optimality in both low dimensions and high dimensions. Experiments in regression, classification, and neural networks show excellent performance of the proposed methodology at the occurrence of gross outliers.

## 1 Introduction

Outliers are bound to occur in real-world big data and may severely affect statistical analysis. Many commonly used methods, including the lasso (Tibshirani, 1996), break down in the presence of a single bad observation. We study the problem in a general setup. Given an arbitrary loss function  $l$ , a response vector  $\mathbf{y} \in \mathbb{R}^n$  and a design matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$ , the coefficient vector  $\boldsymbol{\beta} \in \mathbb{R}^p$  can be estimated by solving the following optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} l(\boldsymbol{\eta}; \mathbf{y}) + \sum_{j=1}^p P(\beta_j) \text{ s.t. } \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \quad (1)$$

where  $l$  measures the discrepancy between the observed response and the prediction, and  $P$  represents a regularizer that can also be a constraint. Neither  $l$  nor  $P$  is necessarily convex. The loss can be, say, a deviance function corresponding to a generalized linear model (GLM), a robust loss like Huber's loss, or the well-known hinge loss or exponential

loss for the purpose of classification. The loss function need not be associated with any probability distribution.

The goal of this work is to robustify (1) for any given  $l$  at the occurrence of a small number of *gross outliers*—we stress that the desired **outlier resistance** should not only accommodate samples with mild deviations from the model assumption, but also handle extreme anomalies with severe outlyingness and leverage. Some frequently used robust losses—Huber’s loss and the  $\ell_1$ -norm loss, in particular—are not resistant to gross outliers, and hence one may want a reinforced robustification in the setup of (1). The nature of the task eventually leads to an  $\ell_0$ -type regularization proposal.

The research of robust statistics undergoes a long history and leads to fruitful outcomes. A standard means to gain robustness is to modify the loss, or equivalently, reweight the samples. We illustrate the idea with robust regression. Instead of minimizing the quadratic loss  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$ , M-estimation (Huber and Ronchetti, 2009; Hampel et al., 2011; Maronna et al., 2006) solves  $\min_{\boldsymbol{\beta}} \sum_i \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$  or  $\mathbf{X}^T \psi(\mathbf{X}\boldsymbol{\beta} - \mathbf{y}) = \mathbf{0}$ , where  $\rho$  is a robust loss and  $\psi = \rho'$ . Let  $w(t) = \psi(t)/t$  and  $\mathbf{r} = \mathbf{X}\boldsymbol{\beta} - \mathbf{y}$ . The score equation can be written as  $\mathbf{X}^T \{w(\mathbf{r}) \circ (\mathbf{X}\boldsymbol{\beta} - \mathbf{y})\} = \mathbf{0}$ , where the multiplicative weights  $w(r_i)$  are applied componentwise. The design of a robust loss thus amounts to that of a weight function. See Loh (2017) and Avella-Medina (2017), for example, for some theoretical properties. We call the above scheme the *multiplicative* robustification, to contrast with the *additive* fashion to be introduced in Section 2. An iterative algorithm called iterative reweighted least squares (IRLS) is popularly used in computation. The most critical and expensive step however lies in the preliminary resistant fit with high-breakdown (Rousseeuw and Yohai, 1984; Rousseeuw, 1985), also the focus of our paper. Some challenges of resistant estimation in methodology, theory, and computation are summarized as follows.

First, when  $l(\boldsymbol{\eta}; \mathbf{y})$  has a more complicated expression than that in terms of  $r_i = y_i - \mathbf{x}_i^T \boldsymbol{\beta}$  only, how to reshape it to guard against gross outliers does not seem straightforward. According to Bianco and Yohai (1996), simply applying  $\rho$  on the deviance residuals of a GLM (Pregibon, 1982) does not ensure consistency and an extra bias correction term must be added into the criterion. Croux and Haesbroeck (2003) limited  $\rho$  to a certain class to guarantee the existence of a finite solution, and included an additional weighting step to get a bounded influence function. We refer to Avella-Medina and Ronchetti (2018) for a robust quasi-likelihood form by use of both a  $\psi$  function and a weight function. To us, another important line of work on the trimmed likelihood estimator (TLE) (Hadi and Luceño, 1997; Vandev and Neykov, 1998; Kurnaz et al., 2018), as an extension of least trimmed squares (LTS) (Rousseeuw, 1985), is most motivating. LTS and TLE give rise to our new regularized form of resistant estimation that is more amenable to both optimization and analysis.

The theoretical challenges faced by today’s robust statisticians are perhaps even more pressing. For example, even under  $n > p$  (let alone in high-dimensional settings), (a) how trimming inflates the risk, (b) how small the universal minimax lower bound could be, and (c) what makes a theoretically sound model selection criterion, are all surprisingly unknown in the literature. We emphasize the *finite-sample* nature of desired statistical theory (e.g., Theorems 4–6), because in real-life data applications, it is often hard to know how large the

sample size should be relative to the number of unknowns to apply asymptotics. Breakdown point provides a useful index of the robustness of a given method, but as pointed out by [Huber and Ronchetti \(2009\)](#), such worst-case studies are not probabilistic and may be too conservative on a particular dataset. In addition to robust analysis, parameter tuning based on asymptotics or breakdown point suffers the same theoretical difficulties.

The last big challenge lies in computation. Resistant fits are much more costly than IRLS or  $\Theta$ -IPOD ([She and Owen, 2011](#)) which takes them as initial points. Due to the inherent nonconvexity, subset sampling is heavily used in LTS-like algorithms to generate multiple starting values ([Rousseeuw and Driessen, 1999](#)). Unfortunately, the number of required samplings grows exponentially in  $p$  and so these procedures already become inaccurate and/or prohibitive for moderate  $p$ . Developing more efficient outlier-resistant algorithms and mitigating the burden of subset sampling are at the core of modern-day robust statistics. Moreover, the fact that the computationally obtained solutions are not necessarily globally optimal poses nontrivial challenges in theoretical analysis.

This paper attempts to address some aforementioned obstacles to gain outlier resistance, with the hope to advance the practice of robust statistics to more sophisticated learning tasks. Our main contributions are threefold. a) A general resistant learning framework is introduced to robustify an arbitrarily given loss, with an “ $\ell_0 + \ell_2$ ” form of regularization to deal with gross outliers in possibly high dimensions. It has a close connection to the method of trimming but includes explicit outlyingness parameters, which in turn facilitates computation, theory, and parameter tuning. b) Although the associated problem is highly nonconvex and nonsmooth, extremely scalable algorithms with implementation ease can be developed with guaranteed convergence. In particular, the statistical error of the sequence of iterates can converge *geometrically* fast. A progressive optimization technique is proposed to relax the regularity condition and alleviate the requirement on the starting point so that on regular datasets the number of data resamplings can be substantially reduced. c) Combined statistical and computational treatments make it possible to perform nonasymptotic analysis beyond the M-estimation setting, and the obtained “A-estimators”, though not necessarily globally or even locally optimal, enjoy minimax rate optimality in both low dimensions and high dimensions.

The rest of the paper is organized as follows. Section 2 introduces our regularized resistant learning framework and compares it to trimming. Section 3 deals with the computational issues by progressive iterative quantile thresholding. Section 4 provides nonasymptotic theoretical support for the obtained estimators under proper regularity conditions. Section 5 reveals some universal minimax lower bounds and investigates the issue of parameter tuning. Section 6 performs extensive simulation studies and real world data analysis. The proof details and more simulation results are given in the appendix.

**Notation.** We use  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T$  to denote the design matrix and  $\mathbf{x}_i$  the  $i$ th sample. Throughout the paper, we use  $l(\boldsymbol{\eta}; \mathbf{y})$  to denote a differentiable loss defined on the systematic component  $\boldsymbol{\eta}$ , and assume that  $l$  is bounded from below and  $\inf_{\boldsymbol{\eta}} l(\boldsymbol{\eta}; \mathbf{y}) \geq 0$  without loss of generality. Sometimes, such as when building the connection to trimming,

we assume that the overall loss takes a sample-additive form,  $l(\boldsymbol{\eta}; \mathbf{y}) = \sum_{i=1}^n l_0(\eta_i; y_i)$ .  $l(\boldsymbol{\eta}; \mathbf{y})$  is said to be  $L$ -strongly smooth or have an  $L$ -Lipschitz continuous gradient if

$$\|\nabla l(\boldsymbol{\eta}_1; \mathbf{y}) - \nabla l(\boldsymbol{\eta}_2; \mathbf{y})\|_2 \leq L\|\boldsymbol{\eta}_1 - \boldsymbol{\eta}_2\|_2, \forall \boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \quad (2)$$

for some  $L > 0$ . For ease of presentation, we frequently use the concatenated notation

$$\bar{\mathbf{X}} = [\mathbf{X}, \mathbf{I}], \quad \bar{\boldsymbol{\beta}} = [\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T]^T,$$

and so  $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} = \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}$ . For any matrix  $\mathbf{A}$ ,  $\|\mathbf{A}\|_2$  is its spectral norm. Given any  $\boldsymbol{\beta} \in \mathbb{R}^p$ ,  $\|\boldsymbol{\beta}\|_0 = \sum_{j=1}^p 1_{\beta_j \neq 0}$ . The hat matrix  $\mathbf{H}$  of  $\mathbf{X}$  is  $\mathbf{X}(\mathbf{X}^T \mathbf{X})^+ \mathbf{X}^T$  with  $^+$  denoting the Moore-Penrose inverse. Given any two vectors  $\boldsymbol{\alpha}, \boldsymbol{\beta}$  of the same size, the inner product  $\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle$  is given by  $\sum \alpha_j \beta_j$ .

Given  $\mathbf{X} \in \mathbb{R}^{n \times p}$  and  $s \leq p$ , we introduce  $0 \leq m_{\mathbf{X}}(s), M_{\mathbf{X}}(s)$  associated with the *restricted isometry property* (Candes and Tao, 2005) by  $m_{\mathbf{X}}(s)\|\boldsymbol{\beta}\|_2^2 \leq \|\mathbf{X}\boldsymbol{\beta}\|_2^2 \leq M_{\mathbf{X}}(s)\|\boldsymbol{\beta}\|_2^2, \forall \boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 \leq s$ . In particular,  $M_{\bar{\mathbf{X}}}(s, o)$  ( $s \leq p, o \leq n$ ) obeying

$$\|\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}\|_2^2 \leq M_{\bar{\mathbf{X}}}(s, o)\|\bar{\boldsymbol{\beta}}\|_2^2, \forall \bar{\boldsymbol{\beta}} : \|\boldsymbol{\beta}\|_0 \leq s, \|\boldsymbol{\gamma}\|_0 \leq o$$

will play an active role in our algorithm design and analysis. Obviously,  $M_{\mathbf{X}}(s) \leq \|\mathbf{X}\|_2^2$  and  $M_{\bar{\mathbf{X}}}(s, o) \leq \|\bar{\mathbf{X}}\|_2^2$ . But since we are primarily interested in  $s, o \ll n$ ,  $\mathbf{X}\boldsymbol{\beta}$  or  $\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}$  for an  $s$ -restricted  $\boldsymbol{\beta}$  or an  $(s, o)$ -restricted  $\bar{\boldsymbol{\beta}}$  involves a much thinner design, and so  $M_{\mathbf{X}}(s), M_{\bar{\mathbf{X}}}(s, o)$  can be way smaller.

We use  $C, c$  and  $L$  to denote universal constants which are not necessarily the same at each occurrence and  $\lesssim$  denotes an inequality that holds up to a multiplicative numerical constant. Finally,  $0 \log 0 = 0$ ,  $a \vee b = \max\{a, b\}$  and  $a \wedge b = \min\{a, b\}$  are adopted.

## 2 $L_0$ -regularization vs. Trimming

We begin with the robust estimation problem (1) without regularization. The customizable loss determines the performance metric, so we prefer *not* to alter its form in the process of robustification. This is possible by use of an additive regularized estimation.

Concretely, re-define  $\boldsymbol{\eta}$  as  $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}$ , where  $\gamma_i$  characterizes the outlyingness of the  $i$ th sample and we allow it to be large. Taking advantage of the sparsity in  $\boldsymbol{\gamma}$ , as outliers are never the norm, we propose an  $\ell_0$ -constrained,  $\ell_2$ -penalized outlier-resistant estimation

$$\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} l(\boldsymbol{\eta}; \mathbf{y}) + \frac{\nu}{2}\|\boldsymbol{\gamma}\|_2^2 \quad \text{s.t. } \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}, \|\boldsymbol{\gamma}\|_0 \leq q, \quad (3)$$

where  $q$  satisfies  $q \leq n/2$  and is assumed to be an integer throughout the paper. Here, the number of unknowns,  $n + p$ , is larger than the number of observations, and the “ $\ell_0 + \ell_2$ ” regularization plays an effective role in dealing with the high-dimensional challenge.

Unlike the popular  $\ell_1$ -norm penalty, the  $\ell_0$  ‘norm’ used in the constraint is indifferent to the magnitude of  $\gamma_i$  and does not result in any undesired bias. An alternative idea

is to utilize a non-convex sparse penalty—in the regression setting, this amounts to M-estimation, the diverse choices of the penalty lead to different robust losses (She and Owen, 2011), and the conclusion extends to regularized  $\beta$ -estimation in possibly high dimensions (She and Chen, 2017, Lemma 2). But (3) has some distinct advantages: the constraint directly controls the maximum number of outliers and  $\|\cdot\|_0$  is arguably the ideal function to enforce sparsity regardless of the severity of aberrant samples. In practice, specifying the value of  $q$ , rather than a penalty parameter  $\lambda$ , is easier and more convenient.

In addition to the sparsity-promoting regularization, we include the  $\ell_2$ -shrinkage to enhance numerical stability and compensate for collinearity in the presence of clustered outliers. We will see in Section 3.1 that with  $\nu$  introduced, the samples with minor deviations can still contribute to the model fitting. Empirically,  $\nu$  is not a sensitive parameter, and we often set  $\nu$  to a mild value (say 1e-4). As long as  $\nu > 0$ , the  $\gamma$ -estimate is finite. But if we force  $\nu = 0$ , (3) has an interesting and important connection to the method of trimming (Rousseeuw, 1985; Hadi and Luceño, 1997; Vandev and Neykov, 1998).

**Theorem 1.** *Assume  $l(\eta; \mathbf{y})$  takes a sample-additive form  $\sum_{i=1}^n l_0(\eta_i; y_i)$  and  $\inf_{\eta} l_0(\eta; y) = 0$  for all  $y \in \mathcal{Y} \subset \mathbb{R}$ .*

i) *Given any minimizer  $(\hat{\beta}, \hat{\gamma})$  of the  $\ell_0$ -constrained joint problem  $\min_{\beta, \gamma} l(\mathbf{X}\beta + \gamma; \mathbf{y})$  s.t.  $\|\gamma\|_0 \leq q$ ,  $\hat{\beta}$  is also a globally optimal solution to the trimmed problem on  $\beta$ :  $\min_{\beta} \sum_{i=1}^{n-q} l_0^{(i)}(\beta)$  with  $l_0^{(1)}(\beta) \leq \dots \leq l_0^{(n)}(\beta)$  the order statistics of  $l_0(\mathbf{x}_i^T \beta; y_i)$ ; conversely, given any  $\hat{\beta}$  that minimizes  $\sum_{i=1}^{n-q} l_0^{(i)}(\beta)$ , there always exists a  $\hat{\gamma} \in \bar{\mathbb{R}}^n$  with  $\bar{\mathbb{R}} = [-\infty, \infty]$  so that  $(\hat{\beta}, \hat{\gamma})$  is a solution to the  $\ell_0$ -constrained joint problem.*

ii) *The same connection holds between the  $\ell_0$ -penalized joint problem  $\min_{\beta, \gamma} l(\mathbf{X}\beta + \gamma; \mathbf{y}) + \tau \|\gamma\|_0$  and the winsorized problem  $\min_{\beta} \sum \tau \wedge l_0(\mathbf{x}_i^T \beta; y_i)$  for any cutoff  $\tau \geq 0$ .*

iii) *Finally, any minimizer  $\hat{\beta}$  of the winsorized optimization problem is a solution to the trimmed problem with  $q = |\{i : l_0(\mathbf{x}_i^T \hat{\beta}; y_i) > \tau\}|$ , and any  $(\hat{\beta}, \hat{\gamma})$  that minimizes the  $\ell_0$ -penalized joint criterion is also an optimal solution to the  $\ell_0$ -constrained joint problem with  $q = \|\hat{\gamma}\|_0$ .*

So in a sense, imposing the  $\ell_0$  regularization on  $\gamma$  amounts to using a trimmed version or setting a cutoff of the original loss on  $\beta$ , both of which can effectively bound the influence of outliers and necessarily result in nonconvexity. The joint  $(\beta, \gamma)$  formulation under  $\ell_0$ -penalization corresponds to winsorizing the loss (which unsurprisingly gives a redescending  $\psi$  in the setup of robust regression), and the  $\ell_0$ -constrained form has a similar power but utilizes a data-adaptive cutoff. It is noteworthy that our additive regularized framework makes it possible to utilize any loss (which can be convex) for resistant estimation.

**Remark 2.1.** *Because of the connections, some breakdown-point and influence-function results of the estimators defined by (3) can be directly obtained from say Alfons et al. (2013) and Öllerer et al. (2015). For example, Proposition 1 of Alfons et al. (2013) can be slightly modified to argue that the breakdown point is greater than  $q/n$  when  $\eta > 0$  and  $l$  is convex under some regularity conditions. Moreover, interested readers may refer to Appendix A.11 for a risk-based breakdown point defined via the Orlicz norm of the effective noise, which takes the randomness into account, cf. (A.76), (A.77). One can then combine the empirical*

process theory and generalized Bregman calculus to perform breakdown point studies for an extended real-valued loss  $l(\boldsymbol{\eta}; \mathbf{y})$  that may not be a function of  $\mathbf{y} - \boldsymbol{\eta}$  as in  $M$ -estimation. See Remark A.5 and Theorem A.5 for details, where the benefit of using a strongly convex  $l$  on the systematic component is shown.

On the other hand, the equivalence between the sets of globally optimal solutions does not preclude practically obtained estimates from being different. For the analysis of locally optimal or alternatively optimal estimators to be presented in later sections, our additive regularized form appears to be advantageous over the trimming form.

Finally, an extension of (3) to high dimensions gives rise to resistant variable selection or resistant shrinkage estimation

$$\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} l(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}; \mathbf{y}) + P(\boldsymbol{\beta}; \lambda) + \frac{\nu}{2} \|\boldsymbol{\gamma}\|_2^2 \text{ s.t. } \|\boldsymbol{\gamma}\|_0 \leq q. \quad (4)$$

$P(\boldsymbol{\beta}; \lambda)$  can be, say, a sparsity-inducing penalty/constraint, or an  $\ell_2$ -type penalty. (4) allows for  $p > n$ . Similar to Theorem 1, in the special case of  $l(\boldsymbol{\eta}; \mathbf{y}) = \|\mathbf{y} - \boldsymbol{\eta}\|_2^2/2$ ,  $P(\boldsymbol{\beta}; \lambda) = \lambda \|\boldsymbol{\beta}\|_1$  and  $\nu = 0$ , (4) can be shown to be equivalent to the sparse LTS (Alfons et al., 2013). In comparison, our explicit introduction of the outlyingness vector  $\boldsymbol{\gamma}$  provides new insights. First, outliers can be directly revealed from the  $\boldsymbol{\gamma}$ -estimate. Second, the formulation eases the design of optimization-based algorithms (Section 3). Third, (3) and (4) enable us to borrow high dimensional statistics tools to investigate the non-asymptotic behavior of resistant estimators (Section 4) and develop a new information criterion for regularization parameter tuning (Section 5).

### 3 Progressive Iterative Quantile-thresholding

This section studies how to solve the computational problems defined in Section 2, where the main obstacle lies in nonconvexity and nonsmoothness. Thanks to the *sparsity* of the problem, obtaining a globally optimal solution may not be necessary in many real world applications. This section develops two classes of iterative algorithms, the BCD type and the MM type, both of which can accommodate a differentiable loss function that is not necessarily convex. More importantly, we propose a progressive optimization technique to relax the condition required to enjoy the best statistical accuracy.

#### 3.1 Outlier-resistant regression

To illustrate the main ideas and techniques, this part studies outlier-resistant regression with  $n > p$ :

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\gamma} \in \mathbb{R}^n} \frac{1}{2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\gamma}\|_2^2 + \frac{\nu}{2} \|\boldsymbol{\gamma}\|_2^2 := f(\boldsymbol{\beta}, \boldsymbol{\gamma}) \text{ s.t. } \|\boldsymbol{\gamma}\|_0 \leq q. \quad (5)$$

Compared with LTS, the inclusion of  $\boldsymbol{\gamma}$  in (5) renders the quadratic loss intact, and so block-wise coordinate descent (**BCD**) can be used for computation. The sub-problem of

$\beta$  is trivial, and given  $\beta$ , the best  $\gamma$  can be obtained with a quantile thresholding operator  $\Theta^\#$  (She et al., 2013). Given any  $s \in \mathbb{R}^n$ , define  $\Theta^\#(s; q, \nu)$  to be a vector  $t \in \mathbb{R}^n$  satisfying  $t_{(j)} = s_{(j)} / (1 + \nu)$  if  $1 \leq j \leq q$ , and 0 otherwise, where  $s_{(1)}, \dots, s_{(n)}$  are the order statistics of  $s_1, \dots, s_n$  satisfying  $|s_{(1)}| \geq \dots \geq |s_{(n)}|$ , and  $t_{(1)}, \dots, t_{(n)}$  are defined similarly. To ensure  $\Theta^\#$  is a function and avoid ambiguity, we assume throughout the paper that in performing  $\Theta^\#(s; q, \nu)$ , either  $|s_{(q)}| > |s_{(q+1)}|$  or  $|s_{(q)}| = |s_{(q+1)}| = 0$  occurs, referred to as the  $\Theta^\#$ -uniqueness assumption. The resultant BCD algorithm proceeds as follows

$$\begin{cases} \gamma^{(t+1)} = \Theta^\#(\mathbf{y} - \mathbf{X}\beta^{(t)}; q, \nu), \\ \beta^{(t+1)} = (\mathbf{X}^T \mathbf{X})^+ \mathbf{X}^T (\mathbf{y} - \gamma^{(t+1)}), \end{cases} \quad (6)$$

or simply

$$\gamma^{(t+1)} = \Theta^\#(\mathbf{H}\gamma^{(t)} + (\mathbf{I} - \mathbf{H})\mathbf{y}; q, \nu), \quad (7)$$

starting with  $\beta^{(0)} \in \mathbb{R}^p$  or  $\gamma^{(1)} \in \mathbb{R}^n$ . (6) and (7) are referred to as iterative quantile-thresholding (IQ) algorithms in the paper. Intuitively, (7) repeatedly thresholds a weighted average of  $\mathbf{y}$  and the current  $\gamma$ -estimate. But the quantile thresholding is no ordinary hard-thresholding due to its *iteration-varying* threshold and concurrent  $\ell_2$  shrinkage. (7) shares some similarity with the celebrated “C-step” designed on the basis of trimming (Rousseeuw and Driessen, 1999). On the other hand, at each iteration C-step only uses a subset of observations to get the coefficient vector, while IQ uses all the samples with adjustment. In a sense, C-step’s subset selection amounts to checking the zero-nonzero pattern of  $\gamma_i^{(t+1)}$  instead of its exact value.

The algorithm also suggests some virtues of the complementary  $\ell_2$ -shrinkage. When  $\nu = 0$ , given the optimal  $\hat{\beta}$ ,  $\hat{\gamma}_i = 0$  or  $y_i - \mathbf{x}_i^T \hat{\beta}$ . Letting  $d_i = 1_{\hat{\gamma}_i \neq 0}$  and  $\mathbf{D} = \text{diag}\{d_i\}$ , from (6),  $\hat{\beta}$  satisfies

$$\mathbf{X}^T (\mathbf{I} - \mathbf{D}) \mathbf{X} \hat{\beta} = \mathbf{X}^T (\mathbf{I} - \mathbf{D}) \mathbf{y}.$$

Because of the binary nature of  $d_i$ , the  $i$ th observation is either kept unaltered or removed completely, which might hurt the estimation accuracy in the presence of a number of mildly outlying observations. When  $\nu > 0$ , the above equation still holds for  $d_i = 1_{\hat{\gamma}_i \neq 0} / (1 + \nu)$ , and the fact that  $\mathbf{I} - \mathbf{D}$  is nonsingular even when  $q$  takes a conservatively large value secures numerical stability and variance reduction. More insights can be gained by transforming  $\mathbf{y} = \mathbf{X}\beta^* + \gamma^* + \epsilon$  to  $\mathbf{y}' = \mathbf{V}^T \gamma^* + \epsilon'$  where  $\mathbf{V}$  is the orthogonal complement to the range of  $\mathbf{X}$ ,  $\mathbf{y}' = \mathbf{V}^T \mathbf{y}$  and  $\epsilon' = \mathbf{V}^T \epsilon$ . Then, during the occurrence of clustered outliers with a large overall leverage, the  $\ell_2$  regularization has a beneficial effect to cope with the coherent Gram matrix  $\mathbf{V}^T \mathbf{V} = \mathbf{I} - \mathbf{H}$ .

Establishing IQ’s rate of convergence is nontrivial since it is a nonconvex optimization algorithm in high dimensions (rigorously speaking,  $n$  observations and  $p + n$  unknowns). We give two results regarding the convergence of its computational error and statistical error.

**Theorem 2.** Consider the algorithm defined by (6) or (7).

a) For any  $T \geq 1$ ,

$$\begin{aligned} \min_{1 \leq t \leq T} \|\mathbf{X}(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t+1)})\|_2^2 &= \min_{1 \leq t \leq T} \|\mathbf{H}(\boldsymbol{\gamma}^{(t)} - \boldsymbol{\gamma}^{(t+1)})\|_2^2 \\ &\leq \frac{1}{T} (\|(\mathbf{I} - \mathbf{H})(\boldsymbol{\gamma}^{(1)} - \mathbf{y})\|_2^2 + \nu \|\boldsymbol{\gamma}^{(1)}\|_2^2). \end{aligned}$$

b) Let  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma}^* + \boldsymbol{\epsilon}$ , where  $\|\boldsymbol{\gamma}^*\|_0 = o^*$ ,  $\boldsymbol{\epsilon}$  is a sub-Gaussian random vector with mean zero and scale bounded by  $\sigma$  (cf. Definition 1 in Appendix A.2). Let  $q = \vartheta o^*$  with  $\vartheta > 1$ . Suppose the design matrix satisfies a restricted isometry property (RIP) for some  $\varepsilon$ :  $0 < \varepsilon < 1$

$$\|\mathbf{H}\boldsymbol{\Delta}\|_2^2 \leq (1 - \varepsilon)\|\boldsymbol{\Delta}\|_2^2, \quad \forall \boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_0 \leq (1 + \vartheta)o^*. \quad (8)$$

Then with probability at least  $1 - C(n - p)^{-c}$ , for any  $t \geq 1$

$$\|\mathbf{H}(\boldsymbol{\gamma}^{(t)} - \boldsymbol{\gamma}^*)\|_2^2 \lesssim \left(\frac{1}{1 + \kappa}\right)^{t-1} \|\mathbf{H}(\boldsymbol{\gamma}^{(1)} - \boldsymbol{\gamma}^*)\|_2^2 + \frac{\nu}{\kappa} \|\boldsymbol{\gamma}^*\|_2^2 + \frac{\sigma^2}{\kappa} \vartheta o^* \log \frac{en}{o^*}, \quad (9)$$

where

$$\kappa = \{\varepsilon - 1/\sqrt{\vartheta} + (1 - 1/\sqrt{\vartheta})\nu\}/(1 - \epsilon), \quad (10)$$

and  $C, c > 0$  are constants. Furthermore, under a slightly stronger noise assumption that  $\epsilon_i$  are independent sub-Gaussian( $0, \sigma^2$ ), the following convergence result holds for all  $t \geq 1$  with probability at least  $1 - C(n - p)^{-c} - C \exp(-cp)$

$$\|\mathbf{X}(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^*)\|_2^2 \lesssim \left(\frac{1}{1 + \kappa}\right)^{t-1} \|\mathbf{X}(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^*)\|_2^2 + \frac{\nu}{\kappa} \|\boldsymbol{\gamma}^*\|_2^2 + p\sigma^2 + \frac{\sigma^2}{\kappa} \vartheta o^* \log \frac{en}{o^*}. \quad (11)$$

The first result in (a) states a *universal* convergence rate of the order  $\mathcal{O}(1/t)$ , and the conclusion is free of any conditions. On the other hand, this rate does not take the model parsimony into full account.

The statistical error convergence shown in (b) is surprising and delightful. Although  $\Theta^\#$  is not a nonexpansive mapping, let alone a contraction, IQ approaches the statistical target *geometrically* fast with high probability, as long as  $\theta$  or  $\nu$  is properly large:

$$\varepsilon + (1 - 1/\sqrt{\vartheta})\nu > 1/\sqrt{\vartheta}. \quad (12)$$

Such a linear convergence result in the discrete nonconvex setup is novel to the best of our knowledge. A positive  $\nu$  can improve the linear rate parameter  $\kappa$ . The error bound remains the same order as long as  $\nu \|\boldsymbol{\gamma}^*\|_2^2$  is dominated by a constant multiple of  $\sigma^2 \vartheta o^* \log \frac{en}{o^*}$ . (The best value of  $\nu$  should be data-adaptive, but fixing it at a mild value already gives satisfactory performance.) The statistical error does not vanish as  $t \rightarrow \infty$  due to the existence of noise, but Section 5 will show that  $p\sigma^2 + o^* \sigma^2 \log(en/o^*)$  is the desired error rate in robust regression. Hence over-optimization seems to be unnecessary.

Of course, to enjoy the fast convergence, a regularity condition on  $\mathbf{H}$  must be assumed. Restricted isometry like (8) is widely used and known to hold in compressed sensing and many high-dimensional sparse problems (Candes and Tao, 2005; Hastie et al., 2015). Intuitively,  $\Delta^T(\mathbf{I} - \mathbf{H})\Delta \geq \varepsilon \|\Delta\|_2^2$ ,  $\forall \Delta : \|\Delta\|_0 \leq (1 + \vartheta)o^*$  means all principal submatrices of size  $(1 + \vartheta)o^*$  of  $\mathbf{I} - \mathbf{H}$  should have eigenvalues greater than or equal to a positive  $\varepsilon$ . The cone to define the restricted eigenvalue shrinks when  $o^*$  is small. It is well known that RIP is much less demanding than the mutual coherence restrictions on  $h_{i,j}(i \neq j)$  (van de Geer and Bühlmann, 2009; Hastie et al., 2015) which are realized by low-leveraged data since  $|h_{i,j}| \leq (h_{ii}h_{jj})^{1/2}$ .

We will show that for strongly convex losses, pursuing *globally* optimal estimators in our additive regularized framework can ultimately remove all regularity conditions, namely, outlier resistance regardless of leverage and outlyingness. But on regular datasets this may be unnecessary. The global minimization comes at a heavy computational cost: multiple random starts have to be used due to the nonconvexity, and experience shows that even when  $p = 50$ , the subset sampling popularly used in robust statistics already becomes ineffective and expensive (Rousseeuw and Hubert, 1997).

So a legitimate question on large data is how to *relax* the condition required for statistical accuracy while maintaining computational efficiency. Theorem 2 sheds new light on the issue. For example, with  $\nu = 0$ ,  $\varepsilon > 1/\sqrt{\vartheta}$  ensures the geometric statistical convergence, and to make it hold more easily, an obvious means is to increase  $\vartheta$ . The larger the value of  $\vartheta$  takes, the bigger  $\varepsilon$  is, and so the smaller the influence of the initial point according to (9) or (11). The adverse effect is the larger error term  $\vartheta o^* \log \frac{en}{o^*}$ . But we can *gradually* tighten the cardinality constraint with a sequence  $Q(t)$  which deceases from  $n$  to  $q$  as  $t$  increases. The resultant algorithm is termed as **Progressive Iterative Quantile-thresholding (PIQ)**. PIQ can substantially improve the mediocre performance of IQ. It is easy to extend Theorem 2 from fixed quantiles to varying quantiles, but deriving a theoretically optimal cooling scheme is challenging. Fortunately, this is not a big issue in practice; various schemes seem to do a decent job, such as  $Q(t) = n - at^2$  (quadratic),  $Q(t) = 2n/(1 + \exp(at))$  (sigmoidal),  $Q(t) = n - a \log t$  (logarithmic) and so on.

Our progressive quantile proposal is innovative in that it is not just another intricate sampling scheme and does not increase the number of initial points. Compared with state-of-the-art algorithms, the tables and figures in Section 6.1 clearly show that it can significantly reduce the need of data resampling (thereby the overall computational time) without sacrificing the statistical accuracy. In addition, we must point out that PIQ works in the *opposite* direction to the class of forward pathwise algorithms and boosting methods (Bühlmann and Yu, 2003; Needell and Tropp, 2009; Zhang, 2013) which all grow a model from the null. Both of our analysis and computer experiments support the backward manner in resistant learning.

An alternative (and perhaps simpler) algorithm can be developed following the principle of majorization-minimization (**MM**) (Hunter and Lange, 2004). The technique is general and will play a vital role in extending BCD-type PIQ to conquer a general loss. Recall the concatenated notation  $\bar{\mathbf{X}} = [\mathbf{X} \ \mathbf{I}]$ ,  $\bar{\boldsymbol{\beta}} = [\boldsymbol{\beta}^T \ \boldsymbol{\gamma}^T]^T$ . Instead of directly minimizing

the original objective function  $f(\bar{\boldsymbol{\beta}}) = \|\mathbf{y} - \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}\|_2^2/2 + \nu\|\boldsymbol{\gamma}\|_2^2/2$ , we construct a surrogate function  $g(\bar{\boldsymbol{\beta}}; \bar{\boldsymbol{\beta}}^-)$  by (joint) linearization

$$g_\rho(\bar{\boldsymbol{\beta}}; \bar{\boldsymbol{\beta}}^-) := \|\mathbf{y} - \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-\|_2^2/2 + \langle \bar{\mathbf{X}}^T(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^- - \mathbf{y}), \bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^- \rangle + \frac{\rho}{2}\|\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-\|_2^2 + \frac{\nu}{2}\|\boldsymbol{\gamma}\|_2^2, \quad (13)$$

where  $\rho$  is the inverse stepsize parameter to be chosen later. Then minimizing  $g$  with respect to  $\bar{\boldsymbol{\beta}}$  yields a sequence of iterates  $\bar{\boldsymbol{\beta}}^{(t+1)} = \arg \min_{\bar{\boldsymbol{\beta}}} g_\rho(\bar{\boldsymbol{\beta}}; \bar{\boldsymbol{\beta}}^{(t)})$  s.t.  $\|\boldsymbol{\gamma}\|_0 \leq q$ . After some algebraic manipulations,  $g$  can be rephrased as

$$\begin{aligned} g_\rho(\bar{\boldsymbol{\beta}}; \bar{\boldsymbol{\beta}}^-) &= \frac{\rho}{2}\|\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^- + \bar{\mathbf{X}}^T(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^- - \mathbf{y})/\rho\|_2^2 + \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-\|_2^2 \\ &\quad - \frac{1}{2\rho}\|\bar{\mathbf{X}}^T(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^- - \mathbf{y})\|_2^2 + \frac{1}{2}\nu\|\boldsymbol{\gamma}\|_2^2 \\ &= \frac{\rho}{2}\|\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^- + \bar{\mathbf{X}}^T(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^- - \mathbf{y})/\rho\|_2^2 + \frac{\nu}{2}\|\boldsymbol{\gamma}\|_2^2 + c(\bar{\boldsymbol{\beta}}^-), \end{aligned} \quad (14)$$

where  $c(\bar{\boldsymbol{\beta}}^-)$  does not depend on  $\boldsymbol{\beta}$ . Noticing the separability in  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$ , the resultant update is given by

$$\begin{cases} \boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{\gamma}^{(t)} - \mathbf{y})/\rho, \\ \boldsymbol{\gamma}^{(t+1)} = \Theta^\#(\boldsymbol{\gamma}^{(t)} - (\mathbf{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{\gamma}^{(t)} - \mathbf{y})/\rho; q, \nu/\rho). \end{cases} \quad (15)$$

As long as the stepsize is properly small, e.g.,  $\rho > \|\bar{\mathbf{X}}\|_2^2$ , (15) is convergent (cf. Theorem 3). A similar conclusion to Theorem 2 can be proved, which supports a progressive scheme with  $Q(t)$  decreasing to  $q$  to relax the requirement of the initial point.

Different from (6), (15) updates  $(\boldsymbol{\beta}, \boldsymbol{\gamma})$  simultaneously and does not involve any matrix inversion or expensive operations. A downside of MM however lies in its smaller step-size, which not only slows down convergence but may (surprisingly) affect statistical accuracy in resistant learning; see Section 4.

## 3.2 Harnessing a general loss

This part uses the optimization techniques developed for resistant regression to deal with a general objective with  $p$  possibly larger than  $n$ :

$$\min_{(\boldsymbol{\beta}, \boldsymbol{\gamma})} l(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}; \mathbf{y}) + P(\varrho\boldsymbol{\beta}; \lambda) + \frac{\nu}{2}\|\boldsymbol{\gamma}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{\gamma}\|_0 \leq q, \quad (16)$$

where  $P$  is often a sparsity-promoting penalty, but can also take the form of a ridge penalty or an  $\ell_0$ -constraint. Here, we add the algorithmic parameter  $\varrho > 0$  to match the scale of the design and aid stepsize selection. With linearization and MM, we can extend (15) to (17), (18) below, applicable to a broad family of  $l$  and  $P$ .

**Theorem 3.** Assume that  $\nabla l(\eta; y)$  is  $L$ -Lipschitz continuous. Given an arbitrary thresholding rule  $\Theta$  with  $\lambda$  as the threshold parameter (cf. Appendix A.3), consider the following algorithm

$$\begin{cases} \boldsymbol{\beta}^{(t+1)} = \Theta(\varrho \boldsymbol{\beta}^{(t)} - \mathbf{X}^T \nabla l(\mathbf{X} \boldsymbol{\beta}^{(t)} + \boldsymbol{\gamma}^{(t)}; \mathbf{y})/\varrho; \lambda)/\varrho, \\ \boldsymbol{\gamma}^{(t+1)} = \Theta^\#(\boldsymbol{\gamma}^{(t)} - \nabla l(\mathbf{X} \boldsymbol{\beta}^{(t)} + \boldsymbol{\gamma}^{(t)}; \mathbf{y})/\varrho^2; q, \nu/\varrho^2), \end{cases} \quad (17)$$

where  $t \geq 0$ ,  $\varrho > 0$ . Let  $f(\bar{\boldsymbol{\beta}}) = l(\bar{\mathbf{X}} \bar{\boldsymbol{\beta}}; \mathbf{y}) + P(\varrho \bar{\boldsymbol{\beta}}; \lambda) + \nu \|\boldsymbol{\gamma}\|_2^2/2$ , where  $P$  is defined based on  $\Theta$ :  $P(\theta; \lambda) - P(0; \lambda) = \int_0^{|\theta|} (\sup\{s : \Theta(s; \lambda) \leq u\} - u) du + q(\theta; \lambda)$  for some  $q(\cdot, \lambda) \geq 0$  satisfying  $q(\Theta(t; \lambda); \lambda) = 0, \forall t \in \mathbb{R}$ . Then, given any starting point  $\bar{\boldsymbol{\beta}}^{(0)}$  and any  $\varrho > 0$ ,  $f(\bar{\boldsymbol{\beta}}^{(t+1)}) \leq f(\bar{\boldsymbol{\beta}}^{(t)}) - (\bar{\boldsymbol{\beta}}^{(t+1)} - \bar{\boldsymbol{\beta}}^{(t)})^T (\varrho^2 \mathbf{I} - L \bar{\mathbf{X}}^T \bar{\mathbf{X}}) (\bar{\boldsymbol{\beta}}^{(t+1)} - \bar{\boldsymbol{\beta}}^{(t)})/2$  and the following convergence rate holds

$$\min_{t \leq T} (\bar{\boldsymbol{\beta}}^{(t+1)} - \bar{\boldsymbol{\beta}}^{(t)})^T (\varrho^2 \mathbf{I} - L \bar{\mathbf{X}}^T \bar{\mathbf{X}}) (\bar{\boldsymbol{\beta}}^{(t+1)} - \bar{\boldsymbol{\beta}}^{(t)}) \leq \frac{2f(\bar{\boldsymbol{\beta}}^{(0)})}{T+1}, \forall T \geq 0. \quad (19)$$

In particular, if  $\varrho^2 > LM_{\bar{\mathbf{X}}}(p, 2q)$  (cf. Section 1) and  $\nu > 0$ , any limit point  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$  of  $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})$  satisfies  $\hat{\boldsymbol{\beta}} = \Theta(\varrho \hat{\boldsymbol{\beta}} - \mathbf{X}^T \nabla l(\mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}}; \mathbf{y})/\varrho; \lambda)/\varrho$  and  $\hat{\boldsymbol{\gamma}} = \Theta^\#(\hat{\boldsymbol{\gamma}} - \nabla l(\mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}}; \mathbf{y})/\varrho^2; q, \nu/\varrho^2)$  provided  $\Theta$  is continuous at  $\varrho \hat{\boldsymbol{\beta}} - \mathbf{X}^T \nabla l(\mathbf{X} \hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}}; \mathbf{y})/\varrho$ .

In robust regression, (17) degenerates to the linear  $\boldsymbol{\beta}$ -update in (15) with  $\rho = \varrho^2$ , but notice the distinctive scaled form of (17) associated with a nonlinear thresholding operator  $\Theta$ . The Lipschitz condition is satisfied by the squared error loss, logistic deviance, and Huber's loss, among many others, but is only used to give a universal stepsize. In fact, using a backtracking line search (Boyd and Vandenberghe, 2004),  $L$  need not be known in implementation; see Remark A.1. The equations satisfied by  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$  actually define a broader class of estimators that will be investigated in Section 4.

In the remaining, we use BCD to tackle (16). The alternating update involves solving two sub-problems:

$$\boldsymbol{\gamma}^{(t+1)} = \arg \min_{\boldsymbol{\gamma}} l(\boldsymbol{\gamma} + \mathbf{X} \boldsymbol{\beta}^{(t)}) + \nu \|\boldsymbol{\gamma}\|_2^2/2 \quad \text{s.t. } \|\boldsymbol{\gamma}\|_0 \leq q, \quad (20)$$

$$\boldsymbol{\beta}^{(t+1)} = \arg \min_{\boldsymbol{\beta}} l(\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\gamma}^{(t+1)}) + P(\varrho \boldsymbol{\beta}; \lambda). \quad (21)$$

The first problem can be addressed using a similar MM trick as in Theorem 3. But there is a more efficient way to directly locate the optimal support of  $\boldsymbol{\gamma}$  when  $l(\boldsymbol{\eta}; \mathbf{y}) = \sum_{i=1}^n l_0(\eta_i; y_i)$ . Specifically, with  $\tilde{l}(t; a, b) = l_0(a+t, b) + \nu t^2/2$ , the criterion becomes  $\sum_{i: \gamma_i=0} l_0(\mathbf{x}_i^T \boldsymbol{\beta}^{(t)}, y_i) + \sum_{i: \gamma_i \neq 0} \tilde{l}(\gamma_i; \mathbf{x}_i^T \boldsymbol{\beta}^{(t)}, y_i)$ . Let  $l_{0,i} = l_0(\mathbf{x}_i^T \boldsymbol{\beta}^{(t)}, y_i)$  and  $\tilde{l}_{\min,i} = \min_t \tilde{l}(t; \mathbf{x}_i^T \boldsymbol{\beta}^{(t)}, y_i)$ . Then the index set  $I$  associated with the  $q$  largest differences  $\delta_i := l_{0,i} - \tilde{l}_{\min,i}$  ( $1 \leq i \leq n$ ) gives the support. When  $\nu = 0$  and  $\tilde{l}_{\min,i} = 0$ , simply sorting the losses  $l_{0,i}$  achieves the goal. Next, one merely needs to solve  $q$  univariate problems  $\min_s \tilde{l}(s; \mathbf{x}_i^T \boldsymbol{\beta}^{(t)}, y_i)$  to get  $\boldsymbol{\gamma}^{(t+1)}$ .

For (21), an iterative MM algorithm similar to (17) can be developed, and we will show that it suffices to performing the update just once,  $\boldsymbol{\beta}^{(t+1)} = \Theta(\varrho \boldsymbol{\beta}^{(t)} - \mathbf{X}^T \nabla l(\mathbf{X} \boldsymbol{\beta}^{(t)} +$

$\gamma^{(t+1)}; \mathbf{y})/\varrho; \lambda)/\varrho$ . Again, we advocate pursuing  $\beta$ -sparsity via the hybrid “ $\ell_0 + \ell_2$ ” regularization in place of (21),

$$\boldsymbol{\beta}^{(t+1)} = \arg \min l(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}^{(t+1)}) + \nu_\beta \|\boldsymbol{\beta}\|_2^2 / 2 \text{ s.t. } \|\boldsymbol{\beta}\|_0 \leq q_\beta.$$

By use of linearization, the resultant progressive quantile-thresholding is given by

$$\boldsymbol{\beta}^{(t+1)} = \Theta^\#(\boldsymbol{\beta}^{(t)} - \mathbf{X}^T \nabla l(\mathbf{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{\gamma}^{(t+1)}; \mathbf{y})/\varrho; q_\beta^{(t)}, \nu_\beta/\varrho),$$

where  $\rho \geq LM_{\mathbf{X}}(2q_\beta)$  (recall the definition of  $M_{\mathbf{X}}$  in Section 1) and the sequence  $q_\beta^{(t)}$  drops to  $q_\beta$  eventually. (One might want to merge the two  $\ell_0$ -constraints into  $\|\bar{\boldsymbol{\beta}}\|_0 \leq q_\beta + q_\gamma$  to simplify the computation, but this would result in a larger statistical error due to the enlarged search space; see the analysis in Section 4.)

At the end of the section, we make a comparison between the MM-type and the BCD-type PIQ algorithms. The first kind has a single-loop structure and maintains low per-iteration complexity. We do observe that MM is faster than BCD in some large- $p$  classification problems (say when  $p > 500$ ), but in the conventional  $n > p$  setup and many other scenarios, BCD is more accurate and efficient. We will mainly use the BCD-type PIQ in applications and analysis unless otherwise specified.

## 4 Statistical Accuracy of A-estimators

A major theoretical challenge brought by the class of estimators from our computational algorithms is their lack of global optimality and stationarity, which significantly differentiates them from conventional **M**-estimators and **Z**-estimators (van der Vaart, 1998). We call such estimators **A**-estimators due to their *alternative* optimality nature. In fact, even the alternative optimality may be merely in a local sense. As far as we know, there is a lack of statistical accuracy studies for such algorithm-driven estimators. In this section, we aim for developing new techniques to reveal tight error rates of A-estimators as well as the comprehensive roles of algorithmic parameters in resistant learning, in both low and high dimensions.

Introducing the notion of noise in the non-likelihood setting is an essential component. We define the *effective noise* associated with a statistical truth  $\bar{\boldsymbol{\beta}}^* = [\boldsymbol{\beta}^{*T}, \boldsymbol{\gamma}^{*T}]^T$  by

$$\boldsymbol{\epsilon} = -\nabla l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*), \quad (22)$$

which is not affected by the regularizer. In a GLM with cumulant function  $b$  and canonical link function  $g = (b')^{-1}$ , the deviance based loss is given by  $l(\boldsymbol{\eta}) = -\langle \mathbf{y}, \boldsymbol{\eta} \rangle + \langle \mathbf{1}, b(\boldsymbol{\eta}) \rangle$ , and so

$$\boldsymbol{\epsilon} = \mathbf{y} - g^{-1}(\mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma}^*) = \mathbf{y} - \mathbb{E}(\mathbf{y}).$$

In general, however, the effective noise jointly determined by the loss and  $\mathbf{y}$  may not be the raw noise on  $\mathbf{y}$ . For example, a robust loss associated with a bounded  $\psi$  function always gives rise to a bounded  $\boldsymbol{\epsilon}$ , thereby sub-Gaussian regardless of the distribution of  $\mathbf{y}$ , making our analysis nonparametric in nature; the same is true for many classification losses.

Unless otherwise specified, we assume that  $\epsilon$  is a *sub-Gaussian* random vector with mean zero and scale bounded by  $\sigma$  (cf. Definition 1 in Appendix A.2). This gives a broad family including Gaussian and bounded random variables; in particular, any Lipschitz  $l$  results in a sub-Gaussian effective noise. However, our analysis is not restricted by sub-Gaussianity; Theorem A.4 in Appendix A.11 gives a universal risk bound when  $\epsilon$  has a general Orlicz norm. Also, we allow  $\gamma_i^*$  to take arbitrarily large values to capture extreme anomalies; another possible way is to assume  $\gamma_i$  are i.i.d. following a certain distribution, but treating  $\gamma^*$  as an  $n$ -dimensional unknown parameter vector (with sparsity) is convenient in detecting outliers and in handling models with non-additive noise.

Our main tool to tackle the diverse form of the loss is the *generalized Bregman function* (GBF) (She et al., 2021): given any differentiable  $l$

$$\Delta_l(\boldsymbol{\alpha}, \boldsymbol{\beta}) := l(\boldsymbol{\alpha}) - l(\boldsymbol{\beta}) - \langle \nabla l(\boldsymbol{\beta}), \boldsymbol{\alpha} - \boldsymbol{\beta} \rangle. \quad (23)$$

If, further,  $l$  is strictly convex,  $\Delta_l(\boldsymbol{\alpha}, \boldsymbol{\beta})$  becomes the standard Bregman divergence  $\mathbf{D}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta})$  (Bregman, 1967). When  $l(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2/2$ ,  $\Delta_l(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \|\boldsymbol{\alpha} - \boldsymbol{\beta}\|_2^2/2$ , abbreviated as  $\mathbf{D}_2(\boldsymbol{\alpha}, \boldsymbol{\beta})$ . In general,  $\Delta_l(\boldsymbol{\alpha}, \boldsymbol{\beta})$  may not be symmetric, and we define its symmetrization by  $\bar{\Delta}_l(\boldsymbol{\alpha}, \boldsymbol{\beta}) := (\Delta_l(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \Delta_l(\boldsymbol{\beta}, \boldsymbol{\alpha}))/2$ .

We work in a general setting with  $p$  possibly larger than  $n$  unless otherwise mentioned. For simplicity, all  $\ell_2$ -shrinkage parameters are set to zero, and we assume  $\nabla l$  is  $L$ -Lipschitz continuous, and take  $L = 1$  without loss of generality.

Let's first consider the advocated doubly constrained form introduced at the end of Section 3 to exemplify the error bounds. As aforementioned, to make the statistical error study more realistic, rather than restricting to the set of globally optimal solutions, one needs to pay particular attention to the A-estimators defined by

$$\left\{ \begin{array}{l} \hat{\boldsymbol{\beta}} \in \arg \min_{\boldsymbol{\beta}: \|\boldsymbol{\beta}\|_0 \leq q_\beta} l(\mathbf{X}\boldsymbol{\beta} + \hat{\boldsymbol{\gamma}}; \mathbf{y}) \end{array} \right. \quad (24a)$$

$$\left\{ \begin{array}{l} \hat{\boldsymbol{\gamma}} \in \arg \min_{\boldsymbol{\gamma}: \|\boldsymbol{\gamma}\|_0 \leq q_\gamma} l(\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\gamma}; \mathbf{y}). \end{array} \right. \quad (24b)$$

But the  $\hat{\boldsymbol{\beta}}$  delivered from the fast PIQ algorithm may not possess the global alternative optimality in (24a). The good news is that according to Theorem A.1 in Appendix A.4, all these global or local A-estimators satisfy

$$\left\{ \begin{array}{l} \hat{\boldsymbol{\beta}} = \Theta^\#(\hat{\boldsymbol{\beta}} - \mathbf{X}^T \nabla l(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}}; \mathbf{y})/\rho; q_\beta) \\ \hat{\boldsymbol{\gamma}} = \Theta^\#(\hat{\boldsymbol{\gamma}} - \nabla l(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}}; \mathbf{y}); q_\gamma), \end{array} \right. \quad (25)$$

for any  $\rho > M_{\mathbf{X}}(2q_\beta)$ . Our statistical accuracy analysis is applicable to the whole class (and MM-type non-global estimators as well, with slight modification). For clarity, define  $o^* = \|\boldsymbol{\gamma}^*\|_0$ ,  $s^* = \|\boldsymbol{\beta}^*\|_0$ .

**Theorem 4.** *Let  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$  be an A-estimator that satisfies (25) for some  $\rho > 0$ ,  $q_\gamma = \vartheta o^*$ ,  $q_\beta = \vartheta s^*$  ( $\vartheta \geq 1$ ) and  $\|\hat{\boldsymbol{\gamma}}\|_0 = q_\gamma$ ,  $\|\hat{\boldsymbol{\beta}}\|_0 = q_\beta$ . Let  $\bar{\mathbf{X}}_\rho = [\mathbf{X}/\sqrt{\rho} \quad \mathbf{I}]$ . Assume that there exists some  $\delta > 0$  such that*

$$(2\bar{\Delta}_l - \delta \mathbf{D}_2)(\bar{\mathbf{X}}_\rho \bar{\boldsymbol{\beta}}, \bar{\mathbf{X}}_\rho \bar{\boldsymbol{\beta}}') \geq \frac{1}{\sqrt{\vartheta}} \mathbf{D}_2(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}') \quad (26)$$

for sparse  $\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}'$  satisfying  $\|\bar{\boldsymbol{\beta}}\|_0 \leq \vartheta s^*, \|\bar{\boldsymbol{\beta}}'\|_0 \leq s^*, \|\boldsymbol{\gamma}\|_0 \leq \vartheta o^*, \|\boldsymbol{\gamma}'\|_0 \leq o^*$ . Then the following error bound holds

$$\mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\hat{\bar{\boldsymbol{\beta}}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*)] \lesssim \frac{1}{\delta^2} \left\{ \vartheta o^* \sigma^2 \log \frac{en}{\vartheta o^*} + \vartheta s^* \sigma^2 \log \frac{ep}{\vartheta s^*} + \sigma^2 \right\}, \quad (27)$$

where  $C, c$  are positive constants.

More generally, if the starting point  $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\gamma}^{(0)})$ :  $\|\boldsymbol{\beta}^{(0)}\|_0 \leq q_\beta, \|\boldsymbol{\gamma}^{(0)}\|_0 \leq q_\gamma$  already satisfies  $\mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^{(0)}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*)] = O(R)\{\vartheta o^* \sigma^2 \log \frac{en}{\vartheta o^*} + \vartheta s^* \sigma^2 \log \frac{ep}{\vartheta s^*} + \sigma^2\}$  with  $+\infty \geq R \geq 1$ , to obtain the same conclusion (27), (26) can be relaxed to

$$\left\{ 2\left(1 - \frac{1}{R}\right)\bar{\Delta}_l + \frac{C}{R(R\delta \vee 1)}\Delta_l - \delta\mathbf{D}_2 \right\}(\bar{\mathbf{X}}\rho\bar{\boldsymbol{\beta}}, \bar{\mathbf{X}}\rho\bar{\boldsymbol{\beta}}') \geq \frac{(1 - 1/R)}{\sqrt{\vartheta}}\mathbf{D}_2(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}') \quad (28)$$

for some constant  $C > 0$ .

The regularity condition is an extension of the classical  $\ell_2$ -form restricted isometry imposed on the augmented design. Indeed, for  $l(\boldsymbol{\eta}; \mathbf{y}) = \|\boldsymbol{\eta} - \mathbf{y}\|_2^2/2$ , the condition can written as  $\|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}\|_2^2 \geq [1/\{(2 - \delta)\sqrt{\vartheta}\}](\rho\|\boldsymbol{\beta}\|_2^2 + \|\boldsymbol{\gamma}\|_2^2)$ ,  $\forall(\boldsymbol{\beta}, \boldsymbol{\gamma}) : \|\boldsymbol{\beta}\|_0 \leq (1 + \vartheta)s^*, \|\boldsymbol{\gamma}\|_0 \leq (1 + \vartheta)o^*$  for some  $\delta > 0$  and large enough  $\vartheta$ . According to the proof, under  $q_\beta = p$ , the term  $\rho\|\boldsymbol{\beta}\|_2^2$  on its right-hand side can be dropped and so the condition degenerates to the ordinary RIP used in Theorem 2, due to the orthogonal decomposition  $\|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}\|_2^2 = \|\mathbf{X}\{\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^+ \mathbf{X}^T\boldsymbol{\gamma}\}\|_2^2 + \|(\mathbf{I} - \mathbf{H})\boldsymbol{\gamma}\|_2^2$ . We have argued in Section 3 that in low dimensions, low leverage implies low mutual coherence which in turn implies RIP. But the plain definition of leverage cannot incorporate outlier scarcity, and is noninformative as  $p > n$ , since  $\mathbf{H} = \mathbf{I}$  when  $\mathbf{X}$  has full row rank. The Bregman-based restricted isometry on the augmented design gives an extension of the notion of leverage to high dimensions and non-quadratic losses.

Theorem 4 gives useful implications regarding the algorithmic parameters as well. For example, from (26), to enjoy the statistical guarantee, the inverse stepsize  $\rho$  must be properly small. The finding is insightful and novel. Throughout the machine learning literature, slow learning rate is recommended when training a nonconvex learner. But our statistical error analysis strongly cautions against using an extremely small step size when nonconvexity occurs. The BCD-type algorithm design appears advantageous in this aspect. Moreover, a high-quality starting point, usually expensive in computation, certainly brings some statistical benefit seen from the generalized regularity condition (28) (consider an extreme case  $R = 1$ ). But simply enlarging  $\vartheta$  gives another promising way to relax (26), and supports the progressive tightening proposal in Section 3.

When  $\vartheta, \delta$  are treated as constants and  $s^* \vee o^* > 0$ , (27) shows that A-estimators can achieve an error bound of the order  $\sigma^2 s^* \log(ep/s^*) + \sigma^2 o^* \log(en/o^*)$ , which reduces to  $p\sigma^2 + \sigma^2 o^* \log(en/o^*)$  when there is no  $\boldsymbol{\beta}$ -sparsity. This error rate involves the number of outliers apart from the number of relevant predictors. But the influence of outliers remains controlled in the sense that the bound does not grow with the magnitude of  $\boldsymbol{\gamma}^*$ . In the situation of relatively few outliers:  $o^* \log(en/o^*) \ll s^* \log(ep/s^*)$ , the proposed method can attain the celebrated rate  $\sigma^2 s^* \log(ep/s^*)$  for (clean) large- $p$  variable selection. On the other

hand, one should be aware that on ‘big dirty’ data with large  $n$  and  $o^*$ ,  $\sigma^2 o^* \log(en/o^*)$  could be the dominant error term.

**Remark 4.1.** *One may wonder whether the overall error rate in Theorem 4 can be further improved by pursuing a global minimizer. Theorem A.2 shows that this is not the case, but the regularity condition gets relaxed. Concretely, the symmetrized GBF  $\bar{\Delta}_l$  in (26) will be replaced by  $\Delta_l$ , and the right-hand side of (26) will be just zero. Hence if the loss  $l$  defined on the systematic component is  $\mu$ -strongly convex (e.g., regression), the error bound holds with  $\delta = \mu$ , free of any RIP or leverage restriction. This shows the inherent soundness of the proposed resistant learning framework.*

**Remark 4.2.** *The  $\ell_2$ -recovery result proved in Theorem 4 is fundamental, from which one can also obtain estimation error bounds. For example, Theorem A.3 shows that under a slightly stronger regularity condition,*

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \sigma^2 \{s^* \log(ep/s^*) + o^* \log(en/o^*)\}/n$$

*holds with high probability. In the classical  $n > p$  setup with no penalty imposed on  $\beta$ , it becomes  $\|\hat{\beta} - \beta^*\|_2^2 \lesssim \{p\sigma^2 + o^*\sigma^2 \log(en/o^*)\}/n$ . See Appendix A.5 for details.*

Our theoretical techniques can cope with the penalized form of (16) as well. For instance, Theorem 5 studies the outlier-resistant lasso (cf. Section 2), where

$$l(\mathbf{X}\beta + \gamma; \mathbf{y}) = \|\mathbf{X}\beta + \gamma - \mathbf{y}\|_2^2/2, \quad P(\beta; \lambda) = \lambda \varrho \|\beta\|_1,$$

with  $\varrho > 0$ . (The theorem proved in Appendix A.6 is actually regarding a general loss and its proof applies to all subadditive penalties.)

**Theorem 5.** *Let  $\lambda = A\sigma\sqrt{\log(ep)}$  with  $A$  a sufficiently large constant. Then if  $(\hat{\beta}, \hat{\gamma})$  is an  $A$ -estimator of resistant lasso for some  $\varrho$  and  $\vartheta$  (cf. Theorem A.1), the risk bound*

$$\mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\bar{\beta}^*, \bar{\mathbf{X}}\hat{\beta})] \lesssim \sigma^2 \vartheta o^* \log \frac{en}{o^*} + K^2 \sigma^2 s^* \log(ep) + \sigma^2 \quad (29)$$

*holds under the assumption that there exists a large enough  $K$  such that  $K\sqrt{J}\|\mathbf{X}\beta + \gamma\|_2 + \varrho\|\beta_{\mathcal{J}^c}\|_1 \geq \|\gamma\|_2^2/(2\lambda\sqrt{\vartheta}) + (1 + \varepsilon)\varrho\|\beta_{\mathcal{J}}\|_1$  for any  $\beta, \gamma$  with  $\|\gamma\|_0 \leq (1 + \vartheta)o^*$ , where  $\varepsilon$  is a positive constant,  $\mathcal{J} = \{j : \beta_j^* \neq 0\}$  and  $J = |\mathcal{J}|$ .*

When  $K$  and  $\vartheta$  are treated as constants, the error rate of the resistant lasso is slightly worse (by a logarithmic term) than (27) for the doubly constrained form.

## 5 Minimax Lower Bounds and Model Comparison

We first derive some universal lower bounds that are satisfied by all estimators. These nonasymptotic results are stated for GLMs. Let  $\eta^* = \bar{\mathbf{X}}\bar{\beta}^*$  and  $y_i|\eta_i^*$  be independent and follow a distribution in the regular exponential family with dispersion  $\sigma^2$  and  $l_0$  be the negative log-likelihood function (cf. Appendix A.7 for details), and consider a general signal class with  $p$  possibly larger than  $n$

$$\mathcal{B}(s^*, o^*) = \{(\beta^*, \gamma^*) : \|\beta^*\|_0 \leq s^*, \|\gamma^*\|_0 \leq o^*\}. \quad (30)$$

**Theorem 6.** *In the regular exponential dispersion family with  $n \geq 2, p \geq 2, 1 \leq o^* \leq n/2, 1 \leq s^* \leq p/2$ , define*

$$P(s^*, o^*) = s^* \log(ep/s^*) + o^* \log(en/o^*). \quad (31)$$

Let  $I(\cdot)$  be any nondecreasing function with  $I(0) = 0, I \not\equiv 0$ . (i) Suppose for some  $\kappa > 0$   $\Delta_l(\mathbf{0}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}})\sigma^2 \leq \kappa \mathbf{D}_2(\mathbf{0}, \bar{\boldsymbol{\beta}}) \forall \bar{\boldsymbol{\beta}} : \|\bar{\boldsymbol{\beta}}\|_0 \leq s^*, \|\boldsymbol{\gamma}\|_0 \leq o^*$ . Then there exist positive constants  $\tilde{c}, c$ , depending on  $I(\cdot)$  only, such that

$$\inf_{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})} \sup_{(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*) \in \mathcal{B}(s^*, o^*)} \mathbb{E}[I(\mathbf{D}_2(\boldsymbol{\beta}^*, \hat{\boldsymbol{\beta}})/\{\tilde{c}\sigma^2 P(s^*, o^*)/\kappa\})] \geq c > 0, \quad (32)$$

where  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$  denotes any estimator of  $(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*)$ . (ii) Suppose  $\Delta_l(\mathbf{0}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}_1)\sigma^2 \leq \bar{\kappa} \mathbf{D}_2(\mathbf{0}, \bar{\boldsymbol{\beta}}_1)$  and  $\underline{\kappa} \mathbf{D}_2(\mathbf{0}, \bar{\boldsymbol{\beta}}_2) \leq \mathbf{D}_2(\mathbf{0}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}_2)$ ,  $\forall \bar{\boldsymbol{\beta}}_i : \|\bar{\boldsymbol{\beta}}_i\|_0 \leq is^*, \|\boldsymbol{\gamma}_i\|_0 \leq io^*, i = 1, 2$ , where  $\underline{\kappa}/\bar{\kappa}$  is a positive constant. Then there exist positive constants  $\tilde{c}, c$  such that

$$\inf_{(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})} \sup_{(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*) \in \mathcal{B}(s^*, o^*)} \mathbb{E}[I(\mathbf{D}_2(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*, \bar{\mathbf{X}}\hat{\boldsymbol{\beta}})/\{\tilde{c}\sigma^2 P(s^*, o^*)\})] \geq c > 0. \quad (33)$$

Finer analysis and lower bounds are presented in Appendix A.7. We illustrate the conclusion for classification. Because the logistic deviance has a  $1/4$ -Lipschitz continuous gradient, the regularity condition in (i) is satisfied for any  $\kappa \geq M_{\bar{\mathbf{X}}}(s^*, o^*)/4$  (cf. Section 1 and Appendix A.3). So when  $\kappa \leq cn$ ,  $\{s^* \log(ep/s^*) + o^* \log(en/o^*)\}/n$ , or  $(p + o^* \log(en/o^*))/n$  when  $s^* = p$ , is the desired minimax lower error rate for all classifiers, both with constant positive probability and in expectation (corresponding to  $I(t) = 1_{t \geq c}$  and  $I(t) = t$ ). A similar conclusion for prediction can be drawn from (ii). To the best of our knowledge, such a kind of results with the usage of GBFs for resistant GLMs is novel in robust statistics.

Together with the upper error bounds (e.g., Theorem 4 and Theorem A.3), PIQ enjoys minimax rate optimality provided that  $q_{\gamma}$  is not set too large relative to  $o^*$ . It is common in practice to directly specify the cardinality bound based on domain knowledge, say,  $\|\boldsymbol{\gamma}\|_0 \leq 0.25n$ , for the sake of  $\boldsymbol{\beta}$ -estimation. On the other hand, if the goal is to identify all outliers to have the best predictive model, one ought to choose the regularization parameters in a more data-adaptive manner. This is regarded as a vital and challenging task in robust statistics (Ronchetti et al., 1997; Müller and Welsh, 2005; Salibian-Barrera and Van Aelst, 2008). Indeed, outliers may render resampling based cross-validation inappropriate, and although there is an abundance of information criteria, which one is sound in theory and reality lacks a clear conclusion. The joint variable selection in high dimensions and the flexible choice of the loss exacerbate the issue. (The  $\ell_2$ -shrinkage parameter is however much less sensitive and can be fixed at a mild value.)

In the general setup where both  $\boldsymbol{\beta}$  and  $\boldsymbol{\gamma}$  are possibly sparse, the upper and lower error bounds suggest a universal function to penalize model complexity

$$P(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \|\boldsymbol{\gamma}\|_0 + \|\boldsymbol{\gamma}\|_0 \log(en/\|\boldsymbol{\gamma}\|_0) + \|\boldsymbol{\beta}\|_0 + \|\boldsymbol{\beta}\|_0 \log(ep/\|\boldsymbol{\beta}\|_0). \quad (34)$$

We refer to the associated criterion as the *predictive information criterion* (PIC). In the case that no variable selection is needed ( $\|\boldsymbol{\beta}\|_0 = p$ ), one can rewrite (34) as

$$P(\boldsymbol{\gamma}) = \|\boldsymbol{\gamma}\|_0 + \|\boldsymbol{\gamma}\|_0 \log(en/\|\boldsymbol{\gamma}\|_0) \quad (35)$$

for pure outlier identification.

(34) has some involved logarithmic terms and is not a simple  $\ell_0$ -type penalty, but it leads to an ideal model comparison criterion with a solid finite-sample theoretical support.

**Theorem 7.** *Assume  $\epsilon$  defined in (22) is a sub-Gaussian random vector with mean zero and scale bounded by a constant and  $\bar{\beta}^* \in \mathcal{M}$  and  $\bar{\beta}^* \neq \mathbf{0}$ . Define  $P(\bar{\beta}) = P(\beta, \gamma) = J(\beta) \log(ep/J(\beta)) + J(\gamma) \log(en/J(\gamma))$ . Assume that there exist constants  $\delta > 0$  and  $A_0 \geq 0$  such that  $(\Delta_l - \delta \mathbf{D}_2)(\bar{\mathbf{X}}\bar{\beta}, \bar{\mathbf{X}}\bar{\beta}') + A_0(P(\bar{\beta}) + P(\bar{\beta}')) \geq 0$ ,  $\forall \bar{\beta}, \bar{\beta}' \in \mathcal{M}$ . Then for a sufficiently large constant  $A$ , any  $(\bar{\beta}, \hat{\gamma})$  that minimizes*

$$l(\bar{\mathbf{X}}\bar{\beta}; \mathbf{y}) + AP(\bar{\beta}) \quad (36)$$

subject to  $\bar{\beta} \in \mathcal{M}$  must satisfy

$$\mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*)] \lesssim s^* \log(ep/s^*) + o^* \log(en/o^*). \quad (37)$$

Theorem 7 achieves the optimal error rate as Theorem 4 and Theorem 5, but a major difference here is that no parameters (like  $q, \lambda$ ) are involved, but just some absolute constants. Moreover, Theorem 7 does not require any RIP or large signal-to-noise ratio conditions.

When the noise distribution has a dispersion parameter  $\sigma^2$ , Theorem 7 still applies, but the penalty in (36) becomes  $A\sigma^2 P(\beta, \gamma)$  with an unknown factor. A preliminary robust scale estimate can be possibly used. But an appealing result for robust regression is that the estimation of  $\sigma$  can be totally bypassed. For clarity, Theorem 8 assumes  $n > p$  and no sparsity in  $\beta$ , in which situation a scale-free form of PIC, motivated by [She and Tran \(2019\)](#), is given by

$$(n - p) \log(\|\mathbf{X}\beta + \gamma - \mathbf{y}\|_2^2) + \alpha_1 \|\gamma\|_0 + \alpha_2 \|\gamma\|_0 \log(en/\|\gamma\|_0), \quad (38)$$

where  $\alpha_1, \alpha_2$  are absolute constants. Again, the soundness of (38) when the outliers are not awfully many places no limit on their leverage or outlyingness. (More generally, using (34) in place of (35) gives a more general form of scale-free PIC for joint variable selection and outlier identification; Theorem 8' proved in [Appendix A.9](#) implies Theorem 8, and applies to  $p > n$ .)

**Theorem 8.** *Let  $\mathbf{y} = \mathbf{X}\beta^* + \gamma^* + \epsilon$ , where  $\mathbf{X}$  has full column rank,  $\epsilon_i$  are independent sub-Gaussian( $0, \sigma^2$ ) and  $\mathbb{E}\epsilon_i^2 = \sigma_i^2 = c_i\sigma^2$  with  $c_i$  some positive constants and  $\sigma^2$  unknown. Define  $l(\mathbf{X}\beta + \gamma; \mathbf{y}) = \|\mathbf{X}\beta + \gamma - \mathbf{y}\|_2^2$  and  $P(\gamma) = \|\gamma\|_0 \log(en/\|\gamma\|_0)$ . Assume  $P(\gamma^*) \leq (n - p)/A_0$  for some constant  $A_0 > 0$ . Let  $\delta(\gamma) = AP(\gamma)/(n - p)$ , where  $A$  is a positive constant satisfying  $A < A_0$ , and so  $\delta(\gamma^*) < 1$ . Then, for sufficiently large values of  $A_0$  and  $A$ , any  $(\beta, \hat{\gamma})$  that minimizes  $\log l(\mathbf{X}\beta + \gamma; \mathbf{y}) + \delta(\gamma)$  subject to  $\delta(\gamma) < 1$  must satisfy  $\mathbf{D}_2(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) \lesssim \sigma^2 o^* \log(en/o^*) + p\sigma^2$  with probability at least  $1 - c(n - p)^{-c'}$  for some constants  $c, c' > 0$ .*

Combining Theorem 6 and Theorem 8 supports the modified BIC proposed by [She and Owen \(2011\)](#) based on empirical studies. However, (38) has some subtle but important differences from BIC, and the resemblance is just a coincidence. First, unlike most information criteria, its derivation and justification do not need an infinite sample size. Second, the factor  $\log(en/\|\boldsymbol{\gamma}\|_0)$ —not exactly  $\log n$  as in BIC that assumes clean data—arises due to the seek for outliers. Third, the general PIC complexity term (34) (cf. Theorem 7 or Theorem 8') has the overall inflation effect *added* to the degrees of freedom, as opposed to the standard multiplicative relation in AIC, BIC, and many others.

In common with most nonasymptotic studies, our theorems do not show tight constants. However, these numerical constants do not depend on a specific problem and in some easier cases can be determined by Monte Carlo experiments—for example, when  $n > p$ , we recommend  $5.5\|\boldsymbol{\gamma}\|_0 + \|\boldsymbol{\gamma}\|_0 \log(en/\|\boldsymbol{\gamma}\|_0)$  in (38) for robust regression. Theoretically deriving the optimal constants requires much finer analysis and we will not pursue in the current paper.

## 6 Experiments

### 6.1 Simulations

In this part, we perform simulation studies to compare PIQ with some popular robust and/or sparse estimation methods in  $p < n$  and  $p > n$  setups. Unless otherwise mentioned, the raw predictor matrix  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]^T \in \mathbb{R}^{n \times p}$  is generated in the default way:  $\mathbf{x}_i \stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ , where  $\boldsymbol{\Sigma}$  is a Toeplitz design covariance matrix with  $\Sigma_{ij} = \rho^{|i-j|}$ . Here, we consider regression and classification models contaminated with highly-leveraged outliers in the following four setups. More experiment results in other settings are in reported in Appendix A.12 due to limited space. Recall  $o^* = \|\boldsymbol{\gamma}^*\|_0, s^* = \|\boldsymbol{\beta}^*\|_0$ .

**Example 1 (n > p, regression):**  $n = 1000, p = 10, \rho = 0.5, \boldsymbol{\beta}^* = [1, 1, 0.5, 0.5, -1.5, -1.5, -1, -1, 1, 1]^T$ . We set  $\boldsymbol{\gamma}^* = [5, \dots, 5, 0, \dots, 0]^T$  with the first  $o^*$  components nonzero, and modify the first  $o^*$  rows of  $\mathbf{X}$  to  $[3, \dots, 3]$ . The response vector is generated according to  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma}^* + \boldsymbol{\epsilon}$  with  $\epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ .

**Example 2 (n > p, classification):**  $n = 1000, p = 10, \rho = 0.5, \boldsymbol{\beta}^* = [3, 3, 1.5, 1.5, 3, 3, -3, -3, 3, 3]^T$ . To introduce outliers, we change the first  $o^*$  rows of  $\mathbf{X}$  to  $[3, \dots, 3]$  so that the first  $o^*$  elements in  $\mathbf{X}\boldsymbol{\beta}^*$  are 45 and set  $\gamma_i^* = -90, 1 \leq i \leq o^*$  and 0 otherwise. The response vector is generated according to the Bernoulli distribution with mean  $1/(1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta}^* - \gamma_i^*))$  for the  $i$ th observation.

**Example 3 (p > n, regression):**  $n = 200, p = 1000, \boldsymbol{\beta}^* = [1, 0.5, 0, 0, -0.5, -1, 0, \dots, 0]^T$  so that  $s^* = 4, \boldsymbol{\gamma}^* = [5, \dots, 5, 0, \dots, 0]^T$ . High-leveraged outliers are introduced in the same way as in Example 1.

**Example 4 (p > n, classification):**  $n = 200, p = 1000, \boldsymbol{\beta}^* = [3, 1.5, 3, 0, \dots, 0]^T$ . Similar

to Example 2, outliers are added, with  $\gamma_i^* = -45$ ,  $1 \leq i \leq o^*$  and 0 otherwise.

The following 11 methods are included for comparison: S-estimator (Rousseeuw and Yohai, 1984), LTS (Rousseeuw, 1985), Bianco and Yohai's robust logistic regression (B-Y) (Bianco and Yohai, 1996), TLE (Hadi and Luceño, 1997), robust quasi-likelihood estimator (QLE) (Cantoni and Ronchetti, 2001), quantile lasso (QL) (Belloni and Chernozhukov, 2011), robust LARS (RLARS) (Khan et al., 2007), sparse LTS (S-LTS) (Alfons et al., 2013), sparse maximum tangent-likelihood estimator (S-MTE) (Qin et al., 2017), elastic-net LTS for classification (enetLTS) (Kurnaz et al., 2018) and PENSE (Freue et al., 2019). To reduce the interference of different regularization parameter tuning schemes, and to make a fair comparison, sparsity parameters and cut-off values for the sake of outlier identification or variable selection are all chosen to yield  $1.5o^*$  or  $1.5s^*$  nonzeros using the true model. The other parameters are set to their default values. In calling PIQ, we use the BCD-type, with  $\nu$  fixed at  $10^{-4}$ . A quadratic cooling schedule  $Q(t) = -at^2 + U$  with  $a = (U - L)/T^2$  ( $0 \leq t \leq T$ ) is applied in regression and a logarithmic cooling  $Q(t) = -a \log t + U$  ( $0 < t \leq T$ ) with  $a = (U - L)/\log T$  is used in classification, where  $L$  is the desired cardinality,  $U = n$  or  $p$  for outlier identification or variable selection, and  $T = 200$ . Owing to the progressive backward optimization, a simple single starting point  $\beta^{(0)} = \mathbf{0}$  already seems satisfactory in PIQ. In each setup, we repeat the experiment 50 times and evaluate the performance of an algorithm using the following metrics. For outlier identification, we report the masking or missing (**M**) rate, as well as the rate of successful joint detection (**JD**); see She and Owen (2011). Concretely, the masking probability is the fraction of undetected true outliers (misses), and the JD is the fraction of simulations with zero miss. For variable selection, we report the false alarm (**FA**) rate (the fraction of spuriously identified variables) in addition to M and JD. In regression experiments, the mean square error on  $\beta$ , denoted by **Err**, is used to evaluate estimation accuracy, while in classification, **Err** refers to the misclassification error on a separate clean testset containing 10,000 observations. The computational time in seconds, denoted by **T**, is also included in the tables or figures to describe the complexity of each algorithm.

Table 1 shows a performance comparison for robust regression and robust classification in Example 1 and Example 2, respectively. To investigate algorithm scalability, we also give a computational time comparison in Figure 1 with respect to the dimension  $p$ . According to the table, with less than 5% outliers, all methods show satisfactory results for regression or classification. But as the percentage of outliers goes up to 10% (say) or more, the performance of many conventional methods degrades significantly. Also, some algorithms have poor numerical stability as  $p$  is just over 40. In comparison, PIQ is much more accurate and robust, and its scalability is evident.

The advantages of PIQ are even more impressive in Table 2 which summarizes the high-dimensional results. In particular, its boost in accuracy and outlier-resistance is not accompanied by a sacrifice in computational time relative to other robust methods. PIQ offered substantial time savings: its computational cost was at most about one fourth of that of the other algorithms.

We also conducted experiments with different correlation strengths, covariance struc-

Table 1: Performance comparison in  $n > p$  settings (Example 1, Example 2 with  $n = 1000, p = 10$ ). The outlier percentage varies from 1% to 20% for regression, and 3% to 15% for classification.

Regression															
$o^* = 10$			$o^* = 50$			$o^* = 100$			$o^* = 150$			$o^* = 200$			
	Err	M	JD		Err	M	JD		Err	M	JD		Err	M	JD
S	0.06	2.0	82	0.05	0.6	78	0.05	2.2	76	0.27	60	26	0.40	77	0
LTS	0.02	2.0	82	0.02	0.8	70	0.03	4.3	64	0.13	38	40	0.29	74	2
RLARS	0.02	2.4	78	0.11	48	8	0.25	78	0	0.26	78	0	0.28	74	0
PENSE	0.02	1.6	86	0.02	0.6	78	0.23	79	0	0.27	80	0	0.29	75	0
<b>PIQ</b>	0.02	1.6	86	0.02	0.4	84	0.03	0.2	88	0.03	0.1	90	0.05	0.1	90

Classification															
$o^* = 30$			$o^* = 60$			$o^* = 90$			$o^* = 120$			$o^* = 150$			
	Err	M	JD		Err	M	JD		Err	M	JD		Err	M	JD
B-Y	0.06	0.1	98	0.14	42	56	0.28	100	0	0.29	100	0	0.30	100	0
QLE	0.06	0	100	0.23	88	12	0.28	100	0	0.29	100	0	0.30	100	0
TLE	0.09	0	100	0.09	0	100	0.08	0	100	0.11	16	84	0.35	100	0
<b>PIQ</b>	0.06	0	100	0.06	0	100	0.07	0	100	0.07	0	100	0.08	0	100

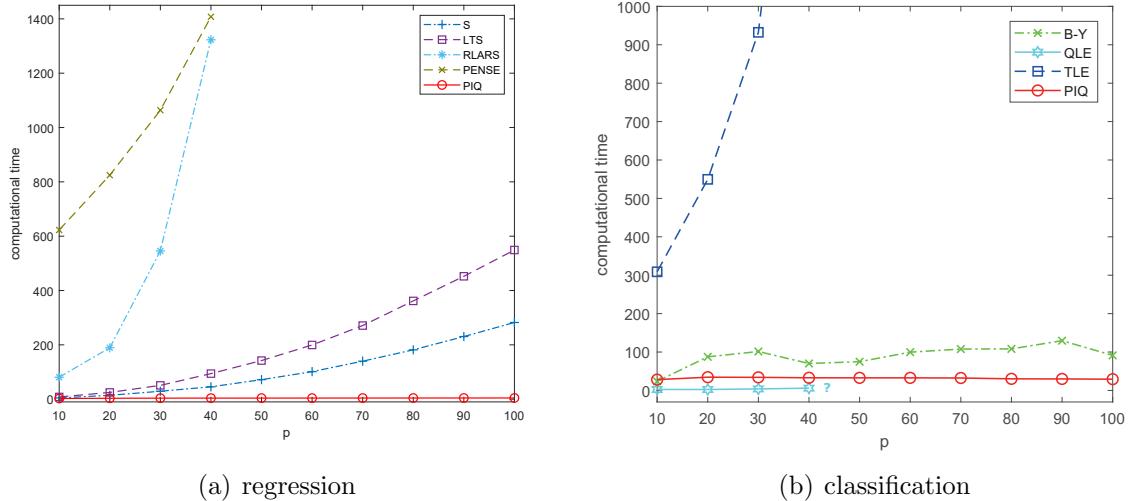


Figure 1: Computational time comparison between different robust methods as  $p$  increases. Left: regression with  $o^* = 0.1n$  and  $n = 1000$  (RLARS and PENSE are too costly compared with the other methods and only parts of their cost curves are shown). Right: classification with  $o^* = 0.05n$  and  $n = 1000$  (when  $p > 40$ , QLE delivered an error message and could not continue).

tures and sparsity levels, and similar conclusions can be drawn from the performance comparison; see Appendix A.12 for details.

Table 2: Performance comparison in  $p > n$  settings (Example 3, Example 4 with  $n = 200, p = 1000$ ). The outlier percentage varies from 5% to 20%.

Regression														
$o^* = 10$							$o^* = 20$							
	Err	$M^\gamma$	$JD^\gamma$	$M^\beta$	$FA^\beta$	$JD^\beta$	<b>T</b>	Err	$M^\gamma$	$JD^\gamma$	$M^\beta$	$FA^\beta$	$JD^\beta$	<b>T</b>
QL	0.37	36	16	19	0.1	40	43	0.71	81	0	25	0.2	24	42
S-MTE	0.44	73	0	14	0.2	54	35	0.64	82	0	18	0.2	48	39
RLARS	0.53	82	4	25	0.2	40	42	1.06	89	0	42	0.3	0	42
S-LTS	0.28	4	80	1	0.1	98	134	0.27	5	74	1	0.2	96	134
<b>PIQ</b>	0.11	2	86	3	0.2	90	4	0.23	5	80	10	0.2	64	4
$o^* = 30$							$o^* = 40$							
	Err	$M^\gamma$	$JD^\gamma$	$M^\beta$	$FA^\beta$	$JD^\beta$	<b>T</b>	Err	$M^\gamma$	$JD^\gamma$	$M^\beta$	$FA^\beta$	$JD^\beta$	<b>T</b>
QL	0.89	81	0	33	0.2	8	45	1.10	76	0	41	0.2	4	50
S-MTE	0.82	79	0	28	0.2	26	42	0.97	75	0	34	0.2	14	54
RLARS	1.52	91	0	58	0.3	0	41	1.55	85	0	57	0.3	0	40
S-LTS	1.24	79	10	29	1.4	16	132	1.52	82	0	46	1.5	0	136
<b>PIQ</b>	0.34	3	78	16	0.3	46	4	0.62	5	76	27	0.3	22	4
Classification														
	$o^* = 10$							$o^* = 20$						
	Err	$M^\gamma$	$JD^\gamma$	$M^\beta$	$FA^\beta$	$JD^\beta$	<b>T</b>	Err	$M^\gamma$	$JD^\gamma$	$M^\beta$	$FA^\beta$	$JD^\beta$	<b>T</b>
enetLTS	12.4	4	96	1.3	0.1	96	62	16.0	22	78	5.3	0.24	86	67
<b>PIQ</b>	11.0	0	100	0	0.1	100	16	11.6	0	100	0	0.1	100	16
	$o^* = 30$							$o^* = 40$						
	Err	$M^\gamma$	$JD^\gamma$	$M^\beta$	$FA^\beta$	$JD^\beta$	<b>T</b>	Err	$M^\gamma$	$JD^\gamma$	$M^\beta$	$FA^\beta$	$JD^\beta$	<b>T</b>
enetLTS	36.2	94	6	43	1.0	16	86	19.8	100	0	66	0.9	0	96
<b>PIQ</b>	12.6	0	100	0.7	0.1	98	14	17.0	8	84	9	0.1	78	14

## 6.2 Robust classification of spam email

The email spam dataset (available at <ftp://ics.uci.edu>) contains 57 predictors including relative frequencies of 54 most commonly occurring words and characters and 3 statistics of capital letters. The total number of emails is 4,601, of which 1,813 are spam ( $y = 1$ ) and 2,788 are non-spam ( $y = 0$ ). We would like to know whether accommodating *mislabeled* emails (if any) could lead to improved spam/non-spam classification.

We applied PIQ with the logistic deviance and the Huberized hinge loss (Chapelle, 2007), denoted by PIQ-L and PIQ-H, respectively, with PIC for tuning. Table 3 compares PIQ with logistic regression, B-Y, TLE, and QLE (cf. Section 6.1) in terms of misclassification error and F1-score (Chinchor, 1992), both averaged over 100 data splits (70% for training and 30% for test). From the table, TLE and QLE failed due to the large sample size. PIQ outperforms the other methods and the choice of the loss does not affect its performance on this dataset.

We then studied the subset of outliers identified by PIQ. Take PIQ-L for instance: there are 73 nonzeros  $\hat{\gamma}_i$ 's, the smallest magnitude being 10.97. To verify if there is indeed a distinction between the outliers ( $o = 1$ ) and clean observations ( $o = 0$ ), we examined 6

Table 3: Email spam data: misclassification error rates and F1-scores.

	Miscls (%)	F1 (%)
Logistic	7.5	90.3
B-Y	7.0	89.4
TLE	—	—
QLE	—	—
PIQ-L	6.8	91.3
PIQ-H	6.8	91.3

indicative features, `free`, `!`, `$`, `hp`, `650` and `cap_avg`. The first three words and signs are often used in spam to attract people’s attention; `hp` and `650` are related to the data creator’s working company and area code; `cap_avg` measures the average length of uninterrupted sequences of capital letters. Given each feature, we made a  $2 \times 2$  table with its rows corresponding to spam and non-spam ( $y = 1, 0$ ) and columns corresponding to outlying and clean ( $o = 1, 0$ ). The cross-product ratio or odds ratio visualizes the strong association between  $y$  and  $o$ , as shown in Figure 2. Indeed, even the odds ratios closer to 1 are either as small as  $\exp(-2.8) = 0.06$  or as large as  $\exp(4.89) = 133$ , demonstrating a strong discrepancy between the outlier subset and the clean subset. We also built an ANOVA model on each feature to test the significance of  $o$ ; all the obtained  $p$ -values are below 1e-10, a clear evidence of the outlyingness.

The data analysis suggests that even on some conventional datasets, mislabeling frequently occurs and guarding against outliers is beneficial. It is fair to say that modern statistical analysis involves more and more large-scale datasets, but with bigger data also come more junk and errors. This could be a severe issue for binary classification, since any inaccuracy in labeling simply means the response goes to the other extreme. Applying a classifier designed under no consideration of gross outliers could be risky in reality, and our resistant learning framework offers an effective fix in this regard.

### 6.3 Robust neural network for Parkinson’s disease detection

Neural networks have attracted a great deal of attention in machine learning. In this experiment, we adapt the PIQ technique to a neural net model and test its performance for Parkinson’s disease (PD) detection. There has been a surge of interest in speech analysis of PD recently, as voice recordings from these patients tend to show typical patterns like aperiodic vibrations. We use the dataset from [Little et al. \(2007\)](#) which collects a range of biomedical voice measurements from normal people and patients with PD, where the 22 predictors for 195 recordings include various measures of vocal fundamental frequency and variation in amplitude, as well as some nonlinear metrics constructed from raw voice recordings. We randomly split the dataset into a training subset (60%) and a test subset (40%).

Previous studies suggest the presence of nonlinear effects and interactions of the features in this challenging classification task ([Little et al. \(2007\)](#), [Sakar et al. \(2013\)](#)). Hence we

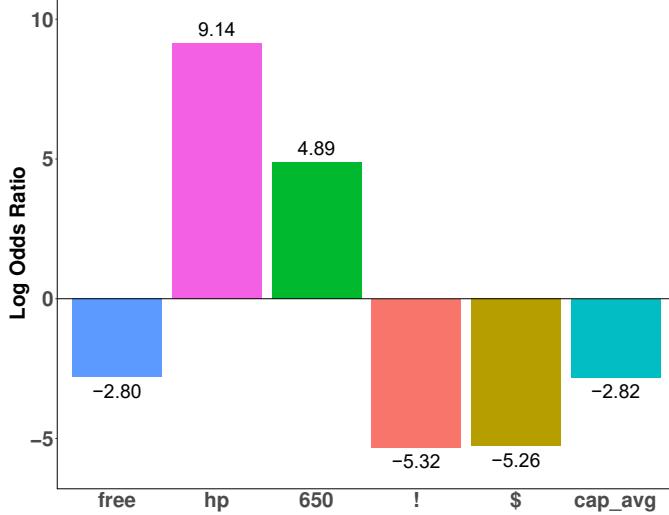


Figure 2: Log odds ratios of six  $2 \times 2$  tables with row variable  $y$  and column variable  $o$ .

designed a feedforward neural network that consists of an input layer, two dense layers with 22 and 12 nodes respectively, and one output layer. The dense layers deploy the popular rectified linear units (ReLU) (Nair and Hinton, 2010), and the output node uses a softmax activation. We trained the network for 200 epochs and attained a misclassification error rate of 9.1% on the test data—in contrast, an SVM with a nonlinear RBF kernel gave 17.9%. (Meanwhile, classical robust classification methods like B-Y, QLE, and TLE all behaved poorly, with the misclassification error rates  $> 25\%$ .)

On the other hand, the noisy signals and potential outliers occurring suggest the need of robustification. But the overall network criterion barely resembles that of regression or logistic regression due to its substantial nonlinearity and hierarchy, and how to reweight the samples in this sophisticated model to limit the influence of outliers sounds tricky. Following the additive robustification scheme, we simply added sample-indicator features into the input layer (cf. Figure 3), accompanied by a group  $\ell_0 + \ell_2$  regularization ( $q = 0.05n$ ,  $\nu = 1e-4$ ) to correct for sample outlyingness in constructing predictive factors. Our optimization-based PIQ algorithm can be seamlessly incorporated into the process of back-propagation. Though a simple modification, the robustified neural network gave an impressive misclassification error rate 6.4%, an improvement of about 30% over that of the vanilla neural network.

## 7 Summary

The work studied how to gain outlier resistance for a given loss that is not necessarily quadratic and may go beyond the likelihood setup. Motivated by the method of trimming, we proposed an  $\ell_0 + \ell_2$  regularized learning framework. We showed that pursuing its (global) minimizers can cope with extreme outliers regardless of leverage and outlyingness, but on large-scale data, there is always a balance between statistical accuracy and computational

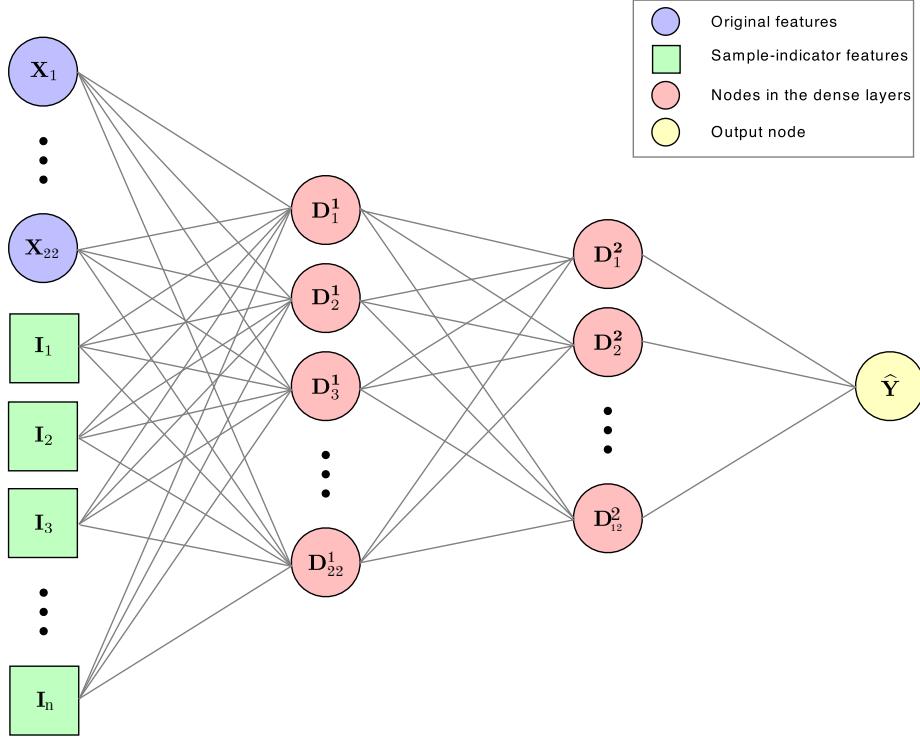


Figure 3: Structure of the robustified neural network on Parkinson’s disease data, where sample-indicator features (denoted by green square nodes) are added to the input layer.

efficiency. Hence a more meaningful and deeper question is how to design a scalable algorithm to alleviate the starting point requirement and the regularity conditions to obtain the desired order of statistical accuracy.

In this work, the interplay between high dimensional statistics and nonconvex optimization led to fruitful results for resistant learning on the learning rate, cardinality control, the choice of regularization parameters, and so on. In particular, a novel progressive backward optimization scheme is brought forward which, unlike subset sampling, is cost effective, and can significantly improve the statistical performance of iterative quantile-thresholding. We hope that the work could raise concern about outliers among (big) data analysts, as well as providing a sensible and universal means to boosting resistance in regression, classification and more sophisticated learning tasks.

## A Proofs

### A.1 Proof of Theorem 1

First, we introduce a lemma to be used for proving the three statements.

**Lemma 1.** *Let  $l_0(\eta; y)$  be a loss function defined on  $\eta \in \mathbb{R}$  with  $y \in \mathcal{Y} \subset \mathbb{R}$  as a parameter and  $\inf_\eta l_0(\eta; y) = L > -\infty$  for all  $y \in \mathcal{Y}$ . Define  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$  and  $l(\boldsymbol{\eta}; \mathbf{y}) = \sum_{i=1}^n l_0(\eta_i; y_i)$ .*

Given any  $\beta$ , let  $l_0^{(1)}(\beta) \leq \dots \leq l_0^{(n)}(\beta)$  be the order statistics of  $l_0(\mathbf{x}_i^T \beta; y_i)$  ( $1 \leq i \leq n$ ). Then

$$\inf_{\|\gamma\|_0 \leq q} l(\mathbf{X}\beta + \gamma; \mathbf{y}) = \sum_{i=1}^{n-q} l_0^{(i)}(\beta). \quad (\text{A.1})$$

Moreover, given any  $\tau \geq 0$ ,

$$\inf_{\gamma} l(\mathbf{X}\beta + \gamma; \mathbf{y}) + \tau \|\gamma\|_0 = \sum_{i=1}^n (\tau + L) \wedge l_0(\mathbf{x}_i^T \beta; y_i). \quad (\text{A.2})$$

For a), let  $(\hat{\beta}, \hat{\gamma})$  be a globally optimal solution to

$$\min_{(\beta, \gamma)} l(\eta; \mathbf{y}) \text{ s.t. } \eta = \mathbf{X}\beta + \gamma, \quad \|\gamma\|_0 \leq q. \quad (\text{A.3})$$

Then for any  $\beta$ ,  $l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y}) \leq l(\mathbf{X}\beta + \hat{\gamma}_\beta; \mathbf{y})$ , where  $\hat{\gamma}_\beta$  is an optimal solution of  $\gamma$  given  $\beta$ . From Lemma 1,  $l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y}) = \sum_{i=1}^{n-q} l_0^{(i)}(\hat{\beta})$ ,  $l(\mathbf{X}\beta + \hat{\gamma}_\beta; \mathbf{y}) = \sum_{i=1}^{n-q} l_0^{(i)}(\beta)$ , and so  $\sum_{i=1}^{n-q} l_0^{(i)}(\hat{\beta}) \leq \sum_{i=1}^{n-q} l_0^{(i)}(\beta)$ .

Now suppose that  $\hat{\beta}$  is a globally optimal solution that minimizes the trimmed criterion on the RHS of (A.1). Then for any  $(\beta, \gamma)$ ,

$$l(\mathbf{X}\beta + \gamma; \mathbf{y}) \geq l(\mathbf{X}\beta + \hat{\gamma}_\beta; \mathbf{y}) = \sum_{i=1}^{n-q} l_0^{(i)}(\beta) \geq \sum_{i=1}^{n-q} l_0^{(i)}(\hat{\beta}) = l(\mathbf{X}\hat{\beta} + \hat{\gamma}_\beta; \mathbf{y}),$$

which means  $(\hat{\beta}; \hat{\gamma}_\beta)$  is a global minimizer of (A.3). The second conclusion for the penalized form and winsorized form can be proved similarly (assuming  $L = 0$ ).

To prove the last statement, we set  $\mathcal{C}(\beta) = \{i : l_0(\mathbf{x}_i^T \beta; y_i) > \tau, 1 \leq i \leq n\}$  and  $c(\beta) = |\mathcal{C}(\beta)|$ . So  $\sum_{i=1}^n \tau \wedge l_0(\mathbf{x}_i^T \beta; y_i) = c(\beta)\tau + \sum_{i=1}^{n-c(\beta)} l_0^{(i)}(\beta)$ .

We claim that for any  $\hat{\beta} \in \arg \min \sum \tau \wedge l_0(\mathbf{x}_i^T \beta; y_i)$  with  $q = c(\hat{\beta})$ , it is also an optimal solution to  $\min \sum_{i=1}^{n-q} l_0^{(i)}(\beta)$ . Prove by contradiction: if  $\sum_{i=1}^{n-q} l_0^{(i)}(\hat{\beta}) > \sum_{i=1}^{n-q} l_0^{(i)}(\beta)$  for some  $\beta$ , then since  $\tau \wedge l_0(\mathbf{x}_i^T \hat{\beta}; y_i) = \tau \geq \tau \wedge l_0(\mathbf{x}_j^T \beta; y_j)$  for all  $i \in \mathcal{C}(\hat{\beta})$ ,  $j : 1 \leq j \leq n$ , we would get  $\sum_{i=1}^n \tau \wedge l_0(\mathbf{x}_i^T \hat{\beta}; y_i) > \sum_{i=1}^n \tau \wedge l_0(\mathbf{x}_i^T \beta; y_i)$ .

Finally, we show that given any  $(\hat{\beta}, \hat{\gamma}) \in \arg \min l(\mathbf{X}\beta + \gamma; \mathbf{y}) + \tau \|\gamma\|_0$  with  $q = \|\hat{\gamma}\|_0$  (where  $\hat{\gamma}_i$  are allowed to take  $\pm\infty$ ), it is also an optimal solution to  $\min_{\beta, \gamma} l(\mathbf{X}\beta + \gamma; \mathbf{y})$  s.t.  $\|\gamma\|_0 \leq q$ . Suppose that there exists  $(\beta, \gamma)$  with  $\|\gamma\|_0 \leq q$  so that  $l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y}) > l(\mathbf{X}\beta + \gamma; \mathbf{y})$ . Then, by considering  $(\beta, \gamma')$  with  $\gamma' = \hat{\gamma}_\beta$  and  $\|\gamma'\|_0 \leq q$  as defined previously,  $\inf_{\|\gamma\|_0 \leq q} l(\mathbf{X}\hat{\beta} + \gamma; \mathbf{y}) = l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y}) > \inf_{\|\gamma\|_0 \leq q} l(\mathbf{X}\beta + \gamma; \mathbf{y})$ , and thus  $\sum_{i=1}^{n-q} l_0^{(i)}(\hat{\beta}) > \sum_{i=1}^{n-q} l_0^{(i)}(\beta)$  from Lemma 1. Using the fact that  $\hat{\gamma}_i \neq 0$  implies  $\tau + l_0(\mathbf{x}_i^T \hat{\beta} + \hat{\gamma}_i; y_i) = \tau$ , and the construction of  $\gamma'$ ,  $l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y}) + \tau \|\hat{\gamma}\|_0 = q\tau + \sum_{i=1}^{n-q} l_0^{(i)}(\hat{\beta})$  and  $l(\mathbf{X}\beta + \gamma'; \mathbf{y}) + \tau \|\gamma'\|_0 = q\tau + \sum_{i=1}^{n-q} l_0^{(i)}(\beta)$ . Summarizing the above gives  $l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y}) > l(\mathbf{X}\beta + \gamma'; \mathbf{y})$ , contradicting the optimality of  $(\hat{\beta}, \hat{\gamma})$ . The proof is complete.

**Proof of Lemma 1** Recall that the loss is assumed bounded from below. Without loss of generality, let  $\inf_{\eta} l_0(\eta; y) = 0$  for all  $y \in \mathcal{Y}$  (otherwise, we can redefine the loss function by  $l_0(\eta; y) - \inf_{\eta} l_0(\eta; y)$ ).

Given  $\beta$ , denote by  $\hat{\gamma}_{\beta}$  a minimizer of  $l(\mathbf{X}\beta + \gamma; \mathbf{y})$  subject to  $\|\gamma\|_0 \leq q$ , whose components can take  $\pm\infty$ . Then if  $\hat{\gamma}_i = 0$ ,  $l_0(\mathbf{x}_i^T \beta + \hat{\gamma}_i; y_i) = l_0(\mathbf{x}_i^T \beta; y_i) \geq 0$ . If  $\hat{\gamma}_i \neq 0$ , by varying its value, which will not violate the constraint, one can argue that  $l_0(\mathbf{x}_i^T \beta + \hat{\gamma}_i; y_i) = \inf_{\eta} l_0(\eta; y_i) = 0$ . Therefore,  $l(\mathbf{X}\beta + \hat{\gamma}_{\beta}; \mathbf{y}) = \sum_{i=1}^{n-q} l_0^{(i)}(\beta)$ . The proof of (A.2) follows similar lines.

## A.2 Proof of Theorem 2

First, we have the following result to argue the (alternative) optimality of  $\gamma^{(t+1)}$  given  $\beta^{(t)}$ , cf. Lemma C.1 in [She et al. \(2013\)](#).

**Lemma 2.** *Given any  $s \in \mathbb{R}^n$ ,  $0 \leq q \leq n$ ,  $\nu \geq 0$ ,  $\hat{\xi} = \Theta^{\#}(s; q, \nu)$  is a global optimal solution to  $\min_{\xi \in \mathbb{R}^n} \|s - \xi\|_2^2/2 + \nu\|\xi\|_2^2/2$  s.t.  $\|\xi\|_0 \leq q$ .*

The below definition of a sub-Gaussian random vector is standard in the literature, see, e.g., [Vershynin \(2012\)](#).

**Definition 1 (Sub-Gaussian random variable/vector).** *We call  $\xi$  a sub-Gaussian( $0, \sigma^2$ ) random variable if and only if  $\xi$  has mean 0 and there exist constants  $C, c > 0$  such that  $\mathbb{P}\{|\xi| \geq t\} \leq Ce^{-ct^2}$ ,  $\forall t > 0$  with the scale (or  $\psi_2$ -norm) of  $\xi$  defined by  $\sigma(\xi) = \inf\{\sigma > 0 : \mathbb{E} \exp(\xi^2/\sigma^2) \leq 2\}$ . More generally,  $\xi \in \mathbb{R}^p$  is called a sub-Gaussian random vector with scale bounded by  $\sigma$  if all one-dimensional marginals  $\langle \xi, \alpha \rangle$  are sub-Gaussian satisfying  $\|\langle \xi, \alpha \rangle\|_{\psi_2} \leq \sigma\|\alpha\|_2$ ,  $\forall \alpha \in \mathbb{R}^p$ .*

a) Using the update in (6), we can represent the optimal  $\beta$  as a function of  $\gamma$ , from which problem (5) is equivalent to

$$\min_{\gamma} f(\gamma) := \frac{1}{2} \|(\mathbf{I} - \mathbf{H})(\gamma - \mathbf{y})\|_2^2 + \frac{\nu}{2} \|\gamma\|_2^2 \text{ s.t. } \|\gamma\|_0 \leq q. \quad (\text{A.4})$$

Let  $l(\gamma) = \|(\mathbf{I} - \mathbf{H})(\gamma - \mathbf{y})\|_2^2/2$ . From Lemma 2, a nice fact of (7) is that

$$\gamma^{(t+1)} = \arg \min_{\|\gamma\|_0 \leq q} g(\gamma; \gamma^{(t)}),$$

where

$$\begin{aligned} g(\gamma; \gamma^-) &= \|(\mathbf{I} - \mathbf{H})(\gamma^- - \mathbf{y})\|_2^2/2 + \langle (\mathbf{I} - \mathbf{H})(\gamma^- - \mathbf{y}), \gamma - \gamma^- \rangle \\ &\quad + \|\gamma - \gamma^-\|_2^2/2 + \nu\|\gamma\|_2^2/2 \\ &= f(\gamma) + (\mathbf{D}_2 - \Delta_l)(\gamma, \gamma^-) \end{aligned}$$

with  $\Delta_l$  and  $\mathbf{D}_2$  introduced in Section 4. By definition,  $g(\gamma^{(t)}; \gamma^{(t)}) \geq g(\gamma^{(t+1)}; \gamma^{(t)})$ , and so we have

$$f(\gamma^{(t)}) - f(\gamma^{(t+1)}) \geq (\mathbf{D}_2 - \Delta_l)(\gamma^{(t+1)}, \gamma^{(t)}). \quad (\text{A.5})$$

Summing up (A.5) over  $t = 1, \dots, T$  gives the conclusion.

b) We begin with a useful lemma. Throughout the proof, given any  $\beta$ , we use  $\mathcal{J}(\beta)$  to denote its support, i.e.,  $\mathcal{J}(\beta) = \{j : \beta_j \neq 0\}$ , and  $J(\beta) = |\mathcal{J}(\beta)|$ . Recall that  $s^* = J(\beta^*)$  and  $o^* = J(\gamma^*)$ .

**Lemma 3.** *Given  $q \leq p$  and  $\nu \geq 0$ , a globally optimal solution to the optimization problem  $\min_{\beta \in \mathbb{R}^p} l(\beta) = \|\mathbf{y} - \beta\|_2^2/2 + \nu\|\beta\|_2^2/2$  s.t.  $\|\beta\|_0 \leq q$  is given by  $\hat{\beta} = \Theta^\#(\mathbf{y}; q, \nu)$ . Let  $\mathcal{J} = \mathcal{J}(\beta)$ ,  $\hat{\mathcal{J}} = \mathcal{J}(\hat{\beta})$  and assume  $J(\hat{\beta}) = q$ . Then, for any  $\beta$  with  $J(\beta) \leq s = q/\vartheta$  and  $\vartheta \geq 1$ ,*

$$l(\beta) - l(\hat{\beta}) \geq \{1 - \mathcal{L}(\mathcal{J}, \hat{\mathcal{J}})\}(1 + \nu)\|\hat{\beta} - \beta\|_2^2/2,$$

where  $\mathcal{L}(\mathcal{J}, \hat{\mathcal{J}}) = (|\mathcal{J} \setminus \hat{\mathcal{J}}|/|\hat{\mathcal{J}} \setminus \mathcal{J}|)^{1/2} \leq (s/q)^{1/2} = \vartheta^{-1/2}$ .

It can be proved by Lemma 9 in [She and Chen \(2017\)](#). By Lemma 3, for any  $\gamma$  satisfying  $\|\gamma\|_0 \leq q$ ,

$$g(\gamma; \gamma^{(t)}) - g(\gamma^{(t+1)}; \gamma^{(t)}) \geq \{1 - \mathcal{L}(\mathcal{J}(\gamma), \mathcal{J}(\gamma^{(t+1)}))\}(1 + \nu)\mathbf{D}_2(\gamma^{(t+1)}, \gamma),$$

and so

$$\begin{aligned} & f(\gamma) + (\mathbf{D}_2 - \Delta_l)(\gamma, \gamma^{(t)}) - f(\gamma^{(t+1)}) - (\mathbf{D}_2 - \Delta_l)(\gamma^{(t+1)}, \gamma^{(t)}) \\ & \geq \{1 - \mathcal{L}(\mathcal{J}(\gamma), \mathcal{J}(\gamma^{(t+1)}))\}(1 + \nu)\mathbf{D}_2(\gamma^{(t+1)}, \gamma), \end{aligned}$$

or

$$\begin{aligned} & f(\gamma) + ((1 + \nu)\mathbf{D}_2 - \Delta_f)(\gamma, \gamma^{(t)}) - f(\gamma^{(t+1)}) - (\mathbf{D}_2 - \Delta_l)(\gamma^{(t+1)}, \gamma^{(t)}) \\ & \geq \{1 - \mathcal{L}(\mathcal{J}(\gamma), \mathcal{J}(\gamma^{(t+1)}))\}(1 + \nu)\mathbf{D}_2(\gamma^{(t+1)}, \gamma). \end{aligned}$$

Substituting  $\gamma^*$  for  $\gamma$  gives

$$\begin{aligned} & \{1 - \mathcal{L}(\mathcal{J}(\gamma^*), \mathcal{J}(\gamma^{(t+1)}))\}(1 + \nu)\mathbf{D}_2(\gamma^{(t+1)}, \gamma^*) + (\mathbf{D}_2 - \Delta_l)(\gamma^{(t+1)}, \gamma^{(t)}) \\ & \leq -\Delta_f(\gamma^{(t+1)}, \gamma^*) - \langle \nabla f(\gamma^*), \gamma^{(t+1)} - \gamma^* \rangle + ((1 + \nu)\mathbf{D}_2 - \Delta_f)(\gamma^*, \gamma^{(t)}). \end{aligned} \quad (\text{A.6})$$

Since  $\|\mathbf{I} - \mathbf{H}\|_2 = 1$ ,  $(\mathbf{D}_2 - \Delta_l)(\gamma^{(t+1)}, \gamma^{(t)}) \geq 0$ . Together with

$$\mathcal{L}(\mathcal{J}(\gamma^*), \mathcal{J}(\gamma^{(t+1)})) \leq 1/\sqrt{\vartheta},$$

we obtain

$$\begin{aligned} & ((1 - 1/\sqrt{\vartheta})(1 + \nu)\mathbf{D}_2 + \Delta_l + \nu\mathbf{D}_2)(\gamma^{(t+1)}, \gamma^*) \\ & \leq (\mathbf{D}_2 - \Delta_l)(\gamma^*, \gamma^{(t)}) + \langle (\mathbf{I} - \mathbf{H})^2(\gamma^* - \mathbf{y}) + \nu\gamma^*, -\gamma^{(t+1)} + \gamma^* \rangle \\ & = (\mathbf{D}_2 - \Delta_l)(\gamma^*, \gamma^{(t)}) + \langle -\mathbf{U}_\perp \mathbf{U}_\perp^T \epsilon + \nu\gamma^*, \gamma^* - \gamma^{(t+1)} \rangle \\ & = (\mathbf{D}_2 - \Delta_l)(\gamma^*, \gamma^{(t)}) + \langle \epsilon', \mathbf{U}_\perp^T(\gamma^{(t+1)} - \gamma^*) \rangle + \nu \langle \gamma^*, \gamma^* - \gamma^{(t+1)} \rangle, \end{aligned} \quad (\text{A.7})$$

where  $\mathbf{X} = [\mathbf{U}, \mathbf{U}_\perp]\mathbf{D}\mathbf{V}^T$  is the singular value decomposition of  $\mathbf{X}$  with  $\mathbf{U}_\perp \in \mathbb{R}^{n \times (n-p)}$  and  $\epsilon' = \mathbf{U}_\perp^T \epsilon$  which is still a sub-Gaussian random vector with mean 0 and scaled bounded by  $\sigma$ . From  $\mathbf{I} - \mathbf{H} = \mathbf{U}_\perp \mathbf{U}_\perp^T$ ,  $\Delta_l(\gamma, \gamma') = \|\mathbf{U}_\perp^T(\gamma - \gamma')\|_2^2/2$ .

The stochastic term  $\langle \epsilon', \mathbf{U}_\perp^T(\gamma^{(t+1)} - \gamma^*) \rangle$  can be bounded using the following lemma.

**Lemma 4.** Let  $\boldsymbol{\epsilon}$  be a sub-Gaussian random vector with mean zero and scale bounded by  $\sigma$ . Given  $\mathbf{X} \in \mathbb{R}^{n \times p}$  and  $0 \leq J \leq p$ , there exist universal constants  $A, C, c > 0$  such that for any  $a > 0$ , the following event

$$\sup_{\boldsymbol{\beta} \in \mathbb{R}^p: \|\boldsymbol{\beta}\|_0 \leq J} \langle \boldsymbol{\epsilon}, \mathbf{X} \boldsymbol{\beta} \rangle - \frac{1}{2a} \|\mathbf{X} \boldsymbol{\beta}\|_2^2 - aA^2\sigma^2 J \log(ep/J) > a\sigma^2 t \quad (\text{A.8})$$

occurs with probability at most  $C \exp(-ct)p^{-cA^2}$  for any  $t \geq 0$ .

The result is a variant of Lemma 4 of [She \(2016\)](#) and can be obtained from its proof. According to [Lemma 4](#), there exist constants  $A, C, c > 0$  such that for any  $a > 0$ ,

$$\langle \boldsymbol{\epsilon}', \mathbf{U}_\perp^T(\boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\gamma}^*) \rangle \leq \frac{1}{a} \boldsymbol{\Delta}_l(\boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\gamma}^*) + aA^2\sigma^2(1 + \vartheta)o^* \log \frac{en}{o^*} \quad (\text{A.9})$$

occurs with probability at least  $1 - C(n - p)^{-c}$ . Since  $\langle \boldsymbol{\gamma}^*, -\boldsymbol{\gamma}^{(t+1)} + \boldsymbol{\gamma}^* \rangle \leq (b/2)\|\boldsymbol{\gamma}^*\|_2^2 + (1/2b)\|\boldsymbol{\gamma}^{(t+1)} - \boldsymbol{\gamma}^*\|_2^2$  for any  $b > 0$ , combining [\(A.7\)](#) and [\(A.9\)](#) gives

$$\begin{aligned} & \{((1 - \frac{1}{\sqrt{\vartheta}}) + (2 - \frac{1}{\sqrt{\vartheta}} - \frac{1}{b})\nu) \mathbf{D}_2 + (1 - \frac{1}{a})\boldsymbol{\Delta}_l\}(\boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\gamma}^*) \\ & \leq (\mathbf{D}_2 - \boldsymbol{\Delta}_l)(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\gamma}^*) + aA^2\sigma^2(1 + \vartheta)o^* \log \frac{en}{o^*} + \frac{b\nu}{2}\|\boldsymbol{\gamma}^*\|_2^2. \end{aligned} \quad (\text{A.10})$$

Let

$$\kappa = \frac{1}{1 - \varepsilon} \left\{ (2 - \frac{1}{a})\varepsilon - \frac{1}{\sqrt{\vartheta}} + (2 - \frac{1}{\sqrt{\vartheta}} - \frac{1}{b})\nu \right\}. \quad (\text{A.11})$$

By the regularity condition,

$$\begin{aligned} & (1 + \kappa)(\mathbf{D}_2 - \boldsymbol{\Delta}_l)(\boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\gamma}^*) \\ & \leq \{((1 - \frac{1}{\sqrt{\vartheta}}) + (2 - \frac{1}{\sqrt{\vartheta}} - \frac{1}{b})\nu) \mathbf{D}_2 + (1 - \frac{1}{a})\boldsymbol{\Delta}_l\}(\boldsymbol{\gamma}^{(t+1)}, \boldsymbol{\gamma}^*) \\ & \leq (\mathbf{D}_2 - \boldsymbol{\Delta}_l)(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\gamma}^*) + aA^2\sigma^2(1 + \vartheta)o^* \log \frac{en}{o^*} + \frac{b\nu}{2}\|\boldsymbol{\gamma}^*\|_2^2. \end{aligned}$$

Finally, by a recursive argument, we have

$$\begin{aligned} & (\mathbf{D}_2 - \boldsymbol{\Delta}_l)(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\gamma}^*) \\ & \leq \left( \frac{1}{\kappa + 1} \right)^{t-1} (\mathbf{D}_2 - \boldsymbol{\Delta}_l)(\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^*) + \frac{a}{\kappa} A^2 \sigma^2 (1 + \vartheta) o^* \log \frac{en}{o^*} + \frac{b\nu}{2\kappa} \|\boldsymbol{\gamma}^*\|_2^2. \end{aligned}$$

Choosing  $a = 1$ ,  $b = 1$ , which ensures  $1 + \kappa \geq 0$ , we obtain the conclusion on the sequence of  $\boldsymbol{\gamma}$ -iterates.

In fact, according to [\(A.11\)](#), we may choose any  $a \geq 1$  and  $b > 0$ , and so the linear convergence rate holds if  $\varepsilon > 1/(2\sqrt{\vartheta}) - (2 - 1/(\sqrt{\vartheta}))\nu$ .

To show the conclusion on the sequence of  $\boldsymbol{\beta}$ -iterates, notice that for any  $t \geq 1$ ,

$$\mathbf{X} \boldsymbol{\beta}^{(t)} - \mathbf{X} \boldsymbol{\beta}^* + \mathbf{H} \boldsymbol{\gamma}^{(t)} - \mathbf{H} \boldsymbol{\gamma}^* = \mathbf{H} \boldsymbol{\epsilon}.$$

and

$$\mathbb{P}[\|\mathbf{H}\boldsymbol{\epsilon}\|_2^2 \geq cr(\mathbf{X})\sigma^2 + t\sigma^2] \leq C \exp(-c[\{t^2/r(\mathbf{X})\} \wedge t]),$$

where  $r(\mathbf{X})$  denotes the rank of  $\mathbf{X}$ . The last inequality can be shown from the Hanson-Wright inequality (Rudelson and Vershynin, 2013) under the independent sub-Gaussian assumption since  $\|\mathbf{H}\|_2 = 1$  and  $\|\mathbf{H}\|_F^2 = r(\mathbf{X})$ . Then applying the Cauchy-Schwartz inequality gives the result (details omitted).

### A.3 Proof of Theorem 3

First, a thresholding function (She and Owen, 2011) is defined as a real-valued function  $\Theta(t; \lambda)$  for  $-\infty < t < \infty$  and  $0 \leq \lambda < \infty$  such that (i)  $\Theta(-t; \lambda) = -\Theta(t; \lambda)$ ; (ii)  $\Theta(t; \lambda) \leq \Theta(t'; \lambda)$  for  $t \leq t'$ ; (iii)  $\lim_{t \rightarrow \infty} \Theta(t; \lambda) = \infty$ ; (iv)  $0 \leq \Theta(t; \lambda) \leq t$  for  $0 \leq t < \infty$ .  $\Theta$  is defined component-wise if either  $t$  or  $\lambda$  is replaced by a vector. Throughout the paper, assume that  $\lambda$  is the threshold parameter.

Let  $f(\bar{\boldsymbol{\beta}}) = \tilde{l}(\bar{\boldsymbol{\beta}}) + P(\varrho\boldsymbol{\beta}; \lambda) + \nu\|\boldsymbol{\gamma}\|_2^2/2 + \iota(\boldsymbol{\gamma}; q)$ , where  $\tilde{l}(\bar{\boldsymbol{\beta}}) = l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}; \mathbf{y})$  and  $\iota(\boldsymbol{\gamma}; q) = 0$  if  $\|\boldsymbol{\gamma}\|_0 \leq q$  and  $+\infty$  otherwise. Let  $g(\bar{\boldsymbol{\beta}}; \bar{\boldsymbol{\beta}}^-) = l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-; \mathbf{y}) + \langle \bar{\mathbf{X}}^T \nabla l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-; \mathbf{y}), \bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^- \rangle + \varrho^2\|\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-\|_2^2/2 + \nu\|\boldsymbol{\gamma}\|_2^2/2 + P(\varrho\boldsymbol{\beta}; \lambda) + \iota(\boldsymbol{\gamma}; q)$ . We drop the dependence on  $\varrho$  for notational simplicity.

First, we claim that for any  $\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}^-$ ,

$$\Delta_{\tilde{l}}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}^-) \leq \varrho^2 \mathbf{D}_2(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}^-) \quad (\text{A.12})$$

provided that  $\varrho \geq \sqrt{L}\|\bar{\mathbf{X}}\|_2$ , where the generalized Bregman notations defined in Section 4 are used. In fact,

$$\begin{aligned} & \Delta_{\tilde{l}}(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}^-) \\ &= \tilde{l}(\bar{\boldsymbol{\beta}}) - \tilde{l}(\bar{\boldsymbol{\beta}}^-) - \langle \nabla \tilde{l}(\bar{\boldsymbol{\beta}}^-), \bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^- \rangle \\ &= \int_0^1 \langle \nabla \tilde{l}(\bar{\boldsymbol{\beta}}^- + t(\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-)), \bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^- \rangle dt - \int_0^1 \langle \nabla \tilde{l}(\bar{\boldsymbol{\beta}}^-), \bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^- \rangle dt \\ &= \int_0^1 \langle \bar{\mathbf{X}}^T \nabla l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^- + t\bar{\mathbf{X}}(\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-)), \bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^- \rangle dt - \int_0^1 \langle \bar{\mathbf{X}}^T \nabla l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-), \bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^- \rangle dt \\ &= \int_0^1 \langle \nabla l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^- + t\bar{\mathbf{X}}(\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-)) - \nabla l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-), \bar{\mathbf{X}}(\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-) \rangle dt \\ &\leq \int_0^1 \|\nabla l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^- + t\bar{\mathbf{X}}(\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-)) - \nabla l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-)\|_2 \|\bar{\mathbf{X}}(\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-)\|_2 dt \\ &\leq \int_0^1 tL\|\bar{\mathbf{X}}\bar{\boldsymbol{\beta}} - \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-\|_2 \|\bar{\mathbf{X}}\bar{\boldsymbol{\beta}} - \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-\|_2 dt \\ &= \frac{L}{2} \|\bar{\mathbf{X}}\bar{\boldsymbol{\beta}} - \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^-\|_2^2 \leq \frac{LM_{\bar{\mathbf{X}}}(p, 2q)}{2} \|\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-\|_2^2 \leq \frac{L\|\bar{\mathbf{X}}\|_2^2}{2} \|\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^-\|_2^2. \end{aligned}$$

Next, we solve  $\min_{\bar{\beta}} g(\bar{\beta}; \bar{\beta}^-)$ . Due to the separability, the problem is equivalent to

$$\begin{aligned} \min_{\beta} \varrho^2 \|\beta - \beta^- + \mathbf{X}^T \nabla l(\bar{\mathbf{X}} \bar{\beta}^-; \mathbf{y}) / \varrho^2\|_2^2 / 2 + P(\varrho \beta; \lambda), \text{ and} \\ \min_{\gamma} \|\gamma - \gamma^- + \nabla l(\bar{\mathbf{X}} \bar{\beta}^-; \mathbf{y}) / \varrho^2\|_2^2 / 2 + (\nu / \varrho^2) \|\gamma\|_2^2 / 2 + \iota(\gamma; q). \end{aligned}$$

With a change of variables, it leads to the MM updates given in the theorem, cf. [She \(2012\)](#). Furthermore, it is easy to show (details omitted) that the construction of  $g$  is equivalent to

$$g(\bar{\beta}; \bar{\beta}^-) = f(\bar{\beta}) + (\varrho^2 \mathbf{D}_2 - \Delta_{\bar{\beta}})(\bar{\beta}, \bar{\beta}^-).$$

Then,  $g(\bar{\beta}; \bar{\beta}^-) - f(\bar{\beta}) = \varrho^2 \mathbf{D}_2(\bar{\beta}, \bar{\beta}^-) - \Delta_{\bar{\beta}}(\bar{\beta}, \bar{\beta}^-)$ . By the optimality of  $\bar{\beta}^{(t+1)}$ , we obtain

$$\varrho^2 \mathbf{D}_2(\bar{\beta}, \bar{\beta}^-) - \Delta_{\bar{\beta}}(\bar{\beta}, \bar{\beta}^-) + f(\bar{\beta}^{(t+1)}) \leq g(\bar{\beta}^{(t+1)}; \bar{\beta}^{(t)}) \leq g(\bar{\beta}^{(t)}; \bar{\beta}^{(t)}) = f(\bar{\beta}^{(t)}). \quad (\text{A.13})$$

It follows that

$$(\varrho^2 \mathbf{D}_2 - \Delta_{\bar{\beta}})(\bar{\beta}^{(t+1)}, \bar{\beta}^{(t)}) \leq f(\bar{\beta}^{(t)}) - f(\bar{\beta}^{(t+1)}). \quad (\text{A.14})$$

Summing up (A.14) over  $t = 0, 1, \dots, T$ , and using  $\Delta_{\bar{\beta}}(\bar{\beta}, \bar{\beta}^-) \leq L \mathbf{D}_2(\bar{\mathbf{X}} \bar{\beta}, \bar{\mathbf{X}} \bar{\beta}^-)$  proved before, we obtain the convergence rate.

Finally, we prove the limit-point equation. Since

$$\nu \|\gamma^{(t)}\|_2^2 / 2 + P(\varrho \beta^{(t)}; \lambda) \leq f(\bar{\beta}^{(t)}) \leq f(\bar{\beta}^{(0)}),$$

the sequence of  $\{\bar{\beta}^{(t)}\}_{t=0}^{\infty}$  is uniformly bounded under  $\nu > 0$ . Moreover,  $\lim_{t \rightarrow \infty} (f(\bar{\beta}^{(t+1)}) - f(\bar{\beta}^{(t)})) = 0$  implies that  $\lim_{t \rightarrow \infty} (\varrho^2 \mathbf{D}_2 - \Delta_{\bar{\beta}})(\bar{\beta}^{(t+1)}, \bar{\beta}^{(t)}) = 0$  and so  $\lim_{t \rightarrow \infty} (\bar{\beta}^{(t+1)} - \bar{\beta}^{(t)}) = 0$  since  $\varrho^2 > L M_{\bar{\mathbf{X}}}(p, 2q)$ . Let  $(\hat{\beta}, \hat{\gamma})$  be any limit point of  $(\beta^{(t)}, \gamma^{(t)})$  satisfying  $\hat{\beta} = \lim_{k \rightarrow \infty} \beta^{(j_k)}$  and  $\hat{\gamma} = \lim_{k \rightarrow \infty} \gamma^{(j_k)}$  for some sequence  $\{j_k, k = 1, 2, \dots\}$ . Then

$$\begin{aligned} \mathbf{0} &= \lim_{k \rightarrow \infty} \{\beta^{(j_k+1)} - \beta^{(j_k)}\} \\ &= \lim_{k \rightarrow \infty} \Theta(\varrho \beta^{(j_k)} - \mathbf{X}^T \nabla l(\mathbf{X} \beta^{(j_k)} + \gamma^{(j_k)}) / \varrho; \lambda) / \varrho - \hat{\beta} \\ &= \Theta(\varrho \hat{\beta} - \mathbf{X}^T \nabla l(\mathbf{X} \hat{\beta} + \hat{\gamma}) / \varrho; \lambda) / \varrho - \hat{\beta} \end{aligned} \quad (\text{A.15})$$

and

$$\begin{aligned} \mathbf{0} &= \lim_{k \rightarrow \infty} \{\gamma^{(j_k+1)} - \gamma^{(j_k)}\} \\ &= \lim_{k \rightarrow \infty} \Theta^{\#}(\gamma^{(j_k)} - \nabla l(\mathbf{X} \beta^{(j_k)} + \gamma^{(j_k)}) / \varrho^2; q, \nu / \varrho^2) - \hat{\gamma} \\ &= \Theta^{\#}(\hat{\gamma} - \nabla l(\mathbf{X} \hat{\beta} + \hat{\gamma}) / \varrho^2; q, \nu / \varrho^2) - \hat{\gamma} \end{aligned} \quad (\text{A.16})$$

due to the continuity assumption of  $\Theta$  and the  $\Theta^{\#}$ -uniqueness assumption.

**Remark A.1.** *The conclusion of  $f(\bar{\beta}^{(t)}) \leq f(\bar{\beta}^{(t+1)})$  holds more generally, as long as  $g(\bar{\beta}^{(t+1)}; \bar{\beta}^{(t)}) \leq g(\bar{\beta}^{(t)}; \bar{\beta}^{(t)})$ . Moreover, the exact value of  $L$  or  $M_{\bar{\mathbf{X}}}(p, 2q)$  need not be known, because we can perform a line search until  $f(\bar{\beta}^{(t+1)}) \leq g(\bar{\beta}^{(t+1)}; \bar{\beta}^{(t)})$  is satisfied.*

## A.4 Thresholding equations of A-estimators

For the penalized form problem  $\min_{\beta, \|\gamma\|_0 \leq q} l(\bar{\mathbf{X}}\bar{\beta}; \mathbf{y}) + P(\varrho\beta; \lambda)$  with  $P$  associated with a thresholding  $\Theta$ :  $P(\theta; \lambda) - P(0; \lambda) = \int_0^{|\theta|} (\sup\{s : \Theta(s; \lambda) \leq u\} - u) du$  as defined in Theorem 3, let  $\mathcal{A}(\lambda, q)$  be the set of all alternative estimators satisfying  $\hat{\beta} \in \arg \min_{\beta} l(\mathbf{X}\beta + \hat{\gamma}; \mathbf{y}) + P(\varrho\beta; \lambda)$  and  $\hat{\gamma} \in \arg \min_{\gamma} l(\mathbf{X}\hat{\beta} + \gamma; \mathbf{y})$  s.t.  $\|\gamma\|_0 \leq q$ .

Similarly, for the  $\ell_0$ -constrained form, let  $\mathcal{A}(q_{\beta}, q_{\gamma})$  be the set of alternative estimators  $(\hat{\beta}, \hat{\gamma})$  satisfying  $\hat{\beta} \in \arg \min_{\beta} l(\mathbf{X}\beta + \hat{\gamma}; \mathbf{y})$  s.t.  $\|\beta\|_0 \leq q_{\beta}$  and  $\hat{\gamma} \in \arg \min_{\gamma} l(\mathbf{X}\hat{\beta} + \gamma; \mathbf{y})$  s.t.  $\|\gamma\|_0 \leq q_{\gamma}$ . (Recall the definition of  $M_{\mathbf{X}}$  in Section 1 and we know  $M_{\mathbf{X}}(q) \leq \|\mathbf{X}\|_2^2$  for any  $q$ .)

**Theorem A.1.** *Under the 1-Lipschitz condition of  $\nabla l$ , (i) any  $(\hat{\beta}, \hat{\gamma}) \in \mathcal{A}(q_{\beta}, q_{\gamma})$  satisfies the following two quantile-thresholding equations:*

$$\begin{cases} \hat{\beta} = \Theta^{\#}(\hat{\beta} - \mathbf{X}^T \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y})/\rho; q_{\beta}) \\ \hat{\gamma} = \Theta^{\#}(\hat{\gamma} - \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y}); q_{\gamma}) \end{cases} \quad (\text{A.17})$$

for any  $\rho > M_{\mathbf{X}}(2q_{\beta})$ , and (ii) any  $(\hat{\beta}, \hat{\gamma}) \in \mathcal{A}(\lambda, q)$  satisfies the mixed thresholding equations

$$\begin{cases} \hat{\beta} = \Theta(\varrho\hat{\beta} - \mathbf{X}^T \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y})/\varrho; \lambda)/\varrho \\ \hat{\gamma} = \Theta^{\#}(\hat{\gamma} - \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y}); q) \end{cases} \quad (\text{A.18})$$

for any  $\varrho > 0$ , provided that  $\Theta$  is continuous at  $\varrho\hat{\beta} - \mathbf{X}^T \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y})/\varrho$ .

These equations are also satisfied by the fixed points of our BCD algorithms seen from the iterate updates given in Section 3.

*Proof.* We prove the result for the doubly constrained form first. Given  $\hat{\beta}$ , construct

$$g(\gamma; \gamma^-) = l(\mathbf{X}\hat{\beta} + \gamma; \mathbf{y}) + (\mathbf{D}_2 - \Delta_{l(\mathbf{X}\hat{\beta} + \cdot)})(\gamma, \gamma^-). \quad (\text{A.19})$$

Under the 1-Lipschitz continuity of  $\nabla l$ , for any  $\gamma, \gamma^-$ ,

$$g(\gamma; \gamma^-) \geq l(\mathbf{X}\hat{\beta} + \gamma; \mathbf{y}).$$

Recall that  $\hat{\gamma}$  is a minimizer of  $l(\mathbf{X}\hat{\beta} + \gamma; \mathbf{y})$  subject to  $\|\gamma\|_0 \leq q_{\gamma}$ . Define

$$\tilde{\gamma} := \Theta^{\#}(\hat{\gamma} - \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma}); q_{\gamma}). \quad (\text{A.20})$$

Then, by use of Lemma 3 and the construction of  $g$ , we have

$$l(\mathbf{X}\hat{\beta} + \tilde{\gamma}; \mathbf{y}) \leq g(\tilde{\gamma}; \hat{\gamma}) \leq g(\hat{\gamma}; \hat{\gamma}) = l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y}), \quad (\text{A.21})$$

which means  $\tilde{\gamma}$  must also be a globally optimal solution to  $\min_{\gamma: \|\gamma\|_0 \leq q_{\gamma}} l(\mathbf{X}\hat{\beta} + \gamma; \mathbf{y})$ . The optimal support is uniquely determined due to the  $\Theta^{\#}$ -uniqueness assumption. Moreover,  $\nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma})$  restricted to  $\mathcal{J} = \{i : \hat{\gamma}_i \neq 0\}$  is a zero vector. It follows from (A.20) that

$$\tilde{\gamma}_i = \hat{\gamma}_i \neq 0, \forall i \in \mathcal{J}$$

and so  $\hat{\gamma} = \Theta^\#(\hat{\gamma} - \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma}); q_\gamma)$ .

Similar to the proof of Theorem 3, given  $\hat{\gamma}$ , any  $\tilde{\beta} \in \arg \min_{\|\beta\|_0 \leq q_\beta} l(\mathbf{X}\beta + \hat{\gamma}; \mathbf{y}) =: \tilde{l}(\beta)$  satisfies (details omitted)

$$\tilde{l}(\tilde{\beta}) + (\rho \mathbf{D}_2 - \Delta_{\tilde{l}})(\tilde{\beta}, \tilde{\beta}) \leq \tilde{l}(\hat{\beta}),$$

where

$$\tilde{\beta} = \Theta^\#(\hat{\beta} - \mathbf{X}^T \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma}; \mathbf{y})/\rho; q_\beta). \quad (\text{A.22})$$

Therefore, for any  $\rho > M_{\mathbf{X}}(2q_\beta)$ ,  $\tilde{l}(\tilde{\beta}) = \tilde{l}(\hat{\beta})$ ,  $(\rho \mathbf{D}_2 - \Delta_{\tilde{l}})(\tilde{\beta}, \hat{\beta}) = 0$ , and so  $\tilde{\beta} = \hat{\beta}$  (cf. Appendix A.3).

To prove the result for the penalized form, it suffices to study the conditions satisfied by a local or coordinate-wise minimizer  $\hat{\beta}$  of  $f_\Theta := l(\mathbf{X}\beta; \mathbf{y}) + \sum_{j=1}^p P_\Theta(|\beta_j|; \lambda)$ . The proof of Theorem 1 in She (2016) can be modified for the purpose. For completeness, we give the details below. Denote the criterion by  $f$  for simplicity and define  $s(u; \lambda) := \Theta^{-1}(u; \lambda) - u$  for  $u \geq 0$ . Let  $\delta f(\beta; \mathbf{h})$  denote the directional derivative of  $f$  at  $\beta$  with increment  $\mathbf{h}$ :

$$\delta f(\beta; \mathbf{h}) = \lim_{\epsilon \rightarrow 0^+} \frac{f(\beta + \epsilon \mathbf{h}) - f(\beta)}{\epsilon}.$$

By the definition of  $P_\Theta$ ,  $\delta f(\beta; \mathbf{h})$  exists for any  $\mathbf{h} \in \mathbb{R}^p$ . Let  $l_0(\beta) = l(\mathbf{X}\beta)$ . Consider the following directional vectors:  $\mathbf{d}_j = [d_1, \dots, d_p]^T$  with  $d_j = \pm 1$  and  $d_{j'} = 0, \forall j' \neq j$ . Then for any  $j$ ,

$$\delta l_0(\beta; \mathbf{d}_j) = d_j \mathbf{x}_j^T \nabla l(\mathbf{X}\beta), \quad (\text{A.23})$$

$$\delta P_\Theta(\beta; \mathbf{d}_j) = \begin{cases} s(|\beta_j|) \text{sgn}(\beta_j) d_j, & \text{if } \beta_j \neq 0, \\ s(|\beta_j|), & \text{if } \beta_j = 0. \end{cases} \quad (\text{A.24})$$

Due to the local coordinatewise optimality of  $\hat{\beta}$ ,  $\delta f(\hat{\beta}; \mathbf{d}_j) \geq 0, \forall j$ . When  $\hat{\beta}_1 \neq 0$ , we obtain  $\mathbf{x}_1^T \nabla l(\mathbf{X}\hat{\beta}) + s(|\hat{\beta}_1|; \lambda) \text{sgn}(\hat{\beta}_1) = 0$ . When  $\hat{\beta}_1 = 0$ ,  $\mathbf{x}_1^T \nabla l(\mathbf{X}\hat{\beta}) + s(|\hat{\beta}_1|; \lambda) \geq 0$  and  $-\mathbf{x}_1^T \nabla l(\mathbf{X}\hat{\beta}) + s(|\hat{\beta}_1|; \lambda) \geq 0$ , i.e.,  $|\mathbf{x}_1^T \nabla l(\mathbf{X}\hat{\beta})| \leq s(|\hat{\beta}_1|; \lambda) = \Theta^{-1}(0; \lambda)$ . To summarize, whenever  $f$  achieves a local coordinate-wise minimum at  $\hat{\beta}$ , we have

$$\hat{\beta}_j \neq 0 \Rightarrow \Theta^{-1}(|\hat{\beta}_j|; \lambda) \text{sgn}(\hat{\beta}_j) = \hat{\beta}_j - \mathbf{x}_j^T \nabla l(\mathbf{X}\hat{\beta}) \quad (\text{A.25})$$

$$\hat{\beta}_j = 0 \Rightarrow \Theta(\mathbf{x}_j^T \nabla l(\mathbf{X}\hat{\beta}); \lambda) = 0. \quad (\text{A.26})$$

When  $\Theta$  is continuous at  $\hat{\beta}_j - \mathbf{x}_j^T \nabla l(\mathbf{X}\hat{\beta})$ , (A.25) implies that  $\hat{\beta}_j = \Theta(\hat{\beta}_j - \mathbf{x}_j^T \nabla l(\mathbf{X}\hat{\beta}); \lambda)$ . Hence  $\hat{\beta}$  must satisfy  $\hat{\beta} = \Theta(\hat{\beta} - \mathbf{X}^T \nabla l(\mathbf{X}\hat{\beta}); \lambda)$ . For the original problem, let  $\beta' = \varrho \beta$  and notice that  $\nabla_{\beta'} l(\mathbf{X}\beta + \hat{\gamma}) = \mathbf{X}^T \nabla l(\mathbf{X}\beta + \hat{\gamma})/\varrho$ . The conclusion thus follows.  $\square$

## A.5 Proof of Theorem 4

Recall that given any  $\beta$ ,  $\mathcal{J}(\beta)$  denotes its support, i.e.,  $\mathcal{J}(\beta) = \{j : \beta_j \neq 0\}$ , and  $J(\beta) = |\mathcal{J}(\beta)|$  and  $s^* = J(\beta^*)$  and  $o^* = J(\gamma^*)$ . Given a matrix  $\mathbf{A}$ , let  $\mathcal{P}_{\mathbf{A}}$  be the orthogonal projection matrix onto the column space of  $\mathbf{A}$  and  $\mathcal{P}_{\mathbf{A}}^\perp$  be its orthogonal complement. The

range of  $\mathbf{A}$  is denoted by  $\mathcal{R}(\mathbf{A})$ . For convenience, we also use  $\mathcal{P}_{\mathbf{A}}$  to stand for  $\mathcal{R}(\mathbf{A})$  or  $\mathcal{R}(\mathcal{P}_{\mathbf{A}})$ . In the proofs,  $[n]$  denotes the set of  $\{1, \dots, n\}$  and we use  $C$  and  $c$  and  $L$  to denote universal constants which are not necessarily the same at each occurrence.

**Lemma 5.** *Any  $(\hat{\beta}, \hat{\gamma})$  satisfying (25) can be re-characterized by*

$$(\hat{\beta}, \hat{\gamma}) \in \arg \min_{(\beta, \gamma)} g(\beta, \gamma; \beta^-, \gamma^-) \Big|_{\beta^- = \hat{\beta}, \gamma^- = \hat{\gamma}} \text{ s.t. } \|\gamma\|_0 \leq q_\gamma, \|\beta\|_0 \leq q_\beta, \quad (\text{A.27})$$

where

$$\begin{aligned} g(\beta, \gamma; \beta^-, \gamma^-) &= l(\mathbf{X}\beta^- + \gamma^-) + \langle \nabla l(\mathbf{X}\beta^- + \gamma^-), \mathbf{X}\beta - \mathbf{X}\beta^- + \gamma - \gamma^- \rangle \\ &\quad + \rho \|\beta - \beta^-\|_2^2/2 + \|\gamma - \gamma^-\|_2^2/2. \end{aligned} \quad (\text{A.28})$$

The proof of the lemma can be shown from Lemma 2 and the details are omitted. A pleasant fact is that (A.27) gives a joint optimization problem, rather than an alternative one. Also, notice that  $g$  may **not** majorize  $l$ .

From (A.28),  $g(\beta, \gamma; \hat{\beta}, \hat{\gamma})$  is equivalent to  $\rho \|\beta - \hat{\beta} + \mathbf{X}^T \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma})/\rho\|_2^2/2 + \|\gamma - \hat{\gamma} + \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma})\|_2^2/2$  up to some additive terms dependent on  $\hat{\beta}, \hat{\gamma}$  only. By  $g(\hat{\beta}, \hat{\gamma}; \hat{\beta}, \hat{\gamma}) \leq g(\beta^*, \gamma^*; \hat{\beta}, \hat{\gamma})$  and Lemma 3,

$$\begin{aligned} &\frac{1}{2} \|\gamma^* - \hat{\gamma} + \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma})\|_2^2 - \frac{1}{2} \|\nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma})\|_2^2 \\ &\quad + \frac{\rho}{2} \|\beta^* - \hat{\beta} + \mathbf{X}^T \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma})/\rho\|_2^2 - \frac{\rho}{2} \|\mathbf{X}^T \nabla l(\mathbf{X}\hat{\beta} + \hat{\gamma})/\rho\|_2^2 \\ &\geq \frac{1}{2} [1 - \mathcal{L}(\mathcal{J}(\gamma^*), \mathcal{J}(\hat{\gamma}))] \|\hat{\gamma} - \gamma^*\|_2^2 + \frac{\rho}{2} [1 - \mathcal{L}(\mathcal{J}(\beta^*), \mathcal{J}(\hat{\beta}))] \|\hat{\beta} - \beta^*\|_2^2. \end{aligned} \quad (\text{A.29})$$

From  $\mathcal{L}(\mathcal{J}(\gamma^*), \mathcal{J}(\hat{\gamma})) \leq 1/\sqrt{\vartheta}$ ,  $\mathcal{L}(\mathcal{J}(\beta^*), \mathcal{J}(\hat{\beta})) \leq 1/\sqrt{\vartheta}$ , and the definition of noise  $\epsilon$ , we can show (details omitted)

$$2\bar{\Delta}_l(\mathbf{X}\hat{\beta} + \hat{\gamma}, \mathbf{X}\beta^* + \gamma^*) \leq \frac{1}{2\sqrt{\vartheta}} (\|\hat{\gamma} - \gamma^*\|_2^2 + \rho \|\hat{\beta} - \beta^*\|_2^2) + \langle \epsilon, \mathbf{X}\hat{\beta} - \mathbf{X}\beta^* + \hat{\gamma} - \gamma^* \rangle. \quad (\text{A.30})$$

Define

$$P(\beta, \gamma) = J(\beta) + J(\beta) \log(ep/J(\beta)) + J(\gamma) + J(\gamma) \log(en/J(\gamma)).$$

Because the function depends on cardinality only, we also denote it by  $P(J(\beta), J(\gamma))$ . We will show that with high probability, the last stochastic term or  $\langle \epsilon, \bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\beta^* \rangle$  in (A.30) can be bounded by the sum of  $\|\bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\beta^*\|_2^2$  and  $\sigma^2 P(\beta^*, \gamma^*) + \sigma^2 P(\hat{\beta}, \hat{\gamma})$ , up to some multiplicative constants.

To achieve the purpose, we decompose  $\bar{\mathbf{X}}(\hat{\beta} - \beta^*) =: \bar{\mathbf{X}}\Delta$  as follows

$$\begin{aligned} \bar{\mathbf{X}}\Delta &= \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}}(\bar{\mathbf{X}}\Delta) + \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}^\perp}(\bar{\mathbf{X}}\Delta) \\ &= \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}}(\bar{\mathbf{X}}\Delta) + \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}^\perp}(\mathbf{X}\hat{\beta} + \hat{\gamma}) - \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}^\perp}(\mathbf{X}\beta^* + \gamma^*) \\ &= \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}}(\bar{\mathbf{X}}\Delta) + \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}^\perp} \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\hat{\beta})}}(\mathbf{X}\hat{\beta} + \hat{\gamma}) \equiv I + II. \end{aligned} \quad (\text{A.31})$$

It is easy to verify that  $\|I\|_2^2 + \|II\|_2^2 = \|\bar{\mathbf{X}}\Delta\|_2^2$ . We use Lemma 6 and Lemma 7 to handle  $\langle \boldsymbol{\epsilon}, I \rangle$  and  $\langle \boldsymbol{\epsilon}, II \rangle$ , respectively. The proof of Lemma 6 is given at the end; the proof of Lemma 7 is similar and omitted.

**Lemma 6.** *Given  $[\mathbf{X}^{(1)}, \mathbf{X}^{(2)}] \in \mathbb{R}^{n \times (p_1+p_2)}$  where  $\mathbf{X}^{(1)} \in \mathbb{R}^{n \times p_1}$ ,  $\mathbf{X}^{(2)} \in \mathbb{R}^{n \times p_2}$ ,  $\text{rank}(\mathbf{X}^{(1)}) = r_1$ ,  $\text{rank}(\mathbf{X}^{(2)}) = r_2$  and  $1 \leq J_1 \leq p_1$ ,  $1 \leq J_2 \leq p_2$ , define  $\Gamma_{J_1, J_2} = \{\boldsymbol{\alpha} \in \mathbb{R}^{p_1+p_2} : \|\boldsymbol{\alpha}\|_2 \leq 1, \boldsymbol{\alpha} \in \mathcal{P}_{[\mathbf{X}^{(1)}_{\mathcal{J}_1}, \mathbf{X}^{(2)}_{\mathcal{J}_2}]} \text{ for some } \mathcal{J}_1 \subset [p_1], \mathcal{J}_2 \subset [p_2] \text{ with } |\mathcal{J}_1| \leq J_1, |\mathcal{J}_2| \leq J_2\}$ . Let  $P'(J_1, J_2) = \sigma^2 \{J_1 \wedge r_1 + J_2 \wedge r_2 + \log \binom{p_1}{J_1} + \log \binom{p_2}{J_2}\}$ . Then for any  $t \geq 0$ ,*

$$\mathbb{P} \left( \sup_{\boldsymbol{\alpha} \in \Gamma_{J_1, J_2}} \langle \boldsymbol{\epsilon}, \boldsymbol{\alpha} \rangle > t\sigma + \sqrt{LP'(J_1, J_2)} \right) \leq C \exp(-ct^2), \quad (\text{A.32})$$

where  $L, C, c$  are universal constant.

**Lemma 7.** *Given  $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}] \in \mathbb{R}^{n \times (p_1+p_2)}$  where  $\mathbf{X}^{(1)} \in \mathbb{R}^{n \times p_1}$ ,  $\mathbf{X}^{(2)} \in \mathbb{R}^{n \times p_2}$ ,  $\text{rank}(\mathbf{X}^{(1)}) = r_1$ ,  $\text{rank}(\mathbf{X}^{(2)}) = r_2$  and  $1 \leq J_1, J'_1 \leq p_1$ ,  $1 \leq J_2, J'_2 \leq p_2$ , let  $\mathcal{P}_{\mathcal{J}_1, \mathcal{J}_2} := \mathcal{P}_{[\mathbf{X}^{(1)}_{\mathcal{J}_1}, \mathbf{X}^{(2)}_{\mathcal{J}_2}]}$  and define  $\Gamma_{J_1, J_2, J'_1, J'_2} = \{\boldsymbol{\alpha} \in \mathbb{R}^{p_1+p_2} : \|\boldsymbol{\alpha}\|_2 \leq 1, \boldsymbol{\alpha} \in \mathcal{R}(\mathcal{P}_{\mathcal{J}_1, \mathcal{J}_2}^\perp \mathcal{P}_{\mathcal{J}'_1, \mathcal{J}'_2}) \text{ for } \mathcal{J}_1, \mathcal{J}'_1 \subset [p_1], \mathcal{J}_2, \mathcal{J}'_2 \subset [p_2] \text{ and } |\mathcal{J}_1| \leq J_1, |\mathcal{J}_2| \leq J_2, |\mathcal{J}'_1| \leq J'_1, |\mathcal{J}'_2| \leq J'_2\}$ . Let  $P''(J_1, J_2, J'_1, J'_2) = \sigma^2 \{J'_1 \wedge r_1 + J'_2 \wedge r_2 + \log \binom{p_1}{J_1} \binom{p_2}{J_2} \binom{p_1}{J'_1} \binom{p_2}{J'_2}\}$ . Then for any  $t \geq 0$ ,*

$$\mathbb{P} \left( \sup_{\boldsymbol{\alpha} \in \Gamma_{J_1, J_2, J'_1, J'_2}} \langle \boldsymbol{\epsilon}, \boldsymbol{\alpha} \rangle > t\sigma + \sqrt{LP''(J_1, J_2, J'_1, J'_2)} \right) \leq C \exp(-ct^2), \quad (\text{A.33})$$

where  $L, C, c$  are universal constant.

Define  $\Gamma_{s,o} = \{\boldsymbol{\Delta} \in \mathcal{P}_{[\mathbf{X}_{\mathcal{J}_1}, \mathbf{I}_{\mathcal{J}_2}]}, \|\boldsymbol{\Delta}\|_2 \leq 1, \mathcal{J}_1 \subset [p], \mathcal{J}_2 \subset [n], |\mathcal{J}_1| \leq s, |\mathcal{J}_2| \leq o\}$ . Given any  $a, b, a' > 0$ , we have

$$\begin{aligned} & \langle \boldsymbol{\epsilon}, \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\boldsymbol{\Delta} \rangle - \frac{1}{a} \|\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\boldsymbol{\Delta}\|_2^2 - bL\sigma^2 P(J(\boldsymbol{\beta}^*), J(\boldsymbol{\gamma}^*)) \\ & \leq \|\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\boldsymbol{\Delta}\|_2 \langle \boldsymbol{\epsilon}, \frac{\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\boldsymbol{\Delta}}{\|\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\boldsymbol{\Delta}\|_2} \rangle - 2\sigma \|\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\boldsymbol{\Delta}\|_2 \sqrt{\frac{b}{a} LP(J(\boldsymbol{\beta}^*), J(\boldsymbol{\gamma}^*))} \\ & \leq \frac{1}{a'} \|\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\boldsymbol{\Delta}\|_2^2 + \frac{a'}{4} \sup_{\boldsymbol{\Delta} \in \Gamma_{s,o}} [\langle \boldsymbol{\epsilon}, \boldsymbol{\Delta} \rangle - 2\sigma \{(b/a)LP(s^*, o^*)\}^{1/2}]_+^2 \end{aligned}$$

by the Cauchy-Schwarz inequality.

Let  $R_{s,o} = \sup_{\boldsymbol{\Delta} \in \Gamma_{s,o}} [\langle \boldsymbol{\epsilon}, \boldsymbol{\Delta} \rangle - 2\sigma \{(b/a)LP(s, o)\}^{1/2}]_+$ . Lemma 6 indicates that  $R_{s,o}$  is bounded above by a constant times  $\sigma^2$  in expectation when choosing  $b/a$  to be a constant greater than  $1/4$ . Note that when  $s = o = 0$ ,  $R_{0,0} = 0$ . When  $s \geq 1, o \geq 1$ , for any  $t \geq 0$ ,

$$\begin{aligned} & \mathbb{P}(R_{s,o} > t\sigma) \\ & \leq \mathbb{P} \left( \sup_{\boldsymbol{\Delta} \in \Gamma_{s,o}} \langle \boldsymbol{\epsilon}, \boldsymbol{\Delta} \rangle - \sqrt{LP(s, o)} > t\sigma + 2\sigma \sqrt{\frac{b}{a} LP(s, o)} - \sigma \sqrt{LP(s, o)} \right) \\ & \leq C \exp(-ct^2) \exp[-c(2\sqrt{b/a} - 1)^2 P(s, o)] \\ & \leq C \exp(-ct^2) (np)^{-c} \leq C \exp(-ct^2), \end{aligned} \quad (\text{A.34})$$

where we used  $\log \binom{p}{J} \leq CJ \log(ep/J)$  and  $C, c > 0$  are constants. Similarly, when  $s = 0$ , we know  $\mathbb{P}(R_{s,o} > t\sigma) \leq C \exp(-ct^2)n^{-c}$  for any  $0 \leq o \leq n$ , and when  $o = 0$ ,  $\mathbb{P}(R_{s,o} > t\sigma) \leq C \exp(-ct^2)p^{-c}$  for any  $0 \leq s \leq p$ . From these tail bounds, we get

$$\mathbb{E}\langle \epsilon, \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\Delta \rangle \leq \mathbb{E}\left[\left(\frac{1}{a} + \frac{1}{a'}\right) \|\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\Delta\|_2^2 + bL\sigma P(J(\beta^*), J(\gamma^*))\right] + a'C\sigma^2.$$

Repeating the treatment using Lemma 7 gives

$$\begin{aligned} \mathbb{E}\langle \epsilon, P_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}}^\perp \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\hat{\beta})}}(\mathbf{X}\hat{\beta} + \hat{\gamma}) \rangle &\leq \mathbb{E}\left[\left(\frac{1}{a} + \frac{1}{a'}\right) \|P_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}}^\perp \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\hat{\beta})}}(\mathbf{X}\hat{\beta} + \hat{\gamma})\|_2^2 \right. \\ &\quad \left. + bL\sigma\{P(J(\hat{\beta}), J(\hat{\gamma})) + P(J(\beta^*), J(\gamma^*))\}\right] + a'C\sigma^2. \end{aligned}$$

Combining the two bounds and using the monotone property of  $P$ , we get for any  $4b > a > 0, a' > 0$ ,

$$\mathbb{E}\langle \epsilon, \bar{\mathbf{X}}\Delta \rangle \leq \mathbb{E}\left(\frac{1}{a} + \frac{1}{a'}\right) \|\bar{\mathbf{X}}\Delta\|_2^2 + 3bL\sigma^2 P(\vartheta s^*, \vartheta o^*) + a'C\sigma^2 \quad (\text{A.35})$$

where  $C, c$  are positive constants. It follows from (A.30), (A.35) and the regularity condition (26) that

$$\mathbb{E}\left[\left(\frac{\delta}{2} - \left(\frac{1}{a} + \frac{1}{a'}\right)\right) \|\bar{\mathbf{X}}\Delta\|_2^2\right] \leq bL'\sigma^2 \vartheta(o^* \log \frac{en}{\vartheta o^*} + s^* \log \frac{ep}{\vartheta s^*}) + Ca'\sigma^2, \quad (\text{A.36})$$

where  $L'$  is a constant. Taking  $a = a' = 8/\delta$  and  $b = 4/\delta$  gives the conclusion.

In the remaining, we prove the more general result under

$$\mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\bar{\beta}^{(0)}, \bar{\mathbf{X}}\bar{\beta}^*)] \leq CM\{\vartheta o^* \sigma^2 \log \frac{en}{\vartheta o^*} + \vartheta s^* \sigma^2 \log \frac{ep}{\vartheta s^*} + \sigma^2\} \quad (\text{A.37})$$

where  $M$  satisfies  $+\infty \geq M \geq 1$  and  $C \geq 0$  is a constant, and

$$\left\{2\left(1 - \frac{1}{M}\right)\bar{\Delta}_l + \frac{C_0}{M(M\delta_0 \vee c_0)}\Delta_l - \delta_0 \mathbf{D}_2\right\}(\bar{\mathbf{X}}\bar{\beta}, \bar{\mathbf{X}}\bar{\beta}') \geq \frac{(1 - 1/M)}{\sqrt{\vartheta}}(\rho \mathbf{D}_2(\beta, \beta') + \mathbf{D}_2(\gamma, \gamma')) \quad (\text{A.38})$$

for all  $\bar{\beta}, \bar{\beta}'$ :  $\|\beta\|_0 \leq \vartheta s^*, \|\beta'\|_0 \leq s^*, \|\gamma\|_0 \leq \vartheta o^*, \|\gamma'\|_0 \leq o^*$  for some  $\delta_0 > 0$  and constants  $C_0, c_0 > 0$ .

First, from the above analysis, we have

$$\mathbb{E}[(2\delta\bar{\Delta}_l - \delta^2\mathbf{D}_2)(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*)] - \mathbb{E}\left[\frac{\delta}{\sqrt{\vartheta}}(\rho \mathbf{D}_2(\hat{\beta}, \beta^*) + \mathbf{D}_2(\hat{\gamma}, \gamma^*))\right] \leq CE. \quad (\text{A.39})$$

where  $E := \vartheta o^* \sigma^2 \log \frac{en}{\vartheta o^*} + \vartheta s^* \sigma^2 \log \frac{ep}{\vartheta s^*} + \sigma^2$ . Next, from  $l(\bar{\mathbf{X}}\hat{\beta}) \leq l(\bar{\mathbf{X}}\bar{\beta}^{(0)})$ , we have

$$\Delta_l(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) \leq \Delta_l(\bar{\mathbf{X}}\bar{\beta}^{(0)}, \bar{\mathbf{X}}\bar{\beta}^*) + \langle \epsilon, \bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\bar{\beta}^* \rangle - \langle \epsilon, \bar{\mathbf{X}}\bar{\beta}^{(0)} - \bar{\mathbf{X}}\bar{\beta}^* \rangle \quad (\text{A.40})$$

from which it follows that for any  $\delta' > 0$ ,

$$\mathbb{E}[(\Delta_l - \delta' \mathbf{D}_2)(\bar{\mathbf{X}}\hat{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*)] \leq (\Delta_l + \frac{1}{M} \mathbf{D}_2)(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^{(0)}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*) + CE(\frac{1}{\delta'} + M). \quad (\text{A.41})$$

Since  $\nabla l$  is  $\text{Lip}(1)$  and  $M \geq 1$ , we obtain

$$\mathbb{E}[(\Delta_l - \delta' \mathbf{D}_2)(\bar{\mathbf{X}}\hat{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*)] \leq CE(\frac{1}{\delta'} + M) \leq \frac{C}{c_1 \wedge c_2} E(\frac{c_1}{\delta'} + c_2 M) \quad (\text{A.42})$$

where  $c_1, c_2 > 0$  are arbitrary constants and recall that  $C$  may not be the same constant at each occurrence. Taking  $\delta^2 = \delta'^2/(c_1 + c_2 M \delta')$ , we get

$$\mathbb{E}[(\frac{\delta^2}{\delta'} \Delta_l - \delta^2 \mathbf{D}_2)(\bar{\mathbf{X}}\hat{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*)] \leq \frac{C}{c_1 \wedge c_2} E. \quad (\text{A.43})$$

Multiplying (A.39) by  $(1 - 1/M)$  and (A.43) by  $1/M$  and adding the two inequalities yield

$$\begin{aligned} & \mathbb{E}[(1 - \frac{1}{M}) \{ 2\bar{\Delta}_l(\bar{\mathbf{X}}\hat{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*) - \frac{\rho \mathbf{D}_2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + \mathbf{D}_2(\hat{\boldsymbol{\gamma}}, \boldsymbol{\gamma}^*)}{\sqrt{\vartheta}} \} \\ & + (\frac{\delta}{M \delta'} \Delta_l - \delta \mathbf{D}_2)(\bar{\mathbf{X}}\hat{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*)] \leq \frac{E}{\delta} (1 + \frac{C}{c_1 \wedge c_2}). \end{aligned}$$

Simple calculation shows

$$\delta'/\delta = c_2 M \delta + \sqrt{c_2^2 M^2 \delta^2 + 4c_1} \leq (2c_2 M \delta) \vee (4c_1)$$

Under (A.38) with  $\delta_0 = 2\delta, C_0 = 1/c_2, c_0 = 4c_1/c_2$ , we get  $\mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\hat{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*)] \lesssim E/\delta_0^2$ . A reparameterization of (A.38) gives the assumed regularity condition.

**Proof of Lemma 6** By definition,  $\{\langle \boldsymbol{\epsilon}, \boldsymbol{\alpha} \rangle : \boldsymbol{\alpha} \in \Gamma_{J_1, J_2}\}$  is a stochastic process with sub-Gaussian increments. The induced metric on  $\Gamma_{J_1, J_2}$  is Euclidean  $d(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = \sigma \|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_2$ .

To bound the metric entropy  $\log \mathcal{N}(\varepsilon, \Gamma_{J_1, J_2}, d)$ , where  $\mathcal{N}(\varepsilon, \Gamma_{J_1, J_2}, d)$  is the smallest cardinality of an  $\varepsilon$ -net that covers  $\Gamma_{J_1, J_2}$  under  $d$ , we notice that  $\boldsymbol{\alpha}$  is always in a  $(r_1 \wedge J_1 + r_2 \wedge J_2)$ -dimensional ball and the number of such ball is at most  $\binom{p_1}{J_1} \binom{p_2}{J_2}$ . By a standard volume argument

$$\begin{aligned} \log \mathcal{N}(\varepsilon, \Gamma_{J_1, J_2}, d) & \leq \log \binom{p_1}{J_1} \binom{p_2}{J_2} \left(\frac{C\sigma}{\varepsilon}\right)^{r_1 \wedge J_1 + r_2 \wedge J_2} \\ & = \log \binom{p_1}{J_1} + \log \binom{p_2}{J_2} + (r_1 \wedge J_1 + r_2 \wedge J_2) \log(C\sigma/\varepsilon), \end{aligned}$$

where  $C$  is a universal constant.

From Dudley's integral bound,

$$\mathbb{P}\left(\sup_{\boldsymbol{\alpha} \in \Gamma_{J_1, J_2}} \langle \boldsymbol{\epsilon}, \boldsymbol{\alpha} \rangle \geq t\sigma + L \int_0^\sigma \sqrt{\log \mathcal{N}(\varepsilon, \Gamma_{J_1, J_2}, d)} d\varepsilon\right) \leq C \exp(-ct^2).$$

The conclusion follows from the Cauchy-Schwarz inequality.  $\square$

**Remark A.2.** Because of the tail bounds, the proof also gives a high-probability form result:

$$\|\bar{\mathbf{X}}\hat{\boldsymbol{\beta}} - \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*\|_2^2 \lesssim \frac{\sigma^2\vartheta}{\delta^2} \left\{ o^* \log \frac{en}{\vartheta o^*} + s^* \log \frac{ep}{\vartheta s^*} \right\}$$

with probability at least  $1 - Cn^{-c(1 \wedge o^*)} p^{-c(1 \wedge s^*)}$  under the same regularity condition.

In addition, if the  $\boldsymbol{\beta}$ -optimization problem is convex, such as in robust regression or logistic regression, the regularity condition can be weakened to  $(2\bar{\Delta}_l - \delta\mathbf{D}_2)(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}, \mathbf{X}\boldsymbol{\beta}' + \boldsymbol{\gamma}') \geq \|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2^2/(4\sqrt{\vartheta})$  for any  $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\beta}', \boldsymbol{\gamma}'$ :  $\|\boldsymbol{\gamma}\|_0 \leq \vartheta o^*$ ,  $\|\boldsymbol{\gamma}'\|_0 \leq o^*$  and some  $\delta > 0$ . It is also worth mentioning that to get the risk bound, (26) can be stated in expectation.

## A.6 Proof of Theorem 5

We prove a more general theorem for a strongly smooth loss  $l$  with  $\nabla l$  1-Lipschitz.

**Theorem 5'.** Consider  $\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} l(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}) + \lambda\varrho\|\boldsymbol{\beta}\|_1$  s.t.  $\|\boldsymbol{\gamma}\|_0 \leq q$ , where  $\varrho > 0$ ,  $\lambda = A\sigma\sqrt{\log(ep)}$  with  $A$  a sufficiently large constant and  $q = \vartheta o^*$  with  $\vartheta \geq 1$ . Then the following inequality holds for any coordinatewise minimum point  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$

$$\mathbb{E}[\|\bar{\mathbf{X}}\hat{\boldsymbol{\beta}} - \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*\|_2^2] \lesssim \sigma^2\vartheta o^* \log(en/o^*) + \sigma^2 K^2 s^* \log(ep) + \sigma^2 \quad (\text{A.44})$$

under the assumption that there exists  $K \geq 0$  such that  $(2\bar{\Delta}_l - \delta\mathbf{D}_2)(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}') + \lambda\varrho\|(\boldsymbol{\beta} - \boldsymbol{\beta}')_{\mathcal{J}^c}\|_1 + K^2\lambda^2 J \geq \|\boldsymbol{\gamma} - \boldsymbol{\gamma}'\|_2^2/(2\sqrt{\vartheta}) + (1 + \varepsilon)\lambda\varrho\|(\boldsymbol{\beta} - \boldsymbol{\beta}')_{\mathcal{J}}\|_1$  holds for any  $\boldsymbol{\beta}, \boldsymbol{\beta}', \boldsymbol{\gamma}, \boldsymbol{\gamma}'$  satisfying  $\|\boldsymbol{\gamma}\|_0 \leq \vartheta o^*$ ,  $\|\boldsymbol{\gamma}'\|_0 \leq o^*$ , where  $\mathcal{J} = \{j : \beta_j^* \neq 0\}$ ,  $J = |\mathcal{J}|$ , and  $\varepsilon, \delta$  are positive constants.

For resistant lasso, where  $l(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}, \mathbf{y}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2$ , the regularity condition is implied by  $K\sqrt{J}\|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}\|_2 + \varrho\|\boldsymbol{\beta}_{\mathcal{J}^c}\|_1 \geq \|\boldsymbol{\gamma}\|_2^2/(2\lambda\sqrt{\vartheta}) + (1 + \varepsilon)\varrho\|\boldsymbol{\beta}_{\mathcal{J}}\|_1$  for some (redefined)  $K \geq 0$ , where  $\|\boldsymbol{\gamma}\|_0 \leq (1 + \vartheta)o^*$  and  $\varepsilon$  is some positive constant, cf. Theorem 5.

*Proof.* First, from the proof of Theorem A.1, any coordinatewise minimum point  $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$  must obey

$$\begin{cases} \boldsymbol{\beta} = \Theta_{\text{soft}}(\boldsymbol{\beta} - \mathbf{X}^T \nabla l(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}; \mathbf{y}); \lambda\varrho), \\ \boldsymbol{\gamma} = \Theta^\#(\boldsymbol{\gamma} - \nabla l(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}; \mathbf{y}); q), \end{cases} \quad (\text{A.45})$$

where  $\Theta_{\text{soft}}(\boldsymbol{\beta}; \lambda) = \text{sign}(\boldsymbol{\beta}) \circ (|\boldsymbol{\beta}| - \lambda)_+$  is the soft-thresholding and  $\circ$  is the component-wise multiplication.

Then, similar to the proof of Lemma 5, we can prove

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) \in \arg \min_{\boldsymbol{\beta}, \boldsymbol{\gamma}} g(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\beta}^-, \boldsymbol{\gamma}^-) \Big|_{\boldsymbol{\beta}^- = \hat{\boldsymbol{\beta}}, \boldsymbol{\gamma}^- = \hat{\boldsymbol{\gamma}}} \text{ s.t. } \|\boldsymbol{\gamma}\|_0 \leq q, \quad (\text{A.46})$$

where  $g(\boldsymbol{\beta}, \boldsymbol{\gamma}; \boldsymbol{\beta}^-, \boldsymbol{\gamma}^-) = l(\mathbf{X}\boldsymbol{\beta}^- + \boldsymbol{\gamma}^-) + \langle \nabla l(\mathbf{X}\boldsymbol{\beta}^- + \boldsymbol{\gamma}^-), \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}^- + \boldsymbol{\gamma} - \boldsymbol{\gamma}^- \rangle + \|\boldsymbol{\beta} - \boldsymbol{\beta}^-\|_2^2/2 + \|\boldsymbol{\gamma} - \boldsymbol{\gamma}^-\|_2^2/2 + \varrho\lambda\|\boldsymbol{\beta}\|_1$ . Then by Lemma 2 in She (2016) and Lemma 3, we have

$$\begin{aligned} & \frac{1}{2}\|\boldsymbol{\gamma}^* - \hat{\boldsymbol{\gamma}} + \nabla l(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}})\|_2^2 - \frac{1}{2}\|\nabla l(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}})\|_2^2 + \lambda\varrho\|\boldsymbol{\beta}^*\|_1 - \lambda\varrho\|\hat{\boldsymbol{\beta}}\|_1 \\ & + \frac{1}{2}\|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} + \mathbf{X}^T \nabla l(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}})\|_2^2 - \frac{1}{2}\|\mathbf{X}^T \nabla l(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}})\|_2^2 \\ & \geq \frac{1}{2}[1 - \mathcal{L}(\mathcal{J}(\boldsymbol{\gamma}^*), \mathcal{J}(\hat{\boldsymbol{\gamma}}))]\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 + \frac{1}{2}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2, \end{aligned} \quad (\text{A.47})$$

or

$$2\bar{\Delta}_l(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}}, \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma}^*) \leq \frac{1}{2\sqrt{\vartheta}}\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 + \lambda\varrho\|\boldsymbol{\beta}^*\|_1 - \lambda\varrho\|\hat{\boldsymbol{\beta}}\|_1 + \langle \boldsymbol{\epsilon}, \bar{\mathbf{X}}\hat{\boldsymbol{\beta}} - \bar{\mathbf{X}}\boldsymbol{\beta}^* \rangle. \quad (\text{A.48})$$

Next we try to bound the stochastic term in a proper way to merge with the  $\boldsymbol{\beta}$ -penalties. Let  $P_H(t; \lambda) = (-t^2/2 + \lambda|t|)1_{|t|<\lambda} + (\lambda^2/2)1_{|t|\geq\lambda}$  and  $P_H(\boldsymbol{\beta}; \lambda) = \sum P_H(\beta_j; \lambda)$ . The following lemma based on combined statistical and computational analysis is useful.

**Lemma 8.** Define  $\Gamma_q = \{(\boldsymbol{\beta}, \boldsymbol{\gamma}) : \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\gamma} \in \mathbb{R}^n, \|\boldsymbol{\gamma}\|_0 \leq q\}$  where  $0 \leq q \leq n$ . Then for any  $\varrho \geq \|\mathbf{X}\|_2$ , there exist universal constants  $A_0, A_1, C, c, c_1, c_2 > 0$  such that for any  $a \geq 2b > 0$ , the following event

$$\sup_{(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Gamma_q} \{2\langle \boldsymbol{\epsilon}, \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} \rangle - \frac{1}{a}\|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}\|_2^2 - \frac{1}{b}P_H(\varrho\boldsymbol{\beta}; \lambda) - aA_0\sigma^2q \log(en/q)\} \geq a\sigma^2t$$

occurs with probability at most  $C \exp(-ct)$ , where  $\lambda = A\sigma\sqrt{\log(ep)}$ ,  $A = \sqrt{ab}A_1$  and  $t \geq 0$ .

From Lemma 8 and  $\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_0 \leq (1 + \vartheta)o^*$ , it is easy to see

$$\begin{aligned} & \langle \boldsymbol{\epsilon}, \bar{\mathbf{X}}\hat{\boldsymbol{\beta}} - \bar{\mathbf{X}}\boldsymbol{\beta}^* + \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^* \rangle \\ & \leq \frac{1}{a}\mathbf{D}_2(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*, \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}) + \frac{1}{2b}P_H(\varrho(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \lambda) + \frac{1}{2}\sigma^2aA_0o^*(1 + \vartheta) \log \frac{en}{o^*} + aR \end{aligned} \quad (\text{A.49})$$

with  $\mathbb{E}R \leq C\sigma^2$ . Let  $\mathcal{J} = \{j : \beta_j^* \neq 0\}$ . Due to the sub-additivity of  $\|\cdot\|_1$  and the fact that  $P_H(t; \lambda) \leq \lambda|t|$ ,

$$\begin{aligned} & \lambda\varrho\|\boldsymbol{\beta}^*\|_1 - \lambda\varrho\|\hat{\boldsymbol{\beta}}\|_1 + \frac{1}{2b}P_H(\varrho(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*); \lambda) \\ & \leq \lambda\varrho\{\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}}\|_1 - \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}^c}\|_1\} + \varepsilon\lambda\varrho\{\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}^c}\|_1 + \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}}\|_1\} \\ & \leq \lambda\varrho\{(1 + \varepsilon)\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}}\|_1 - (1 - \varepsilon)\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}^c}\|_1\} \\ & \leq \lambda\varrho\{(1 + \varepsilon)\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}}\|_1 - \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}^c}\|_1\}, \end{aligned} \quad (\text{A.50})$$

where we set  $b = 1/(2\varepsilon) > 0$ . Combining (A.48), (A.49), and (A.50) gives

$$\begin{aligned} & 2\bar{\Delta}_l(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\gamma}}, \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\gamma}^*) - \frac{1}{2a}\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^* + \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 \\ & \leq \frac{1}{2\sqrt{\vartheta}}\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^*\|_2^2 + \frac{1}{2}\sigma^2aA_0o^*(1 + \vartheta) \log \frac{en}{o^*} \\ & \quad + \lambda\varrho\{(1 + \varepsilon)\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}}\|_1 - \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{\mathcal{J}^c}\|_1\} + aR. \end{aligned} \quad (\text{A.51})$$

Choosing  $a = 2/\delta + 1/\varepsilon$ ,  $A \geq \sqrt{ab}A_1$ , and using the regularity condition, we obtain the conclusion.  $\square$

**Proof of Lemma 8** Define  $l_H(\boldsymbol{\beta}, \boldsymbol{\gamma}, q) = 2\langle \boldsymbol{\epsilon}, \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} \rangle - \|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}\|_2^2/a - P_H(\varrho\boldsymbol{\beta}; \lambda)/b - aA_0\sigma^2q \log(en/q)$ . Let  $P_0(\boldsymbol{\beta}; \lambda) = \lambda^2\|\boldsymbol{\beta}\|_0/2$ . Similarly, define  $l_0(\boldsymbol{\beta}, \boldsymbol{\gamma}, q) = 2\langle \boldsymbol{\epsilon}, \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} \rangle - \|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}\|_2^2/a - P_0(\boldsymbol{\beta}; \lambda)/b - aA_0\sigma^2q \log(en/q)$ . Let  $\mathcal{A}_H = \{\sup_{(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Gamma_q} l_H(\boldsymbol{\beta}, \boldsymbol{\gamma}, q) \geq at\sigma^2\}$  and  $\mathcal{A}_0 = \{\sup_{(\boldsymbol{\beta}, \boldsymbol{\gamma}) \in \Gamma_q} l_0(\boldsymbol{\beta}, \boldsymbol{\gamma}, q) \geq at\sigma^2\}$ . The occurrence of  $\mathcal{A}_H$  implies

$$l_H(\boldsymbol{\beta}^o, \boldsymbol{\gamma}^o, q) \geq at\sigma^2 \quad (\text{A.52})$$

for any  $(\boldsymbol{\beta}^o, \boldsymbol{\gamma}^o)$  that solves

$$\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}: \|\boldsymbol{\gamma}\|_0 \leq q} \|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}\|_2^2/2 - a\langle \boldsymbol{\epsilon}, \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} \rangle + (a/(2b))P_H(\varrho\boldsymbol{\beta}; \lambda). \quad (\text{A.53})$$

**Lemma 9.** *Given any  $\tau \geq \|\mathbf{X}\|_2^2$ , there always exists a globally optimal solution  $\boldsymbol{\beta}^o$  to  $\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2 + \tau P_H(\boldsymbol{\beta}; \lambda)$  such that for any  $j : 1 \leq j \leq p$ , either  $\beta_j^o = 0$  or  $\beta_j^o \geq \lambda\sqrt{\tau}/\|\mathbf{X}\|_2 \geq \lambda$ .*

The lemma can be shown by linearization and the property of  $P_H$ ; see [She \(2012\)](#). From Lemma 9 and  $a/2b \geq 1 \geq \|\mathbf{X}\|_2^2/\varrho^2$ , (A.52) implies the existence of an optimal solution  $(\boldsymbol{\beta}^o, \boldsymbol{\gamma}^o)$  to (A.53) such that  $l_0(\boldsymbol{\beta}^o, \boldsymbol{\gamma}^o, q) = l_H(\boldsymbol{\beta}^o, \boldsymbol{\gamma}^o, q) \geq at\sigma^2$  and so  $\mathcal{A}_H \subset \mathcal{A}_0$ . It suffices to study  $\mathbb{P}(\mathcal{A}_0)$ .

The remaining part follows the lines of the proof of Theorem 4 and is omitted.  $\square$

## A.7 Proof of Theorem 6

Assume the density of  $\mathbf{y} \in \mathcal{Y}^n$  given  $\boldsymbol{\eta}^* = \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*$  is given by

$$p_{\bar{\boldsymbol{\beta}}^*}(\mathbf{y}) = \exp\{[\langle \mathbf{y}, \bar{\boldsymbol{\eta}}^* \rangle - \langle \mathbf{1}, b(\bar{\boldsymbol{\eta}}^*) \rangle]/\sigma^2\} \quad (\text{A.54})$$

with respect to a base measure  $\mu$  defined on  $\mathcal{Y}^n$ . The corresponding loss is  $l(\boldsymbol{\eta}) = \langle \mathbf{1}, l_0(\boldsymbol{\eta}) \rangle$  with  $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} + \gamma_i = \bar{\mathbf{x}}_i^T \bar{\boldsymbol{\beta}}$  and

$$l_0(\eta_i; y_i) = (-y_i\eta_i + b(\eta_i))/\sigma^2. \quad (\text{A.55})$$

Assume the natural parameter space  $\Omega = \{\boldsymbol{\eta} \in \mathbb{R}^n : b(\boldsymbol{\eta}) < +\infty\}$  is open. Then the loss corresponds to a distribution in the regular exponential dispersion family with dispersion  $\sigma^2$  and natural parameter  $\eta_i$ . In the Gaussian case,  $l_0(\eta_i)$  is  $(\eta_i - y_i)^2/(2\sigma^2)$  up to an additive term independent of  $\eta_i$ . Consider a signal class

$$\mathcal{B}(s^*, o^*, M_\beta, M_\gamma) = \{(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*) : \|\boldsymbol{\beta}^*\|_0 \leq s^*, \|\boldsymbol{\gamma}^*\|_0 \leq o^*, \|\boldsymbol{\beta}^*\|_\infty < M_\beta, \|\boldsymbol{\gamma}^*\|_\infty < M_\gamma\} \quad (\text{A.56})$$

where  $0 \leq s^* \leq p$ ,  $0 \leq o^* \leq n$ , and  $+\infty \geq M_\beta, M_\gamma \geq 0$ . The following theorem implies Theorem 6 by setting  $M_\beta = M_\gamma = +\infty$  and assuming  $\underline{\kappa}_\beta/\bar{\kappa}_\beta$  and  $\kappa_\gamma$  are positive constants.

**Theorem 6'.** *In the regular exponential dispersion family with  $n \geq 2, p \geq 2, 1 \leq o^* \leq n/2, 1 \leq s^* \leq p/2$ , define*

$$P_\beta(s^*) = s^* \log(ep/s^*), \quad P_\gamma = o^* \log(en/o^*). \quad (\text{A.57})$$

Let  $I(\cdot)$  be any nondecreasing function with  $I(0) = 0, I \not\equiv 0$ . (i) Suppose for some  $\kappa_\beta, \kappa_\gamma > 0$   $\Delta_l(\mathbf{0}, \mathbf{X}\beta)\sigma^2 \leq \kappa_\beta \mathbf{D}_2(\mathbf{0}, \beta)$ ,  $\forall \beta : \|\beta\|_0 \leq s^*, \|\beta\|_\infty < M_\beta$ , and  $\Delta_l(\mathbf{0}, \gamma)\sigma^2 \leq \kappa_\gamma \mathbf{D}_2(\mathbf{0}, \gamma)$ ,  $\forall \gamma : \|\gamma\|_0 \leq o^*, \|\gamma\|_\infty < M_\gamma$ . Then there exist positive constants  $\tilde{c}, c$ , depending on  $I(\cdot)$  only, such that

$$\inf_{(\bar{\beta}, \hat{\gamma})} \sup_{(\beta^*, \gamma^*) \in \mathcal{B}(s^*, o^*, M_\beta, M_\gamma)} \mathbb{E} \left[ I \left( \mathbf{D}_2(\bar{\beta}^*, \hat{\beta}) / \{ \tilde{c} [\min\{\sigma^2 P_\beta(s^*)/\kappa_\beta, M_\beta^2 s^*\} \right. \right. \\ \left. \left. + \min\{\sigma^2 P_\gamma(o^*)/\kappa_\gamma, M_\gamma^2 o^*\} \} \} \right) \right] \geq c > 0, \quad (\text{A.58})$$

where  $(\hat{\beta}, \hat{\gamma})$  denotes any estimator of  $(\beta^*, \gamma^*)$ . (ii) Suppose  $\Delta_l(\mathbf{0}, \mathbf{X}\beta_1)\sigma^2 \leq \bar{\kappa}_\beta \mathbf{D}_2(\mathbf{0}, \beta_1)$  and  $\underline{\kappa}_\beta \mathbf{D}_2(\mathbf{0}, \beta_2) \leq \mathbf{D}_2(\mathbf{0}, \mathbf{X}\beta_2)$ ,  $\forall \beta_i : \|\beta_i\|_0 \leq i s^*, \|\beta_i\|_\infty < M_\beta, i = 1, 2$ , and  $\Delta_l(\mathbf{0}, \gamma)\sigma^2 \leq \kappa_\gamma \mathbf{D}_2(\mathbf{0}, \gamma)$ ,  $\forall \gamma : \|\gamma\|_0 \leq o^*, \|\gamma\|_\infty < M_\gamma$ . where  $0 \leq \underline{\kappa}_\beta \leq \bar{\kappa}_\beta, 0 \leq \kappa_\gamma$ . Then there exist positive constants  $\tilde{c}, c$  such that

$$\inf_{(\bar{\beta}, \hat{\gamma})} \sup_{(\beta^*, \gamma^*) \in \mathcal{B}(s^*, o^*, M_\beta, M_\gamma)} \mathbb{E} \left[ I \left( \mathbf{D}_2(\bar{\mathbf{X}}\bar{\beta}^*, \bar{\mathbf{X}}\hat{\beta}) / \{ \tilde{c} [\min\{(\underline{\kappa}_\beta/\bar{\kappa}_\beta)\sigma^2 P_\beta(s^*), \underline{\kappa}_\beta M_\beta^2 s^*\} \right. \right. \\ \left. \left. + \min\{\sigma^2 P_\gamma(o^*)/\kappa_\gamma, M_\gamma^2 o^*\} \} \} \right) \right] \geq c > 0. \quad (\text{A.59})$$

*Proof.* First we introduce a lemma (She et al., 2021, Lemma 3(iii)).

**Lemma 10.** For any  $p_{\bar{\beta}_1}, p_{\bar{\beta}_2}$  in the regular exponential dispersion family, the Kullback-Leibler divergence of  $p_{\bar{\beta}_2}$  from  $p_{\bar{\beta}_1}$ , defined by  $\mathcal{K}(p_{\bar{\beta}_1}, p_{\bar{\beta}_2}) = \int p_{\bar{\beta}_1} \log(p_{\bar{\beta}_1}/p_{\bar{\beta}_2}) d\nu$ , satisfies

$$\mathcal{K}(p_{\bar{\beta}_1}, p_{\bar{\beta}_2}) = \Delta_l(\bar{\mathbf{X}}\bar{\beta}_2, \bar{\mathbf{X}}\bar{\beta}_1),$$

where  $\Delta_l(\cdot, \cdot)$  is the generalized Bregman divergence notation introduced in Section 4.

To prove the desired rate for estimation, we make a discussion in two cases.

Case (i)  $\min\{\sigma^2 P_\gamma(o^*)/\kappa_\gamma, M_\gamma^2 o^*\} \leq \min\{\sigma^2 P_\beta(s^*)/\kappa_\beta, M_\beta^2 s^*\}$ . Consider a signal subclass

$$\mathcal{B}^1 = \{\bar{\beta} = [\beta^T, \mathbf{0}^T]^T : \beta_j \in \{0, \tau R\}, \|\beta\|_0 \leq s^*\},$$

where

$$R = \sigma(\log(ep/s^*))^{1/2}/\kappa_\beta^{1/2} \wedge M_\beta$$

and  $1 > \tau > 0$  is a small constant to be chosen later. Clearly,  $\mathcal{B}^1 \in \mathcal{B}(s^*, o^*, M_\beta, M_\gamma)$ . By Stirling's approximation,  $\log |\mathcal{B}^1| \geq \log \binom{p}{s^*} \geq s^* \log(p/s^*) \geq c s^* \log(ep/s^*)$  for some universal constant  $c$ .

Let  $\rho(\bar{\beta}_1, \bar{\beta}_2) = \|\bar{\beta}_1 - \bar{\beta}_2\|_0$ , the Hamming distance between  $\bar{\beta}_1$  and  $\bar{\beta}_2$ . By Lemma A.3 in Rigollet and Tsybakov (2011), there exists a subset  $\mathcal{B}^{10} \subset \mathcal{B}^1$  such that  $\bar{\beta}_0 = \mathbf{0} \in \mathcal{B}^{10}$  and

$$\log |\mathcal{B}^{10}| \geq c_1 s^* \log(ep/s^*), \rho(\bar{\beta}_1, \bar{\beta}_2) \geq c_2 s^*, \forall \bar{\beta}_1, \bar{\beta}_2 \in \mathcal{B}^{10}, \bar{\beta}_1 \neq \bar{\beta}_2$$

for some universal constants  $c_1, c_2 > 0$ . Then

$$\|\bar{\beta}_1 - \bar{\beta}_2\|_2^2 = \tau^2 R^2 \rho(\bar{\beta}_1, \bar{\beta}_2) \geq c_2 \tau^2 R^2 s^* \quad (\text{A.60})$$

for any  $\bar{\beta}_1, \bar{\beta}_2 \in \mathcal{B}^{10}$ ,  $\bar{\beta}_1 \neq \bar{\beta}_2$ .

By Lemma 10 and the regularity condition, for any  $\bar{\beta} \in \mathcal{B}^{10}$ , we have

$$\mathcal{K}(p_{\bar{\beta}}, p_{\bar{\beta}_0}) = \Delta_l(\bar{\mathbf{X}}\bar{\beta}_0, \bar{\mathbf{X}}\bar{\beta}) \leq n\tau^2\kappa_\beta R^2 s^*/\sigma^2.$$

Therefore,

$$\frac{1}{|\mathcal{B}^{10}| - 1} \sum_{\bar{\beta} \in \mathcal{B}^{10} \setminus \{\bar{\beta}_0\}} \mathcal{K}(p_{\bar{\beta}}, p_{\bar{\beta}_0}) \leq \tau^2 \kappa_\beta s^* \log(ep/s^*). \quad (\text{A.61})$$

Combining (A.60) and (A.61) and choosing a sufficiently small value for  $\tau$ , we can apply Theorem 2.7 of [Tsybakov \(2008\)](#) to get the desired lower bound.

*Case (ii)*  $\min\{\sigma^2 P_\gamma(o^*)/\kappa_\gamma, M_\gamma^2 o^*\} \geq \min\{\sigma^2 P_\beta(s^*)/\kappa_\beta, M_\beta^2 s^*\}$ . Define a signal sub-class

$$\mathcal{B}^2 = \{\bar{\beta} = [\mathbf{0}^T, \gamma^T]^T : \gamma_i \in \{0, \tau R\}, \|\gamma\|_0 \leq o^*\},$$

where  $R = \sigma(\log(en/o^*))^{1/2}/\tau^{1/2} \wedge M_\gamma$  and  $1 > \tau > 0$  is a small constant. The afterward treatment is similar to (i). The details are omitted.

The proof for the lower bound of  $\|\bar{\mathbf{X}}\bar{\beta}^* - \hat{\bar{\mathbf{X}}}\hat{\beta}\|_2^2$  follows similar lines.  $\square$

## A.8 Proof of Theorem 7

Let  $J(\gamma) = |\mathcal{J}(\gamma)| = \|\gamma\|_0$  and  $\mathcal{J}(\gamma)$  is the support of  $\gamma$ , i.e.,  $\mathcal{J}(\gamma) = \{i : \gamma_i \neq 0\}$ . The optimality of  $(\hat{\beta}, \hat{\gamma})$  implies that

$$\Delta_l(\mathbf{X}\hat{\beta} + \hat{\gamma}, \mathbf{X}\beta^* + \gamma^*) \leq AP_o(\beta^*, \gamma^*) - AP_o(\hat{\beta}, \hat{\gamma}) + \langle \epsilon, \bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\bar{\beta}^* \rangle. \quad (\text{A.62})$$

The stochastic term  $\langle \epsilon, \bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\bar{\beta}^* \rangle$  can be decomposed and bounded in a similar way as in the proof of Theorem 4. The difference is to use the union bound to show the  $\mathbb{E} \sup_{s \leq p, o \leq n} R_{s,o}^2 \leq C$ . Take the first term  $\langle \epsilon, I \rangle$  from the decomposition as an example:

$$\begin{aligned} & \langle \epsilon, \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\Delta \rangle - \frac{1}{a} \|\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\Delta\|_2^2 - bLP_o(J(\beta^*), J(\gamma^*)) \\ & \leq \frac{1}{a'} \|\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\Delta\|_2^2 + \frac{a'}{4} \sup_{s \leq p, o \leq n} \sup_{\Delta \in \Gamma_{s,o}} [\langle \epsilon, \Delta \rangle - 2\{(b/a)L P_o(s, o)\}^{1/2}]_+^2 \\ & \equiv \frac{1}{a'} \|\mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\beta}^*)}} \bar{\mathbf{X}}\Delta\|_2^2 + \frac{a'}{4} \sup_{s \leq p, o \leq n} R_{s,o}^2. \end{aligned}$$

When  $s = o = 0$ ,  $R_{0,0} = 0$ . When  $s \geq 1$  and  $o \geq 1$ , for any  $t \geq 0$ , if  $4b/a$  is a constant greater than 1,

$$\begin{aligned}
& \mathbb{P}\left(\sup_{1 \leq o \leq n, 1 \leq s \leq p} R_{s,o} > t\sigma\right) \\
& \leq \sum_{s=1}^p \sum_{o=1}^n \mathbb{P}\left(\sup_{\Delta \in \Gamma_{s,o}} \langle \epsilon, \Delta \rangle - \sqrt{LP(s,o)} > t\sigma + 2\sigma \sqrt{\frac{b}{a}LP(s,o)} - \sqrt{LP(s,o)}\right) \\
& \leq C \exp(-ct^2) \sum_{s=1}^p \sum_{o=1}^n \exp[-c(2\sqrt{b/a} - 1)^2 P(s,o)] \\
& \leq C \exp(-ct^2) \exp(-c(\log n + \log p)) \sum_{s=1}^p \sum_{o=1}^n \exp(-c(s+o)) \\
& \leq C \exp(-ct^2)(np)^{-c},
\end{aligned} \tag{A.63}$$

where the last inequality is due to the sum of geometric series. Similarly, when  $s = 0$ ,  $\mathbb{P}(\sup_{0 \leq o \leq n} R_{0,o} > t\sigma) \leq C \exp(-ct^2)n^{-c}$ , and when  $o = 0$ ,  $\mathbb{P}(\sup_{0 \leq s \leq p} R_{s,0} > t\sigma) \leq C \exp(-ct^2)p^{-c}$  for any  $0 \leq s \leq p$ . Hence  $\mathbb{E} \sup_{s,o} R_{s,o}^2 \leq C$ . To summarize, we obtain that for any constants  $a, b, a', b' > 0$  satisfying  $4b > a$ ,

$$\mathbb{E}\langle \epsilon, \bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\bar{\beta} \rangle \leq \left(\frac{1}{a} + \frac{1}{a'}\right) \|\mathbf{X}\hat{\beta} - \mathbf{X}\beta + \hat{\gamma} - \gamma\|_2^2 + 2bL[P(\hat{\beta}, \hat{\gamma}) + P(\beta, \gamma)] + a'C. \tag{A.64}$$

Using the regularity condition and choosing constants  $a, a', b, b'$  and  $A$  sufficiently large such that  $1/a + 1/a' = \delta/4$ ,  $4b > a$  and  $A = A_0 + 2bL$ , we obtain the error bound.

## A.9 Proof of Theorem 8

We prove a general theorem in possibly high dimensions where  $\beta$  is also desired to be sparse.

**Theorem 8'.** *Let  $\mathbf{y} = \mathbf{X}\beta^* + \gamma^* + \epsilon$ , where  $\epsilon_i$  are independent sub-Gaussian( $0, \sigma^2$ ) and  $\mathbb{E}\epsilon_i^2 = \sigma_i^2 = c_i\sigma^2$  with  $c_i$  some positive constants and  $\sigma^2$  unknown. Let  $l_0(\mathbf{X}\beta + \gamma; \mathbf{y}) = \|\mathbf{y} - \mathbf{X}\beta - \gamma\|_2^2/2$ . Define  $P(\beta, \gamma) = J(\beta) \log(ep/J(\beta)) + J(\gamma) \log(en/J(\gamma))$ . Assume the true model is parsimonious in the sense that  $P(\beta^*, \gamma^*) \leq n/A_0$  for some constant  $A_0 > 0$ . Let  $\delta(\beta, \gamma) = AP(\beta, \gamma)/n$  where  $A$  is a positive constant satisfying  $A < A_0$ , and so  $\delta(\beta^*, \gamma^*) < 1$ . Then for sufficiently large values of  $A_0$  and  $A$ , any  $(\hat{\beta}, \hat{\gamma})$  that minimizes  $\log l_0(\mathbf{X}\beta + \gamma; \mathbf{y}) + \delta(\beta, \gamma)$  s.t.  $\delta(\beta, \gamma) < 1$  satisfies*

$$\mathbf{D}_2(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) \lesssim \sigma^2 P(\beta^*, \gamma^*) \tag{A.65}$$

with probability at least  $1 - C'n^{-c'} - Cn^{-c(1 \wedge o^*)}p^{-c(1 \wedge s^*)}$  for some constants  $c, c', C, C' > 0$ .

*Proof.* Let  $h(\beta, \gamma; A) = 1/(n - AP(\beta, \gamma))$ . It follows from  $1/(1 - \delta) \geq \exp(\delta)$  for any  $0 \leq \delta < 1$  and  $\exp(\delta) \geq 1/(1 - \delta/2)$  for any  $0 \leq \delta < 2$  that  $n\|\mathbf{y} - \mathbf{X}\hat{\beta} - \hat{\gamma}\|_2^2 h(\hat{\beta}, \hat{\gamma}; A/2) \leq \|\mathbf{y} - \mathbf{X}\hat{\beta} - \hat{\gamma}\|_2^2 \exp[\delta(\hat{\beta}, \hat{\gamma})] \leq \|\mathbf{y} - \mathbf{X}\beta^* - \gamma^*\|_2^2 \exp[\delta(\beta^*, \gamma^*)] \leq \|\mathbf{y} - \mathbf{X}\beta^* - \gamma^*\|_2^2 h(\beta^*, \gamma^*; A)n$ .

Since  $h(\hat{\beta}, \hat{\gamma}; A/2) > 0$ , we have

$$\|\mathbf{y} - \mathbf{X}\hat{\beta} - \hat{\gamma}\|_2^2 \leq \|\mathbf{y} - \mathbf{X}\beta^* - \gamma^*\|_2^2 h(\beta^*, \gamma^*; A) / h(\hat{\beta}, \hat{\gamma}; A/2).$$

It follows that

$$\begin{aligned} & \|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^* + \hat{\gamma} - \gamma^*\|_2^2 \\ & \leq \|\epsilon\|_2^2 [h(\beta^*, \gamma^*; A) / h(\hat{\beta}, \hat{\gamma}; A/2) - 1] + 2\langle \epsilon, \mathbf{X}\hat{\beta} - \mathbf{X}\beta^* + \hat{\gamma} - \gamma^* \rangle \\ & \leq \frac{A\|\epsilon\|_2^2}{n\sigma^2 - A\sigma^2 P(\beta^*, \gamma^*)} \sigma^2 P(\beta^*, \gamma^*) - \frac{A\|\epsilon\|_2^2}{2n\sigma^2} \sigma^2 P(\hat{\beta}, \hat{\gamma}) + 2\langle \epsilon, \mathbf{X}\hat{\beta} - \mathbf{X}\beta^* + \hat{\gamma} - \gamma^* \rangle. \end{aligned}$$

The proof of Theorem 7 gives a high-probability bound for the stochastic term: for any constant  $a, b, a' > 0$  satisfying  $4b > a$ ,

$$\begin{aligned} & 2\langle \epsilon, \mathbf{X}\hat{\beta} - \mathbf{X}\beta^* + \hat{\gamma} - \gamma^* \rangle \\ & \leq 2\left(\frac{1}{a} + \frac{1}{a'}\right) \|\mathbf{X}\hat{\beta} - \mathbf{X}\beta^* + \hat{\gamma} - \gamma^*\|_2^2 + 4bL\sigma^2 [P(\hat{\beta}, \hat{\gamma}) + P(\beta^*, \gamma^*)] \end{aligned} \tag{A.66}$$

with probability at least  $1 - Cn^{-c(1 \wedge o^*)}p^{-c(1 \wedge s^*)}$  for some  $c, C > 0$ .

Assume  $c_0\sigma^2 \leq \mathbb{E}\epsilon_i^2 \leq C_0\sigma^2$  with  $c_0, C_0$  positive constants and let  $\varepsilon$  and  $\varepsilon'$  be two constants satisfying  $0 < \varepsilon < 1, \varepsilon' > 0$ . On  $\mathcal{A} = \{(c_0 - \varepsilon)n\sigma^2 \leq \|\epsilon\|_2^2 \leq (C_0 + \varepsilon')n\sigma^2\}$ , we have

$$\begin{aligned} & \frac{A\|\epsilon\|_2^2}{n\sigma^2 - A\sigma^2 P(\beta^*, \gamma^*)} \sigma^2 P(\beta^*, \gamma^*) - \frac{A\|\epsilon\|_2^2}{2n\sigma^2} \sigma^2 P(\hat{\beta}, \hat{\gamma}) \\ & \leq \frac{(C_0 + \varepsilon')AA_0}{A_0 - A} \sigma^2 P(\beta^*, \gamma^*) - \frac{(c_0 - \varepsilon)A}{2} \sigma^2 P(\hat{\beta}, \hat{\gamma}). \end{aligned}$$

With  $A_0$  large enough, we can choose  $a, a', b, A$  such that  $1/a + 1/a' < 1/2$ ,  $4b > a$  and  $8bL/c_0(1 - \varepsilon) \leq A$ . From the Hanson-Wright inequality (Rudelson and Vershynin, 2013)  $\mathcal{A}$  occurs with probability at most  $c'_2 \exp(-c_2 n)$ , where  $c_2, c'_2$  are dependent on constants  $\varepsilon, \varepsilon'$ . The conclusion results.  $\square$

**Remark A.3.** For robust regression in low dimensions, applying Theorem 8' to the reduced model  $\mathbf{U}_\perp^T \mathbf{y} = \mathbf{U}_\perp^T \gamma^* + \epsilon'$  gives Theorem 8, where  $\mathbf{U}_\perp \in \mathbb{R}^{n \times (n-r(\mathbf{X}))}$  is the orthogonal complement of  $\mathbf{U}$  that is from the SVD:  $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ , and  $\epsilon' = \mathbf{U}_\perp^T \epsilon$ . The proof follows the same lines; in particular, when applying the Hanson-Wright inequality, note that  $\mathbb{E}[\|\epsilon'\|_2^2] = \text{Tr}\{\text{diag}\{\sigma_i^2\}\mathbf{U}_\perp \mathbf{U}_\perp^T\} \in [c_0(n-r(\mathbf{X}))\sigma^2, C_0(n-r(\mathbf{X}))\sigma^2]$  and  $\|\mathbf{U}_\perp \mathbf{U}_\perp^T\|_F^2 = n-r(\mathbf{X})$  and  $\|\mathbf{U}_\perp \mathbf{U}_\perp^T\|_2 = 1$ .

## A.10 Error bounds of optimal solutions

This part points out that the error rate remains the same if  $(\hat{\beta}, \hat{\gamma})$  is globally optimal, but the regularity condition gets relaxed. For simplicity, we drop the  $\ell_2$  penalty terms.

**Theorem A.2.** Let  $(\hat{\beta}, \hat{\gamma})$  be any globally optimal solution of  $\min_{\|\beta\|_0 \leq q_\beta, \|\gamma\|_0 \leq q_\gamma} l(\mathbf{X}\beta + \hat{\gamma}; \mathbf{y})$  with  $\|\hat{\gamma}\|_0 = q_\gamma$ ,  $\|\hat{\beta}\|_0 = q_\beta$ ,  $q_\gamma = \vartheta o^*$ ,  $q_\beta = \vartheta s^*$  and  $\vartheta \geq 1$ . Assume that there exists some  $\delta > 0$  such that

$$(\Delta_l - \delta \mathbf{D}_2)(\bar{\mathbf{X}}\bar{\beta}, \bar{\mathbf{X}}\bar{\beta}') \geq 0 \quad (\text{A.67})$$

holds for any  $\|\beta\|_0 \leq \vartheta s^*$ ,  $\|\beta'\|_0 \leq s^*$ ,  $\|\gamma\|_0 \leq \vartheta o^*$ ,  $\|\gamma'\|_0 \leq o^*$ . Then the following error bound holds

$$\mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*)] \lesssim \frac{\vartheta}{\delta^2} \sigma^2 (s^* \log \frac{ep}{s^*} + o^* \log \frac{en}{o^*}) + \frac{\sigma^2}{\delta^2}. \quad (\text{A.68})$$

(A.67), apart from replacing  $2\bar{\Delta}_l$  by  $\Delta_l$ , removes the term on the right-hand side of (26). It is easy to see that if  $l$  is  $\mu$ -strongly convex, then (A.67) is true with  $\delta = \mu$ . In particular, when  $l_0(\eta; y) = (\eta - y)^2/2$ , the condition holds trivially ( $\delta = 1$ ).

Under a slightly stronger regularity condition than that used in Theorem 4, we get an estimation error bound.

**Theorem A.3.** Let  $(\hat{\beta}, \hat{\gamma})$  be an  $A$ -estimator satisfying  $\|\hat{\gamma}\|_0 = q_\gamma$ ,  $\|\hat{\beta}\|_0 = q_\beta$  with  $q_\gamma = \vartheta o^*$ ,  $q_\beta = \vartheta s^*$  and  $\vartheta \geq 1$ . Assume  $\beta^* \neq \mathbf{0}$ . Then, with (26) replaced by  $(1 - \epsilon)(2\bar{\Delta}_l - \delta \mathbf{D}_2)(\bar{\mathbf{X}}\bar{\beta}, \bar{\mathbf{X}}\bar{\beta}') \geq \frac{1}{\sqrt{\vartheta}} \mathbf{D}_2(\bar{\beta}, \bar{\beta}')$   $\forall \bar{\beta}, \bar{\beta}'$ :  $\|\beta\|_0 \leq \vartheta s^*$ ,  $\|\beta'\|_0 \leq s^*$ ,  $\|\gamma\|_0 \leq \vartheta o^*$ ,  $\|\gamma'\|_0 \leq o^*$  for some  $\epsilon, \delta > 0$ , with probability at least  $1 - Cp^{-c}$ ,  $\rho \|\hat{\beta} - \beta^*\|_2^2 + \|\hat{\gamma} - \gamma^*\|_2^2 \lesssim \frac{\sigma^2}{\delta \epsilon} \{ \vartheta^{3/2} [s^* \log(ep/s^*) + o^* \log(en/o^*)] \}$ . In particular, we have the estimation error bound

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{\sigma^2 \vartheta^{3/2} [s^* \log(ep/s^*) + o^* \log(en/o^*)]}{\rho} \quad (\text{A.69})$$

with probability  $1 - Cp^{-c}$  under (26) with  $\bar{\mathbf{X}}_\rho$  redefined as  $[\mathbf{X} / \sqrt{\rho(1 + \epsilon)} \quad \mathbf{I}]$ .

When  $\vartheta, \epsilon, \delta$  are constants and  $\rho \geq cn$ , the estimation error bound in (A.69) is of the order  $\sigma^2 \{s^* \log(ep/s^*) + o^* \log(en/o^*)\}/n$ .

*Proof.* To prove Theorem A.2, by definition,  $\hat{\beta} = [\hat{\beta}^T, \hat{\gamma}^T]^T$  satisfies  $l(\bar{\mathbf{X}}\hat{\beta}) \leq l(\bar{\mathbf{X}}\bar{\beta}^*)$  or equivalently,

$$\Delta_l(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) \leq \langle \epsilon, \bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\bar{\beta}^* \rangle. \quad (\text{A.70})$$

Treating the stochastic term in the same way as in the proof of Theorem 4, we obtain

$$\mathbb{E}\langle \epsilon, \bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\bar{\beta}^* \rangle \leq \left( \frac{1}{a} + \frac{1}{a'} \right) \|\bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\bar{\beta}^*\|_2^2 + 3bL\sigma^2 P(q_\beta, q_\gamma) + a'C\sigma^2, \quad (\text{A.71})$$

for any  $a, a', b > 0$  satisfying  $4b > a$ . The regularity condition implies

$$(\delta/2) \|\bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\bar{\beta}^*\|_2^2 \leq \Delta_l(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*). \quad (\text{A.72})$$

Combining the above three inequalities and choosing  $a, a', b$  such that  $a = a' = 8/\delta$  and  $b = 4/\delta$  gives the desired result.

To prove Theorem A.3, recall that from the proof of Theorem 4, we obtain

$$\begin{aligned} 2\bar{\Delta}_l(\mathbf{X}\hat{\beta} + \hat{\gamma}, \mathbf{X}\beta^* + \gamma^*) &\leq \frac{1}{2\sqrt{\vartheta}}(\|\hat{\gamma} - \gamma^*\|_2^2 + \rho\|\hat{\beta} - \beta^*\|_2^2) \\ &\quad + \left(\frac{1}{a} + \frac{1}{a'}\right)\|\bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\beta^*\|_2^2 + 3bL\vartheta\sigma^2P_o(s^*, o^*) \end{aligned}$$

with probability at least  $1 - Cn^{-c(1 \wedge o^*)}p^{-c}$ . Using the regularity condition, and setting  $a = a' = 4/\delta$  and  $b = 2/\delta$ , we get the estimation error bound (A.69) with high probability (details omitted).  $\square$

## A.11 General noise and stochastic breakdown

Given a random variable  $X$ , its Orlicz  $\psi$ -norm is defined by

$$\|X\|_\psi = \inf \left\{ M > 0 : \mathbb{E}\psi\left(\frac{|X|}{M}\right) \leq 1 \right\},$$

where  $\psi$  is a strictly increasing convex function on  $[0, +\infty)$  with  $\psi(0) = 0$ ; see, e.g., [van der Vaart and Wellner \(1996\)](#). Similar to the definition of sub-Gaussian random vectors (cf. Definition 1), we say that a random vector  $\epsilon$  has its  $\psi$ -norm bounded above by  $\sigma$  if

$$\|\langle \epsilon, \alpha \rangle\|_\psi \leq \sigma\|\alpha\|_2$$

for any vector  $\alpha$ . The components of  $\epsilon$  are not required to be independent.

Some commonly used  $\psi$ -norms in statistics are the  $L_p$  norms,  $\psi(x) = x^p$  with  $p \geq 1$ , which means  $|X|$  possesses a finite  $p$ -th moment, and  $\psi_p$  norms,  $\psi(x) = \exp(x^p) - 1$ , covering sub-Gaussian ( $p = 2$ ) and sub-exponential ( $p = 1$ ) random variables. By Markov's inequality, that  $X$  has a finite  $\psi$ -norm  $\sigma$  implies a tail probability bound

$$\mathbb{P}(|X| > t) \leq \frac{2}{\psi(t/\sigma) + 1}, \quad \forall t > 0$$

which encompasses diverse heavy/light tail decays.

In the following theorem, instead of restricting our attention to a sub-Gaussian type effective noise  $\epsilon$  (cf. (22)), which is sensible in many applications especially when a Lipschitz continuous loss is in use, we make a more general assumption that  $\epsilon$  has a finite Orlicz  $\psi$ -norm bounded above by  $\sigma$  and  $\psi(x) \geq cx^2$  (i.e., the second moment exists), together with the regularity condition  $\limsup_{x,y \rightarrow \infty} \psi(x)\psi(y)/\psi(cxy) < \infty$  ([van der Vaart and Wellner, 1996](#)) for some constant  $c > 0$  that is satisfied by say the  $L_p$  norms and  $\psi_p$  norms. (Note that the  $\psi$  here is not the same psi function to define an M-estimator.)

**Theorem A.4.** *Let  $\hat{\beta}$  be a solution to  $\min_{\bar{\beta}: \|\beta\|_0 \leq q_\beta, \|\gamma\|_0 \leq q_\gamma} l(\bar{\mathbf{X}}\bar{\beta}; \mathbf{y})$  with  $0 \leq q_\gamma \leq n, 0 \leq q_\beta \leq p$ . Assume that there exists some  $\delta > 0$  such that  $(\Delta_l - \delta\mathbf{D}_2)(\bar{\mathbf{X}}\bar{\beta}, \bar{\mathbf{X}}\bar{\beta}') \geq 0$  for*

sparse  $\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{\beta}}'$  satisfying  $\|\boldsymbol{\beta}\|_0 \leq q_\beta, \|\boldsymbol{\beta}'\|_0 \leq q_\beta, \|\boldsymbol{\gamma}\|_0 \leq q_\gamma, \|\boldsymbol{\gamma}'\|_0 \leq q_\gamma$ . Then the following oracle inequality holds for any  $\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 \leq q_\beta, \|\boldsymbol{\gamma}\|_0 \leq q_\gamma$

$$\mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\hat{\bar{\boldsymbol{\beta}}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*)] \lesssim \frac{1}{\delta^2} \left\{ \delta \mathbb{E}\Delta_l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*) + \sigma^2 \left[ \psi^{-1} \left( \exp \left\{ c(q_\gamma \log \frac{en}{q_\gamma} + q_\beta \log \frac{ep}{q_\beta}) \right\} \right) \right]^2 \right\}, \quad (\text{A.73})$$

where  $c$  is a positive constant. In particular, in the sub-Gaussian case with  $\psi(x) = \exp(x^2) - 1$ , the RHS becomes  $\frac{1}{\delta^2} \{ \delta \mathbb{E}\Delta_l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*) + \sigma^2 q_\gamma \log(en/q_\gamma) + \sigma^2 q_\beta \log(ep/q_\beta) + \sigma^2 \}$ .

*Proof.* By definition, for any  $\bar{\boldsymbol{\beta}} : \|\boldsymbol{\beta}\|_0 \leq q_\beta, \|\boldsymbol{\gamma}\|_0 \leq q_\gamma, l(\bar{\mathbf{X}}\hat{\bar{\boldsymbol{\beta}}}) \leq l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}})$  and so

$$\Delta_l(\bar{\mathbf{X}}\hat{\bar{\boldsymbol{\beta}}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*) \leq \Delta_l(\bar{\mathbf{X}}\bar{\boldsymbol{\beta}}, \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}^*) + \langle \boldsymbol{\epsilon}, \bar{\mathbf{X}}(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})/\|\bar{\mathbf{X}}(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})\|_2 \rangle \|\bar{\mathbf{X}}(\hat{\bar{\boldsymbol{\beta}}} - \bar{\boldsymbol{\beta}})\|_2. \quad (\text{A.74})$$

Define  $\Gamma(q_\beta, q_\gamma) = \{\boldsymbol{\theta} \in \mathbb{R}^n : \|\boldsymbol{\theta}\|_2 \leq 1, \boldsymbol{\theta} = \bar{\mathbf{X}}\bar{\boldsymbol{\beta}} + \boldsymbol{\gamma} \text{ for some } \|\boldsymbol{\beta}\|_0 \leq q_\beta, \|\boldsymbol{\gamma}\|_0 \leq q_\gamma\}$  or  $\Gamma$  for short when there is no ambiguity. For notational convenience, we bound  $\sup_{\boldsymbol{\theta} \in \Gamma(q_\beta, q_\gamma)} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle$ . The difficulty lies in the possible *divergence* of the entropy integral for a general  $\psi$ . To conquer this, we apply discretization and make use of the finiteness of the number of range spaces defined by  $\Gamma$ .

Recall that  $\boldsymbol{\theta} \in \Gamma(q_\beta, q_\gamma)$  means  $\boldsymbol{\theta} = \bar{\mathbf{X}}\bar{\boldsymbol{\beta}} \in \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\boldsymbol{\beta}})}}$ . Obviously, if  $\bar{\boldsymbol{\beta}}$  has degenerate zeros:  $\|\boldsymbol{\beta}\|_0 < q_\beta$  and/or  $\|\boldsymbol{\gamma}\|_0 < q_\gamma$ , the range of  $\bar{\mathbf{X}}_{\mathcal{J}(\bar{\boldsymbol{\beta}})}$  is always included in a column subspace of  $\bar{\mathbf{X}}$  indexed by exactly  $q_\beta$  columns in  $\bar{\mathbf{X}}$  and  $q_\gamma$  columns in  $\mathbf{I}_{n \times n}$ , and so we just need to focus on the  $\bar{\boldsymbol{\beta}}$ 's with  $J(\boldsymbol{\beta}) = q_\beta$  and  $J(\boldsymbol{\gamma}) = q_\gamma$ . To use an  $\varepsilon$ -net  $\dot{\Gamma}$  to discretize  $\Gamma$  such that for any  $\boldsymbol{\theta} \in \Gamma$ , there exists  $\dot{\boldsymbol{\theta}} \in \dot{\Gamma} \subset \Gamma$  satisfying  $\|\boldsymbol{\theta} - \dot{\boldsymbol{\theta}}\|_2 \leq \varepsilon$ , we include all nondegenerate subspaces and apply a standard volume argument. Then for any  $0 < \varepsilon < 1$ , the covering number  $\mathcal{N}(\varepsilon, \Gamma, \|\cdot\|_2)$  is bounded by

$$\mathcal{N}(\varepsilon, \Gamma, \|\cdot\|_2) \leq \binom{p}{q_\beta} \binom{n}{q_\gamma} \left(\frac{3}{\varepsilon}\right)^{r(\bar{\mathbf{X}}) \wedge (q_\beta + q_\gamma)}.$$

Furthermore, the construction shows that for  $\boldsymbol{\theta} \in \Gamma$  and  $\boldsymbol{\theta} = \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}$  with no degeneracy in  $\bar{\boldsymbol{\beta}}$ ,  $\dot{\boldsymbol{\theta}}$  is also in the subspace determined by  $\mathcal{P} := \mathcal{P}_{\bar{\mathbf{X}}_{\mathcal{J}(\bar{\boldsymbol{\beta}})}}$ . Since  $\mathcal{P}\boldsymbol{\alpha} \in \Gamma$  for any  $\|\boldsymbol{\alpha}\|_2 \leq 1$  we have

$$\begin{aligned} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle &= \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} - \dot{\boldsymbol{\theta}} \rangle + \max_{\dot{\boldsymbol{\theta}} \in \dot{\Gamma}} \langle \boldsymbol{\epsilon}, \dot{\boldsymbol{\theta}} \rangle \\ &= \varepsilon \langle \boldsymbol{\epsilon}, \mathcal{P}(\boldsymbol{\theta} - \dot{\boldsymbol{\theta}})/\varepsilon \rangle + \max_{\dot{\boldsymbol{\theta}} \in \dot{\Gamma}} \langle \boldsymbol{\epsilon}, \dot{\boldsymbol{\theta}} \rangle \\ &\leq \varepsilon \sup_{\boldsymbol{\theta} \in \Gamma} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle + \max_{\dot{\boldsymbol{\theta}} \in \dot{\Gamma}} \langle \boldsymbol{\epsilon}, \dot{\boldsymbol{\theta}} \rangle, \end{aligned}$$

and so  $(1 - \varepsilon) \sup_{\boldsymbol{\theta} \in \Gamma} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle \leq \max_{\dot{\Gamma}} \langle \boldsymbol{\epsilon}, \dot{\boldsymbol{\theta}} \rangle$ . We can take say  $\varepsilon = 0.5$  to turn to a finite-class problem instead of using the entropy integral.

Indeed, by an extension of Massart's finite class lemma (cf. Lemma 2.2.2 in [van der Vaart and Wellner \(1996\)](#)) and Stirling's formula, we obtain

$$\left\| \sup_{\boldsymbol{\theta} \in \Gamma(2q_\beta, 2q_\gamma)} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle \right\|_\psi \leq C \psi^{-1} \left( \exp \left\{ c(q_\gamma \log \frac{en}{q_\gamma} + q_\beta \log \frac{ep}{q_\beta}) \right\} \right) \sigma$$

where  $C, c$  are constants depending on  $\psi$  only.

Based on (A.74) and the regularity condition,

$$\frac{\delta}{4} \|\bar{\mathbf{X}}(\hat{\bar{\beta}} - \bar{\beta}^*)\|_2^2 \leq \Delta_l(\bar{\mathbf{X}}\hat{\bar{\beta}}, \bar{\mathbf{X}}\bar{\beta}^*) - \frac{\delta}{4} \|\bar{\mathbf{X}}(\hat{\bar{\beta}} - \bar{\beta}^*)\|_2^2 \leq \Delta_l(\bar{\mathbf{X}}\bar{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) + \frac{1}{\delta} \left( \sup_{\boldsymbol{\theta} \in \Gamma(2q_\beta, 2q_\gamma)} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle \right)^2.$$

By assumption,  $\{\mathbb{E}[(\sup_{\boldsymbol{\theta} \in \Gamma(2q_\beta, 2q_\gamma)} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle)^2]\}^{1/2} \lesssim \|\sup_{\boldsymbol{\theta} \in \Gamma(2q_\beta, 2q_\gamma)} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle\|_\psi$ . Hence

$$\mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\hat{\bar{\beta}}, \bar{\mathbf{X}}\bar{\beta}^*)] \lesssim \frac{1}{\delta} \mathbb{E}\Delta_l(\bar{\mathbf{X}}\bar{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) + \frac{1}{\delta^2} \{[\sigma^2 \psi^{-1}(\exp\{cq_\gamma \log \frac{en}{q_\gamma} + cq_\beta \log \frac{ep}{q_\beta}\})]^2\}.$$

The proof is complete.  $\square$

**Remark A.4.** *The oracle inequality implies that even if  $\bar{\beta}^*$  has  $\|\beta^*\|_0 \gg q_\beta$ ,  $\|\gamma^*\|_0 \gg q_\gamma$ , as long as  $\bar{\beta}^*$  can be well approximated by some  $\bar{\beta}$ :  $\|\beta\|_0 \leq q_\beta$ ,  $\|\gamma\|_0 \leq q_\gamma$  in the sense that the bias  $\mathbb{E}\Delta_l(\bar{\mathbf{X}}\bar{\beta}, \bar{\mathbf{X}}\bar{\beta}^*)$  is controlled by  $\sigma^2 [\psi^{-1}(\exp\{c(q_\gamma \log \frac{en}{q_\gamma} + q_\beta \log \frac{ep}{q_\beta})\})]^2 / \delta$  up to a multiplicative constant, the prediction risk bound is of the order  $\sigma^2 / \delta^2$  times*

$$\left[ \psi^{-1} \left( \exp \left\{ c(q_\gamma \log \frac{en}{q_\gamma} + q_\beta \log \frac{ep}{q_\beta}) \right\} \right) \right]^2. \quad (\text{A.75})$$

*This applicability to approximately sparse signals is quite useful in reality.*

**Remark A.5.** *Taking a specific  $\bar{\beta} = \bar{\beta}^*$  in (A.73) and assuming the associated regularity condition holds, the error bound depends on  $\gamma^*$  through its support size only, thereby finite regardless of its magnitude or outlyingness. This gives a risk-based breakdown point result that accounts for the randomness of the estimators.*

Concretely, to define a general stochastic breakdown point, we fix the true systematic component  $\mathbf{X}\beta$ , but can freely alter the response  $\mathbf{y}$  in

$$\mathcal{Y}(o) = \{\mathbf{y} : \|\gamma\|_0 \leq o, \|\boldsymbol{\epsilon}\|_\psi < +\infty\} \quad (\text{A.76})$$

with  $o \in \mathbb{N} \cup \{0\}$ , where  $\boldsymbol{\epsilon} = -\nabla l(\mathbf{X}\beta + \gamma; \mathbf{y})$ . Given any estimator  $(\hat{\beta}, \hat{\gamma})$  (that implicitly depends on the data  $\mathbf{X}, \mathbf{y}$ ), its finite-sample breakdown point  $\epsilon^*$  can be defined by

$$\frac{1}{n} \min\{o : \sup_{\mathbf{y} \in \mathcal{Y}(o)} \mathbb{E}[\mathbf{D}_2(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta})] = +\infty\}. \quad (\text{A.77})$$

Then, in the setup of Theorem A.4, the estimator given  $q_\beta, q_\gamma$  has the stochastic breakdown point

$$\epsilon^* \geq (q_\gamma + 1)/n. \quad (\text{A.78})$$

For example, for  $\hat{\beta} = \arg \min_{\bar{\beta}: \|\beta\|_0 \leq q_\beta, \|\gamma\|_0 \leq q_\gamma} l(\bar{\mathbf{X}}\bar{\beta}; \mathbf{y})$  with a quadratic  $l(\boldsymbol{\eta}; \mathbf{y}) = \|\mathbf{y} - \boldsymbol{\eta}\|_2^2 / 2$ , the contamination model associated with  $\mathcal{Y}(o)$  is  $\mathbf{y} = \mathbf{X}\beta + \gamma + \boldsymbol{\epsilon}$  subject to  $\|\gamma\|_0 \leq o$ , where the nonzero components of  $\gamma$  can be arbitrarily large, but even in high dimensions the breakdown point of  $\hat{\beta}$  is no lower than  $(q_\gamma + 1)/n$ . The direct link between the  $\ell_0$ -constraint and breakdown point facilitates parameter tuning.

Statistically speaking, obtaining a precise error rate  $[\psi^{-1}(\exp\{c(q_\gamma \log \frac{en}{q_\gamma} + q_\beta \log \frac{ep}{q_\beta})\})]^2$  in Theorem A.4 is much more informative than simply knowing that the risk is finite for the purpose of breakdown studies. However, it is an interesting question to determine what relaxed conditions  $\epsilon$  should satisfy to guarantee (A.78) for a general *extended* real-valued function  $l : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ . Toward this, we change the  $\mathbf{D}_2$  in (A.77) to the generalized Bregman  $\Delta_l$ :  $\epsilon^* = \min\{o : \sup_{\mathbf{y} \in \mathcal{Y}(o)} \mathbb{E}[\Delta_l(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta})] = +\infty\}/n$ , and make a no-model-ambiguity assumption:  $l$  is differentiable at  $\bar{\mathbf{X}}\bar{\beta}^* \in D = \{\boldsymbol{\eta} : l(\boldsymbol{\eta}) < +\infty\}$  with the gradient  $\nabla l(\bar{\mathbf{X}}\bar{\beta}^*) = -\epsilon$ ,  $\bar{\mathbf{X}}\bar{\beta}^*$  is a finite optimal solution to the *Fenchel conjugate* when  $\boldsymbol{\zeta} = -\epsilon$ :

$$l^*(\boldsymbol{\zeta}) = \sup_{\boldsymbol{\eta} \in \mathbb{R}^n} \langle \boldsymbol{\zeta}, \boldsymbol{\eta} \rangle - l(\boldsymbol{\eta}), \quad (\text{A.79})$$

and the extended real-valued convex function  $l^*$  is differentiable at  $-\epsilon$ . The assumption simply means that  $(\bar{\mathbf{X}}\bar{\beta}^*, -\epsilon)$  makes a conjugate pair. Note that  $l$  need not be overall strictly convex, especially when  $D$  is compact, according to Danskin's min-max theorem (Bertsekas, 1999). Then we have the following conclusion.

**Theorem A.5.** *Under  $\mathbb{E}[\Delta_{l^*}(\epsilon, -\epsilon)] < M < +\infty$  for all  $\boldsymbol{\gamma}^* : \|\boldsymbol{\gamma}^*\|_0 \leq q$ , the  $\Delta_l$ -risk based breakdown point for any finite solution  $\hat{\beta}$  to  $\min_{\bar{\beta} : \|\boldsymbol{\gamma}\|_0 \leq q} l(\bar{\mathbf{X}}\bar{\beta}; \mathbf{y})$  satisfies  $\epsilon^* \geq (q + 1)/n$ .*

In the special case that  $l$  is strongly convex on  $\mathbb{R}^n$  with respect to a certain norm, as long as the effective noise has the second moment, the theorem recovers (A.78) since  $l^*$  can be shown to be strongly smooth with respect to the Euclidean norm (Rockafellar, 1970) (albeit not delivering a concrete error rate as before). Two other notable features apart from the randomness:  $l$  is a general extended real-valued function (not restricted to be a function of  $\mathbf{y} - \boldsymbol{\eta}$  as in M-estimation), and its conjugate plays an important role in bounding the risk; also, we do not need to assume a positive norm penalty which is a key element to the proof of Proposition 1 in Alfons et al. (2013).

*Proof.* First, we show that

$$\Delta_l(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) \leq \Delta_{l^*}(\epsilon, -\epsilon). \quad (\text{A.80})$$

By definition,  $l(\bar{\mathbf{X}}\hat{\beta}) \leq l(\bar{\mathbf{X}}\bar{\beta}^*)$ , from which it follows that  $\Delta_l(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) \leq \langle \epsilon, \bar{\mathbf{X}}\hat{\beta} - \bar{\mathbf{X}}\bar{\beta}^* \rangle$ . Define  $\boldsymbol{\eta}^* = \bar{\mathbf{X}}\bar{\beta}^*$  and

$$h(\boldsymbol{\delta}) = \Delta_l(\boldsymbol{\delta} + \boldsymbol{\eta}^*, \boldsymbol{\eta}^*).$$

By assumption,  $l$  is a proper function and applying Fenchel-Young's inequality gives

$$\Delta_l(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) \leq \langle \epsilon, \boldsymbol{\delta} \rangle|_{\boldsymbol{\delta}=\bar{\mathbf{X}}\hat{\beta}-\bar{\mathbf{X}}\bar{\beta}^*} \leq \frac{1}{c} \Delta_l(\boldsymbol{\delta} + \bar{\mathbf{X}}\bar{\beta}^*, \bar{\mathbf{X}}\bar{\beta}^*)|_{\boldsymbol{\delta}=\bar{\mathbf{X}}\hat{\beta}-\bar{\mathbf{X}}\bar{\beta}^*} + \frac{1}{c} h^*(c\epsilon)$$

or  $(1 - 1/c)\Delta_l(\bar{\mathbf{X}}\hat{\beta}, \bar{\mathbf{X}}\bar{\beta}^*) \leq h^*(c\epsilon)/c$  for any  $c > 0$ .

On the other hand,

$$\begin{aligned}
h^*(\boldsymbol{\zeta}) &= \sup_{\boldsymbol{\delta}} \langle \boldsymbol{\zeta}, \boldsymbol{\delta} \rangle - l(\boldsymbol{\eta}^* + \boldsymbol{\delta}) + l(\boldsymbol{\eta}^*) + \langle \nabla l(\boldsymbol{\eta}^*), \boldsymbol{\delta} \rangle \\
&= \sup_{\boldsymbol{\delta}} \langle \boldsymbol{\zeta} + \nabla l(\boldsymbol{\eta}^*), \boldsymbol{\delta} \rangle - l(\boldsymbol{\eta}^* + \boldsymbol{\delta}) + l(\boldsymbol{\eta}^*) \\
&= \sup_{\boldsymbol{\delta}} \langle \boldsymbol{\zeta} + \nabla l(\boldsymbol{\eta}^*), \boldsymbol{\eta}^* + \boldsymbol{\delta} \rangle - l(\boldsymbol{\eta}^* + \boldsymbol{\delta}) + l(\boldsymbol{\eta}^*) - \langle \boldsymbol{\zeta} + \nabla l(\boldsymbol{\eta}^*), \boldsymbol{\eta}^* \rangle \\
&= l^*(\boldsymbol{\zeta} + \nabla l(\boldsymbol{\eta}^*)) + l(\boldsymbol{\eta}^*) - \langle \boldsymbol{\zeta} + \nabla l(\boldsymbol{\eta}^*), \boldsymbol{\eta}^* \rangle,
\end{aligned}$$

where  $l(\boldsymbol{\eta}^*), \nabla l(\boldsymbol{\eta}^*)$  are known to be finite. Using the optimality of  $\boldsymbol{\eta}^*$ , we have further

$$h^*(\boldsymbol{\zeta}) = l^*(\boldsymbol{\zeta} + \nabla l(\boldsymbol{\eta}^*)) - l^*(-\boldsymbol{\epsilon}) - \langle \boldsymbol{\zeta}, \boldsymbol{\eta}^* \rangle.$$

Moreover, from the assumption and definition (A.79), it is easy to show that  $\boldsymbol{\eta}^* \in \partial l^*(-\boldsymbol{\epsilon})$ , and so  $\nabla l^*(-\boldsymbol{\epsilon}) = \boldsymbol{\eta}^*$ , from which it follows that

$$h^*(\boldsymbol{\zeta}) = \Delta_{l^*}(\boldsymbol{\zeta} - \boldsymbol{\epsilon}, -\boldsymbol{\epsilon}). \quad (\text{A.81})$$

Taking  $c = 2$  gives (A.80). The rest of the lines follow Remark A.5.  $\square$

## A.12 More simulations

In this part, we present more experiment results by varying the correlation strength, covariance structure and sparsity level.

Concretely, in Example 1 with  $n = 1000, p = 10, o^* = 100$ , we changed the correlation strength  $\rho$  from 0.2 to 0.8. In Example 2 with  $n = 1000, p = 10, o^* = 120$ , three possible covariance structures for  $\boldsymbol{\Sigma}$  were included, Toeplitz structure  $[\rho^{|i-j|}]$ , equally correlated structure  $[\rho 1_{i \neq j}]$ , and blocked structure  $\text{diag}\{\boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1\}$ , with two blocks of equal size and the predictors within each block equally correlated with  $\rho = 0.5$ . In Example 3 and Example 4, where  $n = 200, p = 1000, o^* = 30$ , we considered different sparsity levels. The later steps of introducing high-leveraged outliers remain the same. The results are summarized in Table A.1, Table A.2, and Table A.3. The overall comparison evidently supports PIQ as an extremely resistant and efficient method.

Table A.1: Example 1 with different correlation strengths.

	$\rho = 0.2$				$\rho = 0.4$				$\rho = 0.6$				$\rho = 0.8$			
	Err	M	JD	T	Err	M	JD	T	Err	M	JD	T	Err	M	JD	T
S	0.17	49	38	0.1	0.06	7.3	76	0.1	0.08	0.3	80	0.1	0.12	0.3	80	0.1
LTS	0.11	36	44	0.2	0.04	9.1	64	0.2	0.03	0.5	78	0.2	0.06	0.2	82	0.2
RLARS	0.23	84	0	0.9	0.22	81	0	0.9	0.23	74	0	0.7	0.29	59	0	0.7
PENSE	0.25	85	0	13	0.24	82	0	13	0.21	62	16	13	0.11	0.2	88	13
<b>PIQ</b>	0.02	0.20	90	0.1	0.02	0.2	90	0.1	0.02	0.2	90	0.1	0.06	0.2	88	0.1

Table A.2: Example 2 with different covariance structures.

	Toeplitz structure				Blocked structure				Equally correlated			
	Err	M	JD	T	Err	M	JD	T	Err	M	JD	T
B-Y	0.29	100	0	0.6	0.34	97	0	0.8	0.34	97	0	0.7
QLE	0.29	100	0	0.1	0.33	97	0	0.1	0.33	97	0	0.1
TLE	0.11	16	84	6.0	0.07	0	100	7.0	0.07	0	100	6.0
<b>PIQ</b>	0.07	0	100	0.1	0.06	0	100	0.1	0.06	0	100	0.1

Table A.3: Example 3 and Example 4 with different sparsity levels.

	Regression													
	s* = 2					s* = 4								
	Err	M $\gamma$	JD $\gamma$	M $\beta$	FA $\beta$	JD $\beta$	T	Err	M $\gamma$	JD $\gamma$	M $\beta$	FA $\beta$	JD $\beta$	T
QL	0.70	79	0	4	0.2	92	44	0.89	81	0	33	0.2	8	45
S-MTE	0.58	77	0	0	0.2	100	35	0.82	79	0	28	0.2	26	42
RLARS	1.24	88	0	50	0.1	0	40	1.52	91	0	58	0.3	0	41
S-LTS	0.93	62	18	32	1.6	36	134	1.24	79	10	29	1.4	16	132
<b>PIQ</b>	0.15	2	76	3	0.1	94	4	0.34	3	78	16	0.3	46	4
	s* = 6										s* = 8			
	Err	M $\gamma$	JD $\gamma$	M $\beta$	FA $\beta$	JD $\beta$	T	Err	M $\gamma$	JD $\gamma$	M $\beta$	FA $\beta$	JD $\beta$	T
	QL	1.14	81	0	36	0.2	0	51	1.59	86	0	42	0.1	0
S-MTE	0.98	79	0	29	0.2	4	49	1.34	83	0	31	0.2	0	60
RLARS	1.21	85	0	40	0.4	0	41	1.27	83	0	32	0.6	2	41
S-LTS	1.29	77	22	25	1.0	20	134	1.55	90	4	23	1.0	8	144
<b>PIQ</b>	0.64	21	50	20	0.4	18	4	1.25	33	34	31	0.7	0	4
	Classification										s* = 4			
	Err	M $\gamma$	JD $\gamma$	M $\beta$	FA $\beta$	JD $\beta$	T	Err	M $\gamma$	JD $\gamma$	M $\beta$	FA $\beta$	JD $\beta$	T
	enetLTS	41	100	0	36	1.0	28	65	43	96	4	62	1.0	4
<b>PIQ</b>	15	4	96	2	0.1	96	14	18	5	92	15	0.3	48	15
	s* = 6										s* = 8			
	Err	M $\gamma$	JD $\gamma$	M $\beta$	FA $\beta$	JD $\beta$	T	Err	M $\gamma$	JD $\gamma$	M $\beta$	FA $\beta$	JD $\beta$	T
	enetLTS	46	96	4	82	0.8	14	74	47	96	4	88	0.6	0
<b>PIQ</b>	25	65	20	33	0.5	6	15	26	93	4	39	0.7	0	15

## References

- Alfons, A., Croux, C., and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics*, 7(1):226–248.
- Avella-Medina, M. (2017). Influence functions for penalized m-estimators. *Bernoulli*, 23(4B):3178–3196.
- Avella-Medina, M. and Ronchetti, E. (2018). Robust and consistent variable selection in

- high-dimensional generalized linear models. *Biometrika*, 105(1):31–44.
- Belloni, A. and Chernozhukov, V. (2011).  $\ell_1$ -penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130.
- Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific, 2nd edition.
- Bianco, A. M. and Yohai, V. J. (1996). Robust estimation in the logistic regression model. In *Robust statistics, data analysis, and computer intensive methods*, pages 17–34. Springer, New York, NY.
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge.
- Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217.
- Bühlmann, P. and Yu, B. (2003). Boosting with the  $L_2$  loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339.
- Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theor.*, 51(12):4203–4215.
- Cantoni, E. and Ronchetti, E. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, 96:1022–1030.
- Chapelle, O. (2007). Training a support vector machine in the primal. *Neural Computation*, 19(5):1155–1178.
- Chinchor, N. (1992). MUC-4 evaluation metrics. In *Proceedings of the 4th Conference on Message Understanding*, pages 22–29.
- Croux, C. and Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression. *Computational statistics & data analysis*, 44(1):273–295.
- Freue, G. V. C., Kepplinger, D., Salibián-Barrera, M., and Smucler, E. (2019). Robust elastic net estimators for variable selection and identification of proteomic biomarkers. *The Annals of Applied Statistics*, 13(4):2065–2090.
- Hadi, A. S. and Luceño, A. (1997). Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis*, 25(3):251–272.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (2011). *Robust Statistics: The Approach based on Influence Functions*, volume 1. John Wiley & Sons, Hoboken, NJ.

- Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. Chapman and Hall/CRC.
- Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. John Wiley & Sons, Hoboken, NJ.
- Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58(1):30–37.
- Khan, J. A., Van Aelst, S., and Zamar, R. H. (2007). Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102(480):1289–1299.
- Kurnaz, F. S., Hoffmann, I., and Filzmoser, P. (2018). Robust and sparse estimation methods for high-dimensional linear and logistic regression. *Chemometrics and Intelligent Laboratory Systems*, 172:211–222.
- Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A., and Moroz, I. M. (2007). Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. *Biomedical engineering online*, 6(1):23.
- Loh, P.-L. (2017). Statistical consistency and asymptotic normality for high-dimensional robust  $m$ -estimators. *The Annals of Statistics*, 45(2):866–896.
- Maronna, R. A., Martin, D. R., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*, volume 1. John Wiley & Sons.
- Müller, S. and Welsh, A. (2005). Outlier robust model selection in linear regression. *Journal of the American Statistical Association*, 100(472):1297–1310.
- Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814.
- Needell, D. and Tropp, J. A. (2009). CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3):301–321.
- Öllerer, V., Croux, C., and Alfons, A. (2015). The influence function of penalized regression estimators. *Statistics*, 49(4):741–765.
- Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications. *Biometrics*, pages 485–498.
- Qin, Y., Li, S., Li, Y., and Yu, Y. (2017). Penalized maximum tangent likelihood estimation and robust variable selection. *arXiv:1708.05439*.
- Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Annals of Statistics*, 39(2):731–771.

- Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ.
- Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, 92(439):1017–1023.
- Rousseeuw, P. and Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis*, pages 256–272. Springer, New York, NY.
- Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8:283–297.
- Rousseeuw, P. J. and Driessen, K. V. (1999). Computing LTS Regression for Large Data Sets. Technical report, Institute of Mathematical Statistics Bulletin.
- Rousseeuw, P. J. and Hubert, M. (1997). Recent developments in PROGRESS. *Lecture Notes-Monograph Series*, pages 201–214.
- Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electron. Commun. Probab.*, 18:1–9.
- Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., and Kursun, O. (2013). Collection and analysis of a parkinson speech dataset with multiple types of sound recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4):828–834.
- Salibian-Barrera, M. and Van Aelst, S. (2008). Robust model selection using fast and robust bootstrap. *Computational Statistics & Data Analysis*, 52(12):5121–5135.
- She, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics & Data Analysis*, 56(10):2976–2990.
- She, Y. (2016). On the finite-sample analysis of  $\Theta$ -estimators. *Electronic Journal of Statistics*, 10(2):1874–1895.
- She, Y. and Chen, K. (2017). Robust reduced-rank regression. *Biometrika*, 104(3):633–647.
- She, Y. and Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639.
- She, Y. and Tran, H. (2019). On cross-validation for sparse reduced rank regression. *Journal of the Royal Statistical Society: Series B*, 81:145–161.
- She, Y., Wang, J., Li, H., and Wu, D. (2013). Group iterative spectrum thresholding for super-resolution sparse spectral selection. *IEEE Transactions on Signal Processing*, 61(24):6371–6386.
- She, Y., Wang, Z., and Jin, J. (2021). Analysis of generalized Bregman surrogate algorithms for nonsmooth nonconvex statistical learning. *The Annals of Statistics*, 49(6):3434–3459.

- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.
- Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer, New York, NY.
- van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392.
- van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.
- Vandev, D. and Neykov, N. (1998). About regression estimators with high breakdown point. *Statistics: A Journal of Theoretical and Applied Statistics*, 32(2):111–129.
- Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. *Compressed sensing*, pages 210–268.
- Zhang, T. (2013). Multi-stage convex relaxation for feature selection. *Bernoulli*, 19(5B):2277–2293.