Robust Lasso With Missing and Grossly Corrupted Observations

Nam H. Nguyen and Trac D. Tran, *Senior Member, IEEE*

***Abstract*—This paper studies the problem of accurately recovering a k-sparse vector $\beta^\star \in \mathbb{R}^p$ from highly corrupted linear measurements $y = X\beta^\star + e^\star + w$, where $e^\star \in \mathbb{R}^n$ is a sparse error vector whose nonzero entries may be unbounded and w is a stochastic noise term. We propose a so-called extended Lasso optimization which takes into consideration sparse prior information of both $\beta^\star$ and $e^\star$. Our first result shows that the extended Lasso can faithfully recover both the regression as well as the corruption vector. Our analysis relies on the notion of extended restricted eigenvalue for the design matrix X. Our second set of results applies to a general class of Gaussian design matrix X with i.i.d. rows $\mathcal{N}(0, \Sigma)$, for which we can establish a surprising result: the extended Lasso can recover exact signed supports of both $\beta^\star$ and $e^\star$ from only $\Omega(k \log p \log n)$ observations, even when a linear fraction of observations is grossly corrupted. Our analysis also shows that this amount of observations required to achieve exact signed support is indeed optimal.**

***Index Terms*—Compressed sensing, error correction, high-dimensional inference, ℓ_1-minimization, robust recovery, sparse linear regression.**

I. Introduction

ONE of the central problems in statistics is the problem of linear regression in which the goal is to accurately estimate the regression vector $\beta^\star \in \mathbb{R}^p$ from the noisy observations

$$y = X\beta^\star + w \tag{1}$$

where $X \in \mathbb{R}^{n\times p}$ is the measurement or design matrix, and $w \in \mathbb{R}^n$ is the stochastic observation noise vector. A particular situation recently attracted much attention from the research community concerns with the model in which the number of regression variables p is larger than the number of observations n $(p \geq n)$. In such circumstances, without imposing some additional assumptions for this model, it is obvious that the problem is ill-posed, and thus, the linear regression is not consistent. Accordingly, there have been various lines of work on high-dimensional inference based on imposing different types of structure constraints such as sparsity and group sparsity (see, e.g., [1]–[15]). Among them, the most popular model focused on sparsity assumption of the regression vector. To estimate β, a standard method, namely Lasso [1], was proposed to use ℓ_1-penalty as a surrogate function to enforce sparsity constraint.

$$\min_\beta \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \tag{2}$$

where λ is the positive regularization parameter, the ℓ_1-norm of the regression vector is $\|\beta\|_1$, defined as $\|\beta\|_1 \triangleq \sum_{i=1}^p |\beta_i|$, and the ℓ_2-norm in the first summand is conventionally defined as $\|a\|_2 \triangleq (\sum_{i=1}^n a_i^2)^{1/2}$ for any $a \in R^n$.

Within the past few years, there has been numerous studies to understand the ℓ_1-regularization aspect of sparse regression models (see, e.g., [5]–[11]). These works are mainly characterized by the type of the loss functions considered. For instance, authors of [9] and [11] seek to obtain a regression estimate $\widehat{\beta}$ that delivers small prediction error, while others [6], [10], [11] seek to produce a regressor with minimal parameter estimation error, which is measured by the ℓ_2-norm of $(\widehat{\beta} - \beta^\star)$. Another line of work (see, e.g., [5], [8], and [16]) considers the variable selection in which the goal is to obtain an estimate that correctly identifies the support of the true regression vector. To achieve low prediction or parameter estimation loss, it is now well known that it is both sufficient and necessary to impose certain lower bounds on the smallest singular values of the design matrix (see, e.g., [7] and [10]), while the notion of small mutual coherence for the design matrix (see, e.g., [5], [8], and [9]) is required to achieve accurate variable selection.

We notice that all previous work relies on the assumption that the observation noise has bounded energy. Without this assumption, it is very likely that the estimated regressor is either not reliable or we fail to identify the correct support. With this observation in mind, in this paper, we extend the linear model (1) by considering the noise with unbounded energy. It is clear that if all entries of y are corrupted by large errors, then it is impossible to faithfully recover the regression vector $\beta^\star$. However, in many practical applications such as face recognition, acoustic recognition, and dense sensor network, only a portion of the observation vector is contaminated by gross error. Formally, we have the mathematical model

$$y = X\beta^\star + e^\star + w \tag{3}$$

where $e^\star \in \mathbb{R}^n$ is the sparse error whose locations of nonzero entries are unknown and whose magnitudes can be arbitrarily large, whereas w is the conventional noise vector with bounded energy. In this paper, we assume that w has a multivariate Gaussian $\mathcal{N}(0, \sigma^2 I_{n\times n})$ distribution. This model also includes

Manuscript received December 22, 2011; revised August 10, 2012; accepted September 25, 2012. Date of publication December 11, 2012; date of current version March 13, 2013. This work was supported in part by the National Science Foundation under Grants CCF-1117545 and CCF-0728893; in part by the Army Research Office under Grants 58110-MA-II and 60219-MA; and in part by the Office of Naval Research under Grant N102-183-0208. This paper was presented in part at the Annual Conference on Neural Information Processing Systems, Granada, Spain, December 2011.

The authors are with the Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218 USA (e-mail: nam@jhu.edu; trac@jhu.edu).

Communicated by E. Serpedin, Associate Editor for Signal Processing.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2012.2232347

as a special case the missing data problem in which all the entries of y is not fully observed, but some are missing. This problem is particularly important in computer vision and biology applications. If some entries of y are missing, the nonzero entries of $e^\star$ whose locations are associated with the missing entries of the observation vector y have the same values as entries of y but with reverse polarity.

The problems of faithfully recovering data under gross error has gained increasing attentions recently with many interesting practical applications (see, e.g., [17]–[19]) as well as theoretical consideration (see, e.g., [20]–[23]). Another recent line of research on recovering the data from grossly corrupted measurements has been also studied in the context of robust principal component analysis (see, e.g., [24]–[26]). Let us consider several examples as illustrations.

1) *Face recognition.* The model (3) has been proposed by Wright *et al.* [17] in the context of face recognition. In this problem, a face test sample y is assumed to be represented as a linear combination of training faces in the dictionary X. Hence, $y = X\beta$, where β is the coefficient vector used for classification. However, it is often the case that the testing face of interest is occluded by unwanted objects such as glasses, hats, scarfs, etc. These occlusions, which occupy a portion of the test face, can be considered as the sparse error $e^\star$ in the model (3).
2) *Subspace clustering.* An important problem in high-dimensional data analysis is to cluster the data points into multiple subspaces. A recent work of Elhamifar and Vidal [18] show that this problem can be solved by expressing each data point as a sparse linear combination of all other data points. Coefficient vectors recovered from solving the Lasso problems are then employed for clustering. If the data points are represented as a matrix X, then we wish to find a sparse coefficient matrix B such that $X = XB$ and $\mathrm{diag}(B) = 0$. When the data are missing or contaminated by outliers, the authors formulate the problem as $X = XB + E$ and minimize a sum of two ℓ_1-norms with respect to both B and E [18].
3) *Sparse graphical model estimation.* Given a random vector $x \in \mathbb{R}^p$ with unknown covariance matrix Σ, the goal is to estimate Σ or its precision matrix $\Omega = \Sigma^{-1}$ from n independent copies of x: $x_1, \ldots, x_n \in \mathbb{R}^p$. Assuming that the matrix Ω is sparse, Meinshausen and Bühlmann [7] propose to solve the following Lasso problem:

$$\min_B \frac{1}{2n} \|X - XB\|_F^2 + \lambda \|B\|_1 \quad \text{s.t. } \mathrm{diag}(B) = 0,$$

where $X = [x_1^T, \ldots, x_n^T]$ and the Frobenius norm $\|A\|_F \triangleq (\sum_{i,j} A_{ij}^2)^{1/2}$ for any matrix A. The precision matrix Ω can be estimated via the coefficient matrix B. When the data X is partially observed/missing, a more robust method is to take into account the sparsity assumption and minimize

$$\min_{B,E} \frac{1}{2n} \|X - XB - E\|_F^2 + \lambda_b \|B\|_1 + \lambda_e \|E\|_1$$

subject to $\mathrm{diag}(B) = 0$, where E represents partially missing information. Though this problem is quite different from the aforementioned subspace clustering problem, the technical approach is considerably similar.
4) *Sensor network.* In this model, a network of sensors collect measurements of a signal $\beta^\star$ independently by simply projecting $\beta^\star$ onto the row vectors of a sensing matrix X, $y_i = \langle X_i, \beta^\star \rangle$ [27]. The measurements y_i are then sent to the central hub for analysis. However, it is highly likely that a small percentage of sensors might fail to send the measurements correctly and sometimes even report totally irrelevant measurements. Therefore, it is more appropriate to employ the observation model in (3) than the model in (1).

It is worth noticing that in the aforementioned applications, $e^\star$ always plays the role as the sparse (undesired) error. However, in other applications, $e^\star$ might actually contain meaningful information and thus necessary to be recovered. An example of this kind of problem is signal separation, in which $\beta^\star$ and $e^\star$ are considered as two distinct signal components (e.g., video or audio). Furthermore, in applications such as classification and clustering, the assumption that the test sample y is a linear combination of a few training samples in the dictionary (playing the role of the design matrix) X might be violated [28]. The sparse component $e^\star$ can thus be seen as the compensation for the linear regression model mismatch.

Given the observation model (1) and the sparsity assumptions on both regression vector $\beta^\star$ and error $e^\star$, we propose the following convex minimization to estimate the unknown regression vector $\beta^\star$ as well as the error vector $e^\star$:

$$\min_{\beta,e} \frac{1}{2n} \|y - X\beta - e\|_2^2 + \lambda_{n,\beta} \|\beta\|_1 + \lambda_{n,e} \|e\|_1 \qquad (4)$$

where $\lambda_{n,\beta}$ and $\lambda_{n,e}$ are positive regularization parameters. This optimization, which we call *extended Lasso*, can be seen as a generalization of the Lasso program. Indeed, by setting $\lambda_{n,e} = 0$, (6) returns to the standard Lasso. The additional regularization associated with the error e encourages sparsity of the reconstructed vector, where the penalty parameter $\lambda_{n,e}$ controls its sparsity level. In this paper, we focus on the following questions: what are necessary and sufficient conditions for the ambient dimension p, the number of observations n, the sparsity index k of the regression $\beta^\star$, and the fraction of corruption in $e^\star$ so that 1) the extended Lasso is able (or unable) to recover the exact support sets of both $\beta^\star$ and $e^\star$? 2) the extended Lasso is able to recover $\beta^\star$ and $e^\star$ with small prediction error and parameter error? We are particularly interested in understanding the asymptotic situation where the fraction of error gets arbitrarily close to 100%.

In this paper, we assume normalization of the design matrix X. Specifically, we assume the ℓ_2-norm of columns of the matrix X are $\Theta(\sqrt{n})$. Moreover, without loss of generality, we use the following observation model in replacement for the model in (3)

$$y = X\beta^\star + \sqrt{n} e^\star + w. \qquad (5)$$

As we can see, columns of both the design matrix X and the matrix $\sqrt{n}I_{n\times n}$ has the same scale. Thus, this model's change only helps our results in the next sections to be more interpretable. The optimization (4) is now converted to the following problem:

$$\min_{\beta,e} \frac{1}{2n}\left\|y - X\beta - \sqrt{n}e\right\|_2^2 + \lambda_{n,\beta}\left\|\beta\right\|_1 + \lambda_{n,e}\left\|e\right\|_1. \quad (6)$$

Previous Work: The problem of recovering the estimation vector $\beta^\star$ and error $e^\star$ is originally proposed by Wright *et al.* in the appealing paper [17] and analyzed by Wright and Ma [20]. In the absence of the stochastic noise w in the observation model (3), the authors propose to estimate $(\beta^\star, e^\star)$ by solving the following linear program:

$$\min_{\beta,e} \left\|\beta\right\|_1 + \left\|e\right\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}e. \quad (7)$$

From a different viewpoint, in the intriguing paper [29], Lee *et al.* study a general loss function model. To obtain more flexibility in controlling the undesirable influence of the model, they introduce a case-specific parameter vector $e \in \mathbb{R}^n$ for the observation vectors and modify the optimization to take into account this parameter. Interestingly, the model turns out to be coincident with (6) when applying to the linear regression problem with Lasso penalty. Extensive simulations have shown that the model (6) is considerably robust to noise. However, no theoretical analysis is provided in the paper.

In another direction, the problem of robust Lasso under corrupted observations is also carefully investigated by Wang *et al.* [30]. In this appealing paper, instead of using the quadratic loss function as in Lasso, the authors propose to employ LAD-Lasso criterion:

$$\min_{\beta} \left\|y - X\beta\right\|_1 + \sum_{j=1}^{p} \lambda_j |x_j|. \quad (8)$$

This optimization combines the LAD criterion and Lasso penalty, where the first term is designed to be robust to outliers and the second term again promotes the sparse representation of the estimator. However, due to the lack of the quadratic loss that enforces the estimation to be consistence with the observation in ℓ_2-norm sense, this optimization might not guarantee to deliver a solution that satisfies small prediction error.

On the theoretical side, the result of [20] is asymptotic in nature. The analysis reveals an interesting phenomenon that for a class of Gaussian design matrix with i.i.d. columns $\mathcal{N}(\mu, \frac{\nu}{n}I_n)$ obeying $\left\|\mu\right\|_2 = 1$ and $\left\|\mu\right\|_\infty \leq Cn^{-1/2}$, the optimization (7) can recover $(\beta^\star, e^\star)$ precisely with high probability even when the fraction of corruption is arbitrarily close to one. However, the result only holds under rather stringent conditions. In particularly, the authors require the number of observations n grow proportionally with the ambient dimension p, and the sparsity index k is a very small portion of n. These conditions are of course far from the optimal bound in compressed sensing (CS) and statistics literature (recall $k \leq O(n/\log p)$ is sufficient in conventional analysis (see, e.g., [8] and [31]). Our model, though different and not directly comparable to that of [20], shows that while the sparse error can be any fraction of the sample size, the sparsity level of $\beta^\star$ is almost optimal with $k \leq O(n/\log p \log n)$.

Another line of work has also focused on the optimization (7). In both[19] and [21], the authors establish that for Gaussian design matrix X, if $n \geq C(k+s)\log p$ where s is the sparsity level of $e^\star$, then the recovery is exact. This follows from the fact that the combination matrix $[X,\ I]$ obeys the restricted isometry property, a well-known property in CS used to guarantee exact recovery of sparse vectors via ℓ_1-minimization. These results, however, do not allow the fraction of corruption to come close to unity. Also related to our paper is recent work by Studer *et al.* [32], [33] in which the authors establish different results for deterministic design matrix. In recent work, Dalalyan and Keriven [34], [35] investigate an important recovery problem in multiview geometry in which they employ the ℓ_1-minimization together with considering outliers. Their work is interesting, though the formulation is different from ours.

Among the previous work, the most closely related to our current paper are recent results by Li [23] and Nguyen *et al.* [22] in which a positive regularization parameter τ is employed to control the sparsity of $e^\star$. Using different methods, both sets of authors show that as τ is deterministically selected to be $1/\sqrt{\log p}$ and X is a suborthogonal matrix, whose columns are selected uniformly at random from columns of an orthogonal matrix, then the solution of following optimization (9) is exact even a constant fraction of observation is corrupted. Moreover, Li [23] establishes a similar result with Gaussian design matrix in which the number of observations is only on the order of $k\log p$—a level that is known to be optimal in both CS and statistics communities:

$$\min_{\beta,e} \left\|\beta\right\|_1 + \tau\left\|e\right\|_1 \quad \text{s.t.} \quad y = X\beta + \sqrt{n}e. \quad (9)$$

Our Contribution: This paper considers a general setting in which the observations are contaminated by both sparse and dense errors. We allow the corruptions to linearly grow with the number of observations and have arbitrarily large magnitudes. We establish a general scaling of the quadruplet (n, p, k, s) such that the proposed extended Lasso stably recovers both the regression and the corruption vector. Of particular interest to us are the answers to the following questions:

1) First, under what scalings of (n, p, k, s) does the extended Lasso obtain the unique solution with small estimation error?
2) Second, under what scalings of (n, p, k) does the extended Lasso obtain the exact signed support recovery even when almost all observations are corrupted? The signed support recovery implies that the extended Lasso identifies exact locations and signs of nonzero coefficients of both $\beta^\star$ and $e^\star$.
3) Third, under what scalings of (n, p, k, s) that no solution of the extended Lasso specifying the correct signed support exists?

To answer the first question, we introduce a notion of extended restricted eigenvalue (RE) for a matrix $[X,\ I]$ where I is the identity matrix. We show that this property is satisfied for a general class of random Gaussian design matrices. The answers to the last two questions require stricter conditions on the design matrix. In particular, for random Gaussian design matrix with i.i.d. rows $\mathcal{N}(0, \Sigma)$, we rely on two standard assumptions:

invertibility and mutual incoherence. Our analysis in this setting is relied on the elegant technique introduced by Wainwright [8].

If we denote $Z = [X,\ I]$, where I is an identity matrix and $\overline{\beta} = [\beta^{\star^T}, e^{\star^T}]^T$, then the observation vector y is reformulated as $y = Z\overline{\beta} + w$, which is the same as the standard Lasso model. However, previous results (see, e.g., [8] and [10]) applying to random Gaussian design matrices are irrelevant to this setting since Z no longer behaves like a Gaussian matrix. To establish the theoretical analysis, we need a deeper study on the interaction between the Gaussian and identity matrices. By exploiting the fact that the matrix Z consists of two components where one has a special structure, our analysis reveals an interesting phenomenon: extended Lasso can accurately recover both the regressor $\beta^\star$ and the corruption $e^\star$ even when the fraction of corruption is up to 100%. We measure the recoverability of these variables under two criteria: parameter accuracy and feature selection accuracy. Moreover, our analysis can be extended to the situation in which the identity matrix can be replaced by a tight frame D as well as extended to other models such as group Lasso or matrix Lasso with sparse error.

We would like to note that in this paper, several improvements are placed over our previous results presented at the NIPS conference [36]. We require weaker mutual incoherence condition, which is crucial in establishing Theorem 2; unknown constants in [36, Ths. 2 and 3] are replaced by explicit values; and more intensive simulations are conducted to show consistent agreements with theoretical prediction.

Notation: We summarize here some standard notations employed throughout this paper. We reserve T and S as the sparse support of $\beta^\star$ and $e^\star$, respectively. The sparsity levels of $\beta^\star$ and $e^\star$ are denoted as $|T| = k$ and $|S| = s$. We also reserve p and n as the problem and sample sizes. That means $\beta^\star \in \mathbb{R}^p$ and $y \in \mathbb{R}^n$. Given a design matrix $X \in \mathbb{R}^{n\times p}$ and subsets S and T, we use X_{ST} to denote the $|S| \times |T|$ submatrix obtained by extracting those rows indexed by S and columns indexed by T. For a vector $h \in \mathbb{R}^p$, we use the conventional notations for ℓ_1-and ℓ_2-norm of h as $\|h\|_1 = \sum_{i=1}^p |h_i|$ and $\|h\|_2 = (\sum_{i=1}^p h_i^2)^{1/2}$, respectively. For a matrix $X \in \mathbb{R}^{n\times p}$, we denote $\|X\|$ and $\|X\|_\infty$ as the operator norms. In particular, $\|X\|$ is denoted as the spectral norm and $\|X\|_\infty$ as the ℓ_∞/ℓ_∞ operator norm: $\|X\|_\infty = \max_i \sum_{j=1}^p |x_{ij}|$.

We use the notation C_1, C_2, c_1, c_2, etc., to refer to positive constants, whose value may change from line to line. Given two functions f and g, the notation $f(n) = \mathcal{O}(g(n))$ means that there exists a constant $c < +\infty$ such that $f(n) \le cg(n)$; the notation $f(n) = \Omega(g(n))$ means that $f(n) \ge cg(n)$; and the notation $f(n) = \Theta(g(n))$ means that $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$. The symbol $f(n) = o(g(n))$ indicates that $f(n)/g(n) \to 0$.

Organization: The remainder of this paper is structured as follows. Section II provides the main results, detailed discussions and their consequences. Section III performs extensive experiments to validate theoretical results presented in the previous section. Section IV provides analysis of the estimation error, whereas Sections V and VI deliver proofs of the necessary and sufficient conditions for the exact signed support recovery. Several technical aspects of these proofs and some well-known concentration inequalities are presented in Appendix. We conclude the paper in Section VII with more discussion.

II. Main Results

In this section, we provide precise statements for the main results of this paper. In Section II-A, we establish the parameter estimation and provide a deterministic result which is based on the notion of extended RE. We further show that the random Gaussian design matrix satisfies this property with high probability. Section II-B considers feature estimation. We establish conditions for the design matrix such that the solution of the extended Lasso has the exact signed supports.

A. Parameter Estimation

As in conventional Lasso, to obtain a low parameter estimation bound, it is necessary to impose conditions on the design matrix X. In this paper, we introduce the notion of extended RE condition. Let $\mathbb{C}$ be a restricted set, we say that the matrix X satisfies the extended RE assumption over the set $\mathbb{C}$ if there exists some $\kappa_l > 0$ such that

$$\frac{1}{\sqrt{n}}\left\|Xh + \sqrt{n}f\right\|_2 \ge \kappa_l(\|h\|_2 + \|f\|_2) \quad \text{for all } (h,f) \in \mathbb{C} \tag{10}$$

where the restricted set $\mathbb{C}$ of interest is defined with $\lambda \overset{\triangle}{=} \frac{\lambda_{n,e}}{\lambda_{n,\beta}}$ as follows:

$$\mathbb{C} \overset{\triangle}{=} \{(h,f) \in \mathbb{R}^p \times \mathbb{R}^n | \ \|h_{T^c}\|_1 + \lambda\|f_{S^c}\|_1 \le 3\|h_T\|_1 + 3\lambda\|f_S\|_1\}. \tag{11}$$

This assumption is a natural extension of the RE condition and restricted strong convexity considered in [10], [37], and [38]. In the absence of a vector f in the (10) and in the set $\mathbb{C}$, this condition returns to the RE defined in [10]. As discussed in more detail in [10] and [39], RE is among the weakest assumption on the design matrix such that the solution of the Lasso is consistent.

With this assumption at hand, we now state the first theorem.

Theorem 1: Consider the optimal solution $(\widehat{\beta}, \widehat{e})$ to the optimization problem (6) with regularization parameters chosen as

$$\lambda_{n,\beta} = c\frac{2}{\mu}\frac{\|X^*w\|_\infty}{n} \quad \text{and} \quad \lambda_{n,e} = c\frac{2\|w\|_\infty}{\sqrt{n}} \tag{12}$$

for some $\mu \in (0,1]$ and $c \ge 1$. Assuming that the design matrix X obeys the extended RE, then the error set $(h,f) = (\widehat{\beta} - \beta^\star, \widehat{e} - e^\star)$ is bounded by

$$\|h\|_2 + \|f\|_2 \le 3\kappa_l^{-2}\left(\lambda_{n,\beta}\sqrt{k} + \lambda_{n,e}\sqrt{s}\right). \tag{13}$$

There are several interesting observations from this theorem.

1) The error bound naturally split into two components related to the sparsity indices of $\beta^\star$ and $e^\star$. In addition, the error bound contains three quantities: the sparsity indices, regularization parameters, and the extended RE constant. If the terms related to the corruption $e^\star$ are omitted, then we

obtain similar parameter estimation bound as in the standard Lasso (see, e.g., [10] and [38]).

2) The choice of regularization parameters $\lambda_{n,\beta}$ and $\lambda_{n,e}$ can be made explicitly: assuming w is a Gaussian random vector whose entries are $\mathcal{N}(0,\sigma^2)$ and the design matrix has $\sqrt{n}$-normed columns, it is clear that with high probability, $\frac{1}{n}\|X^*w\|_\infty \leq 2\sqrt{\frac{\sigma^2\log p}{n}}$ and $\frac{1}{\sqrt{n}}\|w\|_\infty \leq 2\sqrt{\frac{\sigma^2\log n}{n}}$. Thus, it is sufficient to select $\lambda_{n,\beta} \geq \frac{4}{\mu}\sqrt{\frac{\sigma^2\log p}{n}}$ and $\lambda_{n,e} \geq 4\sqrt{\frac{\sigma^2\log n}{n}}$.
3) At the first glance, the parameter μ does not seem to have any meaningful interpretation and setting $\mu = 1$ seems to be the best selection due to the smallest estimation error it can produce for (17). However, this parameter actually controls the relationship between the sparsity levels of the regression and the sparse error. To see this, we rewrite the optimization (6) in the constrained form

$$\|\beta\|_1 + \lambda\|e\|_1 \quad \text{s.t.} \quad \left\|y - X\beta - \sqrt{n}e\right\|_2 \leq \sigma \tag{14}$$

where $\lambda = \frac{\lambda_{n,e}}{\lambda_{n,\beta}} = \mu\sqrt{n}\frac{\|w\|_\infty}{\|X^*w\|_\infty}$. It has been well known that this constrained optimization is equivalent to (6) if the regularization parameter $\lambda_{n,\beta}$ is chosen correctly. As one can observe, the parameter λ in (14) is used to control the balance between two ℓ_1-norm terms. As the value of λ (or μ) increases, we expect to obtain the solution $(\widehat{\beta},\widehat{e})$ of (14) with sparser $\widehat{e}$ and denser $\widehat{\beta}$. Accordingly, the smaller μ should be imposed when we predict a larger amount of corruptions in the observations, while larger μ is used otherwise. In the following corollary 1, we show that with the appropriate choices of $\lambda_{n,\beta}$ and $\lambda_{n,e}$ as in (16), by setting $\mu = 1/\sqrt{\log n}$, the extended Lasso guarantee to stably recover the regression vector even a linear fraction of the observations is corrupted. The explicit expressions between n, p, k, and μ will be carefully studied in Theorem 2 in Section II-B.

In the following lemma, we show that the extended RE condition actually exists for a large class of random Gaussian design matrix whose rows are i.i.d. zero mean with covariance Σ. Before stating the lemma, let us define some quantities operating on the covariance matrix Σ: $C_{\min} \triangleq \lambda_{\min}(\Sigma)$ is the smallest eigenvalue of Σ; $C_{\max} \triangleq \lambda_{\max}(\Sigma)$ is the largest eigenvalue of Σ; and $\xi(\Sigma) \triangleq \max_i \Sigma_{ii}$ is the maximal entry on the diagonal of the matrix Σ.

Lemma 1: Consider the random Gaussian design matrix whose rows are i.i.d. $\mathcal{N}(0,\Sigma)$. Select

$$\lambda \triangleq \frac{\mu}{\sqrt{\xi(\Sigma)}}\sqrt{\frac{\log n}{\log p}} \tag{15}$$

with $\mu \in [\frac{1}{\sqrt{\log n}}, 1]$. Then with probability greater than $1 - c_1\exp(-c_2 n)$, the matrix X satisfies the extended RE with parameter $\kappa_l = \frac{1}{16}\min\{\sqrt{C_{\min}}, 1\}$, provided that $n \geq C\frac{\xi(\Sigma)C_{\max}}{C_{\min}^2}k\log p$ with $C \geq 144^2$ and $s \leq \min\{C'\frac{n}{\mu^2\log n}, C''\frac{C_{\min}^2}{C_{\max}}n\}$ for some constants C, C' and C''.

We would like to offer a few remarks.

1) The choice of parameter λ is nothing special here. When the design matrix is Gaussian and independent with the Gaussian stochastic noise w, we can easily show that $\frac{1}{n}\|X^*w\|_\infty \leq 2\sqrt{\xi(\Sigma)\frac{\sigma^2\log p}{n}}$ with probability at least $1 - 2\exp(-\log p)$. Therefore, the selection of λ follows from Theorem 1.
2) Our argument do not yield good estimates for the constants C, C', and C'' in the above lemma. The proof delivers the values $C \geq 144^2$, $C' \leq 1/54^2$ and $C'' < 1/6400$. These values can possibly be reduced by more careful analysis and/or by reducing the value of the parameter κ_l.
3) The proof of this lemma, shown in Appendix, boils down to controlling two terms
   a) RE with X.

$$\frac{1}{n}\|Xh\|_2^2 + \|f\|_2^2 \geq \kappa_r(\|h\|_2^2 + \|f\|_2^2) \ \forall\ (h,f) \in \mathbb{C}.$$

   b) *Mutual incoherence.* The column space of the matrix X is incoherent with the column space of the identity matrix. That is, there exists some $\kappa_m > 0$ such that

$$\frac{1}{\sqrt{n}}|\langle Xh, f\rangle| \leq \kappa_m(\|h\|_2 + \|f\|_2)^2 \ \forall\ (h,f) \in \mathbb{C}.$$

   If the incoherence between these two column spaces is sufficiently small such that $4\kappa_m < \kappa_r$, then we can conclude that $\|Xh + f\|_2^2 \geq (\kappa_r - 2\kappa_m)(\|h\|_2 + \|f\|_2)^2$. The small mutual incoherence property is especially important since it provides how the regression separates itself away from the sparse error.
4) To simplify our result, we consider a special case of the uniform Gaussian design, in which $\Sigma = I_{p\times p}$. In this situation, $C_{\min} = C_{\max} = \xi(\Sigma) = 1$. We have the following result which is a corollary of Theorem 1 and Lemma 1.

Corollary 1 (Standard Gaussian Design): Let X be a standard Gaussian design matrix. Consider the optimal solution $(\widehat{\beta},\widehat{e})$ to the optimization problem (6) with regularization parameters chosen as

$$\lambda_{n,\beta} = \frac{4}{\mu}\sqrt{\frac{\sigma^2\log p}{n}} \quad \text{and} \quad \lambda_{n,e} = 4\sqrt{\frac{\sigma^2\log n}{n}} \tag{16}$$

for some $\mu \in [\frac{1}{\sqrt{\log n}}, 1]$. Also, assuming that $n \geq Ck\log p$ and $s \leq \min\{C'\frac{n}{\mu^2\log n}, C''n\}$ for some constants C, C' and C'' defined in Lemma 1. Then with probability greater than $1 - c_1\exp(-c_2 n)$, the error set $(h,f) = (\widehat{\beta} - \beta^\star, \widehat{e} - e^\star)$ is bounded by

$$\|h\|_2 + \|f\|_2 \leq 3\kappa_l^{-2}\left(\frac{1}{\mu}\sqrt{\frac{\sigma^2 k\log p}{n}} + \sqrt{\frac{\sigma^2 s\log n}{n}}\right) \tag{17}$$

with $\kappa_l = \frac{1}{16}$.

Corollary 1 reveals a remarkable result: by setting $\mu = 1/\sqrt{\log n}$, even when the fraction of corruption is linearly proportional with the number of samples n, the extended Lasso (6) is still capable of recovering both coefficient vector $\beta^\star$ and corruption (missing) vector $e^\star$ within a bounded

error (17). Without the dense noise w in the observation model (3) ($\sigma = 0$), the extended Lasso actually recovers the exact solution. This is impossible to achieve with the standard Lasso. Furthermore, if we know in prior that the number of corrupted observations is on the order of $\mathcal{O}(n/\log p)$, then selecting $\mu = 1$ instead of $1/\sqrt{\log n}$ will minimize the estimation error (see (17) of Corollary 1).

B. Feature Selection With Random Gaussian Design

In many applications, the feature selection criterion is more preferred (see, e.g., [5] and [8]). Feature selection refers to the property that the recovered parameter has the same signed support as the true regressor. In general, good feature selection implies good parameter estimation but the reverse direction does not usually hold. In this part, we investigate conditions for the design matrix and the scaling of (n, p, k, s) such that both regression and sparse error vectors satisfy these criterion.

Consider the linear model (3) where X is the Gaussian random design matrix whose rows are i.i.d. zero mean with covariance matrix Σ. It has been well known in the Lasso that in order to obtain feature selection accuracy, the covariance matrix Σ must obey two properties: invertibility and small mutual incoherence restricted on the set T. The first property guarantees that (6) is strictly convex, leading to the unique solution of the convex program, while the second property requires that the separation between two components of Σ, one related to the set T and the other to the set T^c, must be sufficiently small.

1) *Invertibility.* To guarantee uniqueness, we require Σ_{TT} to be invertible. Particularly, let $C_{\min}^T = \lambda_{\min}(\Sigma_{TT})$, we require $C_{\min}^T > 0$.
2) *Mutual incoherence.* For some $\gamma \in (0, 1)$,

$$\left\|\Sigma_{T^cT}^*(\Sigma_{TT})^{-1}\right\|_\infty \leq (1-\gamma). \tag{18}$$

It is worth noting that these two invertibility and mutual incoherence properties are exactly the same as the conditions used to establish the exact signed support recovery in the standard Lasso (see, e.g., [5], [8], and [16]).

Toward the end, we will also elaborate on three other quantities operating on the restricted covariance matrix Σ_{TT}: $C_{\max}^T$, which is defined as the maximum eigenvalue of Σ_{TT}: $C_{\max}^T \triangleq \lambda_{\max}(\Sigma_{TT})$; and $D_{\max}^{T-}$ and $D_{\max}^{T+}$, which are denoted as ℓ_∞-norm of matrices Σ_{TT}^{-1} and Σ_{TT}: $D_{\max}^{T-} \triangleq \left\|(\Sigma_{TT})^{-1}\right\|_\infty$ and $D_{\max}^{T+} \triangleq \|\Sigma_{TT}\|_\infty$. The superscript T implies that these quantities depend on the support T.

Our result also involves two other quantities operating on the conditional covariance matrix of $(X_{T^c}|X_T)$ defined as

$$\Sigma_{T^c|T} \triangleq \Sigma_{T^cT^c} - \Sigma_{T^cT}\Sigma_{TT}^{-1}\Sigma_{TT^c}. \tag{19}$$

They are defined as $\rho_u^T(\Sigma_{T^c|T}) = \max_i(\Sigma_{T^c|T})_{ii}$ and $\rho_l^T(\Sigma_{T^c|T}) = \frac{1}{2}\min_{i\neq j}[(\Sigma_{T^c|T})_{ii} + (\Sigma_{T^c|T})_{jj} - 2(\Sigma_{T^c|T})_{ij}]$, which we often denote with the shorthand notation ρ_u^T and ρ_l^T. Unless otherwise stated, for the rest of this section and the proofs of the following theorems, we remove the superscript out of the predefined quantities to simplify the notations, and from now, we write them as $C_{\min}$, $C_{\max}$, $D_{\max}^+$, $D_{\max}^-$, ρ_u, and ρ_l.

We establish the following result for Gaussian random design whose covariance matrix Σ obeys the two assumptions.

Theorem 2 (Achievability): Given the linear model (3) with random Gaussian design and the covariance matrix Σ satisfying invertibility and incoherence properties for any $\gamma \in (0, 1)$, suppose that we solve the extended Lasso (6) with regularization parameters obeying

$$\lambda_{n,\beta} = \frac{8}{\gamma}\sqrt{\frac{\sigma^2\eta\log n\log p}{n}\max\{\rho_u, D_{\max}^+\}} \tag{20}$$

$$\text{and} \qquad \lambda_{n,e} = 4\sqrt{\frac{\sigma^2\log n}{n}}, \tag{21}$$

for some $\eta \in [\frac{1}{\log n}, 1)$. Assume that the sequence (n, p, k, s) and regularization parameters $\lambda_{n,\beta}$, $\lambda_{n,e}$ satisfy $s \leq \eta n$ and $n > \max\{n_1, n_2\}$, where n_1 and n_2 are defined as

$$n_1 \triangleq \frac{4(1+\epsilon)}{(1-\eta)}\frac{\rho_u}{C_{\min}\gamma^2}k\log(p-k)\left\{\frac{9}{4} + (1-\eta)^2\frac{\sigma^2 C_{\min}}{\lambda_{n,\beta}^2 k}\right\}$$

$$\text{and} \quad n_2 \triangleq 48(1+\epsilon)\frac{\eta}{(1-\eta)^2}\frac{\max\{\rho_u, D_{\max}^+\}}{C_{\min}\gamma^2} \times \left(1 - \frac{2\sigma\sqrt{\log n}}{\lambda_{n,e}\sqrt{n}}\right)^{-2} k\log(p-k)\log n$$

for some $\epsilon \in (0, 1)$. In addition, suppose that $\min_{i\in T}|\beta_i^\star| > f_\beta(\lambda_{n,\beta})$ and $\min_{i\in S}|e_i^\star| > f_e(\lambda_{n,\beta}, \lambda_{n,e})$ where

$$f_\beta(\lambda_{n,\beta}) \triangleq c_1\lambda'_{n,\beta} + 20\sqrt{\frac{\sigma^2\log k}{C_{\min}(n-s)}} \quad \text{and} \tag{22}$$

$$f_e(\lambda_{n,\beta}, \lambda_{n,e}) \triangleq c_2\lambda'_{n,\beta}\sqrt{C_{\max}}\sqrt{\frac{sk + k\sqrt{sk}}{n}} + c_3\lambda_{n,e}$$

$$\text{with } \lambda'_{n,\beta} \triangleq \lambda_{n,\beta}\sqrt{\frac{k\log(p-k)}{(1-\eta)^2 n}}\left\|\Sigma_{TT}^{-1/2}\right\|_\infty^2. \tag{23}$$

Then, the following properties holds with probability greater than $1 - c\exp(-c'\max\{\log n, \log(p-k)\})$:

1) The solution pair $(\widehat{\beta}, \widehat{e})$ of the extended Lasso (6) is unique and has the exact signed support.
2) ℓ_∞-norm bounds: $\left\|\widehat{\beta} - \beta^\star\right\|_\infty \leq f_\beta(\lambda_{n,\beta})$ and $\|\widehat{e} - e^\star\|_\infty \leq f_e(\lambda_{n,\beta}, \lambda_{n,e})$.

There are several interesting observations from the theorem.

1) The first important observation is that the extended Lasso is robust to arbitrarily large and sparse error observation. Under the same invertibility and mutual incoherence assumptions on the covariance matrix Σ as the standard Lasso, the extended Lasso program can recover both the regression vector and error with exact signed supports even when almost all the observations are contaminated by arbitrarily large error with unknown support. What we sacrifice for the corruption robustness is an additional $\log$ factor to the number of samples. We notice that when the error fraction η is in the order of $\frac{1}{\log n}$ or in other word, the number of corrupted observation is $\mathcal{O}(\frac{n}{\log n})$, only $n = \Omega(k\log(p-k))$ samples are sufficient to recover

the exact signed supports of both the regression and sparse error vectors. This can be obtained by replacing $\eta = c\frac{1}{\log n}$ into quantities n_1 and n_2 of Theorem 2 and $n \geq \max\{n_1, n_2\}$.

2) We consider the special case with Gaussian random design in which the covariance matrix $\Sigma = I_{p\times p}$. In this case, entries of X is i.i.d. $\mathcal{N}(0,1)$ and we have quantities $C_{\min} = C_{\max} = D^{+}_{\max} = D^{-}_{\max} = \rho_u = \rho_l = 1$. In addition, the invertibility and mutual incoherence properties are automatically satisfied with $\gamma = 1$. The theorem implies that when the number of errors s is arbitrarily close to n, the number of samples n needed to recover the exact signed supports obeys $\frac{n}{\log n} = \Omega(k\log(p-k))$. Furthermore, Theorem 2 guarantees consistency in element-wise ℓ_∞-norm of the estimated regression at the rate of

$$\left\|\widehat{\beta} - \beta^\star\right\|_\infty = \mathcal{O}\left(\sqrt{\frac{\sigma^2 \log p}{n}}\sqrt{\frac{\eta k \log n \log(p-k)}{n}}\right).$$

As η is chosen to be $1/\sqrt{\log n}$ (equivalent to establish s close to $n/\log n$), the ℓ_∞ error rate is on the order of $\mathcal{O}(\sigma\sqrt{\frac{\log p}{n}})$, which is known to be the same as that of the standard Lasso. On the other hand, if we select η is arbitrarily close to unity—equivalently, s is close to n, and the ℓ_∞ error rate is on the order of $\mathcal{O}(\sigma\sqrt{\frac{\log n \log p}{n}})$. This is naturally interpreted as the more fraction of corruption is on the observations, the higher reconstruction error we expect to get. What interesting is that we draw an explicit connection between the fraction of corruption and the reconstruction error obtained by the extended Lasso optimization.

3) Corollary 1, though interesting, is not able to guarantee stable recovery when the fraction of corruption converges to unity. We show in Theorem 2 that this fraction can come arbitrarily close to unity by sacrificing a factor of $\log n$ for the number of samples. Theorem 2 also implies that there is a significant difference between recovery to obtain small parameter estimation error versus recovery to obtain correct variable selection. When the amount of corrupted observations is linearly proportional to n, recovering the exact signed supports require an increase from $\Omega(k\log p)$ (in Corollary 1) to $\Omega(k\log p\log n)$ samples (in Theorem 2). This behavior is captured similarly by the standard Lasso, as pointed out in the discussion after Corollary 2 of [8].

Our next theorem shows that the number of samples needed to recover accurately the signed support is actually optimal. That is, whenever the rescaled sample size satisfies a certain threshold, regardless of what the regularization parameters $\lambda_{n,\beta}$ and $\lambda_{n,e}$ are selected, no solution of the extended Lasso can correctly identify the signed supports with high probability.

Theorem 3 (Inachievability): Given the linear model (3) with random Gaussian design and the covariance matrix Σ satisfying invertibility and incoherence properties for any $\gamma \in (0,1)$, let $\eta, \delta \in (0,1)$ and the sequence (n,p,k,s) satisfies $s \geq \eta n$ and $n < \max\{n_1, n_2\}$, where n_1 and n_2 are defined as

$$n_1 \triangleq \frac{2(1-\delta)}{(1-\eta)}\frac{\rho_l k\log(p-k)}{C_{\max}(2-\gamma)^2}\left\{\frac{3}{8} + (1-\eta)^2\frac{\sigma^2 C_{\max}}{\lambda^2_{n,\beta}k}\right\}.$$

$$n_2 \triangleq \frac{(1-\delta)}{12}\frac{\eta}{(1-\eta)^2}\frac{\rho_l}{C_{\max}} \times \left(1 + \frac{2\sqrt{\sigma^2\log n}}{\lambda_{n,e}\sqrt{n}}\right)^{-2} k\log(n-s)\log(p-k).$$

Then, with probability tending to unity, no solution pair of the extended Lasso (7) has the correct signed support.

When the covariance matrix of the design matrix X is $\Sigma = I_{p\times p}$, or equivalently, entries of X are i.i.d. Gaussian $\mathcal{N}(0,1)$. In addition, assume that the regularization parameters λ_β, n, and $\lambda_{n,e}$ are chosen from the families of (20) and (21), respectively. That is, $\lambda_{\beta,n} = 8\sqrt{\frac{\sigma^2\eta\log n\log p}{n}}$ and $\lambda_{e,n} = 4\sqrt{\frac{\sigma^2\log n}{n}}$. Then, the theorem implies that the extended Lasso (7) is not able to achieve the correct signed support solution whenever the number of observations is less than

$$n \leq \max\left\{c_1\frac{1}{1-\eta}k\log(p-k),\ c_2\frac{\eta}{(1-\eta)^2}k\log(p-k)\log(1-\eta)n\right\}.$$

III. Illustrative Simulations

In this section, we conduct several simulations to show the recovery capability of the extended Lasso and to verify the conclusions of Theorems 2 and 3. In the first simulation, we compare the performance of the extended Lasso minimization with the original version and demonstrate the robustness of the proposed optimization to outliers. Next, we experimentally draw the phase transition curves with varying sample sizes and show that the results in Theorems 2 and 3 predict well the sharp transition between failure and success recovery of the extended Lasso. The last set of simulations demonstrates the error correction capacity with varying sample sizes and the fractions of error.

1) Comparison: Our first simulation compares the proposed extended Lasso with the conventional Lasso when some observations are corrupted by outliers. In this setup, the design matrix X of size $n\times p$ with $n = 200$ and $p = 512$ is generated from the Gaussian distribution with i.i.d. entries $\mathcal{N}(0,1)$. The regression vector $\beta^\star \in \mathbb{R}^p$ has $k = 30$ nonzero entries with magnitudes fixed to ones. This extreme case is considered more difficult to recover. We corrupt the observation vector $y = X\beta^\star$ by two sources of error: the dense noise w with i.i.d. entries $\mathcal{N}(0, 0.1)$ and the sparse error vector $e^\star$ with the number of nonzero entries $s = 20$. Magnitudes of nonzero entries of $e^\star$ are i.i.d. Gaussian random variables $\mathcal{N}(0, 10)$.

Fig. 1(a) plots the observation vectors y before (blue line) and after (red line) corrupted by the sparse error vector $e^\star$. As can be seen from Fig. 1(b) and (c), the original Lasso fails to recover the regression vector, while the extended Lasso successfully recovers $\beta^\star$. Notably, most of the nonzero coefficients are detected and their magnitudes are close to the true regression.

Furthermore, the extended Lasso is also able to exactly recover the sparse error, as demonstrated in Fig. 1(d). This implies that the locations of corrupted observations are identified. This property is particularly important in applications such as sensor network, where one wants to detect the defected sensors in the network.

2) Phase Transition With Varying Scaling Sample Size: Next, we provide several simulations to illustrate the capability of the extended Lasso in recovering the exact regression signed support when a significant fraction of observations is corrupted by large error. Simulations are performed for a range of parameters (n, p, k, s), where the design matrix X is uniform Gaussian random whose rows are i.i.d. $\mathcal{N}(0, I_{p\times p})$. For each fixed set of (n, p, k, s), we generate sparse vectors $\beta^\star$ and $e^\star$ where locations of nonzero entries are distributed uniformly at random. The dense noise w is selected with i.i.d. Gaussian $\mathcal{N}(0, \sigma)$ entries and with $\sigma = 0.1$ fixed in all experiments. Magnitudes of nonzero entries of $\beta^\star$ are selected to be $2\sigma\sqrt{\log p}$ or $-2\sigma\sqrt{\log p}$ with equal probability. Similarly, we specify the magnitudes of nonzero entries of $e^\star$ to be $2\sigma\sqrt{\log p \log n}$ or $-2\sigma\sqrt{\log p \log n}$ with the same probability.

In our experiments, we consider varying problem sizes $p = \{128, 256, 512\}$ and three types of regression sparsity indices: sublinear sparsity $(k = 0.2p/\log(0.2p))$, linear sparsity $(k = 0.2p)$, and fractional power sparsity $(k = 0.5p^{0.75})$. In all cases, we fixed the error support size $s = n/2$. This implies that half of the observations is corrupted, or in other word, the fraction of corruption $\eta = 1/2$. By this selection, Theorem 2 suggests that the number of samples is required to be $n \geq 2Ck\log(p-k)\log n$ to guarantee exact signed support recovery. We now investigate the success and failure of the extended Lasso in recovering the sparsity pattern by varying the sample size n. In the simulation, we scale n such that

$$\frac{n}{\log n} = 2\theta\frac{\eta}{(1-\eta)^2}k\log(p-k)$$

where the parameter θ varies from 0 to 2 with stepsize 0.1.

In the algorithm, we select $\lambda_{n,\beta} = 2\sqrt{\frac{\sigma^2\log p\log n}{n}}$ and $\lambda_{n,e} = 2\sqrt{\frac{\sigma^2\log n}{n}}$, as suggested by Theorem 2, where the noise level $\sigma = 0.1$ is fixed. The algorithm reports a success if the solution pair has the same signed support as $(\beta^\star, e^\star)$. Here, we declare entries of the solutions $\widehat{\beta}$ and $\widehat{e}$ to be zeros if their values are smaller than the noise level σ. Fig. 2 shows three panels corresponding to three types of regression sparsity levels: (a) sublinear sparsity, (b) linear sparsity, and (c) fractional power sparsity. Each panel depicts three curves associated with three problem sizes and each point on the curves represent the average of 100 trials.

As demonstrated by the simulation results, as the rescaling sample size $\theta \geq 1$ or the number of observations $n \geq 4k\log(p-k)\log n$, our extended Lasso is capable of recovering the exact signed support of both $\beta^\star$ and $e^\star$ even 50% of the observations are contaminated. As $\theta \leq 1$ or the number of observations $n \leq 4k\log(p-k)\log n$, the probability of success starts diving down to zero, implying the failure of the extended Lasso. This observation matches with the results in our Theorem 2 and 3.

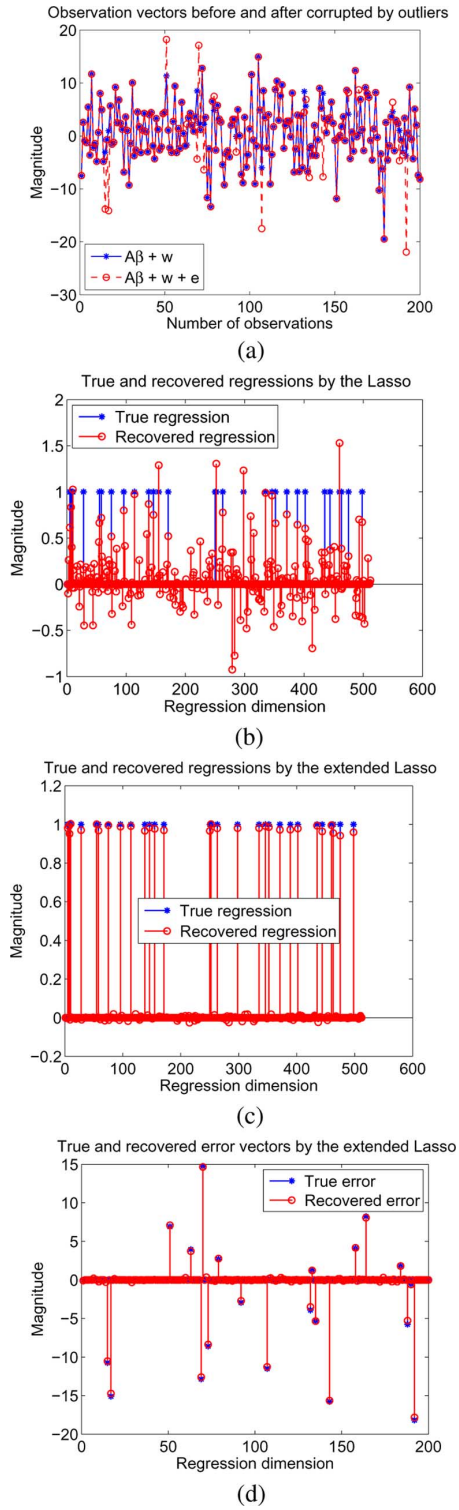


Fig. 1. Compare the original Lasso with the proposed extended Lasso in recovering the regression from corrupted observations. A regression vector $\beta^\star \in \mathbb{R}^p$ with $p = 512$ and sparsity $k = 30$ is generated. Here, nonzero entries have value ones. The design matrix $X \in R^{n\times p}$ with $n = 200$ is Gaussian distributed with entries being i.i.d. $\mathcal{N}(0, 1)$. (a) Observation vectors before and after corrupted by the sparse error $e^\star$. (b) Recovered regression from the original Lasso. (c) Recovered regression from the extended Lasso. (d) Recovered sparse error.

We also notice that from Fig. 2, the transition from failure to success of the extended Lasso gets sharper with larger problem sizes.

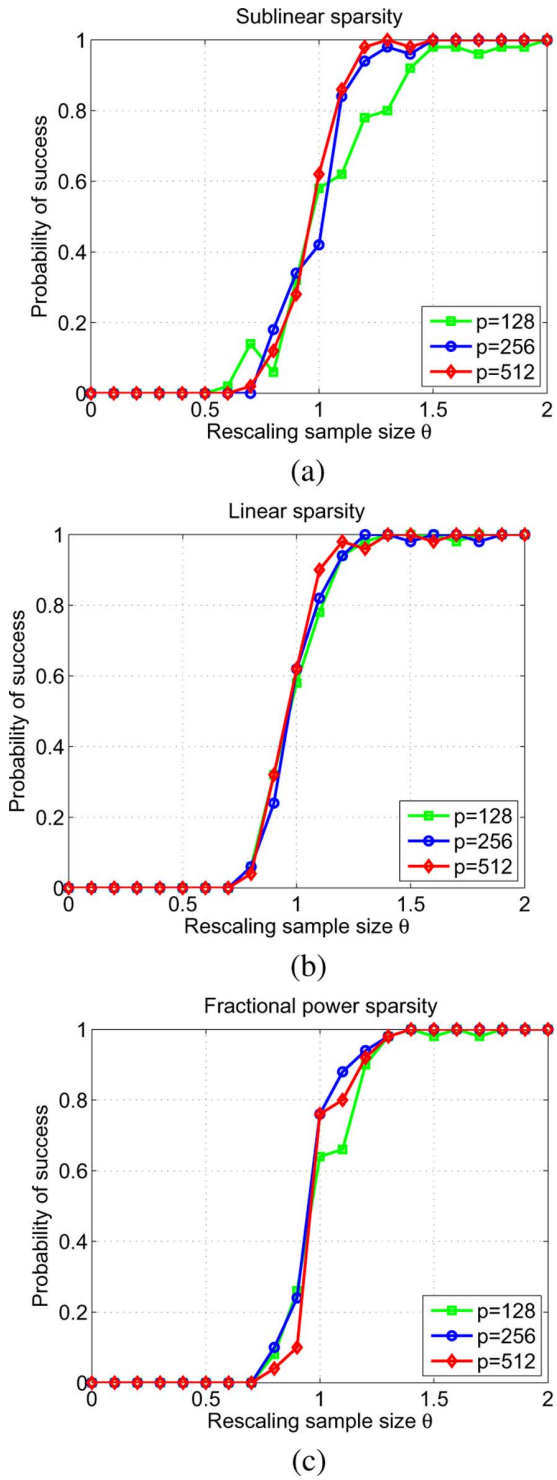


Fig. 2. Probability of success in recovering the signed supports versus the sample size $\theta = \frac{(1-\eta)^2}{\eta}\frac{n}{2k\log(p-k)\log n}$ with $\eta = 1/2$. The extended Lasso algorithm in (6) is performed with the uniform Gaussian design matrix. Three figures correspond to sublinear sparsity index $k = 0.2p/\log(0.2p)$, linear sparsity index $k = 0.1p$ and fractional power sparsity index $k = 0.5p^{0.75}$. Each figures shows three curves, corresponding to the problem size $p = \{128, 256, 512\}$.

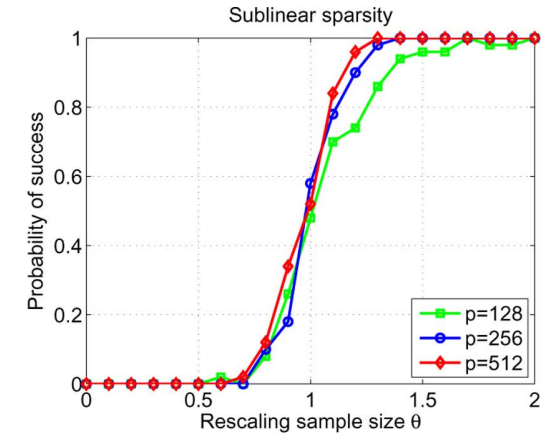
Referees have pointed out that it would be valuable to understand the behaviors of the extended Lasso algorithm with more structured design matrices, rather than the matrix with i.i.d. Gaussian entries. In the next experiment, we focus on this scenario in which simulations are conducted on the nonuniform Gaussian design ensemble with the Toeplitz covariance matrix Σ. The structure of Σ is defined as follows:

$$\Sigma = \begin{pmatrix} 1 & \zeta & \zeta^2 & \dots & \zeta^{p-2} & \zeta^{p-1} \\ \zeta & 1 & \zeta & \zeta^2 & \dots & \zeta^{p-2} \\ \zeta^2 & \zeta & 1 & \zeta & \dots & \zeta^{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \zeta^{p-1} & \zeta^{p-2} & \dots & \zeta^2 & \zeta & 1 \end{pmatrix} \tag{24}$$

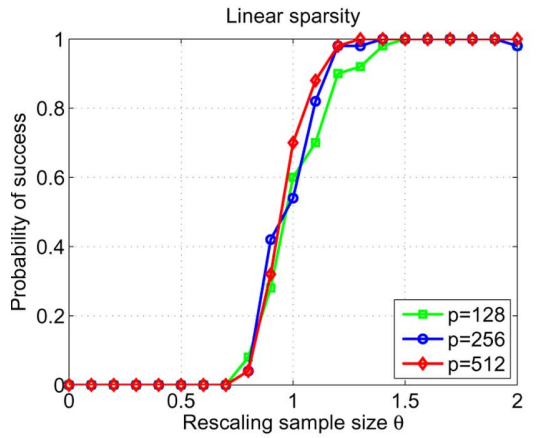
for some $\zeta \in (-1, 1)$. Similar as the previous experiment, three problem sizes $p = \{128, 256, 512\}$ and three types of regression sparsity levels are performed. Again, we fix the fraction of error to be $\eta = 1/2$, or $s = n/2$. Fig. 3 demonstrates three panels of curves according to: (a) sublinear sparsity $k = 0.2p/\log(0.2p)$, (b) linear sparsity $k = 0.2p$, and (c) fractional power sparsity $k = 0.5p^{0.75}$ for one particular value of $\zeta = 0.10$. As expected from our Theorems 2 and 3, the success rate consistently increases with larger sample sizes. Specifically, for values of $n \geq 4c_\Sigma k\log(p-k)\log n$ with the constant c_Σ depending on quantities $\rho_u(\Sigma)$, $C_{\min}(\Sigma)$, and $D^+_{\max}(\Sigma)$ defined in Section II-B, the extended Lasso successfully recovers the signed supports of both $\beta^\star$ and $e^\star$. The extended Lasso gradually fails for values of $n \leq 4c'_\Sigma k\log(p-k)\log n$.

In the previous experiments, the Gaussian design matrices are performed with strong theoretical support. In the next experiment, we use the Bernoulli ensemble as the design matrix and show a similar phase transition behavior as the Gaussian design matrix, predicting a universal law for the extended Lasso minimization in recovering the regression under grossly corrupted observations. Using the same setup as previous experiments and generating the Bernoulli random matrix whose entries receive values ± 1 with equal probability, we plot curves in Fig. 4. As one can see from Figs. 2–4, the phase transition is identical for the two completely different design matrices. This simulation result suggests a similar lower and upper bounds on the number of observations for the Bernoulli design matrix, as in Theorems 2 and 3.

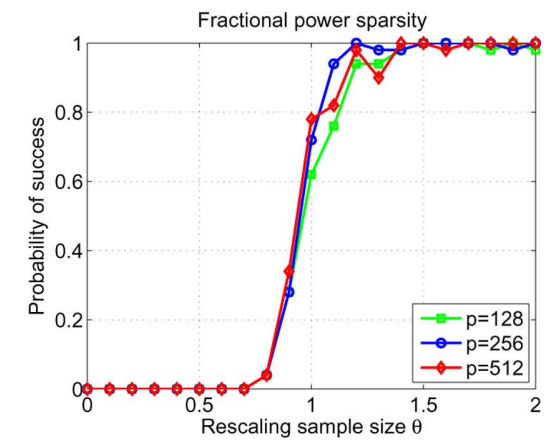
3) Error Correction Capability: While the previous experiments demonstrate the sharp phase transition between success and failure with varying values of sample sizes and fixed the fraction of error, Theorem 2 suggests that the fraction of error can approach one with increasing sample size. In the last experiment, we fix the problem size $p = 2048$ and generate the sparse vector $\beta^\star$ with sparsity level $k = 1$. We run the extended Lasso algorithm with different values of sample size $n = 200$, 400, 800, 1600 and with varying error fraction η from 0.3 to 1. For each fixed set (n, p, k, η), the experiment is performed 100 times. Fig. 5 plots the fraction of error versus the probability of success with varying sample sizes, where in the first panel, curves are drawn with the uniform Gaussian design matrix, the Gaussian design matrix with Toeplitz covariance family is applied to draw curves in the second panel, and the Bernoulli random matrix is applied to draw curves in the last panel. Again, the algorithm claims a success if the signed supports of both $\beta^\star$ and $e^\star$ are exactly recovered. As clearly illustrated in the figure, the extended Lasso algorithm allows more error fractions as the sample sizes increase. This agrees well with the intuition and with the results of Theorem 2.



(a)

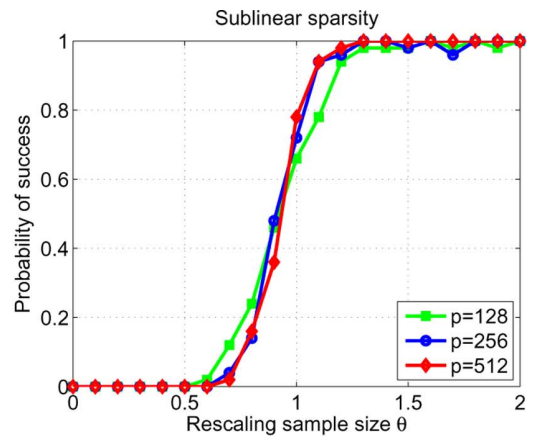


(b)

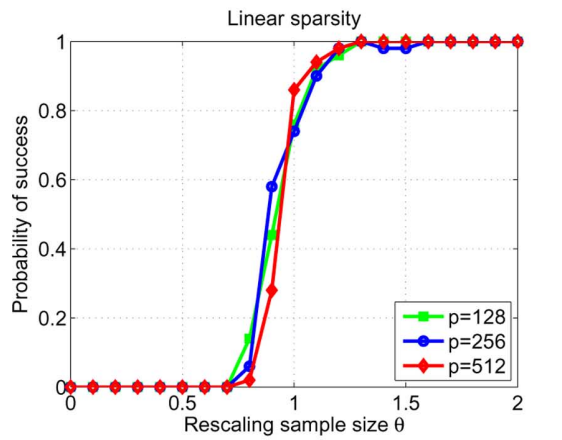


(c)

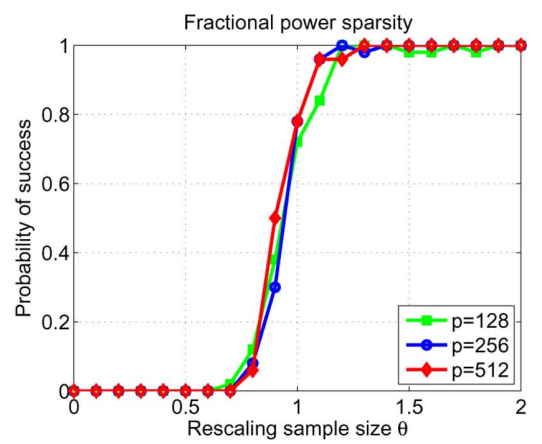
Fig. 3. Probability of success in recovering the signed supports versus the sample size $\theta = \frac{(1-\eta)^2}{\eta}\frac{n}{2k\log(p-k)\log n}$ with $\eta = 1/2$. The extended Lasso algorithm in (6) is performed with the Gaussian design matrix whose rows are $\mathcal{N}(0,\Sigma)$ with the covariance matrix Σ being Toeplitz (24). Three figures correspond to sublinear sparsity index $k = 0.2p/\log(0.2p)$, linear sparsity index $k = 0.1p$ and fractional power sparsity index $k = 0.5p^{0.75}$. Each figures shows three curves, corresponding to the problem size $p = \{128, 256, 512\}$.



(a)



(b)



(c)

Fig. 4. Probability of success in recovering the signed supports versus the sample size $\theta = \frac{(1-\eta)^2}{\eta}\frac{n}{2k\log(p-k)\log n}$ with $\eta = 1/2$. The extended Lasso algorithm in (6) is performed with the Bernoulli design matrix. Three figures correspond to sublinear sparsity index $k = 0.2p/\log(0.2p)$, linear sparsity index $k = 0.1p$ and fractional power sparsity index $k = 0.5p^{0.75}$. Each figures shows three curves, corresponding to the problem size $p = \{128, 256, 512\}$.

IV. Proof of Theorem 1 and Related Results

Proof of Theorem 1: Since $(\widehat{\beta}, \widehat{e})$ is the pair of the optimal solution of (6), we have

$$\frac{1}{2n}\left\|y - X\widehat{\beta} - \sqrt{n}\widehat{e}\right\|_2^2 + \lambda_{n,\beta}\left\|\widehat{\beta}\right\|_1 + \lambda_{n,e}\left\|\widehat{e}\right\|_1 \leq \frac{1}{2n}\left\|y - X\beta^\star - \sqrt{n}e^\star\right\|_2^2 + \lambda_{n,\beta}\left\|\beta^\star\right\|_1 + \lambda_{n,e}\left\|e^\star\right\|_1. \quad (25)$$

From $h = \widehat{\beta} - \beta^\star$ and $f = \widehat{e} - e^\star$, we can easily see that

$$\left\|y - X\widehat{\beta} - \sqrt{n}\widehat{e}\right\|_2^2 = \left\|y - X\beta^\star - \sqrt{n}e^\star\right\|_2^2 - 2\left\langle w, Xh + \sqrt{n}f\right\rangle + \left\|Xh + \sqrt{n}f\right\|_2^2.$$

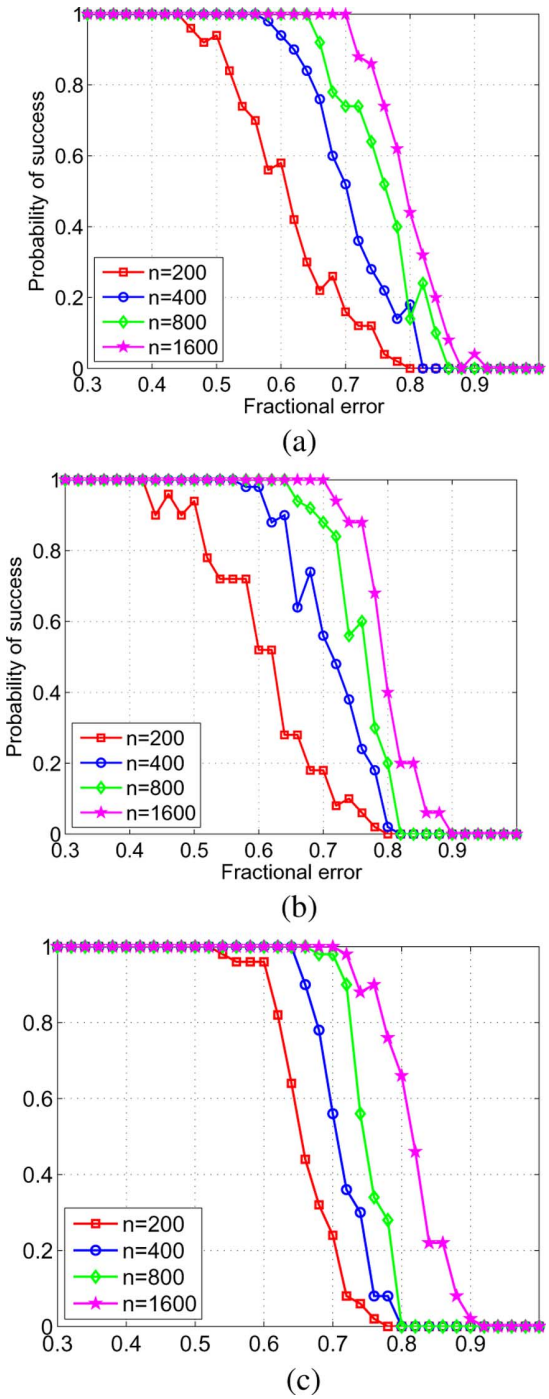


Fig. 5. Probability of success in recovering the signed supports with increasing error fraction and the sample sizes. The problem size $p = 2048$ and the regression sparsity level $k = 1$ are fixed. Error fraction η is varied from 0.4 to 1 with stepsize 0.02. Note that for value of $\eta < 0.4$, the extended Lasso always recover the signed support correctly. (a) Curves plotted with the uniform Gaussian design matrix. (b) Curved plotted with the Gaussian design matrix whose row are i.i.d. $(0, \Sigma)$ and Σ is the Toeplitz covariance matrix in (24) with $\zeta = 0.10$. (c) Curves plotted with the Bernoulli design matrix whose entries are ± 1 with equal probability. As n increases, the fraction of error that can be corrected via the extended Lasso approaches one.

Moreover, it is clear that

$$\begin{aligned}\|\beta^\star\|_1 - \left\|\widehat{\beta}\right\|_1 &= \|\beta^\star\|_1 - \|\beta^\star + h\|_1 \\ &= \|\beta^\star\|_1 - \|\beta^\star + h_T\|_1 - \|h_{T^c}\|_1 \\ &\le \|h_T\|_1 - \|h_{T^c}\|_1\end{aligned}$$

where the last inequality follows from the triangular inequality. Similarly, we can establish the bound with e

$$\|e^\star\|_1 - \|\widehat{e}\|_1 \le \|f_S\|_1 - \|f_{S^c}\|_1 .$$

Putting these pieces into (25), we obtain

$$\begin{aligned}&\frac{1}{2n}\left\|Xh + \sqrt{n}f\right\|_2^2 \\ &\le \frac{1}{n}\left\langle w, Xh + \sqrt{n}f\right\rangle + \lambda_{n,\beta}(\|h_T\|_1 - \|h_{T^c}\|_1) \\ &+ \lambda_{n,e}(\|f_S\|_1 - \|f_{S^c}\|_1) \\ &\le \frac{1}{n}\|X^*w\|_\infty \|h\|_1 + \frac{1}{\sqrt{n}}\|w\|_\infty \|f\|_1 \\ &+ \lambda_{n,\beta}(\|h_T\|_1 - \|h_{T^c}\|_1) + \lambda_{n,e}(\|f_S\|_1 - \|f_{S^c}\|_1) \\ &\le \left(\frac{\|X^*w\|_\infty}{n} + \lambda_{n,\beta}\right)\|h_T\|_1 - \left(\lambda_{n,\beta} - \frac{\|X^*w\|_\infty}{n}\right)\|h_{T^c}\|_1 \\ &+ \left(\frac{\|w\|_\infty}{\sqrt{n}} + \lambda_{n,e}\right)\|f_S\|_1 - \left(\lambda_{n,e} - \frac{\|w\|_\infty}{\sqrt{n}}\right)\|f_{S^c}\|_1 . \quad (26)\end{aligned}$$

By the choices of $\lambda_{n,\beta}$ and $\lambda_{n,e}$ in the theorem, we have $\frac{1}{n}\|X^*w\|_\infty \le \frac{\mu}{2}\lambda_{n,\beta} \le \frac{\lambda_{n,\beta}}{2}$ and $\frac{1}{\sqrt{n}}\|w\|_\infty \le \frac{\lambda_{n,e}}{2}$. Therefore

$$\begin{aligned}\frac{1}{2n}\left\|Xh + \sqrt{n}f\right\|_2^2 &\le \lambda_{n,\beta}\frac{3}{2}\|h_T\|_1 - \frac{\lambda_{n,\beta}}{2}\|h_{T^c}\|_1 \\ &+ \frac{3}{2}\lambda_{n,e}\|f_S\|_1 - \frac{1}{2}\lambda_{n,e}\|f_{S^c}\|_1 .\end{aligned}$$

The left-hand side is greater than zero, thus the error pair (h, f) belongs to the set $\mathbb{C}$ defined in (11). Hence, by the extended RE

$$\begin{aligned}\kappa_l^2(\|h\|_2 + \|f\|_2)^2 &\le 3\lambda_{n,\beta}\|h_T\|_1 + 3\lambda_{n,e}\|f_S\|_1 \\ &\le 3\lambda_{n,\beta}\sqrt{k}\|h\|_2 + \lambda_{n,e}\sqrt{s}\|f\|_2\end{aligned}$$

where the last inequality follows from the crude ℓ_1/ℓ_2 bound: $\|h_T\|_1 \le \sqrt{k}\|h\|_2$. If $\lambda\sqrt{s/k} \le 1$, the right-hand side is upper bounded by $3\lambda_{n,\beta}\sqrt{k}(\|h\|_2 + \|f\|_2)$. On the other hand, it is upper bounded by $3\lambda_{n,e}\sqrt{s}(\|h\|_2 + \|f\|_2)$ if $\lambda\sqrt{s/k} \ge 1$. Combining these pieces together, we conclude

$$\|h\|_2 + \|f\|_2 \le 3\kappa_l^{-2}\max\left\{\lambda_{n,\beta}\sqrt{k}, \lambda_{n,e}\sqrt{s}\right\}$$

which completes our proof. □

Proof of Lemma 1: Decompose $\frac{1}{n}\|Xh + \sqrt{n}f\|_2^2 = \frac{1}{n}\|Xh\|_2^2 + \|f\|_2^2 + \frac{2}{\sqrt{n}}\langle Xh, f\rangle$. In order to lower bound the left-hand side, our main tool is to control the lower bound of each term on the right-hand side.

To establish a lower bound of $\frac{1}{n}\|Xh\|_2^2$, we leverage an appealing result of [37]. This result stated that for any Gaussian random matrix X with i.i.d. $\mathcal{N}(0, \Sigma)$ rows, there exists universal positive constants c_1, c_2 such that the following inequality holds with probability greater than $1 - c_1\exp(-c_2 n)$:

$$\frac{1}{\sqrt{n}}\|Xv\|_2 \ge \frac{\sqrt{C_{\min}}}{4}\|v\|_2 - 9\sqrt{\xi(\Sigma)}\sqrt{\frac{\log p}{n}}\|v\|_1 \quad (27)$$

for $\forall\, v \in \mathbb{R}^p$. Here, we remind the reader of the notation $\xi(\Sigma) = \max_{j=1,\ldots,d}\Sigma_{jj}$ and $C_{\min} = \lambda_{\min}(\Sigma)$.

We now apply this inequality for the error vector h in the set $\mathbb{C}$. Since $h \in \mathbb{C}$, we have

$$\|h\|_1 \leq 4\|h_T\|_1 + 3\lambda\|f_S\|_1 \leq 4\sqrt{k}\|h\|_2 + 3\lambda\sqrt{s}\|f\|_2.$$

Next taking advantage of (27) yields

$$\frac{1}{\sqrt{n}}\|Xh\|_2 \geq \left(\frac{\sqrt{C_{\min}}}{4} - 36\sqrt{\frac{\xi k\log p}{n}}\right)\|h\|_2 - 27\lambda\sqrt{\frac{\xi s\log p}{n}}\|f\|_2$$

where we denote the shorthand notation $\xi \stackrel{\triangle}{=} \xi(\Sigma)$. This inequality leads to

$$\frac{1}{\sqrt{n}}\|Xh\|_2 + \|f\|_2 \geq \left(\frac{\sqrt{C_{\min}}}{4} - 36\sqrt{\frac{\xi k\log p}{n}}\right)\|h\|_2 + \left(1 - 27\lambda\sqrt{\frac{\xi s\log p}{n}}\right)\|f\|_2.$$

From the assumption on n of the lemma, the first term in the bracket of the right-hand side equation is greater than $\sqrt{C_{\min}}/8$. In addition, the assumption on s guarantees that the second term in the bracket is greater than $1/2$. Thus, $\frac{1}{\sqrt{n}}\|Xh\|_2 + \|f\|_2 \geq \frac{\sqrt{C_{\min}}}{8}\|h\|_2 + \frac{1}{2}\|f\|_2$; or equivalently $\frac{1}{n}\|Xh\|_2^2 + \|f\|_2^2 \geq \frac{C_{\min}}{128}\|h\|_2^2 + \frac{1}{8}\|f\|_2^2 \geq \frac{1}{128}(C_{\min}\|h\|_2^2 + \|f\|_2^2)$.

Combining this result with the inequality in the following lemma 2, we conclude that

$$\frac{1}{n}\left\|Xh + \sqrt{n}f\right\|_2^2 \geq \frac{1}{256}(C_{\min}\|h\|_2^2 + \|f\|_2^2) \geq \frac{1}{256}\min\{C_{\min}, 1\}(\|h\|_2^2 + \|f\|_2^2)$$

as claimed. □

Lemma 2: Consider the random Gaussian design matrix X whose rows are i.i.d. $\mathcal{N}(0,\Sigma)$. Assume that $s \leq \min\{\frac{\xi}{C_{\min}}\frac{k\log p}{\mu^2\log n}, C_1\frac{C_{\min}}{C_{\max}}n\}$ and $n \geq C_2\frac{\xi C_{\max}}{C_{\min}^2}k\log p$ for some sufficiently small constant C_1 and sufficiently large constant C_2; then the following inequality holds with probability greater than $1 - \exp(-cn)$:

$$\frac{2}{\sqrt{n}}|\langle Xh, f\rangle| \leq \frac{1}{256}(C_{\min}\|h\|_2^2 + \|f\|_2^2).$$

Proof: Divide the set $\{1,2,\ldots,p\}$ into subset $T_1, T_2, \ldots, T_q$ of size k such that the first set T_1 contains k entries of h indexed by T, the set T_2 contains k largest absolute entries of the vector h_{T^c}, T_3 contains the second k largest absolution entries of h_{T^c}, and so on. By the same strategy, we also divide the set $\{1,2,\ldots,n\}$ into subset $S_1, S_2, \ldots, S_r$ such that the first set S_1 contains s entries of f indexed by S and sets $S_2, S_3, \ldots$ are of size $s' \geq s$.

We now have

$$\frac{1}{\sqrt{n}}|\langle Xh, f\rangle| \leq \sum_{i,j}\frac{1}{\sqrt{n}}|\langle X_{S_iT_j}h_{T_j}, f_{S_i}\rangle| \leq \max_{ij}\frac{1}{\sqrt{n}}\left\|X_{S_iT_j}\right\|\sum_{ij}\|h_{T_j}\|_2\|f_{S_i}\|_2.$$

Notice that the matrix $X_{S_iT_j}$ is the random Gaussian matrix whose rows are $\mathcal{N}(0,\Sigma_{T_jT_j})$. By the random Gaussian matrix concentration in Lemma 14 in Appendix D, we have with probability greater than $1 - 2\exp(-\tau^2 s'/2)$:

$$\left\|X_{S_iT_j}\right\| \leq \left\|\Sigma_{T_jT_j}^{1/2}\right\|\left(\sqrt{k} + \sqrt{s'} + \tau\sqrt{s'}\right).$$

Choosing $\tau = \tau'\sqrt{\frac{n}{s'}}$ and taking the union bound over all possibility of T_j and S_i, we have this inequality holds with probability at least $1 - 2\binom{n}{s}\binom{p}{k}\exp(-\tau'^2 n/2)$. Assuming that $n \geq c_1^{-1}k\log(p/k)$, we have $\binom{p}{k} \leq \left(\frac{ep}{k}\right)^k \leq \exp(c_1 n)$. In addition, assuming $n \geq c_2^{-1}s'\log(n/s')$, we have $\binom{n}{s'} \leq \left(\frac{en}{s'}\right)^s \leq \exp(c_2 n)$. Therefore, with sufficiently small constants c_1 and c_2, we get

$$\max_{ij}\frac{1}{\sqrt{n}}\left\|X_{S_iT_j}\right\| \leq \sqrt{C_{\max}}\left(\sqrt{\frac{k}{n}} + \sqrt{\frac{s'}{n}} + \tau'\right)$$

with probability at least $1 - \exp(-(\tau'^2/2 - c_1 - c_2)n)$ where we recall the definition of $C_{\max} \stackrel{\triangle}{=} \lambda_{\max}(\Sigma)$.

A standard bound in [40] gives us $\sum_{i=3}^q\|h_{T_i}\|_2 \leq k^{-1/2}\|h_{T^c}\|_1$. In addition, since h belongs to the set $\mathbb{C}$ defined in (11), $\|h_{T^c}\|_1 \leq 3\sqrt{k}\|h\|_2 + 3\lambda\sqrt{s}\|f\|_2$. Hence

$$\sum_{i=1}^q\|h_{T_i}\|_2 \leq 2\|h\|_2 + \sum_{i=3}^q\|h_{T_i}\|_2 \leq 5\|h\|_2 + 3\lambda\sqrt{\frac{s}{k}}\|f\|_2.$$

Similar manipulations along with the choice of $s' \geq s$ also yields

$$\sum_{i=3}^r\|f_{S_i}\|_2 \leq s'^{-1/2}\|f_{S^c}\|_1 \leq \frac{3}{\lambda}\sqrt{\frac{k}{s'}}\|h\|_2 + 3\|f\|_2$$

leading to

$$\sum_{i=1}^r\|f_{S_i}\|_2 \leq \frac{3}{\lambda}\sqrt{\frac{k}{s'}}\|h\|_2 + 5\|f\|_2.$$

Hence, with probability greater than $1 - \exp(-(\tau'^2/2 - c_1 - c_2)n)$, $\frac{1}{\sqrt{n}}|\langle Xh, f\rangle|$ is upper bounded by

$$\sqrt{C_{\max}}\left(\sqrt{\frac{k}{n}} + \sqrt{\frac{s'}{n}} + \tau'\right) \times \left(5\|h\|_2 + 3\lambda\sqrt{\frac{s'}{k}}\|f\|_2\right)\left(\frac{3}{\lambda}\sqrt{\frac{k}{s'}}\|h\|_2 + 5\|f\|_2\right).$$

We select $s' \stackrel{\triangle}{=} \frac{\xi}{C_{\min}}\frac{k\log p}{\mu^2\log n}$. This choice of s' leads to $\lambda\sqrt{\frac{s'}{k}} = \sqrt{C_{\min}}$ and $\frac{1}{\lambda}\sqrt{\frac{k}{s'}} = \frac{1}{\sqrt{C_{\min}}}$. Therefore

$$\frac{1}{\sqrt{n}}|\langle Xh, f\rangle| \leq 25\frac{\sqrt{C_{\max}}}{\sqrt{C_{\min}}}\left(\sqrt{\frac{k}{n}} + \sqrt{\frac{s'}{n}} + \tau'\right) \times (\sqrt{C_{\min}}\|h\|_2 + \|f\|_2)^2 \leq \frac{1}{256}(\sqrt{C_{\min}}\|h\|_2 + \|f\|_2)^2$$

for sufficiently small τ and $n \geq C\frac{C_{\max}}{C_{\min}}s'$ with C sufficiently large, $C > 25 \times 256$. □

V. Proof of Theorem 2—Achievability

By the Karush–Kuhn–Tucker (KKT) condition, $\widehat{\beta}$ and $\widehat{e}$ is a pair of solution of (6) if and only if the following set of equations satisfies:

$$-\frac{1}{n}X^*(y - X\widehat{\beta} - \sqrt{n}\widehat{e}) + \lambda_{n,\beta} z^{(\beta)} = 0 \tag{28}$$

$$-\frac{1}{\sqrt{n}}(y - X\widehat{\beta} - \sqrt{n}\widehat{e}) + \lambda_{n,e} z^{(e)} = 0, \tag{29}$$

where $z^{(\beta)}$ and $z^{(e)}$ are elements of the subgradients of the ℓ_1-norm evaluated at $\widehat{\beta}$ and $\widehat{e}$, respectively. It has been well established that $(\widehat{\beta}, \widehat{e})$ is the unique solution to the extended Lasso program if

$$\begin{cases} \frac{1}{n}X_i^*(y - X\widehat{\beta} - \sqrt{n}\widehat{e}) = \lambda_{n,\beta}\operatorname{sgn}(\widehat{\beta}_i), & \text{for } \widehat{\beta}_i \neq 0 \\ |z_i^{(\beta)}| = \frac{1}{n\lambda_{n,\beta}}|X_i^*(y - X\widehat{\beta} - \sqrt{n}\widehat{e})| < 1, & \text{for } \widehat{\beta}_i = 0 \end{cases} \tag{30}$$

and

$$\begin{cases} \frac{1}{\sqrt{n}}(y_i - X_i\widehat{\beta} - \sqrt{n}\widehat{e}_i) = \lambda_{n,e}\operatorname{sgn}(\widehat{e}_i), & \text{for } \widehat{e}_i \neq 0 \\ |z_i^{(e)}| = \frac{1}{\sqrt{n}\lambda_{n,e}}|y_i - X_i\widehat{\beta} - \sqrt{n}\widehat{e}_i| < 1, & \text{for } \widehat{e}_i = 0. \end{cases} \tag{31}$$

We will show that under the assumptions of Theorem 2, the solution pair of the extended Lasso is given by $(\widehat{\beta}, \widehat{e}) = (\beta^\star + h, e^\star + g)$, where $h_{T^c} = 0$, $g_{S^c} = 0$ and

$$\begin{aligned} h_T = &-(X_{S^cT}^* X_{S^cT})^{-1} \\ &\times [X_{S^cT}^* w_{S^c} - \sqrt{n}\lambda_{n,e} X_{ST}^* \operatorname{sgn}(e_S^\star) + n\lambda_{n,\beta}\operatorname{sgn}(\beta_T^\star)] \end{aligned} \tag{32}$$

and

$$\begin{aligned} g_S = &\frac{1}{\sqrt{n}} X_{ST}(X_{S^cT}^* X_{S^cT})^{-1} \\ &\times [X_{S^cT}^* w_{S^c} - \sqrt{n}\lambda_{n,e} X_{ST}^* \operatorname{sgn}(e_S^\star) + n\lambda_{n,\beta}\operatorname{sgn}(\beta_T^\star)] \\ &+ \frac{1}{\sqrt{n}} w_S - \lambda_{n,e}\operatorname{sgn}(e_S^\star). \end{aligned} \tag{33}$$

The expressions of h_T and g_S in the above equations are obtained by solving the KKT conditions (28) and (29) restricted on $\widehat{\beta}_{T^c} = 0$ and $\widehat{e}_{S^c} = 0$ together with setting $z_T^{(\beta)} = \operatorname{sgn}(\beta_T^\star)$ and $z_S^{(e)} = \operatorname{sgn}(e_S^\star)$. We note that due to the conditions of the sample size n and the fraction of errors in Theorem 2, $X_{S^cT}^* X_{S^cT}$ is invertible thanks to the random Gaussian matrix concentration inequalities (see Lemma 14 in Appendix D). Therefore, the expressions of h_T and g_S are valid.

To confirm that $(\widehat{\beta}, \widehat{e})$ is the optimal solution of the extended Lasso (6), in the following sections, we will check that $\widehat{\beta}$ and $\widehat{e}$ chosen above obey conditions (30) and (31). In particular,

1) In Section V-A, we show that $\left\|z_{T^c}^{(\beta)}\right\|_\infty < 1$.
2) In Section V-B, we show that $\left\|z_{S^c}^{(e)}\right\|_\infty < 1$.
3) In Section V-C, we establish that $\|h_T\|_\infty \le f_\beta(\lambda_{n,\beta})$. It then follows from the assumptions of Theorem 2 that $\|h_T\|_\infty < \min_{i\in T}|\beta_i^\star|$ and, therefore, $\operatorname{supp}(\widehat{\beta}_T) = \operatorname{supp}(\beta_T^\star)$ and $\operatorname{sgn}(\widehat{\beta}_T) = \operatorname{sgn}(\beta_T^\star)$.
4) In Section V-D, we establish that $\|g_S\|_\infty \le f_e(\lambda_{n,\beta}, \lambda_{n,e})$. It then follows from the assumptions Theorem 2 that $\|g_S\|_\infty < \min_{i\in S}|e_i^\star|$ and, therefore, $\operatorname{supp}(\widehat{e}_S) = \operatorname{supp}(e_S^\star)$ and $\operatorname{sgn}(\widehat{e}_S) = \operatorname{sgn}(e_S^\star)$.

A. Verify the Upper Bound of $\left\|Z_{T^c}^{(\beta)}\right\|_\infty$

Proof: First, we define a notation which will be used throughout the rest of the paper. Let $\lambda \triangleq \frac{\lambda_{n,e}}{\lambda_{n,\beta}}$. By the definition of $\lambda_{n,\beta}$ and $\lambda_{n,e}$ in (20), we have

$$\lambda = \frac{\gamma}{2\sqrt{\max\{\rho_u, D_{\max}^+\}}}\sqrt{\frac{1}{\eta\log p}} \tag{34}$$

where we introduce another shorthand notation $\rho_u \triangleq \rho_u(\Sigma_{T^c|T})$.

From the expression of $\widehat{\beta} = \beta^\star + h$ and $\widehat{e} = e^\star + g$ with $h_{T^c} = 0$, $g_{S^c} = 0$ and h_T, g_S defined in (32) and (33), we substitute into $z_{T^c}^{(\beta)} = \frac{1}{\lambda_{n,\beta}} X_{T^c}^*(y - X\widehat{\beta} - \widehat{e})$ together with noticing that $X_{T^c}^* X_T - X_{ST^c}^* X_{ST} = X_{S^cT^c}^* X_{S^cT}$, $X_{T^c}^* w - X_{ST^c}^* w_S = X_{S^cT^c}^* w_{S^c}$ to arrive at

$$\begin{aligned} z_{T^c}^{(\beta)} = &\frac{1}{n\lambda_{n,\beta}} X_{S^cT^c}^* \Pi_{S^cT} w_{S^c} \\ &- X_{S^cT^c}^* X_{S^cT}(X_{S^cT}^* X_{S^cT})^{-1} z \\ &+ \frac{1}{\sqrt{n}}\lambda X_{ST^c}^* \operatorname{sgn}(e_S^\star). \end{aligned} \tag{35}$$

Here, we define $\Pi_{S^cT} \triangleq I - X_{S^cT}(X_{S^cT}^* X_{S^cT})^{-1} X_{S^cT}^*$ which is an orthogonal projection onto the column space of X_{S^cT} and $z \triangleq \frac{1}{\sqrt{n}}\lambda X_{ST}^* \operatorname{sgn}(e_S^\star) - \operatorname{sgn}(\beta_T^\star)$.

We can further simplify the expression of $z_{T^c}^{(\beta)}$ by denoting

$$v \triangleq \begin{pmatrix} \frac{1}{\sqrt{n}}\lambda \operatorname{sgn}(e_S^\star) \\ \frac{1}{n\lambda_{n,\beta}}\Pi_{S^cT} w_{S^c} - X_{S^cT}(X_{S^cT}^* X_{S^cT})^{-1} z \end{pmatrix}. \tag{36}$$

Then, we have

$$z_{T^c}^{(\beta)} = [X_{ST^c}^* \quad X_{S^cT^c}^*] v = X_{T^c}^* v. \tag{37}$$

Conditioning on X_T, the matrix $X_{T^c}^*$ can be decomposed into a linear prediction plus a prediction error

$$X_{T^c}^* = \Sigma_{T^cT}\Sigma_{TT}^{-1} X_T^* + E_{T^c}^* \tag{38}$$

where E_{T^c} is a random matrix whose rows are i.i.d. and each row is a $\mathcal{N}(0, \Sigma_{T^c|T})$ Gaussian random vector with $\Sigma_{T^c|T}$ defined in (19). Therefore, $z_{T^c}^{(\beta)}$ consists of two components in which the first is

$$a \triangleq \Sigma_{T^cT}\Sigma_{TT}^{-1} X_T^* v$$

and the second is

$$b \triangleq E_{T^c}^* v. \tag{39}$$

Since Π_{S^cT} is the orthogonal projection onto the space spanned by columns of the matrix X_{S^cT}, we have $X_{S^cT}^* \Pi_{S^cT} = 0$. Thus, a can be simplified as

$$\begin{aligned} a &= \frac{1}{\sqrt{n}}\Sigma_{T^cT}\Sigma_{TT}^{-1}(\lambda X_{ST}^* \operatorname{sgn}(e_S^\star)) - \Sigma_{T^cT}\Sigma_{TT}^{-1} z \\ &= \Sigma_{T^cT}\Sigma_{TT}^{-1}\operatorname{sgn}(\beta_T^\star). \end{aligned} \tag{40}$$

The mutual incoherent assumption in (18) gives us $\|a\|_\infty \leq 1-\gamma$. All that left is to establish the ℓ_∞-norm of the second component: $\|b\|_\infty \leq \gamma$. Denote E_i as the i-th column of the matrix E_{T^c} and condition on X_{S^cT}, the ith coefficient of the vector $b: b_i = \langle E_i, v\rangle$ is a Gaussian random variable with variance $\mathrm{Var}(b_i) = v^*\mathbb{E}E_iE_i^*v \leq \rho_u \|v\|_2^2$ where $\|v\|_2^2$ is quantified as

$$M \stackrel{\triangle}{=} \frac{\lambda^2 s}{n} + \frac{1}{n^2\lambda_{n,\beta}^2}\|\Pi_{S^cT}w_{S^c}\|_2^2 + z^*(X_{S^cT}^*X_{S^cT})^{-1}z. \quad (41)$$

We state two supporting lemmas whose proof are deferred to the end of this section.

Lemma 3: Denote $z = \frac{1}{\sqrt{n}}\lambda X_{ST}^*\operatorname{sgn}(e_S^\star) - \operatorname{sgn}(\beta_T^\star)$. Define the event

$$\mathcal{E}_z \stackrel{\triangle}{=} \left\{ \|z\|_\infty \leq \lambda\sqrt{\frac{D_{\max}^+ s\log p}{n}} + 1 \right\}.$$

Then, $\mathbb{P}(\mathcal{E}_z) \geq 1 - 2\exp(-\log p)$.

Lemma 4: For any $\epsilon \in (0,1)$, define the event $\overline{\mathcal{E}} \stackrel{\triangle}{=} \{M \leq \overline{M}\}$, where

$$\overline{M} \stackrel{\triangle}{=} \frac{1}{n}\lambda^2 s + \left(1 + \max\left\{\epsilon, 4\sqrt{\frac{k}{n-s}}\right\}\right) \times \left(\frac{\sigma^2(n-s)}{n^2\lambda_{n,\beta}^2} + \frac{k\left(1+\lambda\sqrt{\frac{D_{\max}^+ s\log p}{n}}\right)^2}{(n-s)C_{\min}}\right). \quad (42)$$

Then, $\mathbb{P}(\overline{\mathcal{E}}) \geq 1 - c_1\exp(-c_2(n-s)\epsilon^2\})$ for some universal constants $c_1, c_2 > 0$.

Conditioned on the event $\overline{\mathcal{E}}$ defined in Lemma 4, the probability $\mathbb{P}(\max_{i\in T^c}|b_i| \geq \gamma)$ is upper bounded by

$$\mathbb{P}(\max_{i\in T^c}|b_i| \geq \gamma|\overline{\mathcal{E}}) + \exp(-c_2(n-s)).$$

We recall that b_i is a zero-mean Gaussian random variable; thus, the standard Gaussian tail bound in (63) allows us to derive

$$\mathbb{P}(\max_{i\in T^c}|b_i| \geq \gamma|\overline{\mathcal{E}}) \leq 2(p-k)\exp\left(-\frac{\gamma^2}{2\rho_u\overline{M}}\right).$$

This exponential probability decays at the rate of $\exp(-c\log(p-k))$ provided that $\frac{1}{\gamma^2}2\rho_u\overline{M}\log(p-k)$ is strictly less than one. Now we replace the definition of $\overline{M}$ in (42) into this inequality. To do this, we notice that $k \leq n-s$ from the sample size assumption of Theorem 2; thus, we can select $\epsilon \in (0,1)$ such that $4\sqrt{\frac{k}{n-s}} \leq \epsilon$. Following some simple algebra, we find that it is sufficient to have

$$\frac{n-s}{1+\epsilon} > \frac{2\rho_u}{C_{\min}\gamma^2}k\log(p-k) \times \left\{\frac{C_{\min}(n-s)}{(1+\epsilon)k}\frac{\lambda^2 s}{n} + \left(1+\lambda\sqrt{\frac{D_{\max}^+ s\log p}{n}}\right)^2 + \frac{(n-s)^2}{n^2}\frac{\sigma^2 C_{\min}}{\lambda_{n,\beta}^2 k}\right\}.$$

Replace the expression of λ in (34) and $s = \eta n$ and perform some simple algebra, we conclude that the ℓ_∞-norm of $z_{T^c}^{(\beta)}$ is strictly less than one as long as the following bound of the sample size obeys:

$$\frac{n}{2(1+\epsilon)} > \frac{1}{(1-\eta)}\frac{2\rho_u}{C_{\min}\gamma^2}k\log(p-k) \times \left\{\frac{9}{4} + (1-\eta)^2\frac{\sigma^2 C_{\min}}{\lambda_{n,\beta}^2 k}\right\}$$

which matches with the assumption of Theorem 2. □

Proof of Lemma 3: Recall the expression of z in the lemma, we have by the triangular inequality, $\|z\|_\infty \leq \frac{\lambda}{\sqrt{n}}\|X_{ST}^*\operatorname{sgn}(e_S^\star)\|_\infty + 1$. Furthermore, we know that the matrix X_{ST} can be represented as $W_{ST}\Sigma_{TT}^{1/2}$ where $W_{ST} \in \mathbb{R}^{s\times k}$ is the random matrix with i.i.d. zero mean entries and unit variance. Hence

$$\|X_{ST}^*\operatorname{sgn}(e_S^\star)\|_\infty = \left\|\Sigma_{TT}^{1/2}W_{ST}^*\operatorname{sgn}(e_S^\star)\right\|_\infty \leq \sqrt{D_{\max}^+}\|W_{ST}^*\operatorname{sgn}(e_S^\star)\|_\infty$$

where the inequality follows from matrix submultiplicative norm and $\left\|\Sigma_{TT}^{1/2}\right\|_\infty \leq \|\Sigma_{TT}\|_\infty^{1/2} = \sqrt{D_{\max}^+}$.

Consider the random variable $V_i = \langle w_i, \operatorname{sgn}(e_S^\star)\rangle$ where w_i is a column vector of W_{ST}. Recall that each entry of w_i is $\mathcal{N}(0,1)$ and $\|\operatorname{sgn}(e_S^\star)\|_2 = \sqrt{s}$. Hence, V_i is a Gaussian r.v. with variance s. Applying Gaussian tail bound (63) in the Appendix together with taking the union bound yields

$$\mathbb{P}\left(\|W_{ST}^*\operatorname{sgn}(e_S^\star)\|_\infty \geq \tau\right) \leq 2k\exp(-\tau^2/2s).$$

Selecting $\tau = 2\sqrt{s\log p}$ so that the probability exponentially decays to zero. Combining these inequalities completes the proof of Lemma 3. □

Proof of Lemma 4: Since Π_{S^cT} is the orthogonal projection matrix, we have $\|\Pi_{S^cT}w_{S^c}\|_2^2 \leq \|w_{S^c}\|_2^2$. In addition, $\frac{1}{\sigma^2}\|w_{S^c}\|_2^2$ is the χ^2-variate with $(n-s)$ degrees of freedom; thus

$$\mathbb{P}\left(\frac{1}{n^2\lambda_{n,\beta}^2}\|\Pi_{S^cT}w_{S^c}\|_2^2 \geq (1+\epsilon)\frac{\sigma^2(n-s)}{n^2\lambda_{n,\beta}^2}\right) \leq 2\exp\left(-\frac{3(n-s)\epsilon^2}{16}\right).$$

Turning to the last term of M, by the spectral norm bound of the Gaussian random matrix (68), we obtain

$$z^*(X_{S^cT}^*X_{S^cT})^{-1}z \leq \left(1+4\sqrt{\frac{k}{n-s}}\right)\frac{\|z\|_2^2}{(n-s)C_{\min}}$$

with probability greater than $1 - c_1\exp(-c_2(n-s))$. Conditioned on the event $\mathcal{E}_z$ in Lemma 3, we have $\|z\|_2^2 \leq k\|z\|_\infty^2 \leq k\left(1+\lambda\sqrt{\frac{D_{\max}^+ s\log p}{n}}\right)^2$. The proof is completed by combining these bounds. □

B. Verify the Upper Bound of $\left\|Z_{S^c}^{(e)}\right\|_\infty$

Proof: By replacing expressions of $\widehat{\beta}$ and $\widehat{e}$ into $z_{S^c}^{(e)} = \frac{1}{\sqrt{n}\lambda_{n,e}}(y_{S^c} - X_{S^c}\widehat{\beta})$, we get

$$z_{S^c}^{(e)} = \frac{1}{\sqrt{n}\lambda_{n,e}}\Pi_{S^cT}w_{S^c} + \frac{\sqrt{n}}{\lambda}X_{S^cT}(X_{S^cT}^*X_{S^cT})^{-1}z \quad (43)$$

where we use the same notations of Π_{S^cT} and z as in the previous section: $\Pi_{S^cT} \triangleq I - X_{S^cT}(X_{S^cT}^*X_{S^cT})^{-1}X_{S^cT}^*$ and $z \triangleq \frac{1}{\sqrt{n}}\lambda X_{ST}^*\operatorname{sgn}(e_S^\star) - \operatorname{sgn}(\beta_T^\star)$. To show that $\left\|z_{S^c}^{(e)}\right\|_\infty < 1$, we bound ℓ_∞-norm of each term of the sum (43) separately. In particular, we will establish that with probability converging to one, the ℓ_∞-norm of the first term is bounded by $\frac{2\sigma\sqrt{\log n}}{\lambda_{n,e}\sqrt{n}}$ and that of the second term is less than $(1 - \frac{2\sigma\sqrt{\log n}}{\lambda_{n,e}\sqrt{n}})$. The proof is, therefore, completed by the triangular inequality.

We begin by establishing the ℓ_∞-norm of the first term of $z_{S^c}^{(e)}$ in (43):

$$\frac{1}{\sqrt{n}\lambda_{n,e}}\left\|\Pi_{S^cT}w_{S^c}\right\|_\infty = \max_i \frac{1}{\sqrt{n}\lambda_{n,e}}\left|\langle u_i, w_{S^c}\rangle\right|$$

where u_i is a column vector of Π_{S^cT}. Since $\frac{1}{\sqrt{n}\lambda_{n,e}}\langle u_i, w_{S^c}\rangle$ is a sum of Gaussian random variables with zero mean and variance $\frac{\sigma^2}{n\lambda_{n,e}^2}\|u_i\|_2^2$, it can be bounded by the Gaussian tail inequality in (63) in Appendix D. Notice that spectral norm of any orthogonal projection is one, $\|u_i\|_2 \le 1$. We have

$$\mathbb{P}\left(\frac{1}{\sqrt{n}\lambda_{n,e}}\left|\langle u_i, w_{S^c}\rangle\right| \ge \tau\right) \le 2\exp\left(-\frac{n\lambda_{n,e}^2\tau^2}{2\sigma^2}\right).$$

Choose $\tau = \frac{2\sigma\sqrt{\log n}}{\sqrt{n}\lambda_{n,e}}$ and take the union bound over all $|S^c|$ columns of the matrix Π_{S^cT}, we have

$$\mathbb{P}\left(\frac{1}{\sqrt{n}\lambda_{n,e}}\left\|\Pi_{S^cT}w_{S^c}\right\|_\infty \ge \frac{2\sigma\sqrt{\log n}}{\sqrt{n}\lambda_{n,e}}\right) \le 2|S^c|\exp\left(-2\log n\right). \quad (44)$$

Next, we control the upper bound of $\frac{\sqrt{n}}{\lambda}\left\|X_{S^cT}(X_{S^cT}^*X_{S^cT})^{-1}z\right\|_\infty$. The following lemma, whose proof is deferred to Appendix A, establishes this bound.

Lemma 5: Under the assumptions of Theorem 2, for any vector $z \in \mathbb{R}^k$ independent with X_{S^cT}, the following statement holds:

$$\frac{\sqrt{n}}{\lambda}\left\|X_{S^cT}(X_{S^cT}^*X_{S^cT})^{-1}z\right\|_\infty < \frac{2}{3}\left(1 - \frac{2\sigma\sqrt{\log n}}{\lambda_{n,e}\sqrt{n}}\right)\|z\|_\infty$$

with probability greater than $1 - c_1\exp(-c_2\max\{\log(p-k), \log(n-s)\})$.

Since $\operatorname{sgn}(\beta_T^\star)$ and $X_{ST}^*\operatorname{sgn}(e_S^\star)$ are statistically independent with X_{S^cT}, $z \triangleq \frac{1}{\sqrt{n}}\lambda X_{ST}^*\operatorname{sgn}(e_S^\star) - \operatorname{sgn}(\beta_T^\star)$ satisfies the assumption of Lemma 5. Moreover, by Lemma 3 and the definition of λ in (34), we have with high probability

$$\|z\|_\infty \le 1 + \lambda\sqrt{\frac{D_{\max}^+ s\log p}{n}} \le \frac{3}{2}$$

where the last inequality holds from the assumption of Theorem 2. Now, applying Lemma 5 leads to $\frac{\sqrt{n}}{\lambda}\left\|X_{S^cT}(X_{S^cT}^*X_{S^cT})^{-1}z\right\|_\infty \le 1 - \frac{2\sigma\sqrt{\log n}}{\lambda_{n,e}\sqrt{n}}$.

Putting these two bounds together and using the triangular inequality, we conclude that with high probability, $\left\|z_{S^c}^{(e)}\right\|_\infty < 1$ as claimed. □

C. Establish the ℓ_∞ Bound of $\widehat{\beta}_T - \beta_T^\star$

Recalling the formula of $(\widehat{\beta}_T - \beta_T^\star)$ from (32), the triangular inequality yields

$$\left\|\widehat{\beta}_T - \beta_T^\star\right\|_\infty \le \left\|(X_{S^cT}^*X_{S^cT})^{-1}X_{S^cT}^*w_{S^c}\right\|_\infty + n\lambda_{n,\beta}\left\|(X_{S^cT}^*X_{S^cT})^{-1}(\frac{1}{\sqrt{n}}\lambda X_{ST}^*\operatorname{sgn}(e_S^\star) - \operatorname{sgn}(\beta_T^\star))\right\|_\infty \triangleq \mathcal{T}_1 + \mathcal{T}_2. \quad (45)$$

To bound the first quantity, we consider a random vector $u = (\frac{1}{n-s}X_{S^cT}^*X_{S^cT})^{-1}\frac{1}{n-s}X_{S^cT}^*w_{S^c}$ and note that $\mathcal{T}_1 = \|u\|_\infty$. This bound, which is stated below, has been established in [8, eq. (48)]: there exists some numerical constant c such that

$$\mathbb{P}\left(\mathcal{T}_1 \ge 20\sqrt{\frac{\sigma^2\log k}{C_{\min}(n-s)}}\right) \le 4\exp(-c(n-s)). \quad (46)$$

Turning now to the second quantity $\mathcal{T}_2$. We have

$$\mathcal{T}_2 \le \frac{n\lambda_{n,\beta}}{n-s}\left\|\left(\frac{X_{S^cT}^*X_{S^cT}}{n-s}\right)^{-1}z\right\|_\infty$$

where $z \triangleq \frac{1}{\sqrt{n}}\lambda X_{ST}^*\operatorname{sgn}(e_S^\star) - \operatorname{sgn}(\beta_T^\star)$. To bound $\mathcal{T}_2$, we follow similar arguments in [8], Section V-B. We can now state the following lemma, which is modified from [8, Lemma 5].

Lemma 6 [8]: Let $z \in \mathbb{R}^k$ be a fixed nonzero vector and $W \in \mathbb{R}^{n\times k}$ be a random matrix with i.i.d. entries $W_{ij} \sim \mathcal{N}(0,1)$. Then, there exists positive constants c_1 and c_2 such that

$$\mathbb{P}\left(\left\|\left[\left(\frac{W^*W}{n}\right)^{-1} - I_{k\times k}\right]z\right\|_\infty \ge c_1\tau\|z\|_\infty\right) \le 4\exp(-c_2\min\{k, \frac{n\tau^2}{k}\}).$$

Using Lemma 6 with $\tau = \sqrt{\frac{k\log(p-k)}{n}}$ and following similar arguments as in [8], Section V-B, we have a similar probabilistic bound for $\mathcal{T}_2$ as [8, eq. (41)]

$$\mathbb{P}\left(\mathcal{T}_2 \ge c_1\lambda_{n,\beta}\sqrt{\frac{kn\log(p-k)}{(n-s)^2}}\left\|\Sigma_{TT}^{-1/2}\right\|_\infty\left\|\Sigma_{TT}^{-1/2}z\right\|_\infty\right) \le 4\exp(-c_2\min\{k, \log(p-k)\}). \quad (47)$$

Furthermore, Lemma 3 states that $\|z\|_\infty \le 3/2$ with high probability. Conditioning on the event $\mathcal{E} = \{\|z\|_\infty \le 3/2\}$, we have $\left\|\Sigma_{TT}^{-1/2}z\right\|_\infty \le \frac{3}{2}\left\|\Sigma_{TT}^{-1/2}\right\|_\infty$. Thus, (47) leads to

$$\mathbb{P}\left(\mathcal{T}_2 \ge c_1\lambda_{n,\beta}\sqrt{\frac{kn\log(p-k)}{(n-s)^2}}\left\|\Sigma_{TT}^{-1/2}\right\|_\infty^2 |\mathcal{E}\right) \le 4\exp(-c_2\min\{k, \log(p-k)\}).$$

By the total probability rule, $\mathbb{P}(\mathcal{T}_2 \geq \tau) \leq \mathbb{P}(\mathcal{T}_2|\mathcal{E}) + \mathbb{P}(\mathcal{E}^c)$. Therefore, we conclude that with probability greater than $1 - 6\exp(-c_2 \min\{k, \log(p-k)\})$:

$$\mathcal{T}_2 \leq c_2 \lambda_{n,\beta} \sqrt{\frac{k\log(p-k)}{(1-\eta)^2 n}} \left\|\Sigma_{TT}^{-1/2}\right\|_\infty^2. \tag{48}$$

Overall, combining the bound of $\mathcal{T}_2$ with the bound of $\mathcal{T}_1$ in (46) concludes that $\left\|\widehat{\beta}_T - \beta_T^\star\right\|_\infty \leq f_\beta(\lambda_{n,\beta})$ with probability at least $1 - 10\exp(-c_3 \min\{k, \log(p-k)\}$, where $f_\beta(\lambda_{n,\beta})$ is defined in (22).

D. Establish the ℓ_∞ Bound of $\widehat{e}_S - e_S^\star$

Recalling the formula of $\widehat{e}_S - e_S^\star$ in (33) and applying the triangular inequality, we get

$$\begin{aligned}\|\widehat{e}_S - e_S^\star\|_\infty &\leq \frac{1}{\sqrt{n}} \left\|X_{ST}(X_{S^cT}^* X_{S^cT})^{-1} X_{S^cT}^* w_{S^c}\right\|_\infty \\ &\quad + \lambda_{n,\beta}\sqrt{n}\left\|X_{ST}(X_{S^cT}^* X_{S^cT})^{-1} z\right\|_\infty \\ &\quad + \frac{1}{\sqrt{n}}\|w_S\|_\infty + \lambda_{n,e} \\ &\triangleq \mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3 + \lambda_{n,e} \end{aligned} \tag{49}$$

where we again denote $z = \frac{1}{\sqrt{n}}\lambda X_{ST}^* \operatorname{sgn}(e_S^\star) - \operatorname{sgn}(\beta_T^\star)$. We first consider the easiest term $\mathcal{T}_3 = \frac{1}{\sqrt{n}}\|w_S\|_\infty$. Since w_S is a random vector with i.i.d. $\mathcal{N}(0,\sigma^2)$ entries, by Gaussian extreme order statistics [41], $\mathcal{T}_3 \leq 2\sqrt{\frac{\sigma^2 \log s}{n}}$.

Turning to the first term $\mathcal{T}_1$, we define a vector $v \in \mathbb{R}^s$ whose entries are $v_i \triangleq x_i(X_{S^cT}^* X_{S^cT})^{-1} X_{S^cT}^* w_{S^c}$ where x_i is the ith row of the matrix X_{ST} and notice that $\mathcal{T}_1 = \|v\|_\infty$. Conditioned on X_T, it is clear that v_i is a zero mean random variable with variance $\sigma^2 x_i(X_{S^cT}^* X_{S^cT})^{-1} x_i^*$. In addition, we recall that X_T can be represented as $X_T = W_T \Sigma_{TT}^{1/2}$, where W_T is the $n \times k$ standard Gaussian matrix. Thus, $x_i(X_{S^cT}^* X_{S^cT})^{-1} x_i^* = w_i(W_{S^cT}^* W_{S^cT})^{-1} w_i^* \leq \|w_i\|_2^2 \left\|(W_{S^cT}^* W_{S^cT})^{-1}\right\|$, where w_i is the ith row of matrix W_{ST}. In short, v_i is a zero mean random variable with variance at most $\widetilde{\sigma}^2 \triangleq \sigma^2 \|w_i\|_2^2 \left\|(W_{S^cT}^* W_{S^cT})^{-1}\right\|$. Using the concentration result for χ^2-variate, we get $\|w_i\|_2^2 \leq 2k$ with probability at least $1 - \exp(-k/2)$. Furthermore, from random matrix theory (67) in Appendix D, $\left\|(W_{S^cT}^* W_{S^cT})^{-1}\right\| \leq \frac{5}{n-s}$ with probability at least $1 - \exp(-(n-s)/2)$.

Next, let us define the event

$$\mathcal{E} = \left\{\widetilde{\sigma}^2 \geq \frac{10\sigma^2 k}{n-s}\right\}.$$

From the above arguments, we have $\mathbb{P}(\mathcal{E}) \leq \exp(-(n-s+k)/2)$. By the total probability rule, we have

$$\mathbb{P}(\mathcal{T}_1 \geq \tau) \leq \mathbb{P}(\mathcal{T}_1 \geq \tau|\mathcal{E}^c) + \mathbb{P}(\mathcal{E}).$$

Conditioning on $\mathcal{E}^c$, v_i is zero mean Gaussian with variance at most $\frac{10\sigma^2 k}{n-s}$. Thus, by the Gaussian tail bound (63) in Appendix D, we derive

$$\mathbb{P}\left(\max_{i\in S}|v_i| \geq \tau\right) \leq 2s\exp\left(-\frac{(n-s)\tau^2}{10\sigma^2 k}\right).$$

Setting $\tau = \sqrt{\frac{20\sigma^2 k\log(p-k)}{n-s}}$ yields the fact that this probability vanishes at rate $2(p-k)^{-1}$. Overall, we can now conclude that

$$\mathbb{P}\left(\mathcal{T}_1 \geq 11\sqrt{\frac{\sigma^2 k\log(p-k)}{n-s}}\right) \leq 2\exp(-\log(p-k)).$$

It is left to bound $\mathcal{T}_2$. By submultiplicative norm inequality, $\mathcal{T}_2$ is bounded by

$$\lambda_{n,\beta}\sqrt{n}\,\|X_{ST}\|_\infty \left\|(X_{S^cT}^* X_{S^cT})^{-1} z\right\|_\infty.$$

We already established $n\lambda_{n,\beta}\left\|(X_{S^cT}^* X_{S^cT})^{-1} z\right\|_\infty$ in (48). In addition, $\|X_{ST}\|_\infty \leq \sqrt{k}\,\|X_{ST}\|$ where by the matrix theory (68) in Appendix D, $\|X_{ST}^* X_{ST}\| \leq 4C_{\max}(s + \sqrt{sk})$ with high probability. Thus, $\|X_{ST}\|_\infty \leq \sqrt{C_{\max}}(sk + k\sqrt{sk})^{1/2}$.

Overall, combining with the bounds of $\mathcal{T}_1$ and $\mathcal{T}_3$, we conclude that $\left\|\widehat{\beta}_T - \beta_T^\star\right\|_\infty \leq f_\beta(\lambda_{n,\beta})$ with probability at least $1 - 10\exp(-c_3\min\{k, \log(p-k)\}$, where $f_e(\lambda_{n,\beta}, \lambda_{n,e})$ is defined as in (23).

VI. Proof of Theorem 3—Inachievability

Our analysis in this section relies on the notion of primal-dual witness introduced by Wainwright [8]. In particular, we will construct a pair of primal solutions $(\widehat{\beta}, \widehat{e})$ and their dual vectors $(z^{(\beta)}, z^{(e)})$. The extended Lasso (6) fails to correctly identify signed support of the coefficient vector $\beta^\star$ and the error $e^\star$ when the ℓ_∞-norm of either $z_{T^c}^{(\beta)}$ or $z_{S^c}^{(e)}$ exceeds unity with probability approaching one. The primal-dual witness is constructed as follows.

1) First, we obtain the solution pair $(\widehat{\beta}_T, \widehat{e}_S)$ of the following restricted Lasso problem:
$$\min_{\beta,e} \frac{1}{2n}\left\|y_S - X_{ST}\beta_T - \sqrt{n}e_S\right\|_2^2 + \lambda_{n,\beta}\|\beta_T\|_1 + \lambda_{n,e}\|e_S\|_1. \tag{50}$$
We also set $\widehat{\beta}_{T^c} = 0$ and $\widehat{e}_{S^c} = 0$.
2) Second, we select $z_T^{(\beta)}$ and $z_S^{(e)}$ as elements of the subgradients $\left\|\widehat{\beta}\right\|_1$ and $\|\widehat{e}\|_1$, respectively.
3) Third, we solve for vectors $z_{T^c}^{(\beta)}$ and $z_{S^c}^{(e)}$ satisfying the KKT conditions in (28). We then verify whether the dual feasibility conditions of both $\left\|z_{T^c}^{(\beta)}\right\|_\infty < 1$ and $\|\widehat{e}_{S^c}\|_\infty < 1$ are satisfied.
4) Fourth, we check whether the sign consistency $z_T^{(\beta)} = \operatorname{sgn}(\beta_T^\star)$ and $z_S^{(e)} = \operatorname{sgn}(e_S^\star)$ are satisfied.

The following result summarizes the use of the primal-dual witness construction in providing the proof of Theorem 3:

Lemma 7: If either steps 3 or 4 of the primal-dual construction fails, then the extended Lasso fails to recover the correct signed supports of both $\beta^\star$ and $e^\star$.

The proof of this lemma is essentially similar to that of Lemma 2(c) in [8]; thus, we omit the detail here.

In our proof, we assume that $z_T^{(\beta)} = \operatorname{sgn}(\beta_T^\star)$ and $z_S^{(e)} = \operatorname{sgn}(e_S^\star)$; otherwise, the sign consistency would fail. Under these assumptions, it is easy to check that the solution $(\widehat{\beta}_T, \widehat{e}_S)$ of the optimization (50) is expressed in (32)

and (33). Thus, we can derive equations of $z_{T^c}^{(\beta)}$ and $z_{S^c}^{(e)}$, as in (35) and (43).

In the following two sections, we establish the claim by showing that under the conditions of the sample size n and $s = \eta n$ as in Theorem 3, the ℓ_∞-norm of either $z_{T^c}^{(\beta)}$ or $z_{S^c}^{(e)}$ exceeds unity with probability tending to one. It is clear that if the extended Lasso (6) fails to recover signed support vectors with $s = \eta n$, it also fails to do so with $s > \eta n$ since it is easier to solve the extended Lasso when there is less corrupted observations.

A. Lower ℓ_∞-Norm Bound of $z_{T^c}^{(\beta)}$

Recall the expression of $z_{T^c}^{(\beta)}$ in (35) and its simplified form $z_{T^c}^{(\beta)} = a + b$, where b and a are defined in (39) and (40). We already have $\|a\|_\infty \leq 1 - \gamma$ due to the mutual incoherence assumption. It is now sufficient to show that $\max_{i \in T^c} |b_i|$ exceeds $(2 - \gamma)$ with high probability.

Conditioning on X_T and w, the vector b is zero-mean Gaussian with covariance matrix $M\Sigma_{T^c|T}$, where the random scaling form M has the form (41). The following lemma controls the lower bound of this scaling factor. The proof is similar to that of [8, Lemma 6], so we omit the detail here.

Lemma 8: Define the event $\mathcal{E} = \{M > \underline{M}\}$, where $\underline{M}$ is defined in (51), shown at the bottom of the page. Then, $\mathbb{P}(\mathcal{E}) \leq 1 - c_1 \exp(-c_2(n - s))$ for some $c_1, c_2 > 0$.

Following the proof of Theorem 4 in [8], we have the following lower bound: for all $\nu, \epsilon, \tau > 0$

$$\max_{i \in T^c} |b_i| \geq \sqrt{(2 - \nu)\rho_l(\Sigma_{T^c|T})\underline{M} \log(p - k)} - \tau \tag{52}$$

with probability at least $1 - 2\exp\left(-\frac{\tau^2}{2\underline{M}\rho_u}\right)$. Now, using appropriate choices of $\{\tau, \nu, \gamma\}$, it suffices to establish the bound

$$\rho_l(\Sigma_{T^c|T})\underline{M} \log(p - k) \geq \frac{[(2 - \gamma) + \tau]^2}{(2 - \nu)}. \tag{53}$$

We consider two cases.

1) If $\underline{M} \to +\infty$ or $\underline{M} = \Theta(1)$, then we can choose $\tau^2 = \delta \underline{M} \log(p - k)$ for some $\delta > 0$. For δ sufficiently small, we conclude from (52) that with probability converging to one, there exists some constants $c > 0$ such that

$$\max_{i \in T^c} |b_i| \geq c\sqrt{\log(p - k)}$$

which exceeds $(2 - \gamma)$ regardless of the choice of the sample size n.

2) Otherwise, $\underline{M} = o(1)$. This is satisfied only if $k/n = o(1)$, and thus, the second line of the definition of $\underline{M}$ is applied. Now, we can select τ sufficiently small and have a guarantee that $\frac{\tau^2}{2\underline{M}} \to +\infty$. From the definition of $\underline{M}$, one can see that if $\rho_l \frac{\lambda^2 s}{n} \log(p - k) \geq 2$, we can choose τ and ν strictly positive but arbitrarily close to zero such that $\frac{[(2-\gamma)+\tau]^2}{(2-\nu)} < 2$. Thus, (53) obeys regardless of the selection of the sample size n. Consequently, we assume that

$$\lambda < \sqrt{\frac{2n}{\rho_l s \log(p - k)}}. \tag{54}$$

Under this assumption, we can lower bound $\|z\|_2$ as follows:

$$\begin{aligned} \|z\|_2 &= \left\| \operatorname{sgn}(\beta_T^\star) - \frac{1}{\sqrt{n}} \lambda X_{ST}^* \operatorname{sgn}(e_S^\star) \right\|_2 \\ &\geq \|\operatorname{sgn}(\beta_T^\star)\|_2 - \frac{1}{\sqrt{n}} \lambda \|X_{ST}^* \operatorname{sgn}(e_S^\star)\|_2 \\ &\geq \sqrt{k} - \lambda \sqrt{\frac{k}{n}} \|X_{ST}^* \operatorname{sgn}(e_S^\star)\|_\infty . \end{aligned} \tag{55}$$

As shown during the proof of Lemma 3 that $\frac{1}{\sqrt{n}} \|X_{ST}^* \operatorname{sgn}(e_S^\star)\|_\infty \leq \frac{1}{3}\sqrt{\frac{\rho_l s \log(p-k)}{n}}$ with probability greater than $1 - \exp(-\frac{\rho_l}{18 D_{\max}^+} \log p)$, from the above upper bound of λ, we obtain $\frac{\lambda}{\sqrt{n}} \|X_{ST}^* \operatorname{sgn}(e_S^\star)\|_\infty \leq \frac{\sqrt{2}}{3}$. Consequently, we achieve the lower bound with high probability

$$\|z\|_2 \geq \frac{1}{2}\sqrt{k}. \tag{56}$$

Furthermore, for $(n - s)$ sufficiently large, we select a $\epsilon \in (0, 1/2)$ such that $4\sqrt{\frac{k}{n-s}} < \epsilon$. Now, replacing this bound into the second equation of $\underline{M}$ and performing some simple algebra, we can show that inequality (53) is satisfied as long as

$$\begin{aligned} \frac{\rho_l}{C_{\max}} \frac{k \log(p - k)}{(n - s)} \Bigg\{ &\frac{C_{\max} \lambda^2 s(n - s)}{(1 - \epsilon)kn} \\ &+ \frac{1}{4} + \frac{(n - s)^2}{n^2} \frac{\sigma^2 C_{\max}}{\lambda_{n,\beta}^2 k} \Bigg\} \geq \frac{[(2 - \gamma) + \tau]^2}{(2 - \nu)(1 - \epsilon)}. \end{aligned}$$

Replacing the lower bound of λ in (58) and $s = \eta n$ into the above inequality, we can conclude that inequality (53) is satisfied as long as

$$\begin{aligned} \frac{\rho_l}{C_{\max}(2 - \gamma)^2} \frac{2k \log(p - k)}{(n - s)} &\left\{ \frac{3}{8} + (1 - \eta)^2 \frac{\sigma^2 C_{\max}}{\lambda_{n,\beta}^2 k} \right\} \\ &\geq \frac{[(2 - \gamma) + \tau]^2}{(2 - \gamma)^2 (1 - \nu/2)(1 - \epsilon)}. \end{aligned}$$

Under the assumptions of Theorem 3, the right-hand side is strictly greater than one. On the other hand, τ, ν, and ϵ are parameters that can be chosen in $(0, 1/2)$. By selecting these pa-

$$\underline{M} \triangleq \begin{cases} \frac{\lambda^2 s}{n} + c\frac{k}{n-s}, & \text{if } k/n = \Theta(1) \\ \frac{\lambda^2 s}{n} + \left(1 - \max\left\{\epsilon, 4\sqrt{\frac{k}{n-s}}\right\}\right)\left(\frac{\sigma^2(n-s)}{n^2 \lambda_{n,\beta}^2} + \frac{\|z\|_2^2}{(n-s)C_{\max}}\right), & \text{if } k/n = o(1) \end{cases} \tag{51}$$

rameters to be positive but arbitrarily close to zeros, we can set the right-hand side less than one. Therefore, (53) is satisfied.

B. Lower the ℓ_∞-Norm Bound of $z^{(e)}_{S^c}$

Recalling the equation of $z^{(e)}_{S^c}$ in (43), we have

$$z^{(e)}_{S^c} = \frac{1}{\sqrt{n}\lambda_{n,e}}\Pi_{S^cT} w_{S^c} + \frac{\sqrt{n}}{\lambda} X_{S^cT}(X^*_{S^cT}X_{S^cT})^{-1} z$$

where we recall $z = \frac{1}{\sqrt{n}}\lambda X^*_{ST}\operatorname{sgn}(e^\star_S) - \operatorname{sgn}(\beta^\star_T)$. First, notice that Π_{S^cT} is the orthogonal projection onto the column space of the matrix X_{S^cT}. Thus, two terms in the above summation are orthogonal to each other. Therefore, lowering the ℓ_∞-norm of $z^{(e)}_{S^c}$ by its ℓ_2-norm counterpart, we have

$$\begin{aligned}(n-s)\left\|z^{(e)}_{S^c}\right\|^2_\infty &\geq \left\|z^{(e)}_{S^c}\right\|^2_2 \\ &= \frac{1}{n\lambda^2_{n,e}}\left\|\Pi_{S^cT} w_{S^c}\right\|^2_2 + \frac{n}{\lambda^2}\left\|X_{S^cT}(X^*_{S^cT}X_{S^cT})^{-1}z\right\|^2_2.\end{aligned}$$

From this inequality, we have an important observation that both terms in the sum have to be upper bounded by $(n-s)$. Otherwise, $\left\|z^{(e)}_{S^c}\right\|^2_\infty$ is automatically strictly greater than one, regardless of the choice of the sample size n. This observation suggests to us the required lower bound of $\lambda_{n,e}$ and λ:

$$\lambda_{n,e} \geq \frac{1}{\sqrt{n(n-s)}}\left\|\Pi_{S^cT} w_{S^c}\right\|_2$$

and

$$\lambda \geq \sqrt{\frac{n}{n-s}}\left\|X_{S^cT}(X^*_{S^cT}X_{S^cT})^{-1}z\right\|_2.$$

We now explicitly establish the lower bound of these regularization parameters. First, since $\frac{1}{\sigma^2}\left\|\Pi_{S^cT} w_{S^c}\right\|^2_2$ is the χ^2-variate with $n-s$ degrees of freedom, Lemma 13 in Appendix D suggests us that $\frac{1}{\sigma^2}\left\|\Pi_{S^cT} w_{S^c}\right\|^2_2 \geq \frac{1}{2}(n-s)$ with probability at least $1-\exp(-(n-s)/16)$. Consequently, we require

$$\lambda_{n,e} \geq \sqrt{\frac{\sigma^2}{2n}}. \tag{57}$$

Furthermore, we observe that with probability converging to one

$$\begin{aligned}&\left\|X_{S^cT}(X^*_{S^cT}X_{S^cT})^{-1}z\right\|^2_2 \\ &= z^*(X^*_{S^cT}X_{S^cT})^{-1}z \\ &= z^*\Sigma^{-1/2}_{TT}(W^*_{S^cT}W_{S^cT})^{-1}\Sigma^{-1/2}_{TT}z \\ &\geq \left\|\Sigma^{-1/2}_{TT}z\right\|^2_2 \sigma_{\min}((W^*_{S^cT}W_{S^cT})^{-1}) \\ &\geq \frac{1}{2n}C^{-1}_{\max}\left\|z\right\|^2_2\end{aligned}$$

where the second identity follows from the decomposition $X_{S^cT} = \Sigma^{1/2}_{TT}W_{S^cT}$ and the last inequality is due to the Gaussian random matrix inequality (65) in Appendix D. In combination with the lower bound of $\left\|z\right\|_2$, we require

$$\lambda \geq \sqrt{\frac{k}{8C_{\max}(n-s)}}. \tag{58}$$

Turning to establish the lower bound of $\left\|z^{(e)}_{S^c}\right\|_\infty$, we can show that under the assumptions of Theorem 3, this quantity is strictly greater than one. By the triangular inequality, $\left\|z^{(e)}_{S^c}\right\|_\infty \geq \mathcal{T}_1 - \mathcal{T}_2$, where $\mathcal{T}_1$ is quantified as

$$\frac{\sqrt{n}}{\lambda}\left\|X_{S^cT}(X^*_{S^cT}X_{S^cT})^{-1}z\right\|_\infty$$

and the other term is $\mathcal{T}_2 \stackrel{\triangle}{=} \frac{1}{\lambda_{n,e}\sqrt{n}}\left\|\Pi_{S^cT} w_{S^c}\right\|_\infty$. As shown at the beginning of Section V-B, we have the following inequality to hold with probability greater than $1-2\exp(-\log(n-s))$:

$$\mathcal{T}_2 \leq \frac{2\sqrt{\sigma^2\log(n-s)}}{\lambda_{n,e}\sqrt{n}}.$$

It is now left to justify that under the assumption of Theorem 3, $\mathcal{T}_1 > 1 + \frac{2\sqrt{\sigma^2\log n}}{\lambda_{n,e}\sqrt{n}}$. The remainder of this section is devoted to establish this claim. In what follows, we state two important lemmas, which are the main factor in establishing the lower bound of $\mathcal{T}_1$. The proofs of these lemmas are again deferred to the Appendix.

Lemma 9: For any vector $z \in \mathbb{R}^k$ independent with X_{S^cT}, we have with probability greater than $1-\exp(-\log(n-s))$

$$\begin{aligned}&\left\|X_{S^cT}(X^*_{S^cT}X_{S^cT})^{-1}z - \frac{1}{n-s}X_{S^cT}\Sigma^{-1}_{TT}z\right\|_\infty \\ &\qquad\leq 16\frac{\left\|z\right\|_2\sqrt{2k\log(n-s)}}{\sqrt{C_{\min}(n-s)^3}}.\end{aligned}$$

Lemma 10: With probability at least $1-4\exp(-\frac{1}{4}\log(n-s))$:

$$\left\|X_{S^cT}\Sigma^{-1}_{TT}z\right\|_\infty \geq \frac{2\left\|z\right\|_2\sqrt{\log(n-s)}}{3\sqrt{C_{\max}}}.$$

Once these two lemmas are established, we can now show that under the assumptions of Theorem 3, $\mathcal{T}_1 > 1 + \frac{\sqrt{\sigma^2\log n}}{\lambda_{n,e}\sqrt{n}}$ with high probability. By definition, $z = \frac{1}{\sqrt{n}}\lambda X^*_{ST}\operatorname{sgn}(e^\star_S) - \operatorname{sgn}(\beta^\star_T)$, one can see that z is independent from X_{S^cT}. Thus, by Lemmas 9 and 10 and the triangular inequality, we have, with probability at least $1-\exp(-\log(p-k)) - 4\exp(-\frac{1}{4}\log(n-s))$:

$$\begin{aligned}&\left\|X_{S^cT}(X^*_{S^cT}X_{S^cT})^{-1}z\right\|_\infty \\ &\geq \frac{1}{n-s}\left\|X_{S^cT}\Sigma^{-1}_{TT}z\right\|_\infty - 16(1+\epsilon)\frac{\sqrt{k\log(n-s)}}{\sqrt{C_{\min}(n-s)^3}}\left\|z\right\|_2 \\ &\geq \left(\frac{2\sqrt{\log(n-s)}}{3(n-s)\sqrt{C_{\max}}}\right. \\ &\left. - \frac{\sqrt{\log(n-s)}}{(n-s)\sqrt{C_{\max}}}\sqrt{\frac{256(1+\epsilon)^2kC_{\max}}{(n-s)C_{\min}}}\right)\left\|z\right\|_2.\end{aligned} \tag{59}$$

Recall from the previous section that we require the upper bound of λ in (54). Otherwise, $\left\|z^{(\beta)}_{T^c}\right\|_\infty$ is strictly greater than one regardless of the choice of the sample size n. This upper bound of λ leads to the lower bound of $\left\|z\right\|_2$ in (56). Furthermore,

assuming that $n-s \geq c\frac{C_{\max}}{C_{\min}}k$ for some large enough constant c, we achieve

$$\left\|X_{S^cT}(X_{S^cT}^*X_{S^cT})^{-1}z\right\|_\infty \geq \frac{1}{6}(1-\epsilon)\frac{\sqrt{k\log(n-s)}}{(n-s)\sqrt{C_{\max}}}.$$

Therefore, the requirement $\mathcal{T}_1 > 1+\frac{2\sqrt{\sigma^2\log n}}{\lambda_{n,e}\sqrt{n}}$ is equivalent to

$$(n-s)^2 < \frac{(1-\epsilon)}{6}\left(1+\frac{2\sqrt{\sigma^2\log n}}{\lambda_{n,e}\sqrt{n}}\right)^{-2}\frac{kn\log(n-s)}{\lambda^2 C_{\max}}.$$

Replacing the upper bound of λ in (54) and $s=\eta n$, the above inequality or, equivalently, $\left\|z_{S^c}^{(e)}\right\|_\infty > 1$ is satisfied whenever the sample size n obeys

$$n < \frac{(1-\epsilon)}{12}\frac{\eta}{(1-\eta)^2}\frac{\rho_l}{C_{\max}} \times \left(1+\frac{2\sqrt{\sigma^2\log n}}{\lambda_{n,e}\sqrt{n}}\right)^{-2} k\log(n-s)\log(p-k).$$

VII. Conclusion

In this paper, we studied the ℓ_1-constrained minimization problem for sparse linear regression when the observations are grossly corrupted. We proposed the extended Lasso method which is a natural generalization of the Lasso for recovering both the regression and the error vector effectively. Our main contribution was to establish that this recovery is faithful, under both parameter estimation and variable selection criterions, even when the error magnitude is arbitrarily large and the fraction of error is close to unity. Specifically, our first result indicated that the ℓ_2 estimation error is bounded via the introduction of the extended RE condition evaluated on the combination matrix $[X \quad I]$. Our next results considered the exact signed support recovery for a class of random Gaussian design matrices. We showed that the sign consistency is indeed possible even when almost all the observations are significantly corrupted. More interestingly, we established the lower and upper bounds for the sample size such that the extended Lasso succeeds or fails in recovering the supports with high probability. This number of observations is scaled in term of the model dimension p, the sparsity index k, and the fraction error $\eta = s/n$. Notably, all of our results are consistent with that of the standard Lasso in the absence of sparse error.

There are a number of extensions and open questions related to this study. First, our setup can be extended to robust group/multivariate Lasso model. This model has been shown to outperform the conventional Lasso in many practical applications as well as theoretical analysis (see, e.g., [14], [15], [42], and [43]). It would be interesting to obtain the upper and lower bounds of the sample size when a significant fraction of observations is corrupted in this setting. Another interesting direction is to consider a more general situation where both the observations and the data matrix are corrupted/missing. In a recent paper, Loh and Wainwright [44] established the consistency of the Lasso with noisy/corrupted/missing data matrix. Whether similar results would hold for more general setting is an interesting open problem. Finally, although our current work focused exclusively on linear regression, it would be interesting to investigate the sparse additive models (see, e.g., [45] and [46]) under grossly corrupted observations.

Appendix

1) Proof of Lemma 5: Decomposing X_{S^cT} as $X_{S^cT} = W_{S^cT}\Sigma_{TT}$ where $W_{S^cT} \in \mathbb{R}^{(n-s)\times k}$ is the random matrix with i.i.d. normal Gaussian entries, we have $X_{S^cT}(X_{S^cT}^*X_{S^cT})^{-1} = W_{S^cT}(W_{S^cT}^*W_{S^cT})^{-1}\Sigma_{TT}^{-1/2}$. Consider now the compact singular value decomposition of W_{S^cT}

$$W_{S^cT} = UDV^*, \quad U\in\mathbb{R}^{(n-s)\times k} \text{ and } D,V\in\mathbb{R}^{k\times k}.$$

Since W_{S^cT} is a Gaussian random matrix with i.i.d. entries, columns of U are orthogonal vectors selected uniformly at random. We can consider U as a random matrix distributed on the Haar measure. We have

$$X_{S^cT}(X_{S^cT}^*X_{S^cT})^{-1}z = UD^\dagger V^*\Sigma_{TT}^{-1/2}z.$$

Using the random matrix concentration inequality in (66), we have with probability at least $1-e^{-k}$

$$\|W_{S^cT}\| \leq \sqrt{n-s}\left(1+4\sqrt{\frac{k}{n-s}}\right)^{1/2}.$$

In addition, from (67), we have with high probability

$$\left\|(W_{S^cT}^*W_{S^cT})^{-1}\right\| \leq \left(1+4\sqrt{\frac{k}{n-s}}\right)\frac{1}{n-s}.$$

Combining these pieces together, we conclude that

$$\left\|D^\dagger\right\| = \left\|W_{S^cT}(W_{S^cT}^*W_{S^cT})^{-1}\right\| \leq \left(1+4\sqrt{\frac{k}{n-s}}\right)^{3/2}\frac{1}{\sqrt{n-s}} \leq \frac{\sqrt{1+\epsilon}}{\sqrt{n-s}}$$

assuming that k is sufficiently smaller than $(n-s)$.

Next, our goal is to bound

$$\begin{aligned}\left\|UD^\dagger V^*\Sigma_{TT}^{-1/2}z\right\|_\infty &= \max_i |e_i^*UD^\dagger V^*\Sigma_{TT}^{-1/2}z| \\ &= \max_i |\left\langle U^*, D^\dagger V^*\Sigma_{TT}^{-1/2}ze_i^*\right\rangle| \\ &\triangleq \max_i |f_i(U)|\end{aligned}$$

where f_i is the function acting on the random matrix U, $f_i : \mathbb{R}^{|S^c|\times k} \to \mathbb{R}$.

First, we show that $f_i(U)$ is Lipschitz (with respect to the Euclidean norm) with constant at most $\|f_i\|_L = \sqrt{\frac{(1+\epsilon)k}{C_{\min}(n-s)}}\|z\|_\infty$. Indeed, for any given pair $U_1 \quad U_2 \in \mathbb{R}^{|S^c|\times k}$, we have

$$\begin{aligned}|f_i(U_1)-f_i(U_2)| &= |\left\langle U_1-U_2, D^\dagger V^*\Sigma_{TT}^{-1/2}ze_i^*\right\rangle| \\ &\leq \|U_1-U_2\|_F\left\|D^\dagger V^*\Sigma_{TT}^{-1/2}ze_i^*\right\|_F \\ &\leq \|U_1-U_2\|_F\left\|D^\dagger V^*\right\|\left\|\Sigma_{TT}^{-1/2}\right\|\|ze_i^*\|_F \\ &\leq \|U_1-U_2\|_F\frac{\sqrt{1+\epsilon}}{\sqrt{n-s}}\frac{1}{\sqrt{C_{\min}}}\|z\|_2 \\ &\leq \sqrt{\frac{(1+\epsilon)k}{(n-s)C_{\min}}}\|U_1-U_2\|_F\|z\|_\infty.\end{aligned}$$

Since the distribution of U is invariant under the orthogonal transformation $U \mapsto -U$, $f(U)$ is a symmetric random variable and zero is a median. Hence, by the measure of concentration with respect to Haar measure in Lemma 15, we get

$$\begin{aligned}\mathbb{P}\left(f_i(U) \geq \tau\right) &\leq \exp\left(-\frac{\tau^2(n-s)}{8\left\|f_i\right\|_L^2}\right) \\ &= \exp\left(-\frac{C_{\min}(n-s)^2\tau^2}{8(1+\epsilon)k\left\|z\right\|_\infty^2}\right).\end{aligned}$$

Setting $\tau \stackrel{\triangle}{=} \frac{2\lambda}{3\sqrt{n}}\left(1-\frac{2\sigma\sqrt{\log n}}{\lambda_{n,e}\sqrt{n}}\right)\left\|z\right\|_\infty$ and taking the union bound over all $i \in S^c$, we have

$$\begin{aligned}&\mathbb{P}\left(\left\|UD^\dagger V^*\Sigma_{TT}^{-1/2}z\right\|_\infty \geq \frac{\lambda}{2}\left\|z\right\|_\infty\right) \\ &\leq (n-s)\exp\left(-\frac{C_{\min}(n-s)^2\lambda^2}{12(1+\epsilon)nk}\left(1-\frac{2\sigma\sqrt{\log n}}{\lambda_{n,e}\sqrt{n}}\right)^2\right).\end{aligned}$$

This probability vanishes at rate $\exp(-c\log n)$ provided that

$$(n-s)^2 > 12(1+\epsilon)\left(1-\frac{2\sigma\sqrt{\log n}}{\lambda_{n,e}\sqrt{n}}\right)^{-2}\frac{nk\log n}{C_{\min}\lambda^2}.$$

Replacing the expression of λ in (34) and $s = \eta n$, the above condition is equivalent to

$$\begin{aligned}\frac{n}{\log n} \geq C(1+\epsilon)&\frac{\eta}{(1-\eta)^2}\frac{\max\{\rho_u, D_{\max}^+\}}{C_{\min}\gamma^2} \\ &\times\left(1-\frac{2\sigma\sqrt{\log n}}{\lambda_{n,e}\sqrt{n}}\right)^{-2}k\log(p-k)\end{aligned}$$

where C is a numerical constant smaller than 48.

2) Proof of Lemma 9: Recalling the decomposition of X_{S^cT}: $X_{S^cT} = W_{S^cT}\Sigma_{TT}^{1/2}$, we have

$$\begin{aligned}&X_{S^cT}(X_{S^cT}^*X_{S^cT})^{-1}z - \frac{1}{n-s}X_{S^cT}z \\ &= \left(W_{S^cT}(W_{S^cT}^*W_{S^cT})^{-1} - \frac{1}{n-s}W_{S^cT}\right)\Sigma^{-1/2}z.\end{aligned}$$

Notice that W_{S^cT} is an $(n-s)\times k$ matrix with independent Gaussian entries with zero mean and unit variance. Consider now the reduced singular value decomposition of W_{S^cT}

$$W_{S^cT} = UDV^*, \quad U \in \mathbb{R}^{(n-s)\times k} \text{ and } D, V \in \mathbb{R}^{k\times k}.$$

Then, the columns of U are k orthonormal vectors selected uniformly at random. We can think of U as a random matrix distributed on the Haar measure. The above equation is now formulated as

$$\frac{1}{n-s}UD\left[\left(\frac{D^*D}{n-s}\right)^{-1} - I\right]V\Sigma^{-1/2}z =: U\widetilde{D}V\Sigma^{-1/2}z.$$

It is clear that $\left\|\widetilde{D}\right\| \leq \frac{1}{n-s}\left\|W_{S^cT}\right\|\left\|(\frac{W_{S^cT}^*W_{S^cT}}{n-s})^{-1} - I\right\|$. Recalling the random matrix concentration bounds (66) and (67), we have $\left\|\frac{W_{S^cT}}{\sqrt{n-s}}\right\| \leq (1+4\sqrt{\frac{k}{n-s}})^{1/2}$. Therefore

$$\left\|\widetilde{D}\right\| \leq \frac{4\sqrt{k}}{n-s}\left(1+4\sqrt{\frac{k}{n-s}}\right)^{1/2} =: (1+\epsilon)\frac{4\sqrt{k}}{n-s}$$

where we choose $\epsilon \geq 4\sqrt{k/(n-s)}$.

Our goal now is to establish an upper bound of $\left\|U\widetilde{D}V\Sigma^{-1/2}z\right\|_\infty$, which can be rewritten as

$$\begin{aligned}\max_i |e_i^*U\widetilde{D}V\Sigma^{-1/2}z| &= \max_i |\left\langle U, \widetilde{D}V\Sigma^{-1/2}ze_i^*\right\rangle| \\ &\stackrel{\triangle}{=} \max_i f_i(U)\end{aligned}$$

where f_i is a function operating on the random matrix U, $f_i : \mathbb{R}^{(n-s)\times k} \mapsto \mathbb{R}$.

First, we show that $f_i(U)$ is Lipschitz (with respect to the Euclidean norm) with constant at most $\left\|f_i\right\|_L = \frac{4(1+\epsilon)\sqrt{k}}{n-s}\frac{1}{\sqrt{C_{\min}}}\left\|z\right\|_2$. Indeed, for any given pair $U_1\ U_2 \in \mathbb{R}^{|S^c|\times k}$, we have

$$\begin{aligned}|f_i(U_1) - f_i(U_2)| &= |\left\langle U_1 - U_2, \widetilde{D}V^*\Sigma_{TT}^{-1/2}ze_i^*\right\rangle| \\ &\leq \left\|U_1 - U_2\right\|_F\left\|\widetilde{D}V^*\Sigma_{TT}^{-1/2}ze_i^*\right\|_F \\ &\leq \left\|U_1 - U_2\right\|_F\left\|\widetilde{D}V^*\right\|\left\|\Sigma_{TT}^{-1/2}\right\|\left\|ze_i^*\right\|_F \\ &\leq \left\|U_1 - U_2\right\|_F\frac{4(1+\epsilon)\sqrt{k}}{n-s}\frac{1}{\sqrt{C_{\min}}}\left\|z\right\|_2.\end{aligned}$$

Since the distribution of U is invariant under the orthogonal transformation $U \mapsto -U$, $f(U)$ is a symmetric random variable and zero is a median. Hence, by the measure of concentration with respect to Haar measure (Lemma 15), we get

$$\begin{aligned}\mathbb{P}\left(f_i(U) \geq \tau\right) &\leq \exp\left(-\frac{\tau^2(n-s)}{8\left\|f_i\right\|_L^2}\right) \\ &= \exp\left(-\frac{C_{\min}(n-s)^3\tau^2}{128(1+\epsilon)^2k\left\|z\right\|_2^2}\right).\end{aligned}$$

Setting $\tau^2 \stackrel{\triangle}{=} \frac{256(1+\epsilon)^2\left\|z\right\|_2^2k\log(n-s)}{C_{\min}(n-s)^3}$ and taking the union bound over all $i \in S^c$, we have

$$\begin{aligned}\mathbb{P}\left(\left\|UD^\dagger V^*\Sigma_{TT}^{-1/2}z\right\|_\infty \geq \frac{16(1+\epsilon)\left\|z\right\|_2\sqrt{k\log(n-s)}}{\sqrt{C_{\min}(n-s)^3}}\right) \\ \leq \exp(-\log(n-s))\end{aligned}$$

as claimed.

3) Proof of Lemma 10: We have $X_{S^cT} = W_{S^cT}\Sigma_{TT}^{1/2}$, where W_{S^cT} is a standard Gaussian matrix of size $(n-s)\times k$. Thus, $X_{S^cT}\Sigma_{TT}^{-1}z = W_{S^cT}\Sigma_{TT}^{-1/2}z$, which leads to

$$\begin{aligned}\left\|X_{S^cT}\Sigma_{TT}^{-1}z\right\|_\infty &= \max_{i\in S^c}|\left\langle e_i, W_{S^cT}\Sigma_{TT}^{-1/2}z\right\rangle| \\ &=: \max_{i\in S^c}|f_i(W_{S^cT})|\end{aligned}$$

where $e_i \in \mathbb{R}^{(n-s)}$ is the standard vector whose entry at ith location receives unit value and zeros elsewhere. In order to lower the bound of the random variable $\max_i f_i(W_{S^cT})$, the first step is to show that it is sharply concentrated around its expectation.

Lemma 11: For any $\tau > 0$, we have

$$\mathbb{P}\left(|\max_i f_i(W_{S^cT}) - \mathbb{E}\max_i f_i(W_{S^cT})| \geq \tau\right) \leq 4\exp\left(-\frac{\tau^2}{2\left\|\Sigma_{TT}^{-1/2}z\right\|_2^2}\right). \tag{60}$$

Selecting $\tau \stackrel{\triangle}{=} \left\|\Sigma_{TT}^{-1/2}z\right\|_2\sqrt{\frac{1}{2}\log(n-s)}$, we conclude that with probability greater than $1-4\exp(-\frac{1}{4}\log(n-s))$

$$\max_i f_i(W_{S^cT}) \geq \mathbb{E}\max_i f_i(W_{S^cT}) - \tau. \tag{61}$$

At the second step, we need to lower the bound $\mathbb{E}\max_i f_i(W_{S^cT})$. This can be estimated via Sudakov–Fernique inequality [41]. We have

$$\mathbb{E}(f_i(W_{S^cT}) - f_j(W_{S^cT}))^2 = 2z^*\Sigma_{TT}^{-1}z = 2\left\|\Sigma_{TT}^{-1/2}z\right\|_2^2.$$

Consequently, if we denote g_i, $1\leq i\leq (n-s)$ as a sequence of $\mathcal{N}(0, \left\|\Sigma_{TT}^{-1/2}z\right\|_2^2)$ Gaussian random variables, then we have established a lower bound

$$\mathbb{E}(f_i(W_{S^cT}) - f_j(W_{S^cT}))^2 \geq \mathbb{E}(g_i - g_j)^2.$$

Therefore, the Sudakov–Fernique inequality [41] suggests that the maximum over $f(w_i)$ dominates the maximum over g_i. In particular, we have $\mathbb{E}\max_i f_i(W_{S^cT}) \geq \mathbb{E}\max_i g_i$. Moreover, since $\{g_i\}$ are i.i.d. random variables, by the standard bound for Gaussian extreme, for all $\delta > 0$, we have

$$\begin{aligned}\mathbb{E}\max_i f(W_{S^cT}) &\geq \mathbb{E}\max_i g_i \\ &\geq \left\|\Sigma_{TT}^{-1/2}z\right\|_2\sqrt{(2-\delta)\log(n-s)}.\end{aligned}$$

Substituting this expectation bound into (61) yields

$$\begin{aligned}\max_i f_i(W_{S^cT}) &\geq (\sqrt{2-\delta} - \sqrt{1/2})\left\|\Sigma_{TT}^{-1/2}z\right\|_2\sqrt{\log(n-s)} \\ &> \frac{2}{3}\left\|\Sigma_{TT}^{-1/2}z\right\|_2\sqrt{\log(n-s)}\end{aligned}$$

for δ arbitrarily close to zero. Furthermore, using the standard bound $\left\|\Sigma_{TT}^{-1/2}z\right\|_2 \geq \frac{\|z\|_2}{\left\|\Sigma_{TT}^{1/2}\right\|_2} \geq \frac{\|z\|_2}{\sqrt{C_{\max}}}$, we complete the proof.

Proof of Lemma 11. By the standard Gaussian concentration theorems [41], let w be a standard Gaussian measure on $\mathbb{R}^n$ and f be a Lipschitz function with Lipschitz constant $\|f\|_{\text{lip}}$. Then

$$\mathbb{P}(f(w) - \mathbb{E}f(w) \geq \tau) \leq 4\exp(-\tau^2/2\|f\|_{\text{lip}}^2). \tag{62}$$

We now consider the function $f(W_{S^cT}) \stackrel{\triangle}{=} \max_i f_i(W_{S^cT})$ operating on the standard Gaussian matrix W_{S^cT}. We have

$$\begin{aligned}f(W_{S^cT}^1) - f(W_{S^cT}^2) &= \max_i\left\langle e_i, W_{S^cT}^1\Sigma_{TT}^{-1/2}z\right\rangle \\ &\quad - \max_k\left\langle e_k, W_{S^cT}^2\Sigma_{TT}^{-1/2}z\right\rangle \\ &\leq \max_i\left\langle e_i, (W_{ST}^1 - W_{ST}^2)\Sigma_{TT}^{-1/2}z\right\rangle \\ &\leq \left\|\Sigma_{TT}^{-1/2}z\right\|_2\left\|W_{ST}^1 - W_{ST}^2\right\|_F\end{aligned}$$

where the second inequality follows from the Cauchy–Schwartz inequality. Applying (62) with Lipschitz constant $\left\|\Sigma_{TT}^{-1/2}z\right\|_2$ completes our proof. □

4) Some Concentration Inequalities: In this section, we restate some well-known large deviation bounds for ease of reference. The first is a bound of sum of Gaussian random variables.

Lemma 12: Let $Z_1,\ldots,Z_n$ be independent and zero-mean Gaussian random variables with parameters $\sigma_1^2,\ldots,\sigma_n^2$. Then

$$\mathbb{P}\left(|\sum_{i=1}^n Z_i| \geq \tau\right) \leq 2\exp\left(-\frac{\tau^2}{2\sum_{i=1}^n\sigma_i^2}\right).$$

This bound comes directly from a standard Gaussian bound. For a Gaussian variable $Z\sim\mathcal{N}(0,\sigma^2)$, we have with all $\tau > 0$

$$\mathbb{P}(|Z| \geq \tau) \leq 2\exp\left(-\frac{\tau^2}{2\sigma^2}\right). \tag{63}$$

The following tail bounds on the Chi-square variates taken from [47] are useful

Lemma 13: Let X be a centralized χ^2-variate with d degree of freedom. Then, for all $\tau\in(0,1/2)$, we have

$$\begin{aligned}\mathbb{P}\left(X \geq d(1+\tau)\right) &\leq \exp\left(-\frac{3}{16}d\tau^2\right) \\ \mathbb{P}\left(X \leq d(1-\tau)\right) &\leq \exp\left(-\frac{1}{4}d\tau^2\right).\end{aligned}$$

We also recall some well-known concentration inequalities from random matrix theory

Lemma 14: Let $X^{n\times k}$ be a random matrix, whose entries are standard Gaussian random variables. Denote by $\sigma_{\min}$ and

$\sigma_{\max}$ the smallest and largest singular values of X. Then, we have

$$\mathbb{P}\left(1-\sigma_{\min}(X)/\sqrt{n}\geq\sqrt{\frac{k}{n}}+\tau\right)\leq\exp\left(-n\,\tau^2/2\right)$$
$$\mathbb{P}\left(\sigma_{\max}(X)/\sqrt{n}-1\geq\sqrt{\frac{k}{n}}+\tau\right)\leq\exp\left(-n\tau^2/2\right).$$

By setting $\tau=\sqrt{\frac{k}{n}}$, we conclude that with probability at least $1-\exp(-k/2)$:

$$(1-2\sqrt{k/n})^2\leq\sigma_{\min}\left(X^*X/n\right)\leq\sigma_{\max}\left(X^*X/n\right)\leq(2\sqrt{k/n}+1)^2. \quad (64)$$

A consequence of this quantity is another singular value bound for the inverse matrix of X^*X. We have with probability greater than $1-\exp(-k/2)$:

$$\frac{1}{(2\sqrt{k/n}+1)^2}\leq\sigma_{\min}\left((X^*X/n)^{-1}\right)\leq\sigma_{\max}\left((X^*X/n)^{-1}\right)\leq\frac{1}{(1-2\sqrt{k/n})^2}. \quad (65)$$

From the above two set of inequality and assumption that $k\leq n$, we conclude that with probability greater than $1-\exp(-k/2)$:

$$\left\|\frac{X^*X}{n}-I\right\|\leq 4\sqrt{\frac{k}{n}} \quad (66)$$
$$\left\|\left(\frac{X^*X}{n}\right)^{-1}-I\right\|\leq 4\sqrt{\frac{k}{n}}. \quad (67)$$

For random matrices whose rows are i.i.d. and have distribution $\mathcal{N}(0,\Sigma)$, we can achieve a similar spectral norm bound. We have with probability at least $1-\exp(-k/2)$

$$\left\|\frac{X^*X}{n}-\Sigma\right\|\leq 4\sigma_{\max}(\Sigma)\sqrt{\frac{k}{n}} \quad (68)$$
$$\left\|\left(\frac{X^*X}{n}\right)^{-1}-\Sigma^{-1}\right\|\leq\frac{4}{\sigma_{\min}(\Sigma)}\sqrt{\frac{k}{n}}. \quad (69)$$

Finally, the following lemma states an useful concentration inequality on Haar measure [48].

Lemma 15: Support $k<n$ and let $f:\mathbb{R}^{n\times k}\mapsto R$ with Lipschitz norm

$$\|f\|_{\mathrm{L}}=\sup_{X\neq Y}\frac{f(X)-f(Y)}{X-Y}.$$

Then, if U is distributed according to the Haar measure,

$$\mathbb{P}(f(U)\geq\mathrm{median}(f)+\tau)\leq\exp(-\frac{m\tau^2}{8\|f\|_{\mathrm{L}}^2}).$$

ACKNOWLEDGMENT

We would like to thank anonymous reviewers and the Associate Editor for helpful comments and suggestions which significantly improve the quality of the paper.

REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
[2] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, 2007.
[3] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.
[4] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *Ann. Statist.*, vol. 37, pp. 3469–3497, 2009.
[5] P. Zhao and B. Yu, "On model selection consistency of Lasso," *J. Mach. Learn. Res.*, vol. 7, pp. 2541–2563, 2006.
[6] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *Ann. Statist.*, vol. 37, no. 1, pp. 2246–2270, 2009.
[7] N. Meinshausen and P. Bühlmann, "High dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2008.
[8] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
[9] E. J. Candès and Y. Plan, "Near-ideal model selection by l1 minimization," *Ann. Statist.*, vol. 37, pp. 2145–2177, 2009.
[10] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.
[11] T. Zhang, "Some sharp performance bounds for least squares regression with l1 regularization," *Ann. Statist.*, vol. 37, no. 5, pp. 2109–2144, 2009.
[12] F. Bunea, A. Tsybakov, and M. Wegkamp, "Sparsity oracle inequalities for the Lasso," *Elec. J. Statist.*, vol. 1, pp. 169–194, 2007.
[13] F. Bunea, "Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+\ell_2$ penalization," *Elec. J. Statist.*, vol. 2, pp. 1153–1194, 2008.
[14] G. Obozinski, M. J. Wainwright, and M. I. Jordan, "Support union recovery in high-dimensional multivariate regression," *Ann. Statist.*, vol. 39, no. 1, pp. 1–47, 2011.
[15] J. Huang and T. Zhang, "The benefit of group sparsity," *Ann. Statist.*, vol. 38, no. 4, pp. 1978–2004, 2010.
[16] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 1030–1051, Mar. 2006.
[17] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[18] E. Elhamifar and R. Vidal, "Sparse subspace clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami Beach, FL, Jun. 2009, pp. 2790–2797.
[19] J. N. Laska, M. A. Davenport, and R. G. Baraniuk, "Exact signal recovery from sparsely corrupted measurements through the pursuit of justice," in *Asilomar Conf. Signals Syst. Comput.*, Pacific Grove, CA, Nov. 2009, pp. 1556–1560.
[20] J. Wright and Y. Ma, "Dense error correction via l1 minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3540–3560, Jul. 2010.
[21] Z. Li, F. Wu, and J. Wright, "On the systematic measurement matrix for compressed sensing in the presence of gross error," in *Proc. Data Compress. Conf.*, Snowbird, UT, Mar. 2010, pp. 356–365.
[22] N. H. Nguyen and T. D. Tran, Exact recoverability from dense corrupted observations via l_1 minimization Feb. 2011 [Online]. Available: http://arxiv.org/abs/1102.1227
[23] X. Li, Compressed sensing and matrix completion with constant proportion of corruptions Apr. 2011 [Online]. Available: http://arxiv.org/abs/1104.1041

[24] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 1–37, May 2011.
[25] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2010, pp. 2496–2504.
[26] A. Agarwal, S. Negahban, and M. Wainwright, "Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions," in *Proc. 28th Int. Conf. Mach. Learn.*, Bellevue, WA, Jun. 2011, pp. 1129–1136.
[27] J. Haupt, W. Bajwa, M. Rabbat, and R. Nowak, "Compressed sensing for networked data," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 92–101, Mar. 2008.
[28] N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, "Robust multi-sensor classification via joint sparse representation," in *Proc. Int. Conf. Inf. Fusion*, Chicago, IL, Jul. 2011, pp. 1–8.
[29] Y. Lee, S. N. MacEachern, and Y. Jung, "Regularization of case-specific parameters for robustness and efficiency," *Statis. Sci.*, vol. 27, pp. 350–372, 2012.
[30] H. Wang, G. Li, and G. Jiang, "Robust regression shrinkage and consistent variable selection through the LAD-Lasso," *J. Busi. Econ. Statist.*, vol. 25, no. 3, pp. 347–355, Jul. 2007.
[31] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 5406–5425, Feb. 2006.
[32] C. Studer, P. Kuppinger, G. Pope, and H. Bolcskei, "Sparse signal recovery from sparsely corrupted measurements," in *Proc. Int. Symp. Inf. Theory*, St. Pertersburg, Russia, Aug. 2011, pp. 1422–1426.
[33] C. Studer and R. G. Baraniuk, "Stable restoration and separation of approximately sparse signals," Applied Comput. Har. Anal Jul. 2011 [Online]. Available: http://arxiv.org/abs/1107.0420
[34] A. S. Dalalyan and R. Keriven, "L1-penalized robust estimation for a class of inverse problems arising in multiview geometry," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2009, pp. 441–449.
[35] A. S. Dalalyan and R. Keriven, "Robust estimation for an inverse problem arising in multiview geometry," *J. Math. Imag. Vision*, vol. 43, no. 1, pp. 10–23, 2012.
[36] N. H. Nguyen, N. M. Nasrabadi, and T. D. Tran, "Robust Lasso with missing and grossly corrupted observations," presented at the Adv. Neural Inf. Process. Syst., Granada, Spain, Dec. 2011.
[37] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated Gaussian designs," *J. Mach. Learn. Res.*, vol. 11, pp. 2241–2259, 2010.
[38] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," presented at the Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, Dec. 2009.
[39] S. van de Geer and P. Bühlmann, "On the conditions used to prove oracle results for the Lasso," *Elec. J. Statist.*, vol. 3, pp. 1360–1392, 2009.
[40] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.
[41] M. Ledoux and M. Talagrand, *Probability in Banach Space: Isoperimetry and Processes*. New York: Springer-Verlag, 1991.
[42] K. Lounici, M. Pontil, A. Tsybakov, and S. van de Geer, "Taking advantage of sparsity in multi-task learning," in *Proc. Ann. Conf. Learn. Theory*, Montreal, QC, Canada, Jun. 2009, pp. 73–82.
[43] S. Negahban and M. J. Wainwright, "Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞-regularization," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3841–3863, Jun. 2011.
[44] P.-L. Loh and M. J. Wainwright, "High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity," presented at the Adv. Neural Inf. Process. Syst., Granada, Spain, Dec. 2011.
[45] R. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman, "Sparse additive models," *J. Royal Statist. Soc.: Series B*, vol. 71, no. 5, pp. 1009–1030, Nov. 2009.
[46] L. Meier, S. van de Geer, and P. Bühlmann, "High-dimensional additive modeling," *Ann. Statist.*, vol. 37, no. 6B, pp. 3779–3821, 2009.
[47] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1303–1338, 1998.
[48] M. Ledoux, *The Concentration of Measure Phenomenon*. Providence, RI: American Math. Soc., 2001.

Nam H. Nguyen, biography not available at the time of publication.

Trac D. Tran, biography not available at the time of publication.