SCALED MINIMAX OPTIMALITY IN HIGH-DIMENSIONAL LINEAR REGRESSION: A NON-CONVEX ALGORITHMIC REGULARIZATION APPROACH

By Mohamed Ndaoud

University of Southern California

The question of fast convergence in the classical problem of high dimensional linear regression has been extensively studied. Arguably, one of the fastest procedures in practice is Iterative Hard Thresholding (IHT). Still, IHT relies strongly on the knowledge of the true sparsity parameter s. In this paper, we present a novel fast procedure for estimation in the high dimensional linear regression. Taking advantage of the interplay between estimation, support recovery and optimization we achieve both optimal statistical accuracy and fast convergence. The main advantage of our procedure is that it is fully adaptive, making it more practical than state of the art IHT methods. Our procedure achieves optimal statistical accuracy faster than, for instance, classical algorithms for the Lasso. Moreover, we establish sharp optimal results for both estimation and support recovery. As a consequence, we present a new iterative hard thresholding algorithm for high dimensional linear regression that is scaled minimax optimal (achieves the estimation error of the oracle that knows the sparsity pattern if possible), fast and adaptive.

1. Introduction. Datasets with large numbers of features are becoming increasingly available and important in every field of research and innovation. The representation of such data in any coordinate system leads to so called high dimensional data, whose analysis is often associated with phenomena that go beyond classical estimation theory. Further assumptions on the structure of the underlying signal are required in order to make the estimation problem more well-defined. For instance in a problem of high dimensional regression, we may assume that the vector to estimate is sparse. The present work aims to develop an estimation method that can extract information from existing high dimensional structured datasets in a more efficient way.

1.1. *Statement of the problem.* Assume that we observe the vector of measurements $Y \in \mathbf{R}^n$ satisfying

$$Y = X\beta + \sigma\xi, \tag{1}$$

where $X \in \mathbf{R}^{n\times p}$ is a given design or sensing matrix. The noise ξ is a centered 1-subGaussian random vector (i.e. $\forall \lambda \in \mathbf{R}^n, \mathbf{E}(e^{\langle \lambda,\xi\rangle}) \leq e^{\|\lambda\|^2/2}$), and ξ is independent of X. We denote by $\mathbf{P}_\beta$ the distribution of (Y, X) in model (1), and by $\mathbf{E}_\beta$ the corresponding expectation.

For an integer $s \leq p$, we assume that β is s-sparse, that is it has at most s non-zero components, and we denote by S_β its support and η_β the corresponding binary vector. We also assume that components of β cannot be arbitrarily small. This motivates us to define the following set $\Omega^p_{s,a}$ of s-sparse vectors:

$$\Omega_{s,a} = \{\beta \in \mathbf{R}^p : \quad |\beta|_0 \leq s \quad \text{and} \quad |\beta_i| \geq a, \quad \forall i \in S_\beta\},$$

where $a > 0$, β_i are the components of β for $i = 1, \ldots, p$, and $|\beta|_0$ denotes the number of non-zero components of β. The value a characterizes the scale of the signal. In the rest of the paper, we will always denote by β the vector to estimate, while $\hat{\beta}$ will denote the corresponding estimator. Let us denote by ψ the scaled minimax risk

$$\psi(s, a) = \inf_{\hat{\beta}} \sup_{\beta\in\Omega_{s,a}} \mathbf{E}_\beta \left(\|\hat{\beta} - \beta\|^2\right), \tag{2}$$

where the infimum is taken over all possible estimators $\hat{\beta}$ and $\|.\|$ is the Euclidean norm. The risk (2) was introduced in [16] where sharp lower bounds were given under model (1). More precisely, assuming that the design columns are normalized (i.e. $\|X_i\| = \sqrt{n}$ for all $i = 1, \ldots, p$) and that the noise ξ is a standard random Gaussian vector, a combination of theorems 3 and 4 in [16] provides the following sharp lower bounds:

$$\psi(s, a) \geq (1 + o(1))\frac{2\sigma^2 s \log(ep/s)}{n}, \quad \forall a \leq (1 - \varepsilon)\sigma\sqrt{\frac{2\log(ep/s)}{n}}$$

and

$$\psi(s, a) \geq (1 + o(1))\frac{\sigma^2 s}{n}, \quad \forall a \geq (1 + \varepsilon)\sigma\sqrt{\frac{2\log(ep/s)}{n}},$$

that holds for any $0 < \varepsilon \leq 1$ and such that the limit corresponds to $s/p \to 0$. One of the main motivations of the present work is to provide matching upper bounds for the risk ψ under mild assumptions on the design matrix X.

Notation. In the rest of this paper we use the following notation. For any integer n, $[n]$ denotes the set of integers $\{1, \ldots, n\}$. For given sequences a_n and b_n, we say that $a_n = O(b_n)$ (resp $a_n = \Omega(b_n)$) when $a_n \leq cb_n$ (resp $a_n \geq cb_n$) for some absolute constant $c > 0$. We write $a_n \asymp b_n$ if $a_n = O(b_n)$

and $a_n = \Omega(b_n)$ while $a_n = o(b_n)$ corresponds to $a_n/b_n \to 0$ as n goes to infinity. For $\mathbf{x}, \mathbf{y} \in \mathbf{R}^p$, $\|\mathbf{x}\|_\infty$ is the ℓ_∞ norm of $\mathbf{x}$, $\|\mathbf{x}\|$ the Euclidean norm of $\mathbf{x}$, and $\langle \mathbf{x}, \mathbf{y} \rangle$ the corresponding inner product. For a matrix $X \in \mathbf{R}^{n \times p}$, we denote by X_j its j-th column, and $\|X\|_{2,\infty} := \max_{j=1,\dots,p} \|X_j\|$. For $x, y \in \mathbf{R}$, we denote by $x \vee y$ the maximum of x and y and we set $x_+ = x \vee 0$. The notation $\mathbf{1}\{\cdot\}$ stands for the indicator function. For a vector $Z \in \mathbf{R}^p$, $Z_{(i)}$ denotes the i-th non increasing order statistic of Z such that $Z_{(1)} \geq \cdots \geq Z_{(p)}$. For any finite set S, $|S|$ stands for its length. For any $X \in \mathbf{R}^{p \times p}$ and $S, S' \in \{1, \dots, p\}$, X_S will denote the submatrix of X with columns indexed by S, and $X_{S'S}$ the submatrix of X with columns indexed by S and rows indexed by S'. For a vector $M \in \mathbf{R}^p$, M_S is the restriction of M to the set S. For an SDP matrix A, we denote by $\lambda_{\max}$(resp. $\lambda_{\min}$) the largest (resp. lowest) corresponding eigenvalue, and by $\|A\|_F$ its Frobenius norm. $\mathbf{I}_p$ is the identity matrix in $\mathbf{R}^{p \times p}$.

For the sake of readability of the results, we will assume the design columns to be normalized in the rest of this section , i.e. for all $i = 1, \dots, p$ we have $\|X_i\| = \sqrt{n}$.

1.2. *Related literature.* The literature on minimax sparse estimation in high-dimensional linear regression (for both random and orthogonal design) is very rich and its complete overview falls beyond the format of this paper. We mention here only some recent results close to our work. All sharp results are considered in the regime where $\frac{s}{p} \to 0$ and $s \log(ep/s)/n \to 0$.

1. *Discrepancy between the minimax rate $2s\sigma^2 \log(ep/s)/n$ and the oracle rate $\sigma^2 s/n$.* In the last decade, the success of estimators in sparse linear regression has been characterized by the minimax rate $2s\sigma^2 \log(ep/s)/n$, achieved for instance by the popular Lasso. It is well known by practitioners that Lasso suffers from non-negligible bias, and this bias is unavoidable and at least of order $\sigma^2 \frac{s \log(p/s)}{n}$ (cf. [2]). Other convex estimators, such as Slope, suffer the same bias issue as the bias lower bound in [2] applies as well. On the other hand, the oracle least-squares estimator restricted to the support of the true β enjoys the smaller rate $\sigma^2 s/n$. The focus of the present paper is the discrepancy between the well studied minimax rate $2\sigma^2 \log(ep/s)/n$ and the oracle rate $\sigma^2 s/n$, in particular

   > *For which sparse β is it possible to achieve the oracle rate $\sigma^2 s/n$ without the knowledge of the true support ? When possible, how to construct estimators that achieve the oracle rate $\sigma^2 s/n$?*

   When the magnitude of entries of the signal β is large enough we may expect to get a better estimate than the Lasso by removing the associ-

ated bias. In [16], the separation at which the bias can be removed, in the sense that the oracle $\sigma^2 s/n$ can be achieved, is exactly characterized in the case of orthogonal design. This separation is given by the universal separation $a^* = \sigma\sqrt{\frac{2\log(p/s)}{n}}$. In particular, for a larger than a^*, the estimation risk could be as small as $\sigma^2\frac{s}{n}$. For the case of general designs, the same paper shows that $a \geq a^*$ is necessary in order to remove the bias without providing corresponding sufficient conditions. The recent paper [18] provide insight on the precise phase transition when the design is Gaussian and the sparse vector β is binary; however the findings of the present paper reveal that $a = a^*$ is the threshold at which the transition between $2\sigma^2 s\log(ep/s)/n$ and $\sigma^2 s/n$ occurs for the general class of sparse vectors.
A popular approach to avoid the bias present in convex estimators is Iterative Hard Thresholding (IHT) [6] and its variants. IHT is a non-convex counterpart of the gradient descent corresponding to the Lasso [20, 24, 27, 22]. It is shown in [23, 24] that debaising and support recovery is possible, using Gradient Hard Thresholding Pursuit, under the sub-optimal condition $a \asymp \sigma\sqrt{\frac{s\log(ep/s)}{n}}$. Two other approaches for unbiased estimation are either through concave penalization [26, 11] or forward and backward greedy algorithms [25].

When the oracle rate $\sigma^2 s/n$ is achievable information-theoretically, is it possible to achieve this rate in polynomial time under mild assumptions on the design?

2. *Fast convergence and adaptation.* IHT uses explicitly the sparsity parameter, since at each step it keeps the s largest values of the gradient descent. In [15] for instance, IHT is shown to achieve the minimax rate $\sigma^2 s\log(ep/s)/n$ after a number of iterations of order $\log\left(\frac{n\|\beta\|^2}{s\log(ep/s)}\right)$ which corresponds to the expected number of iterations under strong convexity assumptions on the loss. In [1], algorithms to compute the Lasso are shown to converge geometrically under stronger conditions. FoBa [25] achieves unbiased estimation without the knowledge of s, but may take longer to converge. This motivates the following question:

   Is it possible to achieve the minimax rate $2\sigma^2 s\log(ep/s)$ and the oracle rate $\sigma^2 s/n$ using fast iterative algorithms, without the knowledge of s?

3. *Sharpness.* Here we refer to sharpness for the problem of statistical convergence rates where the statistical accuracy matters up to exact multiplicative constants. It is inspired by statistical physics where phase transitions occur at some sharp threshold. The notion of sharp optimality is useful to compare algorithms in practice since the constants generally hide large values that may impact practical implementations. In [21], SLOPE is shown to be sharply minimax optimal on the set of sparse vectors. Similarly, [14] show sharp results of esti-

mation for the debiased Lasso (with a post-processing step) that are adaptively optimal up to a logarithmic factor. Results in both [7] and [14] hold, under the Gaussian design assumption, in the asymptotic where $s/p \to 0$ and $s\log(p)/n \to 0$. To the best of our knowledge, no sharp results for general designs that are not necessarily Gaussian are known.

Is it possible to extend sharp minimax results with beyond Gaussian designs, for instance to sub-Gaussian designs?

In this paper, we shed some light on these issues. Specifically, we address the above questions in what follows.

1.3. *Main contribution.* The present work is mainly devoted to bridging the gap between statistical optimality and optimization in a sharp an adaptive way. The main novelty is a unified framework to analyze simultaneously estimation error, support recovery and optimization speed. Our objective is to build a procedure that would answer positively the questions stated in Section 1.2. The proposed method is an iterative hard thresholding algorithm where the threshold is updated at each step. For $\lambda > 0$, we define the hard thresholding operator $\mathbf{T}_\lambda : \mathbf{R}^p \to \mathbf{R}^p$, such that

$$\forall u \in \mathbf{R}^p, \forall j = 1, \dots, p, \quad \mathbf{T}_\lambda(u)_j = u_j \mathbf{1}\{|u_j| \geq \lambda\}.$$

We consider here a general class of IHT estimators. For a given sequence $(\lambda_m)_m$ of positive numbers, we define the corresponding sequence of estimators $(\hat{\beta}^m)_m$ such that $\hat{\beta}^0 = 0$ and for $m = 1, 2, \dots$

$$\hat{\beta}^m = \mathbf{T}_{\lambda_m}\left(\hat{\beta}^{m-1} + \frac{1}{n}X^\top(Y - X\hat{\beta}^{m-1})\right). \tag{3}$$

This procedure corresponds to a projected gradient descent on a non-convex set. Our thresholding procedure is, in some sense, an interpolation between the two classical thresholds, namely the largest s component for IHT, and $\sigma\sqrt{\frac{2\log(p)}{n}}$ for the LASSO. We start with a large threshold, we then update it geometrically until hitting the statistical universal threshold. This gives further an explicit stopping time of our procedure that may be seen as an early stopping rule for gradient descent. Another perspective about our procedure is that it could be seen as what we describe later as an *iteration selection* procedure. At each step we may see $\hat{\beta}^m$ as an estimator computed using one iteration starting from the one before $\hat{\beta}^{m-1}$. By analogy with a classical model selection criterion, we can select the iteration that is minimax optimal. This leads, in particular, to a faster adaptive procedure compared to model selection since each estimator is simply one iteration of our algorithm. Our contribution can be summarized as follows:

- We derive a new proof strategy to construct adaptive minimax optimal estimators through algorithmic regularization. This combines techniques from non-convex optimization and model selection.
- We propose a fully adaptive variant of IHT that is scaled minimax optimal (i.e. achieves optimal risk of the oracle that knows the sparsity pattern when possible). We also show optimal support recovery results for this procedure. To the best of our knowledge, our conditions improve upon the previously known ones to achieve support recovery for an IHT procedure. When $s\log(p)^3/n \to 0$, optimal conditions for the problem of support recovery in high dimensional linear regression under Gaussian design are provided in [12] using an iterative procedure without sample splitting. Similarly, our methodology does not require sample splitting. Moreover it achieves support recovery for a larger class of designs under a milder condition.
- We establish sharp optimal results under RIP as $\delta \to 0$. This in particular holds for sub-Gaussian designs as $s\log(ep/s)/n \to 0$. To the best of our knowledge, those are the first sharp estimation results to hold beyond Gaussian design.
- As for the optimization part, we use local strong convexity/local smoothness of the loss function (equivalent in our setting to RIP) in order to get fast global convergence as in [1] for convex penalized estimators. Our analysis makes it possible to study algorithms with non-convex penalization for instance the hard thresholding penalization. Using statistical properties of the model, we benefit from both local strong convexity and non-convex penalization in order to provide optimal worst-case computational guarantees of our procedure.

As a consequence of our methodology, we extend results of scaled minimax optimality to the regression model under RIP. In particular we close the gap by showing that

$$\psi(s,a) = (1+o(1))\frac{2\sigma^2 s\log(ep/s)}{n}, \quad \forall a \le (1-\varepsilon)\sigma\sqrt{\frac{2\log(ep/s)}{n}}$$

and

$$\psi(s,a) = (1+o(1))\frac{\sigma^2 s}{n}, \quad \forall a \ge (1+\varepsilon)\sigma\sqrt{\frac{2\log(ep/s)}{n}},$$

that holds for any $\varepsilon > 0$ and such that the limit corresponds to $s/p \to 0$ and $\delta \to 0$. Moreover, the upper bound is achieved through a polynomial time method that is fully adaptive. Some interesting questions arise based on adaptation to parameters on both optimization and statistical sides that we

address in the Conclusion. We summarize our contribution to the problem of minimax scaled sparse estimation below.

	Non asymptotic results	Sharp results $(a \leq (1-\varepsilon)a^*)$	Sharp results $(a \geq (1+\varepsilon)a^*)$
Minimax lower bounds	$C_4\sigma^2\frac{s\log(ep/s)}{n}$ [3]	$2\sigma^2\frac{s\log(ep/s)}{n}$ [16]	$\sigma^2\frac{s}{n}$ [16]
Risk of LASSO (not adaptive to sparsity)	$C_1\sigma^2\frac{s\log(ep/s)}{n}$ [3]	$2\sigma^2\frac{s\log(ep/s)}{n}$ [14]	$2\sigma^2\frac{s\log(ep/s)}{n}$ [2]
Risk of SLOPE (adaptive to sparsity)	$C_2\sigma^2\frac{s\log(ep/s)}{n}$ [3]	$2\sigma^2\frac{s\log(ep/s)}{n}$ [7]	$2\sigma^2\frac{s\log(ep/s)}{n}$ [2]
RISK of adaptive IHT	$C_3\sigma^2\frac{s\log(ep/s)}{n}$ This paper, Theorem 4	$2\sigma^2\frac{s\log(ep/s)}{n}$ this paper, Theorem 5	$\sigma^2\frac{s}{n}$ this paper, Theorem 6

TABLE 1. *Summary of minimax upper and lower bounds for estimation in high dimensional linear regression where* $a^* = \sigma\sqrt{\frac{2\log(ep/s)}{n}}$ *and* $C_1, C_2, C_3, C_4 > 0$ *some absolute constants. Sharp results hold for any* $0 < \varepsilon < 1$.

2. Non-asymptotic minimax sparse estimation: A new proof strategy. Classical non-asymptotic minimax results for sparse estimation in linear regression are proved for minimizers of well defined objective loss functions. In this section, we present and analyze our variant of iterative hard thresholding algorithm, and show similar minimax results. In what follows we assume that the design X satisfies the following condition. For an integer $s = 1, \ldots, p$, define $L_s, m_s > 0$ such that

$$L_s = \max_{|S|=s} \lambda_{\max}(X_S^\top X_S),$$

and

$$m_s = \min_{|S|=s} \lambda_{\min}\left(X_S^\top X_S\right).$$

Set $\delta_s := 1 - \frac{m_s}{L_s}$.

ASSUMPTION 1. *For $0 < c < 1$ and $s \in [p]$, we say that X satisfies RIP(s, c) if*

$$\delta_s \leq c.$$

In the rest of the paper, we assume that $s \leq p/3$ and that X satisfies RIP$(3s,\delta/2)$ for some $0 \leq \delta < 1$. Assumption 1 is equivalent to Restricted Strong Convexity and Restricted Smoothness on the set of s-sparse vectors, where we assume that $\gamma_s := \frac{L_s}{m_s}$, the *condition number*, is bounded. Here are few remarks concerning this assumption with respect to adaptation.

- Although our results should hold for general γ_s, we decided to consider only the case of bounded γ_s, and omit the dependence on γ_s, to make the presentation of our results simpler and also since a fully adaptive procedure would require an upper bound on γ_s .
- Our results hold under a relaxed assumption on the design where there exits some matrix M such that $MX^\top X$ satisfies a condition similar to RIP as in [14]. For instance, in the case of general Gaussian design with full rank covariance Σ, and for M chosen to be Σ^{-1}, the condition would hold under the usual assumption $s \log(ep/s) = O(n)$. Again, this requires knowing M in advance and constrains adaptation.
- For gradient descent step in IHT, the step size depends on γ_s or more precisely on L_s. When no upper bound on γ_s is unknown, there exist adaptive choices of the step size based on the exact line search for instance.

For all above reasons, we decided to only focus on adaptivity with respect to statistical parameters of the problem, namely s, σ and $\|\beta\|$, and to leave the general case with other results for further research.

Our procedure is an iterative hard thresholding algorithm where the threshold is updated at each step. For $\lambda > 0$, define the hard thresholding operator $\mathbf{T}_\lambda : \mathbf{R}^p \to \mathbf{R}^p$, such that

$$\forall u \in \mathbf{R}^p, \forall j = 1, \dots, p, \quad \mathbf{T}_\lambda(u)_j = u_j \mathbf{1}\{|u_j| \geq \lambda\}.$$

Notice that the usual IHT algorithm corresponds to $\mathbf{T}_{u_{(s)}}(u)$ where $u_{(s)}$ is the s largest entry of u, hence the threshold is data-dependent but most importantly sparsity dependent [6]. For a given sequence $(\lambda_m)_m$ of positive numbers, we define the corresponding sequence of estimators $(\hat{\beta}^m)_m$ such that $\hat{\beta}^0 = 0$ and for $m = 1, 2, \dots$

$$\hat{\beta}^m = \mathbf{T}_{\lambda_m}\left(\hat{\beta}^{m-1} + \frac{1}{\|X\|^2_{2,\infty}} X^\top (Y - X\hat{\beta}^{m-1})\right). \tag{4}$$

This procedure corresponds to a projected gradient descent on a non-convex set. The choice of normalizing the gradient by $\|X\|_{2,\infty}$ instead of L_{3s} is due to the fact that L_{3s} is not tractable and that we do not consider adaptivity with respect to optimization parameters (L_{3s}, δ_{3s}) in this work. If δ_{3s} is small enough, it is easy to see that $\|X\|_{2,\infty}$ is a good proxy for L_{3s}. The usual projection step consists in keeping the largest s components. The operator $\mathbf{T}_\lambda$ plays a similar role here. Imposing sparsity at each step of the procedure is crucial in order to benefit from restricted properties of the design. The novelty of our procedure lies in the fact that it implicitly grants the sparsity of our estimator at each step without having to choose exactly s components.

Unlike the analysis of IHT, in previous works, that benefits from local convex properties of the objective function, we choose to directly analyze the non-convex gradient descent algorithm and leverage the structure of both the signal and design in order to get a contraction of the error. We give here the intuition behind our procedure. In what follows we propose a specific choice for the sequence of thresholds $(\lambda_m)_m$ that will allow us to achieve both optimal statistical accuracy and fast convergence of the algorithm in an adaptive way. Let $\lambda_0, \lambda_\infty >$ and $0 < \kappa < 1$ be given constants, we define the sequence $(\lambda_m)_m$ as follows

$$\lambda_m = \kappa^{m/2}\lambda_0 \vee \lambda_\infty, \quad m = 0, 1, \dots . \tag{5}$$

The sequence of thresholds starts at some very large threshold λ_0, then keeps updating it linearly until reaching a final threshold given by λ_∞. A good choice of λ_∞ is given by the universal statistical threshold $\sqrt{\frac{2\sigma^2 \log(ep/s)}{\|X\|_{2,\infty}^2}}$. Our choice of the thresholding sequence is motivated by the following. Observe that

$$\hat{\beta}^m + \frac{1}{\|X\|_{2,\infty}^2} X^\top (Y - X\hat{\beta}^m) = \beta + \underbrace{\left(\frac{1}{\|X\|_{2,\infty}^2} X^\top X - \mathbf{I}_p \right)(\beta - \hat{\beta}^m)}_{\text{optimization error}} + \underbrace{\frac{\sigma}{\|X\|_{2,\infty}^2} X^\top \xi}_{\text{statistical error}} . \tag{6}$$

At each step, we can decompose the estimation error into two parts. An optimization error that may be reduced thanks to the local contraction of the design (RIP), and a statistical error that is unavoidable. While it is well understood that the choice of a threshold of order $\sqrt{\frac{2\sigma^2 \log(ep/s)}{\|X\|_{2,\infty}^2}}$ is optimal in order to control the statistical error, the same threshold does not grant

sparsity of the estimator at first steps. In fact, if the signal β is "well-spread" (i.e. its coordinates share similar magnitude) and $\|\beta\|$ large enough, it may occur that we select too many coordinates at the first step. This lack of sparsity makes it hard to benefit from restricted properties of the design. The alternative choice of a very large threshold grants sparsity but leads to a high statistical error. The usual choice of keeping the largest s-components at each step, is a natural fix. This intuition is similar to the motivation behind the LARS algorithm [10]. Our thresholding procedure is, in some sense, an interpolation between the two classical thresholds, namely the largest s-component for IHT, and $\sqrt{\frac{2\sigma^2\log(ep/s)}{\|X\|_{2,\infty}^2}}$ for LASSO. We start with a threshold large enough, then as we move forward the optimization error gets smaller which allows us to update the threshold without loosing the contraction. Our choice of thresholding sequence gives also an explicit stopping time, as we may stop once the threshold hits the universal statistical threshold. In the rest of the paper, we use the following notation:

$$\Xi := \frac{\sigma}{\|X\|_{2,\infty}^2} X^\top \xi \quad \text{and} \quad \Phi := \left(\frac{1}{\|X\|_{2,\infty}^2} X^\top X - \mathbf{I}_p \right).$$

It is useful to observe that as long as X satisfies RIP($3s$,$\delta/2$), then Φ is a contraction for $3s$-sparse vectors. This is rephrased in the following Lemma that we prove in the Appendix.

LEMMA 1. *If X satisfies RIP($3s$,$\delta/2$) then for all $S \subset \{1,\dots,p\}$ such that $|S| \le 3s$, we have $\lambda_{\max}(\Phi_{SS}) \le \delta$.*

Before stating our results we give a first result that is relevant to the analysis of our algorithm. We draw the reader's attention, that our analysis is fully deterministic since we place ourselves in a well chosen random event that captures the complexity of our model. Namely, we consider the event

$$\mathscr{O} = \left\{ \sum_{i=1}^{s} \Xi_{(i)}^2 \le \frac{10\sigma^2 s \log(ep/s)}{\|X\|_{2,\infty}^2} \right\}.$$

Using Lemma 3 (cf. Appendix), the event $\mathscr{O}$ holds with high probability. Conditionally on the event $\mathscr{O}$, the next Theorem shows that, at each step, the corresponding estimator is $2s$-sparse and the surrogate function of the estimation error, given by $s\lambda_m^2$, decreases exponentially.

THEOREM 1. *Assume that β is s-sparse, that is $|\beta|_0 \le s$ and that X satisfies RIP($3s$,$\delta/2$). We denote by S the support of β. Let $\lambda_0, \lambda_\infty > 0$, $0 <$*

$\kappa < 1$, and define $(\hat{\beta}^m)_m$ and its corresponding thresholding sequence $(\lambda_m)_m$ as in (4)-(5). Assume that $\delta < 1/36 \vee \kappa$, $\|\beta\| \leq \sqrt{s}\lambda_0$ and $\frac{\sigma\sqrt{40\log(ep/s)}}{\|X\|_{2,\infty}} \leq \lambda_\infty$. If $\mathscr{O}$ holds, then for all m, we have

$$|\hat{\beta}^m_{S^c}|_0 \leq s, \tag{7}$$

and

$$\|\hat{\beta}^m - \beta\|^2 \leq 9s\lambda_m^2. \tag{8}$$

Unlike for convex regularized least squares, the objective function $\|\hat{\beta}^m - \beta\|^2$ does not decrease at each step. Indeed, our gradient descent step may be trapped in some saddle points for instance when there is a big gap between coordinates of β. We get around this issue by finding an upper bounding surrogate function that decreases exponentially. This can be viewed as the highlight of our non-convex approach. The sequence λ_m decreases until it reaches the stationary threshold λ_∞. If we tune λ_∞ with the universal statistical threshold then the final estimation error is optimal. In that case, we may stop the algorithm after a number of steps of order $\log\left(\frac{s\lambda_0^2\|X\|_{2,\infty}^2}{\sigma^2 s\log(ep/s)} \vee 2\right)/\log(1/\kappa)$. Notice that we do not need to know the precise value of δ here but only an upper bound, that we set to $1/36$ in Theorem 1. We did not try to optimize this value. We may also set λ_0 as large as possible and our algorithm would still output an estimator that attains the minimax optimal rate. If $s\lambda_0^2$ is much larger than $\|\beta\|^2$, then the result of Theorem 1 is not optimal from an optimization perspective, in the sense that it would take more steps to stop compared to IHT for instance. In order to achieve both optimal statistical accuracy and optimal fast convergence, we need λ_0 to be roughly of the same order as $\frac{\|\beta\|}{\sqrt{s}} \vee \sigma\frac{\sqrt{\log(ep/s)}}{\|X\|_{2,\infty}}$. The optimal choices of λ_0 and the stopping time m depend on $\|\beta\|$. In order to derive minimax optimal results, we need to make these choices adaptive with respect to $\|\beta\|$. We show next how to tune λ_0 and m in order to grant linear convergence of our procedure. Denote by M the vector

$$M := \frac{1}{\|X\|_{2,\infty}^2}X^\top Y = \beta + \Phi\beta + \Xi,$$

and set

$$\hat{\lambda}_0 = \sqrt{\frac{10\sum_{i=1}^s M_{(i)}^2}{s}} \vee \frac{\sigma}{\|X\|_{2,\infty}}\sqrt{40\log(ep/s)}, \tag{9}$$

and

$$\hat{m} = \left\lfloor 2\log\left(\frac{\hat{\lambda}_0^2\|X\|_{2,\infty}^2}{40\sigma^2\log(ep/s)}\right)/\log(1/\kappa)\right\rfloor + 1. \tag{10}$$

PROPOSITION 1. *Let β be s-sparse and let X satisfy RIP($3s$,$\delta/2$). Assume that $\delta \le 1/4$ and that event $\mathscr{O}$ holds. Then*

$$\left(\|\beta\| \vee \frac{\sigma\sqrt{10s\log(ep/s)}}{\|X\|_{2,\infty}}\right) \le \sqrt{s}\hat{\lambda}_0 \le 10\left(\|\beta\| \vee \frac{\sigma\sqrt{10s\log(ep/s)}}{\|X\|_{2,\infty}}\right),$$

and

$$2\log\left(\frac{\|\beta\|^2\|X\|_{2,\infty}^2}{40\sigma^2 s\log(ep/s)} \vee 1/4\right)/\log(1/\kappa) \le \hat{m}-1 \le 2\log\left(\frac{5\|\beta\|^2\|X\|_{2,\infty}^2}{\sigma^2 s\log(ep/s)} \vee 25\right)/\log(1/\kappa).$$

Observe that for large values of $\|\beta\|$, $\hat{\lambda}_0$ is of the same order as $\frac{\|\beta\|}{\sqrt{s}}$ with high probability and that $\hat{m}$ is of order $\log\left(\frac{\|\beta\|^2\|X\|_{2,\infty}^2}{\sigma^2 s\log(ep/s)}\right)/\log(1/\kappa)$. We are now ready to state a result of the minimax optimality of our procedure. We will also pick λ_∞ to be of the same order as the universal threshold

$$\hat{\lambda}_\infty = \frac{\sigma\sqrt{40\log(ep/s)}}{\|X\|_{2,\infty}}. \tag{11}$$

THEOREM 2. *Let $0 < \kappa < 1$. Assume that $\delta \le 1/36 \vee \kappa$ and that X satisfies RIP($3s$,$\delta/2$). Let $\hat{\lambda}_0$ and $\hat{\lambda}_\infty$ given by (9) and (11), $(\lambda_m)_m$ be the corresponding sequence of estimators (5), and $\hat{m}$ be the stopping time (10). Then the following holds*

$$\sup_{|\beta|_0\le s} \mathbf{P}_\beta\left(\|\hat{\beta}^{\hat{m}} - \beta\|^2 \ge \frac{360\sigma^2 s\log(ep/s)}{\|X\|_{2,\infty}^2}\right) \le e^{-c_1 s\log(ep/s)},$$

$$\sup_{|\beta|_0\le s} \mathbf{P}_\beta\left(|\hat{\beta}^{\hat{m}}|_0 \ge 2s\right) \le e^{-c_1 s\log(ep/s)},$$

and

$$\sup_{|\beta|_0\le s} \mathbf{P}_\beta\left(\hat{m} \ge 2\log\left(\frac{5\|\beta\|^2\|X\|_{2,\infty}^2}{\sigma^2 s\log(ep/s)} \vee 25\right)/\log(1/\kappa) + 1\right) \le e^{-c_1 s\log(ep/s)},$$

for some absolute constant $c_1 > 0$.

Theorem 2 shows that $\hat{\beta}^{\hat{m}}$ achieves optimal statistical accuracy in linear time. Notice that $\hat{m}$ depends on $\log(1/\kappa)$ instead of $\log(1/\delta_{3s})$ simply because δ_{3s} is intractable. As we emphasized earlier, optimal results that are adaptive to γ_{3s} (or equivalently δ_{3s}) fall beyond the scope of this paper. Theorem 2 shows minimax optimality of an estimator constructed through a non-convex algorithmic regularization scheme. Similar estimation error is also achieved using the SLOPE estimator as in [3]. The advantage of the proposed estimator is that it achieves the optimal statistical accuracy in linear time.

3. A fully adaptive minimax optimal procedure. The above minimax estimator depends on the statistical parameters of the model s and σ, and this only through the thresholding sequence $(\lambda_m)_m$. In particular, the choices of $\hat{\lambda}_0$, $\hat{\lambda}_\infty$, and $\hat{m}$ depend on s and σ. For estimation of σ, we may consider the sequence of estimators $\hat{\sigma}_m^2$ of σ^2 such that

$$\hat{\sigma}_m^2 = \frac{\|Y - X\hat{\beta}^m\|^2}{n}. \tag{12}$$

In practice, estimator (12) is considered in the Square-Root Lasso [4] among others in order to adapt to the noise level in high-dimensional regression. During the first steps, $\hat{\sigma}_m^2$ may have bad performance but in this case this means that the threshold λ_m is much larger than the universal statistical threshold making precise estimation of σ not necessary. As we get closer to the final threshold the estimation error gets smaller and estimation of σ is improved as long as $n = \Omega(s\log(ep/s))$. The latter condition is sufficient in order to achieve good estimation of σ. It is worth saying that the same condition is not more restrictive than RIP. Indeed, a consequence of Corollary 7.2 in [3] implies that $n = \Omega(s\log(ep/s))$ as long as X satisfies RIP.

Concerning the choice of the initial threshold, recall that Theorem 2 hold if we replace $\hat{\lambda}_0$ by any upper bound, off to running more iterations. Hence, we can replace $\hat{\lambda}_0$ by the adaptive initial threshold

$$\bar{\lambda}_0 = \sqrt{20}|M|_{(1)} \vee \frac{\hat{\sigma}_0}{\|X\|_{2,\infty}}\sqrt{160\log(ep)}. \tag{13}$$

Based on the fact that $u_{(1)} \geq \frac{1}{s}\sum_{i=1}^s u_{(i)}^2$, threshold (13) is indeed an upper bound for the initial choice $\hat{\lambda}_0$ in (9). Notice that the threshold $\bar{\lambda}_0$ in (13) can be as large as $\|\beta\|$ and not $\|\beta\|/\sqrt{s}$ as it was for $\hat{\lambda}_0$. This loss only appears in the number of iterations where we may run our algorithm for $\log(s)$ more steps. We present now two fully adaptive procedures that are minimax optimal.

3.1. *Adaptive early stopping.* We may now define a new adaptive thresholding sequence as follow

$$\lambda_m = \kappa^{m/2}\lambda \vee \frac{\hat{\sigma}_m}{\|X\|_{2,\infty}}\sqrt{160\log(ep)}, \quad m = 0, 1, \dots. \tag{14}$$

Our choice of m is adaptive as well and given by

$$\bar{m} = \inf\left\{ m/\lambda_m \leq \frac{\hat{\sigma}_m}{\|X\|_{2,\infty}}\sqrt{160\log(ep)} \right\} + 1. \tag{15}$$

Observe that the adaptive stopping rule is exactly given by the step when we hit the statistical threshold. We can now state a minimax optimal result corresponding to our fully adaptive procedure.

THEOREM 3. *Let $0 < \kappa < 1$. Assume that $\delta \leq 1/36 \vee \kappa$, that $n > 14000 s\log(ep)$ and that $\mathbf{E}(\xi\xi^\top) = \mathbf{I}_n$. Let $\bar{\lambda}_0$ and $(\hat{\sigma}_m)_m$ be defined as in (13) and (12), $(\lambda_m)_m$ be the corresponding sequence (14) and $\bar{m}$ be the stopping time (15). Then the following holds*

$$\sup_{|\beta|_0 \leq s} \mathbf{P}_\beta\left(\|\hat{\beta}^{\bar{m}} - \beta\|^2 \geq \frac{4000\sigma^2 s\log(ep)}{\|X\|_{2,\infty}^2} \right) \leq e^{-c_1 s\log(ep/s)},$$

$$\sup_{|\beta|_0 \leq s} \mathbf{P}_\beta\left(|\hat{\beta}^{\bar{m}}|_0 \geq 2s \right) \leq e^{-c_1 s\log(ep/s)},$$

and

$$\sup_{|\beta|_0 \leq s} \mathbf{P}_\beta\left(\bar{m} \geq 2\log\left(\frac{10\|\beta\|^2\|X\|_{2,\infty}^2}{\sigma^2\log(ep)} \vee 100 \right) / \log(1/\kappa) + 1 \right) \leq e^{-c_1 s\log(ep/s)},$$

for some absolute $c_1 > 0$.

The adaptive procedure of Theorem 3 is minimax optimal up to a logarithmic factor (replacing $\log(ep/s)$ by $\log(ep)$). Conditions $n = \Omega(s\log(ep))$ and $\mathbf{E}(\xi\xi^\top) = \mathbf{I}_n$ are only required for adaptation to the noise level σ. Finally the number of iterations may be larger by $\log(s)$ compared to analogous non adaptive results. Hence, up to a logarithmic loss in both statistical accuracy and optimization speed, our early stopping procedure is fast, fully adaptive and minimax optimal.

3.2. *Iteration selection.* In order to capture the optimal dependence with respect to s, we rely on a different approach. One of the most popular methods to achieve adaptation is through the lens of model selection introduced in [5]. Given a set of models (or estimators) one picks a good estimator based on some criterion. More concretely, one may think of Cross-Validation where for each regularization parameter λ an estimator is constructed, then the resulting estimators are either aggregated or one of them is chosen based on some criterion. In what follows, we present an adaptive procedure in the same flavor. We like to see our procedure as an iteration selection method instead of model selection. Indeed, we can think of our $m-$th iteration as an estimator corresponding to λ_m. With this analogy in mind, penalized model selection boils down to a penalized iteration selection. Observe that iteration selection is much faster than the classical model selection since each estimator is computed using only one iteration initialized with the previous estimator. More concretely, we construct all iterations corresponding to the thresholding sequence

$$\lambda_m = \kappa^{m/2}\bar{\lambda}_0, \quad \forall m \in [\hat{T}], \tag{16}$$

where $\bar{\lambda}_0$ was defined in (13) and $\hat{T}$ is defined below. The selected iteration is given by

$$\tilde{m} = \arg\min_{m\in[\hat{T}]} \left\{ \frac{1}{n}\|Y - X\hat{\beta}^m\|^2 + \frac{10\hat{\sigma}^2_{\bar{m}}|\hat{\beta}^m|_0 \log(ep/|\hat{\beta}^m|_0)}{n} \right\}, \tag{17}$$

where $\hat{\sigma}_{\bar{m}}$ was defined above. Again the problem of estimation of σ is easier as long as $s\log(ep) = O(n)$ and we may replace $\hat{\sigma}_{\bar{m}}$ by any good estimator of σ. For completeness of our result we decided to stop the search domain over m once the thresholding sequence $(\lambda_m)_m$ is below $\frac{\sigma}{\|X\|_{2,\infty}}$ where solutions are not granted to be sparse anymore. For that reason we set $\hat{T}$ such that

$$\hat{T} = \inf\left\{ m \geq 0 / \lambda_m \leq \frac{4\hat{\sigma}_{\bar{m}}}{\|X\|_{2,\infty}} \right\},$$

where $(\lambda_m)_m$ is defined in (16). We get the following result.

THEOREM 4. *Let $0 < \kappa < 1$. Assume that $\delta \leq 1/36 \vee \kappa$ and that $n > 14000 s\log(ep)$ and that $\mathbf{E}(\xi\xi^\top) = \mathbf{I}_n$. Let $\tilde{m}$ be defined in (17), then the following holds*

$$\sup_{|\beta|_0\leq s} \mathbf{P}_\beta \left(\|\hat{\beta}^{\tilde{m}} - \beta\|^2 \geq \frac{100^2\sigma^2 s\log(ep/s)}{\|X\|^2_{2,\infty}} \right) \leq e^{-c_1 s\log(ep/s)},$$

$$\sup_{|\beta|_0 \le s} \mathbf{P}_\beta \left(|\hat{\beta}^{\tilde{m}}|_0 \ge 3s \right) \le e^{-c_1 s \log(ep/s)},$$

and

$$\sup_{|\beta|_0 \le s} \mathbf{P}_\beta \left(\hat{T} \ge 2 \log \left(\frac{10 \|\beta\|^2 \|X\|_{2,\infty}^2}{\sigma^2} \vee 100 \right) / \log(1/\kappa) + 1 \right) \le e^{-c_1 s \log(ep/s)},$$

for some absolute $c_1 > 0$.

The iteration selection procedure is different compared to the early stopping one since it chooses the best threshold λ_m instead of worrying about tuning the stopping rule. Moreover it achieves the minimax optimal rate of $\sigma^2 s \log(ep/s)/\|X\|_{2,\infty}^2$ adaptively to all parameters under the mild condition $s \log(ep) = O(n)$. Notice also that the number of constructed estimators $\hat{T}$ is small with overwhelming probability. Overall, full optimal adaptation on the statistics side comes with the price of $\log(s \log(ep/s))$ more steps on the optimization side. The our knowledge, Theorem 4 is the first to provide a fast and adaptive procedure that is minimax optimal.

4. Sharp results and scaled minimax optimality. In this section we present sharp minimax results for the problem of estimation in high dimensional linear regression. Moreover, we show that a variant of our estimator is scaled minimax optimal, improving upon regularized convex estimators that provably suffer from an unavoidable bias term. Our final procedure is an IHT algorithm with fixed threshold that is initialized by $\hat{\beta}^{\tilde{m}}$ defined earlier (Theorem 4). By analogy with non-convex optimization, the step where the initialization is constructed plays the role of the first iterations getting to the basin of attraction. For a given $\lambda > 0$, our final procedure $(\tilde{\beta}^m)_m$ is a variant of IHT where $\tilde{\beta}^0 = \hat{\beta}^{\tilde{m}}$ and for all $m \ge 1$

$$\tilde{\beta}^m = \mathbf{T}_\lambda \left(\tilde{\beta}^{m-1} + \frac{1}{\|X\|_{2,\infty}^2} X^\top (Y - X\tilde{\beta}^{m-1}) \right). \tag{18}$$

For any $\epsilon > 0$, define the statistical threshold given by

$$\lambda_\infty^\epsilon = (1 + \sqrt{\epsilon}) \frac{\sigma \sqrt{2 \log(ep/s)}}{\|X\|_{2,\infty}}. \tag{19}$$

REMARK 1.
- *We can replace* $\tilde{\beta}^0$ *by any estimator that is minimax optimal and is at most* $2s$*- sparse. In particular one may choose to initialize our procedure with square-root slope [9]. Our choice of initialization is not only adaptive but is also fast to compute.*

- *Sharp adaptation to σ can be achieved by choosing $\tilde{\sigma}^2 := \frac{1}{n}\|Y - X\hat{\beta}^{\tilde{m}}\|^2$. It is easy to observe that $\tilde{\sigma} = \sigma(1 + o(1))$ with overwhelming probability under the mild assumption $s\log(ep)/n \to 0$. Hence, all our sharp results hold adaptively to σ under the above condition.*

Sharp results are stated under the standard conditions $s/p \to 0$ and $\delta \to 0$ as $p \to \infty$. Indeed, the first condition is relevant to observe a strict change of behaviour between biased and unbiased estimators and the second one corresponds to $s\log(ep/s)/n \to 0$ under Gaussian design. We remind the reader that s, n, δ depend on p as $p \to \infty$. Our next result states that one step is enough to achieve sharp optimal minimax estimation.

THEOREM 5. *Let $\epsilon \in (0,1)$ and $\lambda^{\epsilon}_{\infty}$ given by (19). Assume that $\delta \le \epsilon \vee 1/400^2$. Let $\lambda \ge \lambda^{\epsilon}_{\infty}$ and let $\tilde{\beta}^m$ be the corresponding sequence of estimators (18). Then, for any $m \ge 0$ we have*

$$\lim_{s/p\to 0}\sup_{|\beta|_0\le s} \mathbf{P}_\beta\Big(\|\tilde{\beta}^m - \beta\| \ge (1 + 4\sqrt{\delta} + 100\delta^{m/2} + o(1))\sqrt{s}\lambda\Big) = 0.$$

Theorem 5 is sharp in the sense that when $\delta \to 0$, then for any $\epsilon > 0$, there exists an estimator (depending on ϵ) achieving the asymptotic minimax error of $(1+\epsilon)\frac{2\sigma^2 s\log(ep/s)}{\|X\|^2_{2,\infty}}$ and that estimator corresponds to the threshold $\lambda = \lambda^{\epsilon}_{\infty}$. Replacing $\log(ep/s)$ by $\log(ep)$ in this choice of λ leads to an adaptive nearly sharp minimax optimal procedure. A similar result was shown for the SLOPE estimator in [21] under Gaussian isotropic designs and in [14] for more general Gaussian designs with known covariance. Our sharp results do not require σ to be known nor the design to be Gaussian, since we conduct a deterministic analysis over the design.

Another advantage of our procedure, is that it eliminates the usual bias due to regularization under almost optimal conditions. Namely, as long as the informative signal components are well separated from zero, then our estimator achieves the same rate of estimating an s-sparse vector as if its support were known. The next result proves that our procedure is scaled minimax optimal.

THEOREM 6. *Let $a > 0$ and $\epsilon \in (0,1)$. Assume that conditions of Theorem 5 hold and that $s \to \infty$. If $a \ge \lambda(1 + \sqrt{\epsilon})$, then $\forall m \ge \log(\log(ep/s))$ we have*

$$\lim_{s\to\infty, s/p\to 0}\sup_{\beta\in\Omega_{s,a}} \mathbf{P}_\beta\left(\|\tilde{\beta}^m - \beta\| \ge (1 + 4\sqrt{\delta} + o(1))\frac{\sigma^2 s}{\|X\|^2_{2,\infty}}\right) = 0.$$

Notice first, that if $a = \Omega\left(\frac{\sigma\sqrt{\log(ep/s)}}{\|X\|_{2,\infty}}\right)$, then we can construct an estimator $\tilde{\beta}^m$ achieving the non-asymptotic minimax parametric statistical error of order $\frac{\sigma^2 s}{\|X\|^2_{2,\infty}}$. For the sharp counterpart, observe that for any $\epsilon > 0$, if $\delta \to 0$ and $s \to \infty$, then there exists an estimator $\tilde{\beta}^m$ (depending on ϵ) that achieves the optimal error of $(1 + o(1))\frac{\sigma^2 s}{\|X\|^2_{2,\infty}}$ w.h.p under the condition $a \geq (1+\epsilon)\frac{\sigma\sqrt{2\log(ep/s)}}{\|X\|_{2,\infty}}$. The last condition on a is necessary in order to achieve such a result as shown in [16]. This result shows that $\tilde{\beta}^m$ is sharply scaled minimax optimal as long as $m \geq \log(\log(ep/s))$. Again full adaptation is granted replacing $\log(ep/s)$ by $\log(ep)$ which leads to nearly sharp optimal results.

5. On support recovery. Throughout the paper our proofs are based on simultaneous analysis of both estimation error and variable selection. As a consequence, we can also recover results for support recovery. For a given vector β, we denote by η the corresponding decoder i.e $\eta_i = \mathbf{1}(\beta_i \neq 0)$. We first give a straightforward result for almost full recovery , i.e $\frac{|\hat{\eta}-\eta|}{s} \to 0$, based on the previous section.

THEOREM 7. *Under the conditions of Theorem 6, we get, for all* $m \geq \log(\log(ep/s))$, *that*

$$\lim_{s/p\to 0}\sup_{\beta\in\Omega_{s,a}} \mathbf{P}_\beta\left(\frac{|\tilde{\eta}^m - \eta|}{s} \geq \omega_p\right) = 0,$$

for some $\omega_p \to 0$.

It comes out that $\tilde{\eta}^m$ achieves almost full recovery under the nearly optimal condition

$$a \geq (1+\epsilon)\frac{\sigma\sqrt{2\log(ep/s)}}{\|X\|_{2,\infty}},$$

for any $\epsilon \geq \delta$ as long as $s/p \to 0$. This sufficient condition is moreover optimal as $\delta \to 0$ by reduction to the Gaussian sequence model studied in [8]. These results improve upon state-of-the-art recovery results in compressed sensing, and in particular results of [17], where authors use a two-stage procedure and sample splitting leading them to a strict loss in the sharp constants. Again our results, as opposed to most of the literature of support recovery, do not assume the design to be Gaussian nor sub-Gaussian.

REMARK 2. *All sharp results are stated for given ϵ. The procedure we construct depends on ϵ. This is due to the fact that we do not have access to a sharp upper bound on δ. In compressed sensing under isotropic sub-Gaussian design, we can replace ϵ by $\frac{s\log(ep/s)}{n}$ for instance. In this case, we can achieve sharp optimal results for any level ϵ.*

In order to prove similar results for support recovery we rely on a different proof strategy. Our result for support recovery does not require sample splitting compared to [17] and is more general than [12] since it is fully adaptive. For any $\epsilon > 0$, the statistical threshold for support recovery is given by

$$\mu_\infty^\epsilon = (1+\sqrt{\epsilon})\frac{\sigma\sqrt{2\log(p)}}{\|X\|_{2,\infty}}, \tag{20}$$

Define the least square solution, given the true support S, such that

$$\tilde{\beta}^* = ((X^\top X)_{SS})^{-1}X_S^\top Y.$$

Then the following result holds.

THEOREM 8. *Assume that conditions of Theorem 5 hold, and $a \geq (1+3\sqrt{\epsilon})\frac{\sigma(\sqrt{2\log(p)}+\sqrt{2\log(s)})}{\|X\|_{2,\infty}}$. Let $(\tilde{\beta}^m)_m$ be the sequence of estimators defined in (18) corresponding to $\lambda = \mu_\infty^\epsilon$ defined in (20). Then, we have for all $m \geq 0$ that*

$$\lim_{s/p\to 0}\sup_{\beta\in\Omega_{s,a}} \mathbf{P}_\beta\left(\|\tilde{\beta}^m - \tilde{\beta}^*\|^2 \geq \frac{150^2(10\delta)^m\sigma^2 s\log(ep/s)}{\|X\|_{2,\infty}^2}\right) = 0,$$

As a consequence, we get that for $m \geq \log(s)$

$$\lim_{s/p\to 0}\sup_{\beta\in\Omega_{s,a}} \mathbf{P}_\beta\left(|\tilde{\eta}^m - \eta| > 0\right) = 0.$$

Hence under the minimal separation condition for support recovery our estimator converges to the oracle least square solution $\tilde{\beta}^m \to \tilde{\beta}^*$ in probability even for non vanishing δ. Notice that $\tilde{\beta}^*$ has the same support as β a.s. It turns out that

$$a \geq (1+3\sqrt{\epsilon})\frac{\sigma(\sqrt{2\log(p)}+\sqrt{2\log(s)})}{\|X\|_{2,\infty}},$$

for any $\epsilon > 0$ is sufficient to achieve exact recovery as long as $\delta \to 0$, $s/p \to 0$. This condition is shown to be necessary in [8] under orthogonal design. We also recover the results of [12] for Gaussian design. Moreover, our approach is fully adaptive and holds beyond Gaussian design.

6. Conclusion. In this paper, we have presented a novel non-asymptotic minimax optimal estimation procedure for high dimensional linear regression. Our procedure is moreover fast and fully adaptive. We also provided sharp asymptotic results beyond the Gaussian design assumption. In particular, our procedure is scaled minimax optimal (i.e. unbiased whenever it is possible). Moreover, optimal results for both exact and almost full recovery were established as $\delta \to 0$. We conclude that our procedure has many attractive properties under the high dimensional linear regression model.

As potential extensions of our results, we believe that full adaptation with respect to the optimization parameters δ_s and L_s has its own interest. Moreover, our results in their actual form do not have the optimal dependence in terms of the condition number γ_s, and it would be interesting to generalize them beyond the RIP condition. Finally, another direction of interest is robust estimation through algorithmic regularization, where our strategy may be used to construct robust estimators with some of the desired properties we have in this paper. We leave all these questions for further research.

Acknowledgements. I would like to thank Alexandre Tsybakov and Pierre Bellec for valuable comments on early versions of this manuscript. This work was partially supported by a James H. Zumberge Faculty Research and Innovation Fund at the University of Southern California and by the National Science Foundation grant CCF-1908905.

References.

[1] Agarwal, A., Negahban, S. and Wainwright, M. J. (2010). Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems* 37–45.

[2] Bellec, P. C. (2018). The noise barrier and the large signal bias of the Lasso and other convex estimators. *arXiv preprint arXiv:1804.01230*.

[3] Bellec, P. C., Lecué, G., Tsybakov, A. B. et al. (2018). Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics* **46** 3603–3642.

[4] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806.

[5] Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society* **3** 203–268.

[6] Blumensath, T. and Davies, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis* **27** 265–274.

[7] Bogdan, M., Van Den Berg, E., Sabatti, C., Su, W. and Candès, E. J. (2015). SLOPEadaptive variable selection via convex optimization. *The annals of applied statistics* **9** 1103.

[8] Butucea, C., Ndaoud, M., Stepanova, N. A. and Tsybakov, A. B. (2018). Variable selection with Hamming loss. *The Annals of Statistics* **46** 1837–1875.

[9] Derumigny, A. et al. (2018). Improved bounds for square-root lasso and square-root slope. *Electronic Journal of Statistics* **12** 741–766.

[10] Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004). Least angle regression. *The Annals of statistics* **32** 407–499.
[11] Feng, L. and Zhang, C.-H. (2017). Sorted Concave Penalized Regression. *arXiv preprint arXiv:1712.09941*.
[12] Gao, C. and Zhang, A. Y. (2019). Iterative Algorithm for Discrete Structure Recovery.
[13] Hsu, D., Kakade, S., Zhang, T. et al. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability* **17**.
[14] Javanmard, A. and Montanari, A. (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics* **46** 2593–2622.
[15] Liu, H. and Barber, R. F. (2018). Between hard and soft thresholding: optimal iterative thresholding algorithms. *arXiv preprint arXiv:1804.08841*.
[16] Ndaoud, M. (2018). Interplay of minimax estimation and minimax support recovery under sparsity. *arXiv preprint arXiv:1810.05478*.
[17] Ndaoud, M. and Tsybakov, A. B. (2018). Optimal variable selection and adaptive noisy Compressed Sensing. *arXiv preprint arXiv:1809.03145*.
[18] Reeves, G., Xu, J. and Zadik, I. (2019). The all-or-nothing phenomenon in sparse linear regression. *arXiv preprint arXiv:1903.05046*.
[19] Rudelson, M., Vershynin, R. et al. (2013). Hanson-Wright inequality and subgaussian concentration. *Electronic Communications in Probability* **18**.
[20] Shen, J. and Li, P. (2017). A tight bound of hard thresholding. *The Journal of Machine Learning Research* **18** 7650–7691.
[21] Su, W. and Candes, E. (2016). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics* **44** 1038–1068.
[22] Wang, Y., Zeng, J., Peng, Z., Chang, X. and Xu, Z. (2015). Linear convergence of adaptively iterative thresholding algorithms for compressed sensing. *IEEE Transactions on Signal Processing* **63** 2957–2971.
[23] Yuan, X., Li, P. and Zhang, T. (2016). Exact recovery of hard thresholding pursuit. In *Advances in Neural Information Processing Systems* 3558–3566.
[24] Yuan, X.-T., Li, P. and Zhang, T. (2018). Gradient hard thresholding pursuit. *Journal of Machine Learning Research* **18** 1–43.
[25] Zhang, T. (2011). Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE transactions on information theory* **57** 4689–4708.
[26] Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics* **38** 894–942.
[27] Zhou, P., Yuan, X. and Feng, J. (2018). Efficient stochastic gradient hard thresholding. In *Advances in Neural Information Processing Systems* 1988–1997.

APPENDIX A: PROOFS OF NON-ASYMPTOTIC RESULTS

Proof of Lemma 1. Recall that X satisfies RIP$(3s, \delta/2)$. It is easy to observe that $m_{3s} \leq \|X\|_{2,\infty}^2 \leq L_{3s}$. Hence for $S \in \{1, \ldots, p\}$ such that $|S| \leq 3s$, we have

$$\lambda_{\max}(\Phi_{SS}) \leq \left(1 - \frac{m_{3s}}{L_{3s}}\right) \vee \left(\frac{L_{3s}}{m_{3s}} - 1\right).$$

Since $1 - \frac{m_{3s}}{L_{3s}} \leq \delta/2$, it remains to prove that

$$\frac{L_{3s}}{m_{3s}} - 1 \leq \delta.$$

Using that fact that $m_{3s} \geq (1 - \delta/2)L_{3s}$, it comes that

$$\frac{L_{3s}}{m_{3s}} - 1 \leq \frac{\delta/2}{1 - \delta/2} \leq \delta,$$

since $\delta \leq 1$. □

Proof of Theorem 1. We proceed by induction. For $m = 0$, The result is obvious. We now assume the result true for m and prove it for $m + 1$. In what follows let us denote by H^{m+1} the vector

$$H^{m+1} = \hat{\beta}^m + \frac{1}{\|X\|_{2,\infty}^2} X^\top (Y - X\hat{\beta}^m).$$

Notice that H^{m+1} can be written in the form

$$H^{m+1} = \beta + \Phi(\beta - \hat{\beta}^m) + \Xi,$$

and that

$$\hat{\beta}^{m+1} = \mathbf{T}_{\lambda_{m+1}}(H^{m+1}).$$

We prove the first part of the result reasoning by the absurd. Assume that $|\hat{\beta}^{m+1}_{S^c}|_0 > s$. Then there exists a subset $\tilde{S}$ of S^c such that $|\tilde{S}|_0 = s$ and

$$s\lambda_{m+1}^2 \leq \sum_{i \in \tilde{S}} (H_i^{m+1})^2 \mathbf{1}\{|H_i^{m+1}| \geq \lambda_{m+1}\}.$$

Since $\tilde{S}$ is not supported on S, then we have

$$\sqrt{s}\lambda_{m+1} \leq \sqrt{\sum_{i \in \tilde{S}} \Xi_i^2} + \sqrt{\sum_{i \in \tilde{S}} \langle \Phi_i^\top, \beta - \hat{\beta}^m \rangle^2}.$$

Since β is s-sparse and $|\hat{\beta}^m_{S^c}|_0 \leq s$ then $\beta - \hat{\beta}^m$ is at most $2s$-sparse. Moreover $|\tilde{S}| = s$. Hence using Lemma 1 we have that

$$\sqrt{s}\lambda_{m+1} \leq \sqrt{\sum_{i=1}^{s} \Xi^2_{(i)}} + \delta\|\beta - \hat{\beta}^m\|.$$

Using the induction hypothesis and event $\mathscr{O}$, we get moreover that

$$\begin{aligned}\sqrt{s}\lambda_{m+1} &\leq \frac{\sqrt{10\sigma^2 s\log(ep/s)}}{\|X\|_{2,\infty}} + 3\delta\sqrt{s}\lambda_m \\ &\leq (1/2 + 3\sqrt{\delta})\sqrt{s}\lambda_{m+1} < \sqrt{s}\lambda_{m+1},\end{aligned}$$

as long as $\delta < 1/36$, which is absurd. Hence $|\hat{\beta}^{m+1}_{S^c}|_0 \leq s$. For the second part, observe that $\forall i \in S$,

$$\hat{\beta}^{m+1}_i - \beta_i = -H^{m+1}_i \mathbf{1}\{|H^{m+1}_i| \leq \lambda_{m+1}\} + \Xi_i + \langle \Phi_i^\top, \beta - \hat{\beta}^m\rangle.$$

Then using the same arguments as before we have

$$\|\hat{\beta}^{m+1}_S - \beta\| \leq \sqrt{s}\lambda_{m+1} + \frac{\sqrt{10\sigma^2 s\log(ep/s)}}{\|X\|_{2,\infty}} + \delta\|\hat{\beta}^m - \beta\|.$$

Moreover on S^c, we have

$$\|\hat{\beta}^{m+1}_{S^c}\| \leq \frac{\sqrt{10\sigma^2 s\log(ep/s)}}{\|X\|_{2,\infty}} + \delta\|\hat{\beta}^m - \beta\|.$$

Hence

$$\|\hat{\beta}^{m+1} - \beta\| \leq \sqrt{s}\lambda_{m+1} + 2\frac{\sqrt{10\sigma^2 s\log(ep/s)}}{\|X\|_{2,\infty}} + 2\delta\|\hat{\beta}^m - \beta\|.$$

Using the definition of λ_m and the induction hypothesis, we get that

$$\|\hat{\beta}^{m+1} - \beta\| \leq \sqrt{s}\lambda_{m+1} + \sqrt{s}\lambda_{m+1} + 6\delta\sqrt{s}\lambda_m.$$

We conclude that

$$\|\hat{\beta}^{m+1} - \beta\| \leq \sqrt{s}\lambda_{m+1}(2 + 6\sqrt{\delta}) \leq 3\sqrt{s}\lambda_{m+1}.$$

□

Proof of Proposition 1. Let $\tilde{S}$ be a set of size s then

$$\|M_{\tilde{S}}\| \leq \|\beta_{\tilde{S}}\| + \delta\|\beta\| + \frac{\sigma\sqrt{10s\log(ep/s)}}{\|X\|_{2,\infty}}.$$

Hence

$$\|M_{\tilde{S}}\| \leq \|\beta\|(1+\delta) + \frac{\sigma\sqrt{10s\log(ep/s)}}{\|X\|_{2,\infty}}.$$

Then

$$\sqrt{s}\hat{\lambda}_0 \leq 2(\sqrt{10}(1+\delta)+1)\left(\|\beta\| \vee \frac{\sigma\sqrt{10s\log(ep/s)}}{\|X\|_{2,\infty}}\right).$$

Hence

$$\sqrt{s}\hat{\lambda}_0 \leq 10\left(\|\beta\| \vee \frac{\sigma\sqrt{10s\log(ep/s)}}{\|X\|_{2,\infty}}\right).$$

We also have for the true support S of β that

$$\|M_S\| \geq (1-\delta)\|\beta\| - \frac{\sigma\sqrt{10s\log(ep/s)}}{\|X\|_{2,\infty}}.$$

If $\|\beta\| \leq \frac{\sigma\sqrt{40s\log(ep/s)}}{\|X\|_{2,\infty}}$ the result is trivial. Else $\|\beta\| > \frac{\sigma\sqrt{40s\log(ep/s)}}{\|X\|_{2,\infty}}$, and

$$\sqrt{s}\hat{\lambda}_0 \geq \sqrt{10}(1-\delta-1/2)\|\beta\| \geq \|\beta\|.$$

The result for $\hat{m}$ is straightforward. □

Proof of Theorem 2. Assume that event $\mathscr{O}$ holds, then using Proposition 1, we have $\|\beta\|^2 \leq s\hat{\lambda}_0^2$. We can then apply Theorem 1 and get that

$$\forall m \geq 0, \quad \|\hat{\beta}^m - \beta\|^2 \leq 9\left(s\hat{\lambda}_0^2\kappa^{m/2} \vee \frac{40\sigma^2 s\log(ep/s)}{\|X\|_{2,\infty}^2}\right).$$

With the choice of $\hat{m}$ we get further using Proposition 1 that

$$s\hat{\lambda}_0^2\kappa^{\hat{m}/2} \leq \frac{40\sigma^2 s\log(ep/s)}{\|X\|_{2,\infty}^2}.$$

Hence

$$\|\hat{\beta}^{\hat{m}} - \beta\|^2 \leq 360\frac{\sigma^2 s\log(ep/s)}{\|X\|_{2,\infty}^2}.$$

It follows that

$$\mathbf{P}\left(\|\hat{\beta}^{\hat{m}}-\beta\|^2 \geq 360\frac{\sigma^2}{\|X\|_{2,\infty}^2}s\log(ep/s)\right) \leq \mathbf{P}\left(\sum_{i=1}^{s}\Xi_{(i)}^2 \geq \frac{10\sigma^2 s\log(ep/s)}{\|X\|_{2,\infty}^2}\right).$$

We conclude using Lemma 3. We proceed similarly for the remaining statements using Proposition 1. □

Proof of Theorem 3. For this proof we consider both events $\mathscr{O}$ and $\mathscr{A}$ where

$$\mathscr{A} = \{|\|\xi\| - \sqrt{n}| \leq 1/4\sqrt{n}\}.$$

Using the Hanson-Wright inequality [19] and the condition on n, it is easy to observe that

$$\mathbf{P}(\mathscr{A}) \leq e^{-c_2 s\log(ep/s)},$$

for some absolute $c_2 > 0$. For the rest of the proof we place ourselves on the event $\mathscr{O}\cap\mathscr{A}$ that holds with probability $1-e^{-c_3 s\log(ep/s)}$ for some absolute $c_3 > 0$. From the definition of $\hat{\sigma}_m^2$, observe that, as long as $\beta-\hat{\beta}_m$ is $2s$-sparse, we have

$$|\hat{\sigma}_m - \sigma| \leq \frac{\|X\|_{2,\infty}}{\sqrt{n}}(1+\delta)\|\beta-\hat{\beta}_m\| + \frac{1}{4}\sigma.$$

Hence

$$\frac{\hat{\sigma}_m}{\|X\|_{2,\infty}}\sqrt{160\log(ep)} \leq \frac{3\sqrt{40\log(ep)}}{\sqrt{n}}\|\beta-\hat{\beta}_m\| + \frac{3\sigma}{\|X\|_{2,\infty}}\sqrt{40\log(ep)}. \tag{21}$$

Next, recall that

$$\bar{\lambda}_0 = \sqrt{20}|M|_{(1)} \vee \frac{\hat{\sigma}_0}{\|X\|_{2,\infty}}\sqrt{160\log(ep)}.$$

Since n is large enough compared to $s\log(ep)$ it is easy to see that

$$\bar{\lambda}_0 \leq \sqrt{20}|M|_{(s)} \vee \|\beta\| \vee \frac{12\sigma}{\|X\|_{2,\infty}}\sqrt{10\log(ep)}.$$

Hence using Proposition 1 it comes that

$$\sqrt{2}\|\beta\| \leq \sqrt{s}\bar{\lambda}_0 \leq 20\sqrt{s}\left(\|\beta\| \vee \frac{\sigma\sqrt{10s\log(ep)}}{\|X\|_{2,\infty}}\right). \tag{22}$$

Hence the threshold $\bar{\lambda}_0$ satisfies the condition required to apply Theorem 2. Let $\hat{m}^*$ such that

$$\hat{m}^* = \inf\left\{m/\lambda_m \leq \frac{6\sigma}{\|X\|_{2,\infty}}\sqrt{40\log(ep)}\right\} + 1. \tag{23}$$

We recall that

$$\hat{m} = \inf\left\{m/\lambda_m \leq \frac{\sigma}{\|X\|_{2,\infty}}\sqrt{40\log(ep/s)}\right\} + 1. \tag{24}$$

Observe that $\hat{m}^* \leq \hat{m}$. As long as $m \leq \hat{m}^*$ then

$$\frac{\sigma}{\|X\|_{2,\infty}}\sqrt{40\log(ep)} \leq 1/6\lambda_m,$$

so the first induction steps remain the same as before. Now using (21) we get further

$$\frac{\hat{\sigma}_m}{\|X\|_{2,\infty}}\sqrt{160\log(ep)} \leq \sqrt{\frac{3500s\log(ep)}{n}}\lambda_m + 1/2\lambda_m.$$

Hence for n larger than $14000s\log(ep)$, $\frac{\hat{\sigma}_m}{\|X\|_{2,\infty}}\sqrt{160\log(ep)}$ is strictly smaller than λ_m and $\hat{m}^* < \bar{m}$. After running $\hat{m}^*$ steps we get

$$\|\beta - \hat{\beta}_{\hat{m}^*}\| \leq \frac{18\sigma}{\|X\|_{2,\infty}}\sqrt{40s\log(ep)},$$

and for all $\hat{m}^* \leq m \leq \hat{m}$

$$|\hat{\sigma}_m - \sigma| \leq \sigma/2.$$

It comes out that for all $\hat{m}^* \leq m \leq \hat{m}$

$$\frac{\hat{\sigma}_m}{\|X\|_{2,\infty}}\sqrt{160\log(ep)} \geq \frac{\sigma}{\|X\|_{2,\infty}}\sqrt{40\log(ep)}.$$

Hence a finite number of steps after the first $\hat{m}^*$ are enough to hit the threshold $\frac{\hat{\sigma}_{\bar{m}}}{\|X\|_{2,\infty}}\sqrt{160\log(ep)}$ and stop the algorithm. Observe that $\bar{m} \leq \hat{m}$ and hence $\frac{\hat{\sigma}_{\bar{m}}}{\|X\|_{2,\infty}}\sqrt{160\log(ep)} \geq \hat{\lambda}_\infty$. We finally get, applying Theorem 1, that

$$\|\beta - \hat{\beta}_{\bar{m}}\| \leq \frac{10\sigma}{\|X\|_{2,\infty}}\sqrt{40s\log(ep)},$$

and that $|\hat{\beta}_{\bar{m}}|_0 \leq 2s$. Going back to the definition of $\bar{m}$ and using (22) we get also that

$$\bar{m} \leq 2\left(\log\left(\frac{10\|\beta\|^2\|X\|_{2,\infty}^2}{\sigma^2\log(ep)} \vee 100\right)/\log(1/\kappa)\right).$$

This concludes the proof. □

Proof of Theorem 4. Before proving the result, we recall that with probability $1 - e^{-cs\log(ep/s)}$ we have

$$|\hat{\sigma}_{\bar{m}} - \sigma| \leq \frac{\|X\|_{2,\infty}}{\sqrt{n}}(1+\delta)\|\beta - \hat{\beta}_{\bar{m}}\| + \frac{1}{20}\sigma \leq \sigma\left((1+\delta)\frac{10\sqrt{40}}{14000} + \frac{1}{20}\right) \leq \sigma/10.$$

In what follows we assume that

$$|\hat{\sigma}_{\bar{m}} - \sigma| \leq \sigma/10. \tag{25}$$

Using Theorem 2 and Lemma 5, then we assume moreover that

$$\|\hat{\beta}^{\hat{m}} - \beta\|^2 \leq \frac{360\sigma^2 s\log(ep/s)}{\|X\|_{2,\infty}^2},$$

$$|\hat{\beta}^{\hat{m}}_{S^c}|_0 \leq s,$$

and

$$\left\langle \xi, \frac{X^\top(\beta - \hat{\beta})}{\|X(\beta - \hat{\beta})\|} \right\rangle^2 \leq 7(s + |\hat{\beta}_{S^c}|_0)\log(ep/(s + |\hat{\beta}_{S^c}|_0)),$$

since all those events hold with probability $1 - e^{-cs\log(ep/s)}$. The remainder of the proof is fully deterministic. Based on (25) it is easy to observe that $\hat{m} \leq \hat{T}$. Hence we have that

$$\frac{1}{n}\|Y - X\hat{\beta}^{\tilde{m}}\|^2 + \frac{1000\hat{\sigma}^2_{\tilde{m}}|\hat{\beta}^{\tilde{m}}|_0\log(ep/|\hat{\beta}^{\tilde{m}}|_0)}{n} \leq \frac{1}{n}\|Y - X\hat{\beta}^{\hat{m}}\|^2 + \frac{1000\hat{\sigma}^2_{\hat{m}}|\hat{\beta}^{\hat{m}}|_0\log(ep/|\hat{\beta}^{\hat{m}}|_0)}{n}. \tag{26}$$

The rest of the proof is decomposed in two parts.

- *Show that* $|\hat{\beta}^{\tilde{m}}|_0 \leq 3s$*:*
  Let us assume that $|\hat{\beta}^{\tilde{m}}|_0 > 3s$. On the one hand, we have

$$\begin{aligned}\|Y - X\hat{\beta}^{\tilde{m}}\|^2 &\geq \sigma^2\|\xi\|^2 + \|X(\beta - \hat{\beta}^{\tilde{m}})\|^2 - 2\sigma\left|\left\langle \xi, X(\beta - \hat{\beta}^{\tilde{m}})\right\rangle\right| \\ &\geq \sigma^2\|\xi\|^2 + \|X(\beta - \hat{\beta}^{\tilde{m}})\|^2 - \sqrt{42\sigma^2|\hat{\beta}^{\tilde{m}}|_0\log(3ep/4|\hat{\beta}^{\tilde{m}}|_0)}\|X(\beta - \hat{\beta}^{\tilde{m}})\| \\ &\geq \sigma^2\|\xi\|^2 + \|X(\beta - \hat{\beta}^{\tilde{m}})\|^2/2 - 21\sigma^2|\hat{\beta}^{\tilde{m}}|_0\log(ep/|\hat{\beta}^{\tilde{m}}|_0).\end{aligned}$$

  It comes out that

$$\frac{1}{n}\|Y - X\hat{\beta}^{\tilde{m}}\|^2 + \frac{1000\hat{\sigma}^2_{\tilde{m}}|\hat{\beta}^{\tilde{m}}|_0\log(ep/|\hat{\beta}^{\tilde{m}}|_0)}{n} \geq \frac{\sigma^2\|\xi\|^2}{n} + \frac{900\hat{\sigma}^2_{\tilde{m}}|\hat{\beta}^{\tilde{m}}|_0\log(ep/|\hat{\beta}^{\tilde{m}}|_0)}{n}.$$

On the other hand, we have

$$\begin{aligned}\|Y - X\hat{\beta}^{\bar{m}}\|^2 &\leq \sigma^2\|\xi\|^2 + \|X(\beta - \hat{\beta}^{\bar{m}})\|^2 + 2\sigma\left|\left\langle \xi, X(\beta - \hat{\beta}^{\bar{m}})\right\rangle\right| \\ &\leq \sigma^2\|\xi\|^2 + \|X(\beta - \hat{\beta}^{\bar{m}})\|^2 + \sqrt{8\sigma s\log(ep/2s)}\|X(\beta - \hat{\beta}^{\bar{m}})\| \\ &\leq \sigma^2\|\xi\|^2 + 3\|X(\beta - \hat{\beta}^{\tilde{m}})\|^2/2 + 4\sigma^2 s\log(ep/2s).\end{aligned}$$

It comes out that

$$\frac{1}{n}\|Y - X\hat{\beta}^{\bar{m}}\|^2 + \frac{1000\hat{\sigma}^2_{\bar{m}}|\hat{\beta}^{\bar{m}}|_0\log(ep/|\hat{\beta}^{\bar{m}}|_0)}{n} \leq \frac{\sigma^2\|\xi\|^2}{n} + \frac{2500\hat{\sigma}^2_{\bar{m}}s\log(ep/2s)}{n}.$$

Going back to (26), we conclude that

$$\frac{900|\hat{\beta}^{\tilde{m}}|_0\log(ep/|\hat{\beta}^{\tilde{m}}|_0)}{n} \leq \frac{2500s\log(ep/2s)}{n}.$$

The last equation does not hold as long as s/p is small enough. As a consequence, we have $|\hat{\beta}^{\tilde{m}}|_0 \leq 3s$.

- *Show that* $\|\hat{\beta}^{\tilde{m}} - \beta\| \leq 100\sigma\sqrt{s\log(ep/s)}/\|X\|_{2,\infty}$*:* Using the above equations we get also that

$$\|X(\beta - \hat{\beta}^{\tilde{m}})\|^2/(2n) \leq \frac{2500\hat{\sigma}^2_{\bar{m}}s\log(ep/es)}{n}.$$

Since $\hat{\beta}^{\tilde{m}}$ is at most $3s$-sparse, then we conclude using RIP that

$$\|\beta - \hat{\beta}^{\tilde{m}}\|^2 \leq \frac{100^2\sigma^2 s\log(ep/s)}{\|X\|^2_{2,\infty}}.$$

The result corresponding to $\hat{T}$ is straightforward based on (25).

□

APPENDIX B: PROOFS OF ASYMPTOTIC RESULTS

Without loss of generality we will assume in the next proofs that $\|\tilde{\beta}^0 - \beta\|^2 \leq \frac{2.10^4\sigma^2 s\log(ep/s)}{\|X\|^2_{2,\infty}}$ and $|\tilde{\beta}^0|_{S^c} \leq s$. Indeed according to Theorem 4 both statements hold with high probability. In order to alleviate notations we assume that $\sigma = \|X\|_{2,\infty}$.

Proof of Theorem 5. Let $\epsilon > 0$. Observe that

$$\mathbf{P}\left(\sum_{i=1}^{p}\Xi_i^2\mathbf{1}\{|\Xi_i| \geq \lambda\} \geq \frac{s}{\log\log(ep/s)}\right)$$

$$\leq \frac{\log\log(ep/s)}{s}\sum_{i=1}^{p}\mathbf{E}\left(\Xi_i^2\mathbf{1}\{|\Xi_i|\geq\lambda\}\right).$$

Hence using Lemma 4, we get

$$\mathbf{P}\left(\sum_{i=1}^{p}\Xi_i^2\mathbf{1}\{|\Xi_i|\geq\lambda_\infty^0(1+\sqrt{\epsilon}/2)\}\geq\frac{s}{\log\log(ep/s)}\right)=o(1).$$

Using Lemma 2, we get also that

$$\mathbf{P}\left(\|\Xi_S\|^2\geq s\sqrt{\log(ep/s)}\right)=o(1).$$

As a consequence, we assume, in what follows, that $\sum_{i=1}^{p}\Xi_i^2\mathbf{1}\{|\Xi_i|\geq\lambda_\infty^0(1+\sqrt{\epsilon}/2)\}\leq\frac{s}{\log\log(ep/s)}$ and $\|\Xi_S\|^2\leq s\sqrt{\log(ep/s)}$. We show first that $\forall m\geq 0$ we have

$$|\tilde{\beta}_{S^c}^m|_0\leq s,$$

and

$$\|\tilde{\beta}^m-\beta\|\leq\sqrt{s}(\lambda+2\log(ep/s)^{1/4})(1+4\sqrt{\delta}+100\delta^{m/2}).$$

For $m=0$ the result is immediate based on the assumption on λ. Assume that result holds for m and we prove it for $m+1$. On S we have

$$|\tilde{\beta}_i^{m+1}-\beta_i|\leq\lambda+|\Xi_i|+|\langle\Phi_i^\top,\beta-\tilde{\beta}^m\rangle|.$$

Hence

$$\|\beta-\tilde{\beta}_S^{m+1}\|\leq\sqrt{s}(\lambda+\log(ep/s)^{1/4})+\delta\|\beta-\tilde{\beta}^m\|.$$

On S^c, we show first that $|\hat{\beta}_{S^c}^{m+1}|_0\leq s$. By absurd, assuming this is not the case, then

$$\begin{aligned}\sqrt{s}\lambda\leq&\sqrt{\sum_{i\in\tilde{S}}|\Xi_i|^2\mathbf{1}\{|\Xi_i|\geq\lambda_\infty^{\epsilon/4}\}}\\&+\sqrt{\sum_{i\in\tilde{S}}|\Xi_i|^2\mathbf{1}\{|\Xi_i|\leq\lambda_\infty^{\epsilon/4},|\langle\Phi_i,\beta-\tilde{\beta}^m\rangle|\leq\sqrt{\epsilon}\lambda_\infty^0/2\}}+\delta\|\beta-\tilde{\beta}^m\|\end{aligned}$$

for some $\tilde{S}$ that is s-sparse. It comes out that

$$\sqrt{s}\lambda\leq\sqrt{\frac{s}{\log\log(ep/s)}}+2\left(1+\frac{1}{\sqrt{\epsilon}}\right)\delta\|\beta-\tilde{\beta}^m\|.$$

Using the induction hypothesis

$$\sqrt{s}\lambda\leq\sqrt{\frac{s}{\log\log(ep/s)}}$$

$$+ 2\sqrt{\delta}(\sqrt{\delta}+1)\sqrt{s}(\lambda + 2\log(ep/s)^{1/4})(1+4\sqrt{\delta}+100\delta^{m/2}).$$

Since $\delta \leq 1/400^2$ and p/s large enough the above statement can not hold. Next we have on S^c

$$|\tilde{\beta}_i^{m+1}| \leq |\Xi_i|\mathbf{1}\{|H_i^{m+1}| \geq \lambda\} + |\langle \Phi_i^\top, \beta - \tilde{\beta}^m\rangle|.$$

Since S_{m+1}^c is s-sparse then arguing as above we get

$$\|\tilde{\beta}_{S^c}^{m+1}\| \leq 2\left(1+\frac{1}{\sqrt{\epsilon}}\right)\delta\|\beta - \tilde{\beta}^m\| + \sqrt{\frac{s}{\log\log(ep/s)}}.$$

Hence

$$\|\tilde{\beta}^{m+1} - \beta\| \leq \sqrt{s}(\lambda + 2\log(ep/s)^{1/4}) + \sqrt{\delta}(3+2\sqrt{\delta})\|\beta - \tilde{\beta}^m\|.$$

Using the induction hypothesis

$$\|\tilde{\beta}^{m+1} - \beta\| \leq$$
$$\sqrt{s}(\lambda + 2\log(ep/s)^{1/4})(1+\sqrt{\delta}(3+2\sqrt{\delta})(1+4\sqrt{\delta}+100\delta^{m/2})).$$

Since $\sqrt{\delta}(3+2\sqrt{\delta})(1+4\sqrt{\delta}+100\delta^{m/2}) \leq 4\sqrt{\delta}+100\delta^{(m+1)/2}$ then the result follows. As a consequence, as $s/p \to 0$,we have for $m \geq 0$

$$\|\tilde{\beta}^m - \beta\| \leq (1+4\sqrt{\delta}+100\delta^{m/2})\sqrt{s}\lambda.$$

This concludes the proof. □

Proof of Theorem 6. Let $\epsilon > 0$. Following the same steps as in the proof of Theorem 5, we assume, in what follows, that $\sum_{i=1}^p \Xi_i^2\mathbf{1}\{|\Xi_i| \geq \lambda_\infty^0(1+\sqrt{\epsilon}/2)\} \leq \frac{s}{\log\log(ep/s)}$, $\sum_{i\in S}\mathbf{1}\{|\Xi_i| \geq \lambda_\infty^0\sqrt{\epsilon}/2\} \leq \frac{s}{2\log(ep/s)\log\log(ep/s)}$ and $\|\Xi_S\|^2 \leq s + \log(s)$. Indeed, sing Lemma 4, we get

$$\mathbf{P}\left(\sum_{i=1}^p \Xi_i^2\mathbf{1}\{|\Xi_i| \geq \lambda_\infty^0(1+\sqrt{\epsilon}/2)\} \geq \frac{s}{\log\log(ep/s)}\right) = o(1),$$

and using Markov inequality

$$\mathbf{P}\left(\sum_{i\in S}\mathbf{1}\{|\Xi_i| \geq \lambda_\infty^0\sqrt{\epsilon}/2\} \geq \frac{s}{2\log(ep/s)\log\log(ep/s)}\right) = o(1).$$

Using Lemma 2, we get also that

$$\mathbf{P}\left(\|\Xi_S\|^2 \geq s + \log(s)\right) = o(1),$$

as $s \to \infty$.

We show that $\forall m \geq 0$ we have

$$|\tilde{\beta}^m_{S^c}|_0 \leq s,$$

and

$$\|\tilde{\beta}^m - \beta\| \leq 100\sqrt{s}\lambda\delta^{m/2} + \left(\sqrt{s + \log(s)} + \sqrt{\frac{4s}{\log\log(ep/s)}}\right)(1 + 4\sqrt{\delta}).$$

For $m = 0$ the result is immediate based on the assumption on λ. Assume the result is true for m and we prove it for $m + 1$. On S we have

$$\begin{aligned}|\tilde{\beta}^{m+1}_i - \beta_i| \leq \lambda \mathbf{1}\{|\beta_i + \Xi_i + \langle \Phi_i^\top, \beta - \tilde{\beta}^m\rangle| \leq \lambda\} \\ + |\Xi_i| + |\langle \Phi_i^\top, \beta - \tilde{\beta}^m\rangle|.\end{aligned}$$

Moreover

$$\begin{aligned}\mathbf{1}\{|\beta_i + \Xi_i + \langle \Phi_i^\top, \beta - \tilde{\beta}^m\rangle| \leq \lambda\} &\leq \mathbf{1}\{|\Xi_i| + |\langle \Phi_i^\top, \beta - \hat{\beta}^m| \geq (|\beta_i| - \lambda)\} \\ &\leq \mathbf{1}\{|\Xi_i| \geq \sqrt{\epsilon}\lambda/2\} + \mathbf{1}\{|\langle \Phi_i^\top, \beta - \hat{\beta}^m\rangle| \geq \sqrt{\epsilon}\lambda/2\}.\end{aligned}$$

Hence

$$\|\beta - \tilde{\beta}^{m+1}_S\| \leq \sqrt{\frac{s}{\log\log(ep/s)}} + \sqrt{s + \log(s)} + \delta\|\beta - \tilde{\beta}^m\|\left(1 + \frac{2}{\sqrt{\epsilon}}\right).$$

On S^c, we show first that $|\tilde{\beta}^{m+1}_{S^c}|_0 \leq s$. By absurd assume this is not the case then

$$\begin{aligned}\sqrt{s}\lambda \leq &\sqrt{\sum_{i\in\tilde{S}} |\Xi_i|^2 \mathbf{1}\{|\Xi_i| \geq \lambda_\infty^{\epsilon/4}\}} \\ &+ \sqrt{\sum_{i\in\tilde{S}} |\Xi_i|^2 \mathbf{1}\{|\Xi_i| \leq \lambda_\infty^{\epsilon/4}, |\langle \Phi_i, \beta - \tilde{\beta}^m\rangle| \leq \sqrt{\epsilon}\lambda^0_\infty/2\}} + \delta\|\beta - \tilde{\beta}^m\|\end{aligned}$$

for some $\tilde{S}$ that is s-sparse. It comes out that

$$\sqrt{s}\lambda \leq \sqrt{\frac{s}{\log\log(ep/s)}} + 2\left(1 + \frac{1}{\sqrt{\epsilon}}\right)\delta\|\beta - \tilde{\beta}^m\|.$$

Using the induction hypothesis

$$\sqrt{s}\lambda \leq \sqrt{\frac{s}{\log\log(ep/s)}} + 2\sqrt{\delta}(\sqrt{\delta} + 1)\|\beta - \tilde{\beta}^m\|.$$

Since $\delta \leq 1/400^2$ and both s and p/s are large enough the above statement can not hold. Next we have on S^c

$$|\tilde{\beta}_i^{m+1}| \leq |\Xi_i| \mathbf{1}\{|H_i^{m+1}| \geq \lambda\} + |\langle \Phi_i^\top, \beta - \tilde{\beta}^m \rangle|.$$

Since S_{m+1}^c is s-sparse then arguing as above we get

$$\|\tilde{\beta}_{S^c}^{m+1}\| \leq 2\left(1 + \frac{1}{\sqrt{\epsilon}}\right)\delta \|\beta - \tilde{\beta}^m\| + \sqrt{\frac{s}{\log\log(ep/s)}}.$$

Hence

$$\|\tilde{\beta}^{m+1} - \beta\| \leq \sqrt{\frac{4s}{\log\log(ep/s)}} + \sqrt{s + \log(s)} + \sqrt{\delta}(3 + 2\sqrt{\delta})\|\beta - \tilde{\beta}^m\|.$$

We conclude using the induction hypothesis since $\sqrt{\delta}(3 + 2\sqrt{\delta})(1 + 4\sqrt{\delta} + 100\delta^{m/2}) \leq 4\sqrt{\delta} + 100\delta^{(m+1)/2}$. As a consequence, as $s/p \to 0$ and $s \to \infty$, we have for $m \geq 0$ that

$$\|\tilde{\beta}^m - \beta\| \leq 100\sqrt{s}\lambda\delta^{m/2} + \sqrt{s}(1 + 4\sqrt{\delta} + o(1)).$$

Observing that for $m \geq \log\log(p/s)$ we have $100\sqrt{s}\lambda\delta^{m/2} = o(\sqrt{s})$ concludes the proof. □

Proof of Theorem 7. Similarly to the proof of Theorem 6 we assume that $\sum_{i=1}^p \Xi_i^2 \mathbf{1}\{|\Xi_i| \geq \lambda_\infty^0(1+\sqrt{\epsilon}/2)\} \leq \frac{s}{\log\log(ep/s)}$, $\sum_{i\in S} \mathbf{1}\{|\Xi_i| \geq \lambda_\infty^0\sqrt{\epsilon}/2\} \leq \frac{s}{2\log(ep/s)\log\log(ep/s)}$ and $\|\Xi_S\|^2 \leq s + \log(s) + \log(p/s)$. Observe that $\|\Xi_S\|^2 \leq s + \log(s) + \log(p/s)$ holds with high probability even for fixed s. It comes out that

$$\begin{aligned}\sum_{i\in S} |\tilde{\eta}_i^m - \eta_i| &= \sum_{i\in S} \mathbf{1}\{|H_i^m| \leq \lambda\} \\ &\leq \sum_{i\in S} \mathbf{1}\{|\Xi_i| \geq \lambda\sqrt{\epsilon}/2\} + \sum_{i\in S} \mathbf{1}\{|\Phi_i^\top, \beta - \tilde{\beta}^m| \geq \lambda\sqrt{\epsilon}/2\}.\end{aligned}$$

Hence using Theorem 6 we get, for $m \geq \log\log(ep/s)$, that

$$\sum_{i\in S} |\tilde{\eta}_i^m - \eta_i| = o(s).$$

Moreover on S^c we have that $|\tilde{\eta}^m|_0 \leq s$ and

$$\sum_{i\in S^c} |\tilde{\eta}_i^m - \eta_i| = \sum_{i\in S^c} \mathbf{1}\{|H_i^m| \geq \lambda_\infty^\epsilon\}$$

$$\leq \sum_{i\in S^c} \mathbf{1}\{|\Xi_i| \geq \lambda_\infty(1+\sqrt{\epsilon}/2)\}$$
$$+ \sum_{i\in S^c} \mathbf{1}\{|\Phi_i^\top, \beta - \tilde{\beta}^m| \geq \lambda_\infty\sqrt{\epsilon}/2\}.$$

Hence using again Theorem 6 we get, for $m \geq \log\log(ep/s)$, that

$$\sum_{i\in S^c} |\tilde{\eta}_i^m - \eta_i| = o(s).$$

The result follows immediatly. □

Proof of Theorem 8. So far we have used the following decomposition

$$H^{m+1} = \beta + \Phi(\beta - \hat{\beta}^m) + \Xi.$$

Using the oracle vector $\tilde{\beta}^*$, we get another decomposition

$$H^{m+1} = \tilde{\beta}^* + \Phi(\tilde{\beta}^* - \hat{\beta}^m) + \tilde{\Xi},$$

where $\tilde{\Xi}_S = 0$ and $\tilde{\Xi}_{S^c} = \frac{1}{\|X\|_{2,\infty}} X_{S^c}^\top \left(\mathbf{I}_n - X_S((X^\top X)_{SS})^{-1} X_S^\top\right)\xi$. Our goal is to estimate $\tilde{\beta}^*$. Observe that the noise $\tilde{\Xi}$ is zero on S and each of its coordinates is 1-subGaussian on S^c. Without loss of generality we may assume that $\|\tilde{\beta}^0 - \tilde{\beta}^*\| \leq 150\sqrt{s\log(ep/s)}$ since $\|\beta - \tilde{\beta}^*\| \leq 50\sqrt{s\log(ep/s)}$ with high probability.

Let $\epsilon > 0$. We assume, in what follows, that $\forall i = 1, \ldots, p, |\tilde{\Xi}_i| \leq \mu_\infty^{\epsilon/4}$. This actually holds with high probability since each coordinate is 1-subGaussian. Similarly, we also assume that $\|\tilde{\beta}^* - \beta\|_\infty \leq \sqrt{2\log(s)} + \sqrt{2\epsilon\log(p)}$. In particular, for all $i \in S$, $|\tilde{\beta}_i^*| \geq (1+2\sqrt{\epsilon})\sqrt{2\log(p)}$.

We show that $\forall m \geq 0$ we have

$$|\tilde{\beta}_{S^c}^m|_0 \leq s,$$

and

$$\|\tilde{\beta}^m - \tilde{\beta}^*\| \leq 150\sqrt{s\log(ep/s)}(10\delta)^{m/2}.$$

For $m = 0$ the result is immediate. Assume the result is true for m and we prove it for $m+1$. On S we have

$$|\tilde{\beta}_i^{m+1} - \tilde{\beta}_i^*| \leq \mu_\infty^\epsilon \mathbf{1}\{|\tilde{\beta}_i^* + \langle\Phi_i^\top, \tilde{\beta}^* - \tilde{\beta}^m\rangle| \leq \mu_\infty^\epsilon\} + |\langle\Phi_i^\top, \tilde{\beta}^* - \tilde{\beta}^m\rangle|.$$

Moreover

$$\mathbf{1}\{|\tilde{\beta}_i^* + \langle\Phi_i^\top, \tilde{\beta}^* - \tilde{\beta}^m\rangle| \leq \mu_\infty^\epsilon\} \leq \mathbf{1}\{|\langle\Phi_i^\top, \tilde{\beta}^* - \hat{\beta}^m| \geq (|\tilde{\beta}_i^*| - \mu_\infty^\epsilon)\}$$

$$\leq \mathbf{1}\{|\langle \Phi_i^\top, \tilde{\beta}^* - \hat{\beta}^m \rangle| \geq \sqrt{\epsilon}\mu_\infty^0\}.$$

Hence

$$\|\tilde{\beta}^* - \tilde{\beta}_S^{m+1}\| \leq \delta \|\tilde{\beta}^* - \tilde{\beta}^m\| \left(2 + \frac{1}{\sqrt{\epsilon}}\right).$$

On S^c, we show first that $|\tilde{\beta}_{S^c}^{m+1}|_0 \leq s$. By absurd assume this is not the case then

$$\sqrt{s}\mu_\infty^\epsilon \leq \sqrt{\sum_{i\in\tilde{S}} |\tilde{\Xi}_i|^2 \mathbf{1}\{|\tilde{\Xi}_i| \geq \mu_\infty^{\epsilon/4}\}}$$
$$+ \sqrt{\sum_{i\in\tilde{S}} |\tilde{\Xi}_i|^2 \mathbf{1}\{|\tilde{\Xi}_i| \leq \mu_\infty^{\epsilon/4}, |\langle \Phi_i, \tilde{\beta}^* - \tilde{\beta}^m \rangle| \leq \sqrt{\epsilon}\mu_\infty^0/2\}} + \delta \|\tilde{\beta}^* - \tilde{\beta}^m\|$$

for some $\tilde{S}$ that is s-sparse. It comes out that

$$\sqrt{s}\mu_\infty^\epsilon \leq 2\left(1 + \frac{1}{\sqrt{\epsilon}}\right)\delta \|\tilde{\beta}^* - \tilde{\beta}^m\|.$$

Using the induction hypothesis

$$\sqrt{s}\mu_\infty^\epsilon \leq 2\sqrt{\delta}(\sqrt{\delta} + 1)\|\tilde{\beta}^* - \tilde{\beta}^m\|.$$

Since $\delta \leq 1/400^2$ and p/s is large enough the above statement can not hold. Next we have on S^c

$$|\tilde{\beta}_i^{m+1}| \leq |\tilde{\Xi}_i| \mathbf{1}\{|H_i^{m+1}| \geq \mu_\infty^\epsilon\} + |\langle \Phi_i^\top, \tilde{\beta}^* - \tilde{\beta}^m \rangle|.$$

Since S_{m+1}^c is s-sparse then arguing as above we get

$$\|\tilde{\beta}_{S^c}^{m+1}\| \leq 2\left(1 + \frac{1}{\sqrt{\epsilon}}\right)\delta \|\tilde{\beta}^* - \tilde{\beta}^m\|.$$

Hence

$$\|\tilde{\beta}^{m+1} - \tilde{\beta}^*\| \leq \sqrt{\delta}(3 + 4\sqrt{\delta})\|\tilde{\beta}^* - \tilde{\beta}^m\|.$$

We conclude using the induction hypothesis since $\sqrt{\delta}(3 + 4\sqrt{\delta})(10\delta)^{m/2} \leq (10\delta)^{(m+1)/2}$. As a consequence, as $s/p \to 0$, we have for $m \geq 0$ that

$$\|\tilde{\beta}^m - \tilde{\beta}^*\| \leq 150\sqrt{s\log(ep/s)}(10\delta)^{m/2}.$$

In particular, we have for $m \geq \log(s)$, that

$$\|\tilde{\beta}^m - \tilde{\beta}^*\|_\infty \leq \mu_\infty^\epsilon/2,$$

since δ is small enough. Based on the separation condition on $\tilde{\beta}^*$, it is now clear that $\tilde{\beta}^m$ and $\tilde{\beta}^*$ share the same support. Hence we conclude that

$$|\tilde{\eta}^m - \eta| = 0.$$

□

APPENDIX C: TECHNICAL LEMMAS

LEMMA 2. *Assume that ξ is a centered 1-subGaussian random vector. Then, for all $S \subset \{1, \dots, p\}$ such that $|S| \leq s$, we have*

$$\forall t \geq 0, \quad \mathbf{P}\left(\frac{1}{\|X\|_{2,\infty}^4} \sum_{i \in S} \left(X_i^\top \xi \right)^2 \geq \frac{s+t}{\|X\|_{2,\infty}^2} \right) \leq e^{-(t/8) \wedge (t^2/(64s))}.$$

PROOF. Observe that

$$\sum_{i \in S} \left(X_i^\top \xi \right)^2 = \|X_S^\top \xi\|^2.$$

Then using Theorem 2.1 in [13] we get that

$$\mathbf{P}\left(\|X_S^\top \xi\|^2 \geq Tr(X_S X_S^\top) + 2\|X_S X_S^\top\|_F \sqrt{t} + 2\lambda_{\max}(X_S X_S^\top) t \right) \leq e^{-t}.$$

We have that

$$Tr(X_S X_S^\top) \leq s\|X\|_{2,\infty}^2.$$

Moreover $\lambda_{\max}(X_S X_S^\top) \leq 2\|X\|_{2,\infty}^2$ (using RIP) and $\|X_S X_S^\top\|_F \leq 2\sqrt{s}\|X\|_{2,\infty}^2$ since the rank of $X_S X_S^\top$ is at most s. Hence we conclude using the above concentration inequality that

$$\mathbf{P}\left(\frac{1}{\|X\|_{2,\infty}^4} \sum_{i \in S} \left(X_i^\top \xi \right)^2 \geq \frac{s + 4\sqrt{st} + 4t}{\|X\|_{2,\infty}^2} \right) \leq e^{-t}.$$

The final bound is obtained by considering $u = 8\sqrt{t}(\sqrt{s} \vee \sqrt{t})$ and equivalently $t = (u/8) \wedge (u^2/(64s))$. □

LEMMA 3. *There exists $c > 0$ such that*

$$\mathbf{P}\left(\sum_{i=1}^{s} \Xi_{(i)}^2 \geq \frac{10\sigma^2 s \log(ep/s)}{\|X\|_{2,\infty}^2} \right) \leq e^{-cs\log(ep/s)}.$$

PROOF. Using Lemma 2, and the union bound since

$$\binom{p}{s} \leq \left(\frac{ep}{s} \right)^s \leq e^{s\log(ep/s)}.$$

The result follows as long as $p \geq s$. □

LEMMA 4. *Let Ξ defined as in Section 2, then we have*

$$\forall t \geq 0, \quad \mathbf{E}(\Xi_i^2 \mathbf{1}\{|\Xi_i| \geq t\}) \leq 2\left(t^2 + \sigma^2/\|X\|_{2,\infty}^2\right) e^{-t^2\|X\|_{2,\infty}^2/(2\sigma^2)}.$$

PROOF. Let $t \geq 0$. We have

$$\begin{aligned}
\mathbf{E}(\Xi_i^2 \mathbf{1}\{|\Xi_i| \geq t\}) &= \int_0^\infty \mathbf{P}(\Xi_i^2 \mathbf{1}\{|\Xi_i| \geq t\} \geq u) du. \\
&= \int_0^{t^2} \mathbf{P}(|\Xi_i| \geq t) du + \int_{t^2}^\infty \mathbf{P}(\Xi_i^2 \geq u) du \\
&= t^2 \mathbf{P}(|\Xi_i| \geq t) + 2 \int_t^\infty u \mathbf{P}(|\Xi_i| \geq u) du \\
&\leq 2 \left(t^2 + \sigma^2 / \|X\|_{2,\infty}^2\right) e^{-t^2 \|X\|_{2,\infty}^2 / (2\sigma^2)}.
\end{aligned}$$

□

LEMMA 5. *Let β be a s-sparse vector and S its support. Then*

$$\mathbf{P}_\beta \left(\sup_{\hat{\beta}} \left\langle \xi, \frac{X^\top(\beta - \hat{\beta})}{\|X(\beta - \hat{\beta})\|} \right\rangle^2 \geq 7(s + |\hat{\beta}_{S^c}|_0) \log(ep/(s + |\hat{\beta}_{S^c}|_0)) \right) \leq e^{-cs \log(ep/s)}.$$

PROOF. Observe that $\beta - \hat{\beta}$ is has sparsity at most $s + |\hat{\beta}_{S^c}|_0$. Let π be the orthogonal projector onto the span of columns of X indexed by the support of $\beta - \hat{\beta}$. It comes out that

$$\left| \left\langle \xi, \frac{X^\top(\beta - \hat{\beta})}{\|X(\beta - \hat{\beta})\|} \right\rangle \right| \leq \|\pi \xi\|.$$

Since π has rank at most $s + |\hat{\beta}_{S^c}|_0$, then using Theorem 2.1 in [13] we get that for all $t \geq 0$ and for fixed π

$$\mathbf{P}\left(\|\pi \xi\|^2 \geq s + |\hat{\beta}_{S^c}|_0 + 3t \right) \leq e^{-t}.$$

It comes out that

$$\mathbf{P}\left(\|\pi \xi\|^2 \geq 7(s + |\hat{\beta}_{S^c}|_0) \log(ep/(s + |\hat{\beta}_{S^c}|)) \right) \leq e^{-2s \log(ep/s)}.$$

Hence

$$\mathbf{P}_\beta \left(\sup_{|\hat{\beta}_{S^c}|_0 = v} \|\pi \xi\|^2 \geq 7(s + v) \log(ep/(s + v)) \right) \leq e^{-s \log(ep/s)}.$$

We conclude using a union bound over all values of v. □

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CA 90089
E-MAIL: ndaoud@usc.edu