Minimax Rates for High-dimensional Double Sparse Structure over $\ell_q$-balls

Zhifan Li$^1$, Yanhang Zhang$^1$, and Jianxin Yin$^{1, 2}$

$^1$School of Statistics, Renmin University of China
$^2$Center for Applied Statistics and School of Statistics, Renmin University of China, jyin@ruc.edu.cn

Abstract

In this paper, we focus on the high-dimensional double sparse structure, where the parameter of interest simultaneously encourages group-wise sparsity and element-wise sparsity in each group. Combining Gilbert-V arshamov bound and its variants, we develop a novel lower bound technique for the metric entropy of the parameter space, which is well suited for the double sparse structure over $\ell_q$-balls for $q \in [0, 1]$. We prove the lower bounds on estimation error in an information theoretical manner, which is based on our proposed lower bound technique and Fano’s inequality. The matching upper bounds are also established, whose proof follows from a direct analysis of the constrained least-squares estimators and results on empirical processes. Moreover, we extend the results over $\ell_q$-balls into the double sparse regression model and establish its minimax rate on the estimation error. Finally, we develop the DSIHT (Double Sparse Iterative Hard Thresholding) algorithm and show its optimality in the minimax sense for solving the double sparse linear regression.

Key words : double sparsity, Gaussian location model, Gilbert-V arshamov bound, iterative hard thresholding, minimax rates.

1 Introduction

Consider a Gaussian location model (GLM)

$$Y = \theta^* + Z,$$  \hspace{1cm} (1)

where $Y \in \mathbb{R}^{d \times m}$ is the observation, $\theta^* = (\theta_1, \cdots, \theta_m) \in \mathbb{R}^{d \times m}$ is the parameter of interest, $\theta_i \in \mathbb{R}^{d}$, $i \in [m]$ and $Z \in \mathbb{R}^{d \times m}$ is the error term whose entries are independently

\footnote{This work is supported by National Key Research and Development Program of China, Grant/Award Number: 2020YFC2004900.}
drawn from $N(0, \frac{\sigma^2}{n})$. Here $n$ refers to the sample size, and $\sigma$ refers to the noise level. The goal of model (1) is to infer the parameter $\theta^*$ given the observation $Y$. GLM is fundamentally important for a variety of statistical problems, such as wavelet Gaussian regression [Donoho and Johnstone 1994, 1995, 1998], model selection [Birgé and Massart 2001], [Barron et al. 1999], [Wu and Zhou 2013], and multiple testing problem under sparsity assumption [Abramovich et al. 2006, 2007].

Nowadays the research based on imposing various types of sparse structures is an active line of the high-dimensional inference, for example, the element-wise sparsity [Tibshirani 1996], [Zhang 2010], [Raskutti et al. 2011], [Yuan et al. 2018], [Zhu et al. 2020] and group-wise sparsity [Yuan and Lin 2006], [Zhao et al. 2009], [Lounici et al. 2011]. In addition, many researchers consider a wide range of assumptions of the sparse structure, including the cases of hard sparsity and soft sparsity [Donoho and Johnstone 1994], [Raskutti et al. 2011]. Specifically, hard sparsity means the parameter of interest has an exact number of non-zero elements, while soft sparsity requires the absolute magnitude to decay at a specific rate, but most of the signals may be non-zero. A widely used approach to capture the notion of these two types of sparsity is $\ell_q$-balls for $q \in [0, 1]$. Suppose that $\eta$ is a $p$-length vector. The $\ell_0$-ball, i.e., the hard sparsity case, is defined as

$$B^p_0(s_0) = \{ \eta \in \mathbb{R}^p | \| \eta \|_0 = \sum_{i=1}^{p} I(\eta_i \neq 0) \leq s_0 \} ,$$

where $s_0$ is the exact sparsity level and $I(\cdot)$ stands for the indicator function. On the other hand, for $q \in (0, 1]$, i.e., the soft sparsity case, the $\ell_q$-ball is defined as

$$B^p_q(R_q) = \{ \eta \in \mathbb{R}^p | \| \eta \|_q = \sum_{i=1}^{p} |\eta_i|^q \leq R_q \} ,$$

where $R_{0}^{1/q}$ is the corresponding radius of the $\ell_q$-ball. For the case $q = 0$, we require the sparsity level $s_0 = R_0$ to simplify the notation. In this paper, we focus our attention on the double sparse structure, that is, simultaneous element-wise and group-wise sparsity. We define double sparsity as follows:

**Definition 1 (Double sparsity)** Given $q \in [0, 1]$, $\theta \in \mathbb{R}^{d \times m}$ is called $(s, R_q)$-double sparse when the index set of the columns of $\theta$ belongs to an $\ell_0$-ball while the $j$-th column $\theta_j$ belongs to an $\ell_q$-ball, that is,

$$\| \theta \|_{0,2} = \sum_{j=1}^{m} I(\| \theta_j \|_2 \neq 0) \leq s \, , \, \theta_j \in B^d_q(R_q) \text{ for } j \in [m].$$

Intuitively, the $m$ columns of $\theta$ naturally separate the entries of $\theta$ into $m$ non-overlapping groups, where each group contains $d$ variables. Double sparsity encourages sparsity across
and within the groups simultaneously. Here denote the corresponding parameter spaces of (2) for \( q = 0 \) and \( q \in (0, 1] \) as \( \Theta_0^{m,d}(s, s_0) \) and \( \Theta_q^{m,d}(s, R_q) \), respectively. In particular, denote the support of \( \theta \in \Theta_0^{m,d}(s, s_0) \) as \( S^{m,d}(s, s_0) \). The parameter spaces can be considered as the Khatri-Rao product of two spaces \( \Gamma \) and \( \Lambda \). Let \( \Gamma = \{ \gamma : \gamma \in \{0, 1\}^m \text{ and } \|\gamma\|_0 \leq s \} \) and \( \Lambda \subseteq B^m \) for a given finite set \( B \subset \mathbb{R}^d \setminus \{0_{1 \times d}\} \). Define the Khatri-Rao product of \( \Gamma \) and \( \Lambda \) as \( \Gamma \odot \Lambda = \{ \gamma_1 \otimes \lambda_1 | \ldots | \gamma_m \otimes \lambda_m : \gamma \in \Gamma \text{ and } \lambda \in \Lambda \} \), where \( \otimes \) denotes the Kronecker product of two vectors. The specific forms of \( \Theta_0^{m,d}(s, s_0) \) and \( \Theta_q^{m,d}(s, R_q) \) are presented as

\[
\Theta_0^{m,d}(s, s_0) = \Gamma \odot \Lambda \text{ for } B = B_0^d(s_0) \setminus \{0_{1 \times d}\}, \\
\Theta_q^{m,d}(s, R_q) = \Gamma \odot \Lambda \text{ for } B = B_q^d(R_q) \setminus \{0_{1 \times d}\}.
\]

Here the index set \( \gamma \in \Gamma \) records the locations of the non-zero columns of \( \theta \) while \( \lambda \in \Lambda \) determines the entries of \( \theta \).

1.1 Related literature

Recently, the applications of the double sparse structure have attracted enormous attention in various fields. For instance in genome-wide association studies (GWAS), the genes belonging to the same pathway can be considered as a group, and it is a widely held view that only a small part of these groups contain casual single nucleotide polymorphisms (SNPs) \cite{Silver2013}. Additionally, classification \cite{Rao2015, Huo2020}, climate data analysis \cite{Zhang2020} and multi-attribute graph estimation \cite{Tugnait2021} can be viewed as the prototypes of applications among many others.

In light of the wide applications of the double sparse structure, various computationally-feasible methods have been studied. For example, \cite{Huang2009, Breheny2009, Huang2012} proposed a framework called bi-level selection to develop efficient algorithms for solving the linear regression under double sparsity. \cite{Simon2013} extended the ordinary Lasso \cite{Tibshirani1996} and group lasso \cite{Yuan2006} to the double sparse case. They considered combining these two types of penalties simultaneously to achieve the double sparsity of the solutions and called this approach as sparse group lasso. Some remarkable approaches have been developed to accelerate the convergence of sparse group lasso \cite{Zhang2020, Ida2019}. However, although the methods mentioned above have demonstrated promising performance in practice \cite{Zhang2020, Breheny2009, Chatterjee2012}, the comprehensive theoretical analysis is still an urgent issue to be solved. Recently, \cite{Tony2022} conducted a detailed statistical analysis for sparse group lasso. They established the minimax lower bounds of the estimation error for the double sparse linear regression. Moreover, they proved that the sparse group lasso achieved an optimal minimax rate of convergence in a range of moderate settings. However, they focused their attention on the double sparse structure over \( \ell_0 \)-ball
rather than under the general sparsity such as the soft sparsity. For this theoretical study, a notable and comprehensive work is Raskutti et al. [2011], in which they investigated the minimax rates of estimation error for high-dimensional linear regression model over $\ell_q$-balls.

Developing efficient methods for estimating the high-dimensional sparse structure still remains as a hot topic in the fields of statistics and machine learning, the most widely studied of which are lasso Tibshirani [1996] and its variants Yuan and Lin [2006], Simon et al. [2013], Jacob et al. [2009], Friedman et al. [2010]. However, these convex estimators suffer from the unavoidable bias of estimation. To remedy this drawback, a line of research concerns the non-convex methods such as Iterative Hard Thresholding (IHT). IHT and its variants have drawn increasing attention for solving high-dimensional linear regression (the dimensionality $p \gg$ the sample size $n$) Blumensath and Davies [2009], Jain et al. [2014], Ndaoud [2020]. Given the sparsity level $s$, IHT first applies a gradient descent step to the parameter and then keeps the $s$ largest absolute values at each step. Recently, Liu and Foygel Barber [2020] showed that IHT achieves the minimax rate $\sigma^2 s \log(ep/s)/n$ after a number of iterations of order $\log(\frac{n\|\beta^*\|^2}{s \log(ep/s)})$, which demonstrates the fast convergence of IHT. Ndaoud [2020] proposed a fully adaptive version of the ordinary IHT which is scaled minimax optimal, that is, achieving the oracle risk that the prior knowledge of sparsity pattern is known Ndaoud [2019], Ndaoud and Tsybakov [2020]. The proposed procedure updates the threshold geometrically at each step until hitting the universal statistical threshold, rather than keeping $s$ components non-zero as the ordinary IHT does. Following this thought, they proposed an iteration selection procedure that determines the appropriate iteration, which is minimax optimal. Motivated by this idea, we extend this IHT-style method to apply to the high-dimensional double sparse linear regression and investigate the corresponding statistical guarantees.

1.2 Our contributions

The main contributions of our work can be summarized as follows:

- Utilizing a combination of $q$-ary and binary Gilbert-Varshamov bounds, we develop a novel approach to construct the packing set of the parameter space (Lemma 3). It can be applied to various estimation problems for the simultaneously structured models, i.e., the parameter of interest possesses multiple sparse structures simultaneously.

- By virtue of this novel technique, we establish the estimation minimax lower bounds, which involves a sum of two quantities: a term involving the complexity of identifying the non-zero columns (groups) and a term that corresponds to the complexity of estimating parameters over each group.

- We extend the framework of the double sparse structure from $\ell_0$-ball to $\ell_q$-balls for $q \in [0, 1]$, which fills the void of statistical limits of the double sparse structure over $\ell_q$-
balls. Notably, the analysis of $\ell_0$-ball and $\ell_q$-ball can be unified under the construction of the entropy number of $\ell_q$-ball. Moreover, we establish the corresponding minimax optimal rates of the estimation error by direct analyzing the constrained estimators.

- We propose a novel double sparse iterative hard thresholding (DSIHT) algorithm to solve the high-dimensional double sparse linear regression. Moreover, we prove that our proposed method achieves optimality in the minimax sense.

1.3 Organization of the paper

The remainder of this paper is organized as follows. After a brief discussion on notations and preliminaries in Section 2.1, the lower bounds on $\ell_2$-loss over $\ell_q$-balls are presented in Section 2.2 and the matching upper bounds are shown in Section 2.3. The double sparse linear regression, an application of the double sparse structure, is studied in Section 3. We first prove the optimal minimax rate of estimation error in Section 3.1. Then, we develop the DSIHT algorithm and prove its optimality in the minimax sense in Section 3.2. In Section 4, we provide the proof sketches of our main theorems. All technical lemmas and the detailed proofs can be found in Appendix.

2 Main Results

2.1 Notations and preliminaries

In the rest of this paper, we use the following notations. For the 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 positive constant $c$. We write $a_n \asymp b_n$ if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$. Denote $[m]$ as the set $\{1, 2, \ldots, m\}$. Denote $I_p$ as the $p \times p$ identity matrix. Denote $x \vee y$ as the minimum of $x$ and $y$. For a set $A$, denote $|A|$ as the cardinality of $A$. For a vector $\eta$, denote $\|\eta\|$ as its Euclidean norm. For a matrix $\theta$, denote $\|\theta\|_F$ as its Frobenius norm.

To evaluate the performance of an estimator $\hat{\theta}$, it is common to define a loss function $L(\cdot, \cdot) : \mathbb{R}^{d \times m} \times \mathbb{R}^{d \times m} \rightarrow \mathbb{R}$ and analyze the loss $L(\hat{\theta}, \theta^*)$. We define the loss function $L(\cdot, \cdot)$ as

\[
L(\hat{\theta}, \theta^*) = \|\hat{\theta} - \theta^*\|_F^2 \quad \text{for all } \hat{\theta}, \theta^* \in \mathbb{R}^{d \times m}.
\]

In this paper, we call the estimation error measured by the $F$-norm as the $\ell_2$-loss for the matrix. This section is devoted to establishing the lower and upper bounds on the following minimax risk:

\[
\mathcal{M}(\Theta_{0}^{m,d}(s, s_0)) = \min_{\hat{\theta}} \max_{\theta^* \in \Theta_{0}^{m,d}(s, s_0)} \|\hat{\theta} - \theta^*\|_F^2.
\]

\[
\mathcal{M}(\Theta_{q}^{m,d}(s, R_q)) = \min_{\hat{\theta}} \max_{\theta^* \in \Theta_{q}^{m,d}(s, R_q)} \|\hat{\theta} - \theta^*\|_F^2.
\]
The concept of covering and packing numbers play an important role in our remaining analysis.

**Definition 2 (Covering and Packing Numbers, Raskutti et al. [2011])** Consider a compact metric space consisting of a set $S$ and a metric $\rho : S \times S \to \mathbb{R}_+$

- An $\epsilon$-covering of $S$ with respect to the metric $\rho$ is a collection $\{\theta^1, \ldots, \theta^N\} \subset S$ such that for all $\theta \in S$, there exists some $i \in \{1, \ldots, N\}$ with $\rho(\theta, \theta^i) \leq \epsilon$. The $\epsilon$-covering number $N(\epsilon; S, \rho)$ is the cardinality of the smallest $\epsilon$-covering.

- A $\delta$-packing of $S$ with respect to the metric $\rho$ is a collection $\{\theta^1, \ldots, \theta^M\} \subset S$ such that $\rho(\theta^i, \theta^j) > \delta$ for all distinct $i, j$. The $\delta$-packing number $M(\delta; S, \rho)$ is the cardinality of the largest $\delta$-packing.

Covering and packing numbers provide essentially the same measure of the massiveness of a set. In particular, the relation between covering number and packing number is described as $M(\epsilon; S, \rho) \leq N(\epsilon; S, \rho) \leq M(\epsilon/2; S, \rho)$. These two quantities exhibit the same scaling behavior as $\epsilon \to 0$. Additionally, the logarithm of the covering number $\log N(\epsilon; S, \rho)$ is known as the metric entropy of $S$ with respect to $\rho$.

**Definition 3 (entropy number of $\ell_q$ ball Triebel [2010], Kühn [2001])** Consider a quasi-Banach space consisting of a compact set $S$ and a quasi-metric $\rho$. $N(\epsilon; S, \rho)$ denotes the covering number with radius $\epsilon$. For $k = 1, 2, \ldots$ the dyadic entropy number is defined as

$$\epsilon_k(S, \rho) := \inf\{\epsilon > 0 | N(\epsilon; S, \rho) \leq 2^{k-1}\}.$$ 

Consider a $p$-dimensional vector space. Suppose $S$ is a $\ell_q$ unit-ball and $\rho$ is the metric induced by $\ell_2$-norm. Then, we have the following results:

$$\epsilon_k(B_q^p(1, \| \cdot \|_2) \simeq \begin{cases} 1 & \text{for } 1 \leq k \leq \log p \\ \left(\frac{\log(\frac{p}{k} + 1)}{k}\right)^{\frac{1-\frac{1}{q}}{2}} & \text{for } \log p \leq k \leq p \\ 2^{-\frac{1}{q} - k - \frac{1}{2}} & \text{for } k \geq p \end{cases} \quad (3a)$$

$$\epsilon_k(B_q^p(1, \| \cdot \|_2) \simeq \begin{cases} 1 & \text{for } 1 \leq k \leq \log p \\ \left(\frac{\log(\frac{p}{k} + 1)}{k}\right)^{\frac{1-\frac{1}{q}}{2}} & \text{for } \log p \leq k \leq p \\ 2^{-\frac{1}{q} - k - \frac{1}{2}} & \text{for } k \geq p \end{cases} \quad (3b)$$

$$\epsilon_k(B_q^p(1, \| \cdot \|_2) \simeq \begin{cases} 1 & \text{for } 1 \leq k \leq \log p \\ \left(\frac{\log(\frac{p}{k} + 1)}{k}\right)^{\frac{1-\frac{1}{q}}{2}} & \text{for } \log p \leq k \leq p \\ 2^{-\frac{1}{q} - k - \frac{1}{2}} & \text{for } k \geq p \end{cases} \quad (3c)$$

By analogy with Raskutti et al. [2011], for the case $q = 0$, we require that $m \geq 4s \geq c_1$ and $d \geq 4s_0 \geq c_2$ as well. For $q \in (0, 1]$, we add some reasonable assumptions that for some constants $C_1, C_2 > 0$ and $\delta \in (0, 1)$, which requires that the triple $(n, d, R_q)$ satisfies

$$\frac{d}{R_q n^{3/2}} \geq C_1 d^\delta \geq C_2. \quad (4)$$

**Remark 1** We clarify the assumption (4) in two aspects:
(1) Our interest is high-dimensional regime where both \( m \) and \( d \) are much larger than \( n \).

(2) Assumption (4) matches the rate of the entropy number in (3b), which makes sense in high-dimensional sparsity. Therefore, we can avoid the trivial situations i.e., (3a) and (3c).

2.2 Lower Bounds on \( \ell_2 \)-loss

Our first main result establishes the minimax lower bounds of estimation error over \( \ell_q \)-balls.

**Theorem 1** Consider model (1) under the double sparse structure.

(a) Case \( q = 0 \): The lower bound of estimation error holds that

\[
\mathcal{M}(\Theta_{0}^{m,d}(s, s_0)) \geq C_\ell \frac{\sigma^2}{n} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})
\]

with probability greater than \( \frac{1}{2} \) for some positive constant \( C_\ell \).

(b) Case \( q \in (0, 1] \): The lower bound of estimation error holds that

\[
\mathcal{M}(\Theta_{q}^{m,d}(s, R_q)) \geq C_\ell \left\{ \frac{\sigma^2}{n} s \log \frac{em}{s} + sR_q \left( \frac{\sigma^2}{n} \log d \right)^{1-\frac{2}{q}} \right\}
\]

with probability greater than \( \frac{3}{8} \) for some positive constant \( C_\ell \).

Theorem 1 shows that the estimation error of \( \theta^* \) consists of two parts under the double sparse structure. Intuitively speaking, the first term, \( s \log \frac{em}{s} \), in the lower bounds corresponds to the complexity of capturing \( s \) non-zero columns while the second term in (5) or (6) is related to the complexity of estimating parameter over \( \ell_q \)-balls.

Moreover, when \( m = s = 1 \), the lower bounds in Theorem 1 reduce to \( \Omega(s_0 \log \frac{ed}{s_0}) \) and \( \Omega(R_q(\frac{\sigma^2}{n} \log d)^{1-\frac{2}{q}}) \), which are consistent with the lower bounds for recovery of sparse vectors over \( \ell_q \)-balls [Donoho and Johnstone 1994, Raskutti et al. 2011]. By setting \( d = s_0 \), Theorem 1(a) recovers the lower bound for estimating the non-overlapping group structure [Lounici et al. 2011].

**Remark 2** Part (a) of Theorem 1 derives the same minimax lower bounds of estimation error as [Tony Cai et al. 2022]. Notably, the parameter space \( \Theta_{0}^{m,d}(s, s_0) \) is slightly different from [Tony Cai et al. 2022], where [Tony Cai et al. 2022] constrains the number of the overall nonzero elements rather than the nonzero elements in each group. In this work, we show that it is convenient to extend the lower bound technique from \( \ell_0 \)-ball to \( \ell_q \)-ball over the parameter space \( \Theta_{q}^{m,d}(s, R_q) \).
2.3 Upper Bounds on $\ell_2$-loss

In this section, we turn to the analysis of the corresponding upper bounds on the $\ell_2$-norm minimax rate over $\ell_q$-balls. Here we consider the constrained least-squares estimators over the parameter spaces $\Theta_{q}^{m,d}(s, R_q)$:

$$\hat{\theta}_q \in \arg \min_{\theta \in \Theta_{q}^{m,d}(s, R_q)} \|y - \theta\|^2_F. \tag{7}$$

**Theorem 2** Consider model (1) under the double sparse structure.

(a) Case $q = 0$: The upper bound of the estimation error holds that

$$\mathcal{M}(\Theta_0^{m,d}(s, s_0)) \leq C_u \frac{\sigma^2 n}{s} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0}) \tag{8}$$

with probability greater than $1 - C_1 \exp\{-C_2 (s \log { \frac{em}{s}} + ss_0 \log \frac{ed}{s_0})\}$ for some positive constants $C_u, C_1$ and $C_2$.

(b) Case $q \in (0, 1]$: The upper bound of the estimation error holds that

$$\mathcal{M}(\Theta_q^{m,d}(s, R_q)) \leq C_u \left\{ \frac{\sigma^2 n}{s} \right\} (s \log \frac{em}{s} + sR_q (\frac{\sigma^2 n}{s} \log d)^{1 - \frac{2}{q}}) \tag{9}$$

with probability greater than $1 - C_1 \exp\{-C_2 (s \log \frac{em}{s} + sR_q (\frac{\sigma^2 n}{s} \log d)^{1 - \frac{2}{q}})\}$ for some positive constants $C_u, C_1$ and $C_2$.

Theorem 2 establishes the matching upper bounds of the estimation error, which implies that the lower bounds in Theorem 1 are rate-sharp. Additionally, Theorem 2 yields that the constrained estimators defined in (7) are rate-optimal. The results of Theorem 1 and 2 together show the minimax optimal rate up to constant factors.

**Remark 3** The results of Theorem 1 and 2 together show the minimax optimal rate up to constant factors. In specific, the minimax rate for $q = 0$ scales as

$$\mathcal{M}(\Theta_0^{m,d}(s, s_0)) \asymp \left\{ \frac{\sigma^2 n}{s} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0}) \right\},$$

and for $q \in (0, 1]$, the minimax rate scales as

$$\mathcal{M}(\Theta_q^{m,d}(s, R_q)) \asymp \left\{ \frac{\sigma^2 n}{s} \right\} (s \log \frac{em}{s} + sR_q (\frac{\sigma^2 n}{s} \log d)^{1 - \frac{2}{q}}).$$
3 Application to Linear Regression Model

In this section, we consider the double sparse linear regression model by Tony Cai et al. [2022], Zhou [2021], which serves as a direct application of our proposed sparse structure. Consider the linear regression model with \( n \) independent observations

\[ Y = X\beta^* + \xi, \quad (10) \]

where \( Y \in \mathbb{R}^n \) is the response variable, \( X \in \mathbb{R}^{n \times p} \) is the design matrix and the random error term \( \xi \in \mathbb{R}^n \) is a sub-Gaussian vector with parameter \( \sigma^2 \). The parameter of interest \( \beta^* \in \mathbb{R}^p \) can be divided into \( m \) predefined non-overlapping groups, each consisting of \( d \) variables. This implies that \( p = m \times d \). The coefficient \( \beta^* \) can be reshaped into a matrix form \( \theta^* \) as in (1), where each group in \( \beta^* \) corresponds to one column of \( \theta^* \), i.e., \( \beta^* = \text{Vec}(\theta^*) \). Without loss of generality, we assume the columns of \( X \) are normalized, i.e., \( \|X_j\|_2 = \sqrt{n} \) for all \( j \in [p] \). On the other hand, we denote \( \Pi \) as the inverse operator of \( \text{Vec}(\cdot) \), i.e., \( \theta^* = \Pi(\beta^*) \). We call \( \beta \) is \( (s, s_0) \)-sparse if \( \Pi(\beta) \in \Theta_0^{m, d}(s, s_0) \). To avoid confusion, denote \( X_{.,kg} \in \mathbb{R}^n \) as the \( (d(g-1) + k) \)-th column of \( X \), reshaped as the \( k \)-th variable in the \( g \)-th group and \( X_{l,kg} \) as the \( l \)-th observation of \( X_{.,kg} \).

Obviously, model (1) is a special case of model (10) with the choice \( X = \mathbf{I}_n \). As a consequence, the results of section 2 can be extended with some additional mild conditions on the design matrix \( X \). In the rest of this section, we first extend the minimax rates for the estimation error (11) and (12) to the double sparse linear regression model. Then, we develop a DSIHT algorithm and prove its optimality in the minimax sense.

3.1 Estimation Error on \( \ell_2 \)-loss

We extend the sparse eigenvalue condition and restricted eigenvalue condition in Raskutti et al. [2011] to the double sparse linear regression, which is stated as the following two assumptions.

**Assumption 1 (Double Sparse Eigenvalue Condition)** For \( \beta \in \mathbb{R}^p \), we assume

(a) Assume that there exists a positive constant \( \tau_\ell < \infty \) such that

\[
\frac{1}{\sqrt{n}} \|X\beta\|_2 \leq \tau_\ell \|\beta\|_2 \quad \text{for all} \quad \Pi(\beta) \in \Theta_0^{m, d}(2s, 2s_0).
\]

(b) Assume that there exists a positive constant \( \tau_u > 0 \) such that

\[
\frac{1}{\sqrt{n}} \|X\beta\|_2 \geq \tau_u \|\beta\|_2 \quad \text{for all} \quad \Pi(\beta) \in \Theta_0^{m, d}(2s, 2s_0).
\]
Assumption 1 is a popular tool for analyzing high-dimensional linear regression. It gives the range of the spectrum of the sub-matrices of $X$. Another widely used assumption is the restricted isometry property (RIP, Candes and Tao [2005]). RIP requires that the constants $\tau_\alpha$ and $\tau_\ell$ are close to one. In contrast, the constants in Assumption 1 can be arbitrarily small and large, respectively. In this regard, Assumption 1 is a less stringent assumption compared to the RIP.

**Assumption 2 (Restricted Eigenvalue Condition)** For $\beta \in \mathbb{R}^p$ and $q \in (0, 1]$, there exists a constant $\tau_{q,\ell} > 0$ and a function $h_{\ell}(R_q, n, s, d)$ such that

$$\frac{\|X\beta\|_2^2}{\sqrt{n}} \geq \tau_{q,\ell} (\|\beta\|_2 - h_{\ell}(R_q, n, s, d)),$$

for all $\Pi(\beta) \in \Theta_{q,d}^m(2s, 2R_q)$ where $h_{\ell}(R_q, n, s, d) := sR_q \left(\frac{\log d}{n}\right)^{1-q/2}$.

Assumption 2 is essential in establishing upper bounds on the estimation error. It is similar to Assumption 1(b), where the only difference is the slack term $h_{\ell}(R_q, n, s, d)$. In the high-dimensional settings, i.e., $p > n$, $X$ must have a nontrivial null space, which leads to unsatisfaction of Assumption 2 with $h_{\ell}(R_q, n, s, d) = 0$. Raskutti et al. [2010] showed that an appropriate choice of slack term guarantees that the restricted eigenvalue condition is satisfied with high probability in high-dimensional settings.

Combined with the technique used in Section 2, we directly obtain the results of the estimation error bounds for the high-dimensional double sparse linear regression.

**Corollary 1** Consider model (10) for a given design matrix $X \in \mathbb{R}^{n \times p}$.

(a) Case $q = 0$: Suppose Assumption 1 holds. Then, the minimax rate scales as

$$\min_{\hat{\beta}} \max_{\Pi(\beta^*) \in \Theta_0^{m,d}(s,s_0)} \|\hat{\beta} - \beta^*\|_2^2 \asymp \left\{ \frac{\sigma^2}{n} \left( s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0} \right) \right\}, \quad (11)$$

(b) Case $q \in (0, 1]$: Suppose Assumption 2 holds and $\sigma R_q \left(\frac{\log d}{n}\right)^{1-\frac{q}{2}} < C_1$. Then, the minimax rate scales as

$$\min_{\hat{\beta}} \max_{\Pi(\beta^*) \in \Theta_q^{m,d}(s,R_q)} \|\hat{\beta} - \beta^*\|_2^2 \asymp \left\{ \frac{\sigma^2}{n} \left( s \log \frac{em}{s} + sR_q \left(\frac{\sigma^2}{n} \log d\right)^{1-\frac{q}{2}} \right) \right\}. \quad (12)$$

### 3.2 Double Sparse Iterative Hard Thresholding Algorithm

Extending the classic IHT-style procedure is a constructive approach to solving the double sparse linear regression. In the beginning of this section, we define a novel thresholding operator $T_\lambda : \mathbb{R}^{d \times m} \to \mathbb{R}^{d \times m}$. Then, we develop the corresponding statistical guarantees, which demonstrate that our proposed method is rate optimal. We specify it in two steps:
Step 1 (Entrywise Condition Checking): define an element-wise hard thresholding operator $T^{(1)}_\lambda(\cdot)$ as

$$[T^{(1)}_\lambda(U)]_{ij} = U_{ij} I\{|U_{ij}| \geq \lambda\}, \quad \forall U \in \mathbb{R}^{d \times m}.$$  

The operator $T^{(1)}_\lambda(\cdot)$ preserves the signal whose absolute magnitude is not less than $\lambda$.

Step 2 (Matrix Condition Checking): We define the operator $T^{(2)}_\lambda(\cdot)$ as the following two operations.

**Column condition:** select the columns of $U$ which satisfy the following condition and denote the index set of the selected columns as

$$\mathcal{J} = \{j \in [m] | \sum_{i=1}^{d} U_{ij}^2 \geq s_0 \cdot \lambda^2\}.$$  

(13)

**Row condition:** identify the maximum $i$ that satisfies the following condition and denote it as $i_{\text{max}}$:

$$i_{\text{max}} = \max \{i \in [d] | \sum_{j=1}^{m} U_{ij}^2 \geq s \cdot \lambda^2\},$$  

(14)

where $U_{(ij)}$ is the $i$-th non-increasing order statistic of the $j$-th column of $U$. Combining the column condition and row condition, we define the active set $\mathcal{A} = \{(i, j) | \sum_{k=1}^{d} I(|U_{kj}| \geq |U_{ij}|) \leq i_{\text{max}}, j \in \mathcal{J}\}$. Overall, the definition of operator $T^{(2)}_\lambda(\cdot)$ is

$$[T^{(2)}_\lambda(U)]_{ij} = \begin{cases} U_{ij}, & \text{if } (i, j) \in \mathcal{A} \\ 0, & \text{if } (i, j) \in \mathcal{A}^c \end{cases}.$$  

Combining these two steps, we obtain the operator $T_\lambda(\cdot) = T^{(2)}_\lambda \circ T^{(1)}_\lambda(\cdot)$. Intuitively speaking, step 1 applies an element-wise hard thresholding operator, which preserves the significant signal and erases the weak signal of the matrix $U$. Step 2 plays a significant role in $T_\lambda(\cdot)$. It uses a small rectangle to further locate the area of the important signal of $U$. The motivation of $T_\lambda(\cdot)$ can refer to Section 4.3 for more details.

With the definition of $T_\lambda(\cdot)$, we present the iterative hard thresholding algorithm for solving the double sparse linear regression as Algorithm 1.

In Algorithm 1, the contraction factor $0 < \kappa < 1$ is a step size that makes the tuning sequence $\{\lambda_t\}$ decrease exponentially. $\lambda_0$ is a sufficiently large value that guarantees the sparsity of the estimator. $\lambda_\infty$ is a threshold used to terminate the iterations of Algorithm 1.

An appropriate choice of $\lambda_\infty$ is roughly of the same order as $\sqrt{\frac{\sigma^2 (\frac{1}{n} \log(\frac{s\log(n)}{s}) + \log(\frac{sd}{s})})}{n}$, which is given by the following theorem.

To conduct the theoretical analysis, we decompose the estimation error at each step.
Algorithm 1 Double Sparse IHT (DSIHT) algorithm

Require: $X, Y, \beta_0, \kappa, \lambda_0, \lambda_\infty$.

1: Initialize $t = 0$.
2: while $\lambda_t \geq \lambda_\infty$, do
3: $\beta_{t+1} = T_{\lambda_t}(\beta_t + \frac{1}{n}X^T(Y - X\beta_t))$.
4: $\lambda_{t+1} = \sqrt{\kappa \lambda_t}$.
5: $t = t + 1$.
6: end while

Ensure: $\hat{\beta} = \beta_{t-1}$.

into two parts:

$$H^{t+1} = \beta_t + \frac{1}{n}X^T(Y - X\beta_t)$$

$$= \beta^* + \left(\frac{1}{n}X^TX - I_p\right)(\beta^* - \beta_t) + \frac{1}{n}X^T\xi. \quad (15)$$

To simplify the notations, we denote $\Xi = \Pi(\frac{1}{n}X^T\xi)$ and $\Phi = \frac{1}{n}X^TX - I_p$. We can reshape $p$-dimensional vector into $d \times m$ matrix by defining $\Xi_{ij} = \frac{1}{n}X_{i,j}\xi$. In what follows, we provide an essential condition, which we call double sparse RIP (DSRIP) condition.

Definition 4 (double sparse RIP condition) Given integers $s \in [m]$ and $s_0 \in [d]$, define $L_S$ and $U_S$ such that

$$U_S = \max_{S \in \mathcal{S}_{m,d}(s,s_0)} \lambda_{\max}(X_S^TX_S),$$

$$L_S = \min_{S \in \mathcal{S}_{m,d}(s,s_0)} \lambda_{\min}(X_S^TX_S).$$

Set $\delta_S = 1 - \frac{L_S}{U_S}$. We say that the design matrix $X$ satisfies $\text{DSRIP}(s, s_0, c), 0 < c < 1$ if $\delta_S \leq c$.

To capture the complexity of model (11), we need the following lemma, which makes the whole IHT algorithm a deterministic procedure.

Lemma 1 For a constant $C > 0$, the event

$$\mathcal{E} := \left\{ \forall S \in \mathcal{S}_{m,d}(s, s_0), \sum_{(i,j) \in S} \Xi_{ij}^2 \leq \frac{10\sigma^2 s \left(\log(\frac{s}{\sigma}) + s_0 \log(\frac{s_0}{s})\right)}{n} \right\}$$

holds with probability greater than $1 - \exp\left\{-C \left(s \log(\frac{s}{\sigma}) + ss_0 \log(\frac{s_0}{s})\right)\right\}$. 

12
Based on Lemma 1, the following theorem shows that the corresponding estimator is $(2s, 2s_0)$-sparse. In addition, the upper bound of estimation error decreases exponentially at each step.

**Theorem 3** Assume that $\beta^*$ is $(s, s_0)$-sparse and $X$ satisfies DSRIP$(3s, 3s_0, \delta/2)$. Let $\lambda_0, \lambda_\infty > 0$ and $0 < \kappa < 1$. Assume that $\delta < \frac{7-4\sqrt{3}}{4} \land \kappa, \|\beta^*\|_2 \leq \sqrt{ss_0}\lambda_0$ and

$$\lambda_\infty \geq \sqrt{\frac{40\sigma^2 \left( \frac{1}{s_0} \log \left( \frac{ss_0}{\sigma^2} \right) + \log \left( \frac{2d}{s_0} \right) \right)}{n}}.$$  

We apply Algorithm 1 and obtain a solution sequence $\{\beta_t\}, t = 1, 2, \ldots$. Denote $S^*$ as the support of $\beta^*$ and $S_t$ as the support of $\beta_t$. Then with probability at least $1 - \exp \left\{ -C \left( s \log \left( \frac{ss_0}{\sigma^2} \right) + ss_0 \log \left( \frac{2d}{s_0} \right) \right) \right\}$, for all $t$, we have

$$S_t \cap (S^*)^c \in \mathcal{S}^{m,d}(s, s_0) \quad (16)$$

and

$$\|\beta_t - \beta^*\|_2 \leq (2 + \sqrt{3})\sqrt{ss_0}\lambda_t. \quad (17)$$

Result (16) shows that the solution sequence $\{\beta_t\}$ can be controlled in a $(2s, 2s_0)$-sparse set, which guarantees the sparsity of the final estimator $\hat{\beta}$. In addition, since the nonconvexity of IHT-style method, the $\ell_2$ error $\|\beta_t - \beta^*\|_2$ can not be guaranteed to decrease at each step. Indeed, a common approach to get around this issue is constructing an upper bound that decreases at each step[2018, Zhu et al. 2020]. In Algorithm 1, sequence $\{\lambda_t\}$ decreases exponentially until it reaches the threshold $\lambda_\infty$. Given an appropriate choice of $\lambda_\infty$, the upper bound (17) achieves optimality in the minimax sense.

**Remark 4** The constant factor $2 + \sqrt{3}$ in (17) is from some technical details, and we give more interpretations in the appendix. Additionally, we ignore the optimality of $\delta$ in DSRIP as Ndaoud [2020] does. As Theorem 3 shows, if $\lambda \approx \lambda_\infty$ holds after finite steps, Algorithm 1 will output a minimax optimal estimator up to some constant factor.

**Theorem 3** is an variant of Ndaoud [2020] under double sparsity. Besides the minimax optimality and IHT computational efficiency, the work of Ndaoud [2020] benefits the fully adaptive procedure of the classical IHT, which preserves minimax optimality without the knowledge of sparsity $s$ and variance $\sigma^2$. We leave it to the future that our proposed procedure should be adaptive to the triple $(\sigma, s, s_0)$.

**Remark 5** The double sparse structure in Definition 1 is slightly different from Tony Cai et al. [2022], Zhou [2021]. Given $\beta \in \mathbb{R}^p$ with $m$ non-overlapping groups, i.e., $\beta = ((\beta_1)^T, (\beta_2)^T, \ldots, (\beta_m)^T)$. For some positive integers $s$ and $s'$, Tony Cai et al. [2022],
Zhou [2021] defined the double sparsity as 

$$\|\beta\|_{0,2} = \sum_{j=1}^{m} I(\|\beta_j\|_2 \neq 0) \leq s, \quad \|\beta\|_0 \leq s'.$$  

(18)

assumes the whole $\beta$ belongs to an $\ell_0$-ball with radius $s'$. In contrast, we assume that each group can be covered by an $\ell_0$ ball with a radius $s_0$. We refer to (18) as heterogeneous double sparsity, and the support sets is defined as $\mathcal{H}S^{m,d}(s,s')$. According to the structure (18), we modify the operator $T_\lambda(\cdot)$ by removing the row condition checking and preserving the remaining parts. Then, we obtain the modified version of Algorithm 4. Notably, if the DSRIP and Lemma 1 hold for $\mathcal{H}S^{m,d}(s,s')$, the following corollary gives the corresponding results for the modified DSIHT:

**Corollary 2** Let $\lambda_0, \lambda_\infty > 0$ and $0 < \kappa < 1$. Assume that $\delta < \frac{3-2\sqrt{2}}{2} \wedge \kappa$, $\|\beta^*\|_2 \leq \sqrt{ss_0} \lambda_0$ and 

$$\lambda_\infty \geq \sqrt{\frac{40\sigma^2 \left(\frac{1}{s_0} \log(e^2) + \log(e^d/s_0)\right)}{n}}.$$  

We apply modified DSIHT and obtain a solution sequence $\{\beta_t\}, t = 1, 2, \cdots$. Denote $S^*$ as the support of $\beta^*$ and $S_t$ as the support of $\beta_t$. Then for all $t$, we have 

$$S_t \cap (S^*)^c \in \mathcal{H}S^{m,d}(s,s')$$

and 

$$\|\beta_t - \beta^*\|_2 \leq (2 + \sqrt{2}) \sqrt{ss_0} \lambda_t.$$  

4 Key Points of Proofs of Main Theorems

In this section, we briefly provide high-level proof sketches of the main technical results. The detailed proofs are postponed to the appendix.

4.1 Proof Sketch of Theorem 1

The proofs of lower bounds generally follow an information-theoretic method based on Fano’s inequality [Thomas and Joy 2006], [Yang and Barron 1999], [Yu 1997]. The technical details are slightly different between cases $q = 0$ and $q \in (0,1]$. Concretely speaking, for $q = 0$, the proof is divided into two main steps:

(1) First, we construct the maximum packing set of the parameter space $\Theta_0^{m,d}(s,s_0)$ by coding theory [Thomas and Joy 2006]. Our novel technique uses the combination of multi-ary and 2-ary Gilbert-Vashamov bound (Lemma 2 in appendix) to establish
the bounds of the packing number $M$ of $\Theta^{m,d}_0(s, s_0)$. For two distinct elements $\theta^i, \theta^j$ in the packing set $\{\theta^1, \theta^2, \ldots, \theta^M\}$, the supports of $\theta^i$ and $\theta^j$ either share the same column index set or not. We give more details about the construction of the packing set in Lemma 3 in the appendix.

(2) The next step is to derive a lower bound on $P[B \neq \tilde{\beta}]$ by the Fano’s inequality [Thomas and Joy, 2006]. We obtain

$$P[B \neq \tilde{\beta}] \geq 1 - \frac{I(Y; B) + \log 2}{\log M},$$

(19)

where the estimator $\tilde{\beta}$ takes values in the packing set, $M$ is the packing number obtained in the first step, and $I(Y; B)$ is the mutual information between random parameter $B$ in the packing set and observation $Y$. We derive the upper bound of mutual information following the classical way in [Thomas and Joy, 2006].

On the other hand, for $q \in (0, 1]$, Raskutti et al. [2011] derives the lower bound for $\ell_q$-ball by Yang-Barron Fano [Yang and Barron, 1999]. In their framework, a sharp result of metric entropy of $\ell_q$-ball [Triebel, 2010, Kühn, 2001] plays an essential role. In our work, we do not follow this idea because our metric entropy consists of an $\ell_0$ part and an $\ell_q$ part, which are incompatible in some sense. In other words, it is difficult to choose a pair of appropriate covering and packing radii that matches the Yang-Barron Fano simultaneously. We use the following alternative techniques to overcome this difficulty:

(1) Inspired by Gao et al. [2015], when the lower bound consists of two parts, we can construct two corresponding parameter subspaces, respectively. Each of the parameter subspaces aims to specify the corresponding risk. Then, we combine these two parts by union bound.

(2) To clarify the parameter subspace corresponding to the $\ell_q$ part of the risk, direct application of the metric entropy results is not straightforward. Inspired by the proofs in Kühn [2001], we use the techniques for lower bound of (dyadic) entropy number to construct our least favor distribution here, which is a substitution to techniques in Raskutti et al. [2011].

4.2 Proof Sketch of Theorem 2

The proof of Theorem 2 mainly focuses on the direct analysis of the constrained least-squares estimators in (7). It makes use of the classical upper bound technique for high-dimensional linear regression [Raskutti et al., 2011], which can be summarized as two main steps:
(1) From the definition of \((7)\), we have 
\[ \|Y - \hat{\theta}_q\|_F^2 \leq \|Y - \theta^*\|_F^2. \]
By some simple algebras, we obtain the basic inequality
\[ \|\hat{\theta}_q - \theta^*\|_F^2 \leq 2\langle Z, \hat{\theta}_q - \theta^*\rangle_F = 2\text{tr}(Z^\top(\hat{\theta}_q - \theta^*)), \quad (20) \]
where \(\text{tr}(\cdot)\) is the trace of a matrix. Obviously, \(\hat{\theta}_q - \theta^* \in \Theta_q^{m,d}(2s, 2R_q)\).

(2) This step uses the chaining and peeling techniques from empirical process theory \cite{Geer2000}, which helps us upper bound the right-hand side of \((20)\). Moreover, some results of the covering number are required to derive the inequalities for the chaining results. More details can be found in Lemma 3 in the appendix.

### 4.3 Proof Sketch of Theorem 3

This proof of Theorem 3 is based on mathematical induction. Assume \((16)\) and \((17)\) hold for \(t\). We first prove \((16)\) holds for \(t + 1\) by contradiction. Then, based on the conclusion of \((16)\), we prove \((17)\) holds for \(t + 1\).

More specifically, if \((16)\) fails for \(t + 1\), it implies that \(S_{t+1}\) cannot contain in a \(s_0 \times s\)-shape region which belongs to \((S^*)^c\). In the following part, we provide a toy but illuminating example of operator \(T_\lambda(\cdot)\). In Fig. 1 we mark \(S^*\) as green region. Given a \(s_0 \times s\)-shape set \(\tilde{S} \in (S^*)^c\), we mark it as red region, and we mark the support set of \(T_{\lambda_{t+1}}(H^{t+1})\) as blue region. Obviously, \(S_t \cap (S^*)^c \notin S^{m,d}(s, s_0)\) for this case. We analyze the columns of the blue region in two situations:

- For the column that belongs to red region, e.g., the 5-th column in Fig. 1 by the column condition of step 2, we have \(\sum_{i=1}^{d}[T_{\lambda_{t+1}}(H^{t+1})]^2_{ij} \geq s_0 \cdot \lambda^2\).
- For the column that exceeds red region, e.g., the first column in Fig. 1, recall that step 1 guarantees that each entry of \(S_{t+1}\) is no less than \(\lambda_{t+1}\), which implies that \(\sum_{i=1}^{s_0}[T_{\lambda_{t+1}}(H^{t+1})]^2_{ij} \geq s_0 \cdot \lambda^2\).

Combining these two parts, We can show that \(ss_0\lambda_{t+1}^2 \leq \sum_{(i,j) \in \tilde{S}}[T_{\lambda_{t+1}}(H^{t+1})]_{ij}\). Combining \((15)\) and \(\tilde{S} \subseteq (S^*)^c\), we have
\[ \sqrt{\sum_{(i,j) \in \tilde{S}}[T_{\lambda_{t+1}}(H^{t+1})]_{ij}} \leq \sqrt{\sum_{(i,j) \in \tilde{S}} \xi_{ij}^2} + \sqrt{\sum_{(i,j) \in \tilde{S}} \langle \Phi_{ij}^T, \beta^* - \beta_t \rangle^2}. \]

The first term can be bounded by event \(E\) and \(\lambda_\infty\), and the second term can be controlled by the DSRIP. In particular, an appropriate range of \(\delta\) derives that \(ss_0\lambda_{t+1}^2 > \sum_{(i,j) \in \tilde{S}}[T_{\lambda_{t+1}}(H^{t+1})]_{ij}\), which leads contradiction with assumption. For \((17)\), it follows a direct analysis from \((16)\) and hypothesis induction.
Figure 1: Illustrative example of operator $T_\lambda(\cdot)$: Considering the worst case where we cannot obtain any true parameter, we still want to control the shape of the support set of $T_\lambda(H)$. The green region stands for $S^*$. In $(S^*)^c$, our aim is to make the support set of $T_\lambda(H)$ a subset of the Red region. By step 1 of the operator $T_\lambda$, we could obtain a Blue region, which can not cover the Red region by Lemma 1 but it may not be a subset of Red. Step 2 helps us "prune" the Blue a subset of $s_0 \times s$-shape.

5 Conclusion

In this paper, we establish the minimax rates for the double sparse structure over $\ell_q$-balls, where $q \in [0, 1]$. These results fill the blank of the statistical guarantees for the double sparse structure under soft sparsity. To establish the minimax lower bounds, we develop a novel technique based on the Gilbert-Varshamov bounds, which helps us give succinct proofs. The proposed technique is of independent interest and can facilitate the analysis of other problems. Then, coupled with Fano’s inequality, we derive the lower bounds of the estimation error. On the other hand, we make use of the chaining and peeling results in the empirical process to obtain the matching upper bounds, which shows that the lower bounds are rate-sharp. Finally, we extend the results to the high-dimensional double sparse linear regression. To capture the double sparse structure, we develop a novel hard thresholding operator and obtain the DSIHT algorithm. We provide a modified version of our proposed DSIHT algorithm for the heterogeneous double sparsity. We show that our proposed method achieves optimality in the minimax sense. The fully adaptive version of our method is not explored in this paper, and we leave it as future work.

A Appendix

In the appendix, we provide additional lemmas and proofs that have been omitted from the main text. To simplify the notations, we denote $C_1, C_2, C_3$ and other similar forms as some positive constants that can differ on different occurrences.
A.1 Technical Lemmas

We first provide some Technical lemmas frequently used in the proof of our main theorems.

**Lemma 2 (Gilbert-Varshamov bound [Gilbert 1952])** Denote $A_q(n, d)$ as the maximum possible size of a $q$-ary code $Q$ with length $n$ and minimum Hamming distance $d$. Then

$$A_q(n, d) \geq \frac{q^n}{\sum_{j=0}^{d-1} \left(\begin{array}{c} n \\ j \end{array}\right)(q - 1)^j}.$$  \hspace{1cm} (21)

We use the 2-ary Gilbert-Varshamov bound on a Hamming distance sphere which is widely used in the high-dimensional problem. Denote $S_k = \{x \in \{0, 1\}^m | \rho_H(x) = k\}$ as the Hamming ball with radius $k$. Then the $\rho$-packing number of $S_k$ is lower bounded as

$$M(\rho; S_k, || \cdot ||_H) \geq \left(\begin{array}{c} m \\ k \end{array}\right).$$

Equivalently,

$$\log M(C_1k; S_k, || \cdot ||_H) \geq C_2k \log \frac{em}{k}.$$  \hspace{1cm} (22)

The proof of Lemma 2 can see [Wu 2017] for more details.

**Lemma 3 (Lower bounds for the packing number)** The packing number of the parameter space $\Theta_0^{m,d}(s, s_0)$ with Hamming distance $\frac{ss_0}{4}$ is lower bounded as

$$M(\frac{ss_0}{4}; \Theta_0^{m,d}(s, s_0), || \cdot ||_H) \geq \exp\left(\frac{1}{4} s \log \frac{em}{s} \right) \exp\left(\frac{ss_0}{4} \log \frac{sd}{s_0}\right).$$

**Proof 1** The proof of Lemma 3 contains four major steps.

**Step 1:** From (22), $\exists$ a packing set $\tilde{\Gamma} \subset \Gamma$ and $|\tilde{\Gamma}| \geq \exp\left(\frac{s}{4} \log \frac{em}{s}\right), \forall \gamma^1, \gamma^2 \in \tilde{\Gamma}$, we have $\rho_H(\gamma^1, \gamma^2) \geq \frac{ss_0}{4}$.

**Step 2:** From (22), $\exists$ a packing set $\tilde{B} \subset B$ and $|\tilde{B}| \geq \exp\left(\frac{ss_0}{2} \log \frac{sd}{s_0}\right), \forall b^1, b^2 \in \tilde{B}$, we have $\rho_H(b^1, b^2) \geq \frac{ss_0}{4}$.

**Step 3:** We consider the maximum possible size of a $|\tilde{B}|$-ary code with length $s$. From (21), we have

$$A_{|\tilde{B}|}(s, \frac{s}{2}) = \frac{|\tilde{B}|^s}{\sum_{i=0}^{s-1} \left(\begin{array}{c} s \\ i \end{array}\right)(|\tilde{B}| - 1)^i} \geq \frac{|\tilde{B}|^s}{\left(\begin{array}{c} s \\ \frac{s}{2} \end{array}\right)} \geq \exp\left(\frac{ss_0}{4} \log \frac{sd}{s_0}\right),$$

where the last inequality follows from step 2. Here, we obtain a packing set $Q$ of the $|\tilde{B}|$-ary code with length $s$. For $Q^1, Q^2 \in Q$, we have $\rho_H(Q^1, Q^2) \geq \frac{ss_0}{4}$.

Steps 1-3 can be seen as a procedure to construct a packing set $\tilde{\Theta}_0^{m,d}(s, s_0) \subset \Theta_0^{m,d}(s, s_0)$. Intuitively speaking, the elements of $\tilde{\Gamma}$ determine the locations of the non-zero columns.
of \( \theta \in \tilde{\Theta}_0^{m,d}(s, s_0) \), while \( \tilde{B} \) determines the candidates for each non-zero column. After localizing the positions of the non-zero columns, we determine the choices of these columns by a code \( Q \) from \( Q \) in step 3. Then, according to \( Q \), we select the corresponding columns from \( \tilde{B} \) to fill the matrix \( \theta \).

**Step 4:** In step 4, we show that \( \tilde{\Theta}_0^{m,d}(s, s_0) \) is a \( \frac{ss_0}{4} \)-packing set of \( \Theta_0^{m,d}(s, s_0) \) and derive the lower bound for the cardinality of \( \tilde{\Theta}_0^{m,d}(s, s_0) \). We split our analysis into the following two cases.

Given \( \theta^i \neq \theta^j \in \tilde{\Theta}_0^{m,d}(s, s_0) \), denote the indices of the non-zero columns of \( \theta^i \) and \( \theta^j \) as \( \gamma^i, \gamma^j \in \Gamma \), respectively. On one hand, if \( \gamma^i \neq \gamma^j \), we have \( \rho_H(\gamma^i, \gamma^j) \geq \frac{s}{4} \) from step 1. Note that both \( \theta^i \) and \( \theta^j \) belong to \( B_0^d(s_0) \). Therefore, we have \( \rho_H(\theta^i, \theta^j) \geq \frac{ss_0}{4} \). On the other hand, when \( \gamma^i = \gamma^j \), from step 3, we know that \( \theta^i, \theta^j \) have at least \( \frac{s}{2} \) different columns by \( Q \). Additionally, step 2 implies that the Hamming distance of the different columns in \( B \) is at least \( \frac{s}{2} \). Consequently, we have \( \rho_H(\theta^i, \theta^j) \geq \frac{ss_0}{4} \).

Combining steps 1-3, we have

\[
M \left( \frac{ss_0}{4}; \Theta_0^{m,d}(s, s_0), \| \cdot \|_H \right) \geq |\tilde{\Theta}_0^{m,d}(s, s_0)| \geq \exp \left( \frac{1}{4} s \log \frac{em}{s} \right) \exp \left( \frac{ss_0}{4} \log \frac{ed}{s_0} \right).
\]

Note that \( \left( \frac{s}{2} \right) \) is negligible due to the slower rate compared to the first multiplier. \( \square \)

**Lemma 4** (Upper bounds for the covering number)

(a) For \( q = 0 \),

\[
\log N(\varepsilon; \Theta_0^{m,d}(s, s_0), \| \cdot \|_F) \leq s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0}.
\]

(b) For \( q \in (0, 1] \) and for all \( \varepsilon \in [\sqrt{s}C_qR_q^\frac{1}{2}\left( \frac{\log d}{d} \right)^{\frac{q}{2}}, \sqrt{s}R_q^{\frac{1}{2}}] \),

\[
\log N(\varepsilon; \Theta_0^{m,d}(s, R_q), \| \cdot \|_F) \leq s \log \frac{em}{s} + s(C_q\frac{sR_q^\frac{1}{2}}{\varepsilon^\frac{1}{2}})^{\frac{q}{2}} \log d.
\]

**Proof 2** We first prove case (b), and case (a) can be proved similarly. For case (b), define \( \tilde{B}_q^d(R_q) \) as the \( \frac{\varepsilon}{\sqrt{s}} \)-covering set of \( B_q^d(R_q) \) with respect to \( \| \cdot \|_2 \). Given \( \theta \in \Theta_q^{m,d}(s, R_q) \), for the \( j \)-th column \( \theta_j \), we can find a vector \( b^j \in \tilde{B}_q^d(R_q) \) that satisfies \( \| b^j - \theta_j \|_2 \leq \frac{\varepsilon^2}{s} \). Therefore, we have

\[
N(\varepsilon; \Theta_q^{m,d}(s, R_q), \| \cdot \|_F) \leq \left( \frac{m}{s} \right)^s \left( N(\frac{\varepsilon}{\sqrt{s}}; B_q^d(R_q), \| \cdot \|_2) \right)^s.
\]
Following from (3b), we have
\[ \epsilon_k(B_d^q(1)) = \frac{\epsilon}{\sqrt{s}} \leq \left( \frac{\log(d/k) + 1}{k} \right)^{\frac{1}{q} - \frac{1}{2}} \leq \left( \frac{\log d}{k} \right)^{\frac{1}{q} - \frac{1}{2}}. \]

Inverting this inequality for \( k = \log N(\frac{\epsilon}{\sqrt{s}}; B_d^q(1), \| \cdot \|_2) \) and allowing for a ball radius \( R_1^q \) yields
\[ \log N(\frac{\epsilon}{\sqrt{s}}; B_d^q(1), \| \cdot \|_2) \leq (C_q \frac{sR_1^2}{\epsilon^2})^{\frac{q}{2}} s \log d. \]

The conditions on the range of \( \epsilon \) guarantee that \( k \in [\log d, d] \). Combining these results, we have
\[ \log N(\epsilon; \Theta^m_d(s, R_q), \| \cdot \|_F) \leq s \log \frac{em}{s} + s(C_q \frac{sR_1^2}{\epsilon^2})^{\frac{q}{2}} \log d. \]

Case (a) can be derived in a similar way as case (b), in which we just need to replace the covering number of \( B_d^q(R_q) \) by \( B_0^d(s_0) \).

A.2 Proof of Theorem 1

Proof 3 For case (a), consider the \( \frac{ss_0}{4} \)-packing set \( \tilde{\Theta}^m_d(s, s_0) = \{ \theta^1, \ldots, \theta^M \} \) defined in Lemma 3, where \( M \) is the cardinality of \( \tilde{\Theta}^m_d(s, s_0) \). Since under \( \ell_0 \)-ball, the values of the non-zero elements of \( \theta \) are not crucial. So we set all the non-zero elements of \( \theta \in \tilde{\Theta}^m_d(s, s_0) \) equal to 1. Let \( \eta^i = \theta^i \delta \), where \( \delta \) is a parameter that will be determined below. For any \( \eta^i \neq \eta^j \), since each \( \theta \in \tilde{\Theta}^m_d(s, s_0) \) has at most \( ss_0 \) elements, we have
\[ \| \eta^i - \eta^j \|_F \leq 2ss_0\delta^2, \forall i, j \in [M] . \] (23)

On the other hand, from the construction of \( \tilde{\Theta}^m_d(s, s_0) \), we have
\[ \| \eta^i - \eta^j \|_F \geq \frac{1}{4} ss_0\delta^2, \forall i, j \in [M] . \] (24)

From the property of the mutual information [Wu (2017)], we obtain that \( I(y; B) \) is upper
bounded by

\[
I(y; B) \leq \frac{1}{M} \sum_{i \neq j} KL(\eta^i || \eta^j) \\
= \frac{1}{M} \sum_{i \neq j} \frac{n}{2\sigma^2} \|\eta^i - \eta^j\|_F^2 \\
\leq \frac{n}{\sigma^2 ss_0 \delta^2},
\]

where \( KL(\cdot || \cdot) \) denotes the Kullback-Leibler divergence, and the last inequality follows from (23). Combining Fano’s inequality [Thomas and Joy 2006] and (25), we have

\[
P(B \neq \hat{\eta}) \geq 1 - \frac{n ss_0 \delta^2 + \log 2}{\log M},
\]

where \( B \) is the random vector uniformly distributed over the packing set \( \tilde{\Theta}_{m,d}^{m,d}(s, s_0) \). To guarantee that \( P(B \neq \hat{\eta}) \geq \frac{1}{2} \), it suffices to choose \( \delta = \frac{1}{2} \sqrt{\left( \frac{1}{8} s \log \frac{em}{s} + \frac{ss_0}{4} \log \frac{ed}{s_0} \right) \frac{\sigma^2}{ss_0}} \). Substituting into equation (24), and by Lemma 3, we have

\[
P\left( M(\Theta_{0,d}^{m,d}(s, s_0)) \geq \frac{\sigma^2}{64n} \left( s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0} \right) \right) \geq \frac{1}{2},
\]

which completes the proof of (3).

For case (b), define \( B = \{ b : b \in \delta \times \{0,1\}^d \text{ and } \|b\|_0 \leq s_0 \} \), where \( \delta \) satisfies that \( \delta^q s_0 = R_q \). Obviously, \( B \subseteq B_{q}(R_{q}) \). Given \( B \), define \( \tilde{\Theta}_{q,d}^{m,d}(s, R_{q}) \) as

\[
\tilde{\Theta}_{q,d}^{m,d}(s, R_{q}) = \tilde{\Gamma} \odot \Lambda \text{ for } \lambda \in \Lambda \subseteq B^m,
\]

where \( \tilde{\Gamma} \) is the same as the definition in Lemma 3. Here we consider splitting the proof of case (b) into two cases. The first case focuses on the parameter space in that the element has the same entries with different indices of the non-zero columns. The other case studies the parameter space that the element has the same indices of the non-zero columns but the entries are different. At last, we combine these two cases by union bounds.

**Case 1:** given any \( \lambda \in B^m \), consider the parameter space

\[
\tilde{\Theta}_{q,d}^{m,d}(s, R_{q}, \lambda) = \{ [\gamma_1 \otimes \lambda_1] \ldots [\gamma_m \otimes \lambda_m] : \gamma \in \tilde{\Gamma} \}.
\]

Denote the cardinality of \( \tilde{\Theta}_{q,d}^{m,d}(s, R_{q}, \lambda) \) as \( M_\lambda \). For any \( \theta^i \neq \theta^j \in \tilde{\Theta}_{q,d}^{m,d}(s, R_{q}, \lambda) \), denote the corresponding indices of the non-zero columns as \( \gamma^i \) and \( \gamma^j \), respectively. From the construction of \( \tilde{\Gamma} \), if \( \gamma^i \neq \gamma^j \), we have \( \rho_H(\gamma^i, \gamma^j) \geq \frac{s}{4} \). Therefore, for \( i, j \in [M_\lambda] \), we have

\[
\|\theta^i - \theta^j\|_F^2 \geq \frac{1}{4} ss_0 \delta^2,
\]

(26)
which implies that $\Theta_{q}^{m,d}(s, R_q, \lambda)$ is a $\frac{1}{4} s s_0 \delta^2$-packing set of $\Theta_{q}^{m,d}(s, R_q)$. On the other hand, $\gamma^i$ and $\gamma^j$ differ in at most $2s$ positions. Therefore,

$$\|\theta^i - \theta^j\|_F^2 \leq 2ss_0\delta^2. \quad (27)$$

Similar to (25), we have

$$I(y; B) \leq \frac{1}{(M\lambda)^2} \sum_{i \neq j} KL(\theta^i\|\theta^j)$$

$$= \frac{1}{(M\lambda)^2} \sum_{i \neq j} \frac{n}{2\sigma^2} \|\theta^i - \theta^j\|_F^2$$

$$\leq \frac{n}{\sigma^2} ss_0\delta^2, \quad (28)$$

where the last inequality follows from (27). Combining Fano’s inequality and (28), we have

$$P(B \neq \hat{\theta}) \geq 1 - \frac{n}{\sigma^2} ss_0\delta^2 + \log 2 \frac{em}{s},$$

where $B$ is the random vector uniformly distributed over the packing set $\tilde{\Theta}_{q}^{m,d}(s, R_q, \lambda)$. Note that $\delta^2 s_0 = R_q$. It suffices to choose $\delta = (\frac{s^2 \log \frac{em}{s}}{16nR_q})^{\frac{1}{2}}$ to guarantee $P(B \neq \hat{\theta}) \geq \frac{1}{2}$.

Substituting into (26), we have

$$P \left( \min_{\hat{\theta}} \max_{\theta \in \tilde{\Theta}_{q}^{m,d}(s, R_q, \lambda)} \|\hat{\theta} - \theta\|_F^2 \geq \frac{\sigma^2}{64n} s \log \frac{em}{s} \right) \geq \frac{1}{2}. \quad (29)$$

**Case 2:** given $\gamma \in \tilde{\Gamma}$, consider the parameter space

$$\tilde{\Theta}_{q}^{m,d}(s, R_q, \gamma) = \{[\gamma_1 \otimes \lambda_1| \ldots |\gamma_m \otimes \lambda_m] : \lambda \in \Lambda\}.$$

Denote the cardinality of $\tilde{\Theta}_{q}^{m,d}(s, R_q, \gamma)$ as $M_{\gamma}$. Similar to steps 2-3 in Lemma 3, we have

$$\|\theta^i - \theta^j\|_F^2 \geq \frac{1}{4} ss_0\delta^2, \quad i, j \in [M_{\gamma}],$$

which implies that $\tilde{\Theta}_{q}^{m,d}(s, R_q, \gamma)$ is a $\frac{1}{4} ss_0\delta^2$-packing set of $\Theta_{q}^{m,d}(s, R_q)$, and

$$|\tilde{\Theta}_{q}^{m,d}(s, R_q, \gamma)| \geq \frac{1}{4} ss_0 \log \frac{ed}{s_0}. $$

On the other hand, we have

$$\|\theta^i - \theta^j\|_F^2 \leq 2ss_0\delta^2, \quad i, j \in [M_{\gamma}].$$
Combining Fano’s inequality and a similar derivation of (28), we have

\[ P(B \neq \hat{\theta}) \geq 1 - \frac{n\sigma^2 s s_0 \delta^2 + \log 2}{\frac{1}{4} s s_0 \log \frac{ed}{s_0}}, \]

where \( B \) is the random vector uniformly distributed over the packing set \( \widetilde{\Theta}_m(d, s, R_q, \gamma) \). To guarantee \( P(B \neq \hat{\theta}) \geq \frac{7}{8} \), assume \( s_0 = d^v \) for some constant \( v \in (0, 1] \). It suffices to choose \( \delta = (\frac{1-v}{4n} \log d)^{\frac{1}{2}} \) from (4). Then, we obtain \( s_0 = R_q(\frac{(1-v)\sigma^2}{4n} \log d)^{-\frac{1}{2}} \). As a result, we have

\[ P(\min_{\theta} \max_{\hat{\theta} \in \Theta_m(d, s, R_q, \lambda)} \| \hat{\theta} - \theta \|^2_F \geq \frac{1}{4n} s R_q(\frac{(1-v)\sigma^2}{4n} \log d)^{1-\frac{1}{2}}) \geq \frac{7}{8}. \] (30)

We combine the risk in (29) and (30) by union bound. For any \( \theta \in \Theta_m(d, s, R_q) \),

\[ P(\| \hat{\theta} - \theta \|^2_F \geq \frac{1}{4n} s R_q(\frac{(1-v)\sigma^2}{4n} \log d)^{1-\frac{1}{2}}) \geq 1 - P(\| \hat{\theta} - \theta \|^2_F \leq \frac{1}{4n} s R_q(\frac{(1-v)\sigma^2}{4n} \log d)^{1-\frac{1}{2}}) - P(\| \hat{\theta} - \theta \|^2_F \leq \frac{\sigma^2}{64n} s \log \frac{em}{s}) \]

\[ = P(\| \hat{\theta} - \theta \|^2_F \geq \frac{1}{4n} s R_q(\frac{(1-v)\sigma^2}{4n} \log d)^{1-\frac{1}{2}}) + P(\| \hat{\theta} - \theta \|^2_F \geq \frac{\sigma^2}{64n} s \log \frac{em}{s}) - 1. \]

Taking \( \sup \) on both sides, we have

\[ \sup_{\theta \in \Theta_m(d, s, R_q)} P(\| \hat{\theta} - \theta \|^2_F \geq \frac{1}{4n} s R_q(\frac{(1-v)\sigma^2}{4n} \log d)^{1-\frac{1}{2}} + \frac{\sigma^2}{64n} s \log \frac{em}{s}) \]

\[ \geq \sup_{\theta \in \tilde{\Theta}_m(d, s, R_q, \lambda)} P(\| \hat{\theta} - \theta \|^2_F \geq \frac{1}{4n} s R_q(\frac{(1-v)\sigma^2}{4n} \log d)^{1-\frac{1}{2}}) + \]

\[ \sup_{\theta \in \tilde{\Theta}_m(d, s, R_q, \gamma)} P(\| \hat{\theta} - \theta \|^2_F \geq \frac{\sigma^2}{64n} s \log \frac{em}{s}) - 1 \]

\[ \geq \frac{7}{8} + \frac{1}{2} - 1 = \frac{3}{8}. \]

Therefore,

\[ P(\mathcal{M}(\Theta_m(d, s, R_q)) \geq \frac{\sigma^2}{64n} s \log \frac{em}{s} + \frac{1}{4n} s R_q(\frac{(1-v)\sigma^2}{4n} \log d)^{1-\frac{1}{2}}) \geq \frac{3}{8} \]

which completes the proof of (6).
A.3 Proof of Theorem 2

Before formally providing the proof of Theorem 2, we provide some useful lemmas, the peeling results in Raskutti et al. [2011]. Define the function \( f(v; X) \), where \( v \in \mathbb{R}^d \) is the vector to be optimized over, and \( X \) is some random vector. We are interested in the constrained problem \( \sup_{\rho(v) \leq r, v \in A} f(v; X) \), where \( \rho : \mathbb{R}^d \to \mathbb{R}^+ \) is some increasing constraint function, and \( A \) is a nonempty set. The goal is to bound the probability of the event defined by

\[
Z := \{ X \in \mathbb{R}^{n \times d} | \exists v \in A \text{ s.t. } f(v; X) \geq 2g(\rho(v)) \}
\]

where \( g : \mathbb{R} \to \mathbb{R}^+ \) is some strictly increasing function.

Lemma 5 (Peeling, Lemma 9 of Raskutti et al. [2011]) Suppose \( g(r) \geq \mu, \forall r \geq 0 \). There exists some constant \( c > 0 \) such that for all \( r > 0 \), we have the tail bound

\[
P \left[ \sup_{v \in A, \rho(v) \leq r} f(v; X) \geq g(r) \right] \leq 2 \exp \left( -ca_n g(r) \right)
\]

for some \( a_n > 0 \). Then, we have

\[
P[Z] \leq \frac{2 \exp(-4ca_n \mu)}{1 - \exp(-4ca_n \mu)}.
\]

In addition, we also need the double sparsity version of Lemma 6 of Raskutti et al. [2011]. Denote \( \tilde{S}((\Theta_0^{m,d}(2s,2s_0), r) = \Theta_0^{m,d}(2s,2s_0) \cap \{ \theta \in \mathbb{R}^{d \times m} | \| \theta \|_F^2 \leq r \} \).

Lemma 6 Assume that there exists a positive constant \( \tau_u > 0 \) such that

\[
\frac{1}{\sqrt{n}} \| X \beta \|_2 \leq \tau_u \| \beta \|_2 \text{ for all } \Pi(\beta) \in \Theta_0^{m,d}(2s,2s_0),
\]

where \( X \) is an \( n \times p \) design matrix. Then there exists some constants \( C_1, C_2 > 0 \), such that for any \( r > 0 \), we have

\[
\sup_{\Pi(\beta) \in \tilde{S}(\Theta_0^{m,d}(2s,2s_0), r)} | z^T X \beta | \leq 6 \sigma r \tau_u \sqrt{\frac{1}{n} \left( s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0} \right)}
\]

with probability greater than \( 1 - C_1 \exp\{-C_2(s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})\} \).

Lemma 6 is a direct consequence of Lemma 6 of Raskutti et al. [2011], in which we replace the covering number of \( \ell_0 \)-ball by the covering number of \( \Theta_0^{m,d}(2s,2s_0) \) in Lemma 4 (a). Notably, in model (1), the design matrix \( X \) simplifies to \( I_n \) and \( \tau_u = 1 \).

Proof 4 we consider the constrained estimator in (7). Note that

\[
\| Y - \hat{\theta}_q \|_F^2 \leq \| Y - \theta^* \|_F^2.
\]
By some simple algebras, we have

\[ \|\hat{\theta}_q - \theta^*\|^2_F \leq 2(Z, \hat{\theta}_q - \theta^*)_F = 2\text{tr}(Z^\top (\hat{\theta}_q - \theta^*)), \]

where \(\text{tr}()\) is trace of a matrix. To simplify the proof, denote \(\eta = \text{Vec}(\theta), \hat{\eta} = \text{Vec}(\hat{\theta}), \hat{\eta}_q = \text{Vec}(\hat{\theta}_q), \eta^* = \text{Vec}(\theta^*), y = \text{Vec}(Y)\) and \(z = \text{Vec}(Z)\). Therefore, \((31)\) is equivalent to

\[ \|\hat{\eta}_q - \eta^*\|^2 \leq 2|z^\top (\hat{\eta}_q - \eta^*)| \]

For case (a), since \(\hat{\theta}_0, \theta^* \in \Theta_0^{m,d}(s, s_0)\), we have \(\Pi(\hat{\eta}_0 - \eta^*) \in \Theta_0^{m,d}(2s, 2s_0)\). From Lemma \(\boxed{\ref{lemma}}\) for any \(r > 0\), we have

\[ \sup_{\Pi(\eta) \in \tilde{S}(\Theta_0^{m,d}(2s, 2s_0), r)} |z^\top \eta| \leq 6\sigma r \sqrt{\frac{1}{n} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})} \]

with probability greater than \(1 - C_1 \exp\{-C_2(s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})\}\).

Consider the event \(Z\) that there exists some \(\eta\) satisfying \(\Pi(\eta) \in \Theta_0^{m,d}(2s, 2s_0)\) such that

\[ |z^\top \eta| \geq 6\sigma \|\eta\|_2 \sqrt{\frac{1}{n} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})}. \]

Following from Lemma \(\boxed{\ref{lemma}}\), we have

\[ P[Z] \leq \frac{2 \exp\{-C_3(s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})\}}{1 - \exp\{-C_3(s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})\}}. \]

This claim follows from Lemma \(\boxed{\ref{lemma}}\) by choosing the function \(f(v; X) = |z^\top \eta|\), the set \(A = \Theta_0^{m,d}(2s, 2s_0)\), the sequence \(a_n = n\), and the functions \(\rho(v) = \|v\|_2\), and \(g(r) = 6\sigma r \sqrt{\frac{1}{n} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})}\). For any \(r \geq \sqrt{\frac{1}{n} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})}\), we are guaranteed that \(g(r) \geq \frac{2}{n} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})\) and \(\mu = \frac{2}{n} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})\), so that the lemma may be applied.

Consequently, by \((32)\) and replacing \(\eta\) in \((33)\) by \(\hat{\eta}_0 - \eta^*\), we have

\[ \|\hat{\theta}_0 - \theta^*\|^2_F = \|\hat{\eta}_0 - \eta^*\|^2_2 \leq C_u \frac{\sigma^2}{n} (s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0}) \]

with probability greater than \(1 - C_1 \exp\{-C_2(s \log \frac{em}{s} + ss_0 \log \frac{ed}{s_0})\}\). This completes the proof for \((8)\).

For case (b), since \(\hat{\theta}_q, \theta^* \in \Theta_q^{m,d}(s, R_q)\), we have \(\Pi(\hat{\eta}_q - \eta^*) \in \Theta_q^{m,d}(2s, 2R_q)\). Denote \(\tilde{S}(\Theta_q^{m,d}(2s, 2R_q), r) = \Theta_q^{m,d}(2s, 2R_q) \cap \{\theta \in \mathbb{R}^{d \times m} \|\theta\|^2_F \leq r\}\) and \(W(\Theta_q^{m,d}(2s, 2R_q), r) = \)

25
\[
\sup_{\Pi_0(\eta)} \| \eta \|_i \leq C_2 n R_q^2 \left( \log d \right)^{\frac{\frac{4}{3} - q}{2}}. 
\] (34)

Since the situation \( \log \frac{em}{s} \geq C_2 n R_q^2 \left( \log d \right)^{\frac{\frac{4}{3} - q}{2}} \) leads to a zero estimator, which is trivial.

Next, we want to construct a constant \( \delta \) that satisfies
\[
\sqrt{n} \delta \geq C_1 r 
\] (35)

and
\[
C_2 \sqrt{n} \delta \geq \int_{\frac{\delta}{16}}^r \sqrt{\log N \left( t; \tilde{S}(\Theta_q^{m,d}(2s, 2R_q), r), \| \cdot \|_F \right)} dt := J(r, \delta). 
\] (36)

As long as \( \| z \|_2^2 \leq 16 \), Lemma 3.2 in Geer et al. [2000] guarantees that
\[
P \left( W(\Theta_q^{m,d}(2s, 2R_q), r) \geq \delta, \| z \|_2^2 \leq 16 \right) \leq C_3 \exp(-C_4 \frac{n \delta^2}{r^2}). 
\]

Note that each entry of \( z \) draws from \( N(0, \frac{\sigma^2}{n}) \). By the tail bounds of \( \chi^2 \) random variables in Raskutti et al. [2011], we have \( P(\| z \|_2^2 \geq 16) \leq C_5 \exp(-C_6 n) \). Consequently, we have
\[
P \left( W(\Theta_q^{m,d}(2s, 2R_q), r) \geq \delta \right) \leq C_3 \exp(-C_4 \frac{n \delta^2}{r^2}) + C_5 \exp(-C_6 n). 
\]

Next, we construct \( \delta \) to satisfy conditions (35) and (36). Let \( \delta = r \left( \sqrt{\frac{s \log \frac{en}{s}}{n}} + \omega \right) \), where \( \omega > 0 \). Obviously, (35) holds. Turning to (36), based on the condition (34), we set \( r = \Omega \left( \sqrt{\frac{s \log \frac{en}{s}}{n}} + \sqrt{s R_q \left( \log d \right)^{\frac{\frac{4}{3} - q}{2}}} \right) \), and it is straightforward to verify that \( \left( \frac{\delta}{16}, r \right) \) lies in the range of \( \varepsilon \) in Lemma 4. Combined with the part (b) of Lemma 4, we have
\[
J(r, \delta) = \int_{\frac{\delta}{16}}^r \sqrt{\log N \left( t; \tilde{S}(\Theta_q^{m,d}(2s, 2R_q), r) \right)} dt 
\leq \int_0^r \sqrt{2s \log \frac{em}{s} \left( 2s R_q \right)^{\frac{4}{3} - q}} \log d \ dt 
\leq \sqrt{2s \log \frac{em}{s} \left( 2s R_q \right)^{\frac{1}{3} - \frac{q}{2}}} \sqrt{\log d} \left( 1 - \frac{q}{2} \right). \]
With our choice of $\delta$, we have

$$
\frac{J(r, \delta)}{\sqrt{n\delta}} \leq \sqrt{2s \log \frac{en}{s} r + \sqrt{2(sR_q) \frac{1}{2} \sqrt{\log dr^{1 - \frac{3}{2} q}}} \cdot \frac{\sqrt{n\delta}}{r \sqrt{s \log \frac{en}{s} + r \sqrt{n\omega}}}.
$$

Setting $\omega = \sqrt{2(sR_q)^{\frac{1}{2} q} \sqrt{\log r^{1 - \frac{3}{2} q}}}$, we obtain $\frac{J(r, \delta)}{\sqrt{n\delta}} \leq \sqrt{2}$, which implies that (36) holds.

Note that $r = \Omega(\sqrt{s \log \frac{en}{s} n + \sqrt{sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}}}})$, we have $\delta = \left[\sqrt{s \log \frac{en}{s} n + \sqrt{sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}}}}\right] r$.

Therefore, with probability at least $1 - C_3 \exp\left[-C_4 n(\frac{s \log \frac{en}{s} n + sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}})}\right]$, we obtain $J(r, \delta) \sqrt{n\delta} \leq \sqrt{2}$, which implies that (36) holds.

We define the event $Z$ as there exists some $\eta$ such that $\Pi(\eta) \in \Theta_m^m(s, R_q)$ so that

$$
|z^T \eta| \geq C_u \|\eta\|_2 \left(\sqrt{\frac{s \log \frac{en}{s}}{n} + \sqrt{sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}}}\right).
$$

From the Lemma 9 of Raskutti et al. [2011], we have

$$
P(Z) \leq \frac{2 \exp\left(-C_7 n(\frac{s \log \frac{en}{s}}{n} + sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}})}\right)}{1 - \exp\left(-C_7 n(\frac{s \log \frac{en}{s}}{n} + sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}})}\right). \tag{37}
$$

This claim follows from Lemma 9 by choosing the function $f(v, X) = |z^T \eta|$, the set $A = \Theta_m^m(2s, 2R_q)$, the sequence $a_n = n$, and the functions $\rho(v) = \|v\|_2$, and $g(r) = r \left(\sqrt{\frac{s \log \frac{en}{s}}{n} + \sqrt{sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}}}\right)$.

Combining these results and (32), with probability at least (37), we have

$$
\|\hat{\eta} - \eta^*\|_2 \leq C_u \|\hat{\eta} - \eta^*\|_2 \left(\sqrt{\frac{s \log \frac{en}{s}}{n} + \sqrt{sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}}}\right).
$$

Consequently,

$$
\|\hat{\eta} - \eta^*\|_2 \leq C_u \left(\frac{s \log \frac{en}{s}}{n} + sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}}\right).
$$

By the definition of $\Pi(\cdot)$, we have

$$
\|\hat{\theta} - \theta^*\|_2 \leq C_u \left(\frac{s \log \frac{en}{s}}{n} + sR_q(\frac{\log d}{n})^{\frac{1}{2} - \frac{q}{4}}\right).
$$
with probability greater than $1 - C_3 \exp\{-C_4 n\left(\frac{s \log \frac{sn}{n}}{n} + s R_q\left(\frac{\sigma^2 v}{4m} \log d\right)^{1-q}\right)\}$, which completes the proof for (9). □

### A.4 Proof of Corollary 1

Minimax rates for double sparsity linear regression are some straight corollaries of the Gaussian location model above when we add design matrix assumptions [12] to it. The intuition of the design matrix assumptions we used here is similar to Raskutti et al. [2011]. Therefore, we need to provide a brief sketch because the main technique here is the same as the proofs of Theorem 1 and Theorem 2.

(a) Upper bound: We denote $\hat{\Delta} = \hat{\beta} - \beta^*$. When $q = 0$, by the definition of maximum likelihood estimator:

$$\hat{\beta} \in \arg \min_{\Pi(\beta) \in \Theta_{m,d}^{m,d}(s, s_0)} \|y - X\beta\|^2 \tag{38}$$

we have $\Pi(\hat{\Delta}) \in \Theta_{m,d}^{m,d}(2s, 2s_0)$. Therefore, by assumption [1]

$$\|\hat{\Delta}\| \leq \frac{1}{\tau_\ell} \|X\hat{\beta} - X\beta\| \leq \sup_{\Pi(\Delta) \in \Theta_{m,d}^{m,d}(2s, 2s_0)} \langle \xi, X^T \hat{\Delta} \rangle$$

The remaining proof is the same as proof [4] by using Lemma 6.

When $q \in (0, 1]$, we restrict our discussion under the condition that $\|\hat{\Delta}\| = \Omega(h_\ell (R_q, n, s, d))$. By Restricted Eigenvalue Condition [2] we conclude the similar inequality:

$$\|\hat{\Delta}\| \leq \frac{1}{\tau_{q,\ell}} \|X\hat{\beta} - X\beta\| \leq \sup_{\Pi(\Delta) \in \Theta_{m,d}^{m,d}(2s, 2R_q)} \langle \xi, X^T \hat{\Delta} \rangle$$

The remaining proof is the same as proof [4] Case (b).

(b) Lower bound:

First, by Lemma 1 of Raskutti et al. [2011], restricted eigenvalue condition helps us to avoid low-rank design matrix situation.

In linear regression parameter estimation here, we use the same packing set construction as in the proof [A.2]. The only difference is that we calculate the upper bound of mutual information in Fano’s inequality:

$$I(y, B) \leq \frac{n}{2\sigma^2} \|X(\beta^i - \beta^j)\|^2 \leq \frac{n}{\sigma^2} ss_0 \tau^2 u^2 \delta^2$$

28
So it is the similar to the proof A.2 except we let \( \delta' \) here be equal to \( \tau u \delta \), where \( \delta \) is defined in proof A.2.

A.5 Proof of Theorem 3

Lemma 7 (Lemma 2 of Ndaoud [2020]) For a given subset \( S \subset [d] \times [m] \) such that \( |S| \leq ss_0 \), we have

\[
P \left( \frac{1}{n} \sum_{(i,j) \in S} (X^T \xi) \leq \sigma^2 (ss_0 + t) \right) e^{-\left( t/2 \right)} \left( 1/4 \right), \forall t \geq 0.
\]

Based on Lemma 7, Lemma 1 is proved by union bound and calculating that:

\[
|S^{m,d}(s, s_0)| \leq \left( \frac{m}{s} \right) \left\{ \left( \frac{d}{s_0} \right) \right\}.
\]

Proof 5 We proceed with the proof of Theorem 3 by induction. The result is trivial for \( t = 0 \). We assume (16) and (17) both hold for \( t \). Then, based on these, we prove that similar results hold for \( t + 1 \). Denote

\[
H^{t+1} = \beta^* + \left( \frac{1}{n} X^T X - I_q \right) (\beta^* - \beta_t) + \frac{1}{n} X^T \xi = \beta^* + \Phi(\beta^* - \beta_t) + \Xi.
\]

Therefore,

\[
\beta_{t+1} = T_{\lambda_{t+1}}(H^{t+1}).
\]

We prove (16) by contradiction. Assume that (16) is wrong for \( t + 1 \), which implies that when we consider the support \((S^*)^c\), there exists at least \( s \) columns satisfying the condition (13) or there exists at least \( s_0 \)-ordering rows satisfying the condition (14). Both of these can lead to the following result, that is, there exists a subset \( \tilde{S} \) of \((S^*)^c\) satisfying that

\[
ss_0 \lambda_{t+1}^2 \leq \sum_{(i,j) \in \tilde{S}} |T_{\lambda_{t+1}}(H^{t+1})|_{ij}
\]

This statement has been elaborated in Section 4.3. Since \( \beta_{ij}^* = 0 \) for \((i,j) \in \tilde{S}\). Then

\[
\sqrt{ss_0 \lambda_{t+1}} \leq \sqrt{\sum_{(i,j) \in S} \Xi_{ij}^2} + \sqrt{\sum_{(i,j) \in \tilde{S}} (\Phi_{ij}^T \beta^* - \beta_t)^2}.
\]

Note that \( \beta^* \) is \((s, s_0)\)-sparse and \( S_t \cap (S^*)^c \in S^{m,d}(s, s_0) \) by induction. Combining the
event $\mathcal{E}$ in Lemma 7, we have that
\[
\sqrt{ss_0\lambda_{t+1}} \leq \sqrt{\frac{10\sigma^2 \left(s \log\left(\frac{ep}{s}\right) + ss_0 \log\left(\frac{ed}{s_0}\right)\right)}{n}} + (2 + \sqrt{3})\delta \sqrt{ss_0\lambda_t} \\
\leq (\frac{1}{2} + (2 + \sqrt{3})\sqrt{\delta})\sqrt{ss_0\lambda_{t+1}} \\
< \sqrt{ss_0\lambda_{t+1}}
\]
as long as $\delta < \frac{7 - \sqrt{3}}{4}$, which is absurd. Therefore, (16) has been proved. Here the second follows from the order of $\lambda_\infty$ and $\lambda_\infty \leq \lambda_t$. For the second part, by (16) and induction hypothesis, we have
\[
\|\hat{\beta}_{t+1}(S^*)\|_2 \leq \sqrt{\frac{10\sigma^2 \left(s \log\left(\frac{ep}{s}\right) + ss_0 \log\left(\frac{ed}{s_0}\right)\right)}{n}} + \delta\|\hat{\beta}_t - \beta^*\|_2. \tag{40}
\]
Moreover, we consider the support set $S' := S^* \cap S_{t+1}^c$, which is consisted of the real parameters that we lose. Because $S^*$ is a $(s, s_0)$-sparse set, and by the meaning of average, the average of entries $(H_{ij}^{(t+1)})^2$ is not larger than $\lambda_{t+1}^2$, and in every condition (entrywise, column, row), we lose no more than $s \cdot s_0$ number of entries. By (39), we have the following result:
\[
\|\hat{\beta}_{t+1} - \beta^*\|_2 \leq \sqrt{3ss_0\lambda_{t+1}} + \delta\|\hat{\beta}_t - \beta^*\|_2 + \frac{10\sigma^2 \left(s \log\left(\frac{ep}{s}\right) + ss_0 \log\left(\frac{ed}{s_0}\right)\right)}{n}. \tag{41}
\]
Combining (40) and (41), we conclude that
\[
\|\hat{\beta}_{t+1} - \beta^*\|_2 \leq \sqrt{3} \cdot \sqrt{ss_0\lambda_{t+1}} + \sqrt{ss_0\lambda_{t+1} + 2\delta\|\hat{\beta}_t - \beta^*\|_2} \\
\leq \sqrt{ss_0\lambda_{t+1}}(\sqrt{3} + 1 + 2(2 + \sqrt{3})\sqrt{\delta}) \\
\leq (2 + \sqrt{3})\sqrt{ss_0\lambda_{t+1}},
\]
where the second inequality follows from $\delta < \kappa$ and $\lambda_{t+1} = \sqrt{\kappa}\lambda_t$.

\section*{References}

David L. Donoho and Iain M. Johnstone. Minimax risk over $\ell_p$-balls for $\ell_p$-error. \textit{Probability Theory and Related Fields}, 99(2):277–303, 1994.
David L Donoho and Iain M Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

David L Donoho and Iain M Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.

Lucien Birgé and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

Zheyang Wu and Harrison H Zhou. Model selection and sharp asymptotic minimaxity. *Probability Theory and Related Fields*, 156(1):165–191, 2013.

Felix Abramovich, Yoav Benjamini, David L Donoho, and Iain M Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34(2):584–653, 2006.

Felix Abramovich, Vadim Grinshtein, and Marianna Pensky. On optimality of bayesian estimation in the normal means problem. *The Annals of Statistics*, 35(5):2261–2286, 2007.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi:https://doi.org/10.1111/j.2517-6161.1996.tb02080.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 04 2010. doi:10.1214/09-AOS729. URL https://doi.org/10.1214/09-AOS729

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls. *IEEE transactions on information theory*, 57(10):6976–6994, 2011.

Xiao-Tong Yuan, Ping Li, and Tong Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18(166):1–43, 2018. URL http://jmlr.org/papers/v18/14-415.html

Junxian Zhu, Canhong Wen, Jin Zhu, Heping Zhang, and Xueqin Wang. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences*, 117(52):33117–33123, 2020. ISSN 0027-8424. doi:10.1073/pnas.2014241117. URL https://www.pnas.org/content/117/52/33117
Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. doi:10.1111/j.1467-9868.2005.00532.x URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.x.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

Karim Lounici, Massimiliano Pontil, Sara van de Geer, and Alexandre B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *The Annals of Statistics*, 39(4):2164 – 2204, 2011. doi:10.1214/11-AOS896 URL https://doi.org/10.1214/11-AOS896.

Matt Silver, Peng Chen, Ruoying Li, Ching-Yu Cheng, Tien-Yin Wong, E-Shyong Tai, Yik-Ying Teo, and Giovanni Montana. Pathways-driven sparse regression identifies pathways and genes associated with high-density lipoprotein cholesterol in two asian cohorts. *PLoS genetics*, 9(11):e1003939, 2013.

Nikhil Rao, Robert Nowak, Christopher Cox, and Timothy Rogers. Classification with the sparse group lasso. *IEEE Transactions on Signal Processing*, 64(2):448–463, 2015.

Yanhao Huo, Lihui Xin, Chuanze Kang, Minghui Wang, Qin Ma, and Bin Yu. Sgl-svm: a novel method for tumor classification via support vector machine with sparse group lasso. *Journal of Theoretical Biology*, 486:110098, 2020.

Yangjing Zhang, Ning Zhang, Defeng Sun, and Kim-Chuan Toh. An efficient hessian based algorithm for solving large-scale sparse group lasso problems. *Mathematical Programming*, 179(1):223–263, 2020.

Jitendra K Tugnait. Sparse-group lasso for graph learning from multi-attribute data. *IEEE Transactions on Signal Processing*, 69:1771–1786, 2021.

Jian Huang, Shuangge Ma, Huiliang Xie, and Cun-Hui Zhang. A group bridge approach for variable selection. *Biometrika*, 96(2):339–355, 2009.

Patrick Breheny and Jian Huang. Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3):369, 2009.

Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 27(4), 2012.

Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245, 2013.
Yasutoshi Ida, Yasuhiro Fujiwara, and Hisashi Kashima. Fast sparse group lasso. *Advances in Neural Information Processing Systems*, 32, 2019.

Soumyadeep Chatterjee, Karsten Steinhaeuser, Arindam Banerjee, Snigdhansu Chatterjee, and Auroop Ganguly. Sparse group lasso: Consistency and climate applications. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 47–58. SIAM, 2012.

T. Tony Cai, Anru R. Zhang, and Yuchen Zhou. Sparse group lasso: Optimal sample complexity, convergence rate, and statistical inference. *IEEE Transactions on Information Theory*, pages 1–1, 2022. doi:10.1109/TIT.2022.3175455.

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440, 2009.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. *Advances in neural information processing systems*, 27, 2014.

Mohamed Ndaoud. Scaled minimax optimality in high-dimensional linear regression: A non-convex algorithmic regularization approach. *arXiv preprint arXiv:2008.12236*, 2020.

Haoyang Liu and Rina Foygel Barber. Between hard and soft thresholding: optimal iterative thresholding algorithms. *Information and Inference: A Journal of the IMA*, 9(4):899–933, 2020.

Mohamed Ndaoud. Interplay of minimax estimation and minimax support recovery under sparsity. In *Algorithmic Learning Theory*, pages 647–668. PMLR, 2019.

Mohamed Ndaoud and Alexandre B. Tsybakov. Optimal variable selection and adaptive noisy compressed sensing. *IEEE Transactions on Information Theory*, 66(4):2517–2532, 2020. doi:10.1109/TIT.2020.2965738.

Hans Triebel. *Fractals and spectra: related to Fourier analysis and function spaces*. Springer Science & Business Media, 2010.

Thomas Kühn. A lower estimate for entropy numbers. *Journal of Approximation Theory*, 110(1):120–124, 2001.
Yuchen Zhou. *High-Dimensional Inference for Low-Dimensional Structures: Double Sparse Vectors and Low-Rank Tensors*. The University of Wisconsin-Madison, 2021.

E.J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005. doi:10.1109/TIT.2005.858979

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

MTCAJ Thomas and A Thomas Joy. *Elements of information theory*. Wiley-Interscience, 2006.

Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.

Chao Gao, Yu Lu, and Harrison H Zhou. Rate-optimal graphon estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015.

Sara A Geer, Sara van de Geer, and D Williams. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522, 1952.

Yihong Wu. Lecture notes on information-theoretic methods for high-dimensional statistics. *Lecture Notes for ECE598YW (UIUC)*, 16, 2017.