Journal of Machine Learning Research 17 (2016) 1-18 Submitted 1/16; Revised 4/16; Published 9/16

A Note on the Sample Complexity of the Er-SpUD Algorithm by Spielman, Wang and Wright for Exact Recovery of Sparsely Used Dictionaries

Radosław Adamczak R.Adamczak@mimuw.edu.pl
Institute of Mathematics
University of Warsaw
Banacha 2
02-097 Warszawa
Poland

Editor: Sara van de Geer

Abstract

We consider the problem of recovering an invertible $n\times n$ matrix A and a sparse $n\times p$ random matrix X based on the observation of $Y = AX$ (up to a scaling and permutation of columns of A and rows of X). Using only elementary tools from the theory of empirical processes we show that a version of the Er-SpUD algorithm by Spielman, Wang and Wright with high probability recovers A and X exactly, provided that $p \geq Cn\log n$, which is optimal up to the constant C.

Keywords: sparse dictionaries, Er-SpUD algorithm, ℓ_1 minimization, exact recovery, sample complexity

1. Introduction

Recovery of sparsely-used dictionaries has recently attracted considerable attention in connection to applications in machine learning, signal processing or computational neuroscience. In particular, two important fields of applications are *dictionary learning* (Olshausen and Field, 1996; Kreutz-Delgado et al., 2003; Bruckstein et al., 2009; Rubinstein et al., 2010; Yang et al., 2010) and *blind source separation* (Zibulevsky and Pearlmutter, 2001; Georgiev et al., 2005). We do not discuss these applications and refer the Reader to the aforesaid articles for details.

Among many approaches to this problem a particularly successful one has been presented by Spielman, Wang, and Wright (2012a,b), who considered the noiseless-invertible case:

The main problem:

Consider an invertible $n \times n$ matrix A and a random $n \times p$ sparse matrix X. Denote $Y = AX$. The objective is to reconstruct A and X (up to scaling and permutation of columns of A and rows of X) based on the observable data Y.

Spielman, Wang, and Wright (2012a,b) provide an algorithm which with high probability successfully recovers the matrices A and X up to rescaling and permutation of the columns

of A and rows of X, provided that X is a sparse random matrix satisfying the following probabilistic assumptions.

Probabilistic model specification

$$X = [X_{ij}]_{1\leq i\leq n, 1\leq j\leq p},$$

where

$$X_{ij} = \chi_{ij} R_{ij}$$

and

- χ_{ij}, R_{ij} are independent random variables,
- χ_{ij} are Bernoulli distributed: $\mathbb{P}(\chi_{ij} = 1) = 1 - \mathbb{P}(\chi_{ij} = 0) = \theta$,
- R_{ij} are i.i.d., with mean zero and satisfy

$$\mu := \mathbb{E}|R_{ij}| \geq 1/10,$$
$$\forall_{t>0}\ \mathbb{P}(|R_{ij}| \geq t) \leq 2e^{-t^2/2}.$$

Following Spielman, Wang, and Wright (2012a) we will say that matrices satisfying the above assumptions follow the Bernoulli-Subgaussian model with parameter θ.

We remark that the constant $1/10$ above is of no importance and has been chosen following Spielman, Wang, and Wright (2012a) and Luh and Vu (2016).

The approach of Spielman, Wang and Wright consists of two steps. At the first step (given by the Er-SpUD algorithm we describe below) one gathers $p/2$ candidates for the rows of X. The second, greedy step (Greedy algorithm, also described below) selects from the candidates the set of n sparsest vectors, which form a matrix of rank n.

The algorithms work as follows:

ER-SpUD(DC): Exact Recovery of Sparsely-Used Dictionaries using the sum of two columns of Y as constraint vectors.

1. `Randomly pair the columns of` Y `into` $p/2$ `groups` $g_j = \{Ye_{j_1}, Ye_{j_2}\}$.
2. `For` $j = 1, \ldots, p/2$
   `Let` $r_j = Ye_{j_1} + Ye_{j_2}$`, where` $g_j = \{Ye_{j_1}, Ye_{j_2}\}$.
   `Solve` $\min_w \|w^T Y\|_1$ `subject to` $r_j^T w = 1$`, and set` $s_j = w^T Y$.

Above we use the convention that if $r_j = 0$ (which happens with nonzero probability), and as a consequence the minimization problem has no solution, then we skip the corresponding step of the algorithm.

The second stage, described below, is run on the set S of vectors s_i returned at the first stage (for notational simplicity we relabel them if $r_j = 0$ for some j). We use the standard notation that $\|x\|_0$ denotes the number of nonzero coordinates of a vector x.

Greedy: A Greedy Algorithm to Reconstruct X and A.

1. REQUIRE: $S = \{s_1, \ldots, s_T\} \subseteq \mathbb{R}^p$.
2. For $i = 1, \ldots, n$
   REPEAT
   $l \leftarrow \text{argmin}_{s_l \in S} \|s_l\|_0$, breaking ties arbitrarily
   $x_i = s_l$, $S = S \setminus \{s_l\}$
   UNTIL $rank([x_1, \ldots, x_i]) = i$
3. Set $X = [x_1, \ldots, x_n]^T$ and $A = YY^T(XY^T)^{-1}$.

Spielman, Wang, and Wright (2012a) have proved that there exist positive constants C, α, such that if

$$\frac{2}{n} \leq \theta \leq \frac{\alpha}{\sqrt{n}}$$

and $p \geq Cn^2 \log^2 n$, then the ER-SpUD algorithm successfully recovers the matrices A, X with probability at least $1 - \frac{1}{Cp^{10}}$. Note that the equation $Y = A'X'$ still holds if we set $A' = A\Pi\Lambda$ and $X' = \Lambda^{-1}\Pi^T X$ for some permutation matrix Π and a nonsingular diagonal matrix Λ. Therefore, by recovery we mean that nonzero multiples of all the rows of X are among the set $\{s_1, \ldots, s_{p/2}\}$ produced by the ER-SpUD(DC) algorithm. In (Spielman, Wang, and Wright, 2012a) it is also proved that if $\mathbb{P}(R_{ij} = 0) = 0$, then for $p > Cn \log n$, with probability $1 - C'n \exp(-c\theta p)$ for any matrices A', X' such that $Y = A'X'$ and $\max_i \|e_i^T X'\|_0 \leq \max_i \|e_i^T X\|_0$ there exists a permutation matrix Π and a nonsingular diagonal matrix Λ such that $A' = A\Pi\Lambda$, $X' = \Lambda^{-1}\Pi^T X$. In fact, Spielman, Wang, and Wright (2012a) prove that with the above probability any row of X is nonzero and has at most $(10/9)\theta p$ nonzero entries, whereas any linear combination of two or more rows of X has at least $(11/9)\theta p$ nonzero entries.

In particular it follows that the Greedy algorithm will extract from the set $\{s_1, \ldots, s_T\}$ multiples of all n rows of X (note that all s_j's are in the row space of Y and thus also in the row space of X). Since X is with high probability of rank n, one easily shows that one can recover A by the formula used in the 3rd step of the algorithm. We remark that Luh and Vu (2016) obtained the same results concerning sparsity of linear combinations of rows of X without the assumptions about the symmetry of the variables R_{ij}.

Note also that for θ of the order n^{-1}, $p = Cn \log n$ is necessary for uniqueness of the solution in the sense described above, otherwise with significant probability some of the rows of X may be zero, which means that some columns of A do not influence the matrix Y.

In (Spielman, Wang, and Wright, 2012a) it is also proved that if $p > Cn\log n$, $\theta > C'\sqrt{\frac{\log n}{n}}$, then with high probability the ER-SpUD algorithm does not recover any of the rows of X.

Spielman, Wang and Wright have conjectured that their algorithm works with high probability provided that $p > Cn\log n$ (which, as mentioned above is required for well-posedness of the problem).

In this note we will consider a modified version of the algorithm with a slightly different first stage. Namely, instead of using only $p/2$ pairs of columns of Y, we will use all $\binom{p}{2}$ pairs. For fixed p it clearly increases the time complexity of the algorithm (which however remains polynomial), but the advantage of this modification is the possibility of proving that it requires only $p = Cn\log n$ to recover X and A with high probability, which as explained above is optimal. More specifically, we will consider the following algorithm.

Modified ER-SpUD(DC): Exact Recovery of Sparsely-Used Dictionaries using the sum of two columns of Y as constraint vectors.

`For` $i = 1, \dots, p-1$
`For` $j = i+1, \dots, p$
`Let` $r_{ij} = Ye_i + Ye_j$
`Solve` $\min_w \|w^T Y\|_1$ `subject to` $r_{ij}^T w = 1$`, and set` $s_{ij} = w^T Y$`.`

The final step of the recovery algorithm is again a greedy selection of the sparsest vectors among the candidates collected at the first step. As before, under the assumption $\mathbb{P}(R_{ij} = 0) = 0$, the greedy procedure successfully recovers X and A, provided that multiples of all the rows of X are present among the input set S.

The main result of this note is

Theorem 1 *There exist absolute constants $C, \alpha \in (0, \infty)$ such that if*

$$\frac{2}{n} \leq \theta \leq \frac{\alpha}{\sqrt{n}}$$

and X follows the Bernoulli-Subgaussian model with parameter θ, then for $p \geq Cn\log n$, with probability at least $1 - 1/p$ the modified ER-SpUD(DC) algorithm successfully recovers all the rows of X, i.e., multiples of all the rows of X are present among the vectors s_{ij} returned by the algorithm.

Remark on the single column algorithm Spielman, Wang, and Wright (2012a,b) proposed also a version of the Er-SpUD algorithm, which instead of sums of two columns of Y as the vectors r_j in the constraints $r_j^T w = 1$ of the optimization problem, chooses simply consecutive columns of Y. They prove that such a version of the algorithm performs well under the assumption that the random variables X_{ij} are i.i.d. standard Gaussian variables, $2/n \leq \theta \leq \alpha/\sqrt{n\log n}$ and $p > Cn^2\log^2 n$ (α, C are again universal constants). We remark that by using our approach in combination with the original arguments of Spielman, Wang, and Wright (2012a) one can prove that this algorithm works for $p > Cn\log^3 n$. To this end

one needs to prove a counterpart of our Lemma 3 (see below) with the vectors b_{ij} replaced just by the columns of the matrix X and combine it with Lemma 12 of Spielman, Wang, and Wright (2012a) (Lemma 6 below) in exactly the same way as in Section B.3. of (Spielman, Wang, and Wright, 2012a) (with $\gamma \simeq 1/\log n$). The needed counterpart of Lemma 3 can be obtained just by formal changes from the proof we present below. The factor $\log^3 n$ (instead of $\log n$) is related to the use of Lemma 12 and is a consequence of the fact that one takes γ depending on n).

Remarks on recent developments Very recently Sun, Qing, and Wright (2015) proposed an algorithm with polynomial sample complexity, which recovers well conditioned dictionaries under the assumption that the variables R_{ij} are i.i.d. standard Gaussian and $\theta \leq 1/2$, thus allowing for the first time for a linear number of nonzero entries per column of the matrix X. Their novel approach is based on non-convex optimization. The sample complexity of the algorithms in (Sun, Qing, and Wright, 2015) is however higher then for the Er-SpUD algorithm; as mentioned by the Authors, numerical simulations suggest that it is at least $p = \Omega(n^2 \log n)$ even in the case of orthogonal matrix A. Sun, Qing, and Wright (2015) conjecture that algorithms with sample complexity $p = O(n \log n)$ should be possible also for large θ.

As for the complexity of the Er-SpUD algorithm (in its original version), the recent article (Luh and Vu, 2016) contains a claim that it works for $p > Cn \log^4 n$, which differs from the number of samples conjectured by Spielman, Wang, and Wright (2012a) just by a polylogarithmic factor. However, as pointed out very recently (after the submission of the first version of this article) by Błasiok and Nelson (2016), the argument of Luh and Vu (2016) contains certain inaccuracies. Moreover, Błasiok and Nelson have proved that if the variables R_{ij} are Rademachers, than for the original version of the Er-SpUD algorithm to work one needs $p \geq n^{1.99}$, which shows that the result of Luh and Vu (2016) and in fact the original conjecture do not hold. Błasiok and Nelson also propose a modified version of the algorithm (in the same spirit as in this article) and prove that it works with probability $1 - \varepsilon$ for $p > Cn \log(Cn/\varepsilon)$, thus obtaining an independent proof of our main result. We remark that while certain aspects of the analysis are common for (Błasiok and Nelson, 2016) and the present article, the main technical ingredient (i.e., bounding the empirical process involved in the estimates) is approached differently. While Proposition 2 below is based on the contraction principle, Błasiok and Nelson (2016) rely on the generic chaining (majorizing measure) method, see (Talagrand, 2014). Let us also remark that it seems that by combining the inequality for empirical processes obtained by Luh and Vu (2016) with the approach of this paper or of (Błasiok and Nelson, 2016) one can prove a weaker result, namely that the modified version of the algorithm works for $p > Cn \log^4 n$.

2. Proof of Theorem 1

We will follow the general approach proposed by Spielman, Wang, and Wright (2012a). The main new part of the argument is an improved bound on the sample complexity for empirical approximation of first moments of arbitrary marginals of the columns of the matrix X, given in Proposition 2 below. So as not to reproduce technical and lengthy parts of the original proof, we organize this section as follows. First, we present the crucial Proposition 2 and provide a brief discussion of its mathematical content. Next, we present an overview of the

main steps in the proof scheme of Spielman, Wang, and Wright (2012a). For parts of the proof not related to Proposition 2 or to the modification of the algorithm considered here, we only indicate the relevant statements from (Spielman, Wang, and Wright, 2012a), while for parts involving the use of Proposition 2 and for the conclusion of the proof we provide the full argument. Proposition 2 is proved in Section 3.

Below by $e_1, \ldots, e_N$ we will denote the standard basis in $\mathbb{R}^N$ for various choices of N (in particular for $N = n$ and $N = p$). The value of N will be clear from the context and so this should not lead to ambiguity.

By B_1^n we will denote the unit ball in the space ℓ_1^n, i.e., $B_1^n = \{x \in \mathbb{R}^n \colon \|x\|_1 \leq 1\}$, where for $x = (x(1), \ldots, x(n))$, $\|x\|_1 = \sum_{i=1}^n |x(i)|$. The coordinates of a vector x will be denoted by $x(i)$ or if it does not interfere with other notation (e.g., for indexed families of vectors) simply by x_i. Again, the meaning of the notation will be clear from the context. If Y is a random variable and $q > 0$, we denote $\|Y\|_q = (\mathbb{E}|Y|^q)^{1/q}$.

Proposition 2 *Let $U_1, U_2, \ldots, U_p, \chi_1, \ldots, \chi_p$ be independent random vectors in $\mathbb{R}^n$. Assume that for some constant M and all $1 \leq i \leq p$, $1 \leq j \leq n$,*

$$\mathbb{E}e^{|U_i(j)|/M} \leq 2 \tag{1}$$

and

$$\mathbb{P}(\chi_i(j) = 1) = 1 - \mathbb{P}(\chi_i(j) = 0) = \theta.$$

Define the random vectors $Z_1, \ldots, Z_p$ with the equality $Z_i(j) = U_i(j)\chi_i(j)$ for $1 \leq i \leq p$, $1 \leq j \leq n$ and consider the random variable

$$W := \sup_{x \in B_1^n} \Big| \frac{1}{p} \sum_{i=1}^p (|x^T Z_i| - \mathbb{E}|x^T Z_i|) \Big|. \tag{2}$$

Then, for some universal constant C and every $q \geq \max(2, \log n)$,

$$\|W\|_q \leq \frac{C}{p}(\sqrt{p\theta q} + q)M \tag{3}$$

and as a consequence

$$\mathbb{P}\Big(W \geq \frac{Ce}{p}(\sqrt{p\theta q} + q)M\Big) \leq e^{-q}. \tag{4}$$

The above proposition can be considered a quantitative version of the uniform law of large numbers for linear functionals $x^T Z$ indexed by the unit sphere in the space ℓ_1^n. As such it is a classical object of study in the theory of empirical processes. The proof we give uses only Bernstein's inequality, see e.g., (van der Vaart and Wellner, 1996), and Talagrand's contraction principle (Ledoux and Talagrand, 1991), which in a somewhat similar context was applied e.g., by Mendelson (2008); Adamczak et al. (2010).

Let us also remark that in the above proposition we do not require independence between components of the random vectors U_i or χ_i for fixed i, but just independence between the random vectors $U_i, \chi_i, i = 1, \ldots, p$.

2.1 Main Steps of the Proof of Theorem 1

As announced, we will now present an outline of the proof of Theorem 1, indicating which steps differ from the original argument in (Spielman, Wang, and Wright, 2012a).

Step 1. A change of variables.

Recall that r_{ij} are sums of two columns of the matrix Y. At the first step of the proof, instead of looking at the original optimization problem

$$\text{minimize } \|w^T Y\|_1 \text{ subject to } r_{ij}^T w = 1 \tag{5}$$

one performs a change of variables $z = A^T w$, $b_{ij} = A^{-1} r_{ij}$, arriving at the optimization problem

$$\text{minimize } \|z^T X\|_1 \text{ subject to } b_{ij}^T z = 1. \tag{6}$$

Note that one cannot solve (6) since it involves the unknown matrices X and A. The goal of the subsequent steps is to prove that with probability sufficiently separated from zero the solution z_* of (6) is a multiple of one of the basis vectors $e_1, \ldots, e_n$, say $z_* = \lambda e_k$. This means that $w_*^T Y = z_*^T X = \lambda e_k^T X$, i.e., (5) recovers the k-th row of X up to scaling. In combination with a coupon collector phenomenon this will allow to conclude that if p is sufficiently large, then all the rows will be recovered (this is the content of Step 4).

Step 2. If $0 < |(\operatorname{supp} Xe_i) \cup (\operatorname{supp} Xe_j)| < 1/(8\theta)$, then $\operatorname{supp}(z_*) \subseteq (\operatorname{supp} Xe_i) \cup (\operatorname{supp} Xe_j)$.

At this step we prove the following lemma, which is a counterpart of Lemma 11 in (Spielman et al., 2012a). It is weaker in that we do not consider arbitrary vectors b_{ij}, but only sums of two distinct columns of X (which is enough for the application in the proof of Theorem 1). On the other hand it works already for $p > Cn \log n$ and not for $p > Cn^2 \log n$ as the original lemma from (Spielman, Wang, and Wright, 2012a).

Lemma 3 *For $1 \le i < j \le p$, define $b_{ij} = Xe_i + Xe_j$, $I_{ij} = (\operatorname{supp} Xe_i) \cup (\operatorname{supp} Xe_j)$. There exist numerical constants $C, \alpha > 0$ such that if $2/n \le \theta \le \alpha/\sqrt{n}$ and $p > Cn \log n$, then with probability at least $1 - p^{-2}$ the random matrix X has the following property:*

(P1) For every $1 \le i < j \le p$, if $0 < |I_{ij}| \le 1/(8\theta)$ then every solution z_* to the optimization problem (6) satisfies $\operatorname{supp} z_* \subseteq I_{ij}$.

Before we pass to the presentation of auxiliary facts needed in the proof of the above lemma, let us indicate briefly the two main observations behind the lemma, not present in (Spielman, Wang, and Wright, 2012a). The first one is Proposition 2, which allows to prove the technical Lemma 5 below. The second one is the fact that due to independence of the entries of the matrix we do not need to use the union bound over all possible locations of nonzero coefficients of Xe_i and Xe_j, instead we can condition on the disjoint events that $(\operatorname{supp} Xe_i) \cup (\operatorname{supp} Xe_j) = I$ (where I ranges over nonempty subsets of $[n]$ with $|I| \le 1/(8\theta)$), estimate appropriate conditional probabilities and integrate the result over I.

To prove Lemma 3, one first shows a counterpart of Lemma 16 in (Spielman, Wang, and Wright, 2012a).

Lemma 4 *For any $1 \leq j \leq p$, if $Z = (\chi_{1j}R_{1j}, \ldots, \chi_{nj}R_{nj})$, then for all $v \in \mathbb{R}^n$,*

$$\mathbb{E}|v^T Z| \geq \frac{\mu}{8}\sqrt{\frac{\theta}{n}}\|v\|_1.$$

Proof Let $\varepsilon_1, \ldots, \varepsilon_n$ be a sequence of i.i.d. Rademacher variables, independent of $\{\chi_{ij}, R_{ij}\}$. By standard symmetrization inequalities, see e.g., Lemma 6.3. in (Ledoux and Talagrand, 1991),

$$\mathbb{E}|v^T R| = \mathbb{E}\Big|\sum_{i=1}^n v_i\chi_{ij}R_{ij}\Big| \geq \frac{1}{2}\mathbb{E}\Big|\sum_{i=1}^n v_i\varepsilon_i\chi_{ij}R_{ij}\Big|.$$

The random variables $\varepsilon_i R_{ji}$ are symmetric and $\mathbb{E}|\varepsilon_i R_{ij}| = \mu$, so by Lemma 16 from (Spielman, Wang, and Wright, 2012a), the right-hand side above is bounded from below by $\frac{\mu}{8}\sqrt{\frac{\theta}{n}}\|v\|_1$. ■

The next lemma is an improvement of Lemma 17 in (Spielman, Wang, and Wright, 2012a), which is crucial for obtaining Lemma 3.

Lemma 5 *There exists an absolute constant C, such that the following holds for $p > Cn\log n$. Let $J \subseteq \{1, \ldots, p\}$ be a fixed subset of size $|J| \leq \frac{p}{4}$. Let X_J be the submatrix of X, obtained by a restriction of X to the columns indexed by J. With probability at least $1 - p^{-8}$, for any $v \in \mathbb{R}^n$,*

$$\|v^T X\|_1 - 2\|v^T X_J\|_1 > \frac{p\mu}{32}\sqrt{\frac{\theta}{n}}\|v\|_1.$$

Proof Note first that by increasing the set J, we increase $\|v^T X_J\|_1$, so without loss of generality we can assume that $|J| = \lfloor p/4 \rfloor$. Apply Proposition 2 with the vectors $U_j = (R_{1j}, \ldots, R_{nj})$ and $\chi_j = (\chi_{1j}, \ldots, \chi_{nj})$ and $q = 8\log p$. Note that our integrability assumptions on R_{ij} imply (1) with M being a universal constant. Therefore, for some absolute constant C and $p \geq Cn\log n$, with probability at least $1 - p^{-8}$ we have

$$\sup_{v\in B_1^n}\Big|\|v^T X\|_1 - \mathbb{E}\|v^T X\|_1\Big| \leq C(\sqrt{p\theta\log p} + \log p) \leq 2C\sqrt{p\theta\log p},$$
$$\sup_{v\in B_1^n}\Big|\|v^T X_J\|_1 - \mathbb{E}\|v^T X_J\|_1\Big| \leq 2C\sqrt{p\theta\log p},$$

where we used that for C sufficiently large, $p/\log p \geq n \geq 1/\theta$.

Thus, by homogeneity, with probability at least $1 - p^{-8}$, for all $v \in \mathbb{R}^n$,

$$\Big|\|v^T X\|_1 - \mathbb{E}\|v^T X\|_1\Big| \leq 2C\sqrt{\theta p\log p}\|v\|_1,$$
$$\Big|\|v^T X_J\|_1 - \mathbb{E}\|v^T X_J\|_1\Big| \leq 2C\sqrt{\theta p\log p}\|v\|_1.$$

In particular this means that (using the notation of Proposition 2)

$$\|v^T X\|_1 \geq \mathbb{E}\|v^T X\|_1 - 2C\sqrt{\theta p\log p}\|v\|_1 = p\mathbb{E}|v^T Z_1| - 2C\sqrt{\theta p\log p}\|v\|_1,$$
$$2\|v^T X_J\|_1 \leq 2\mathbb{E}\|v^T X_J\|_1 + 4C\sqrt{\theta p\log p}\|v\|_1 = 2|J|\mathbb{E}|v^T Z_1| + 4C\sqrt{\theta p\log p}\|v\|_1,$$

and so

$$\|v^T X\|_1 - 2\|v^T X_J\|_1 \geq (p - 2|J|)\mathbb{E}|v^T Z_1| - 6C\sqrt{\theta p \log p}\|v\|_1.$$

Now, by Lemma 4 and the assumed bound on the cardinality of J, we get

$$\|v^T X\|_1 - 2\|v^T X_J\|_1 \geq \Big(\frac{p\mu}{16}\sqrt{\frac{\theta}{n}} - 6C\sqrt{\theta p \log p}\Big)\|v\|_1 > \frac{p\mu}{32}\sqrt{\frac{\theta}{n}}\|v\|_1$$

for $p > C' n \log n$, where C' is another absolute constant. ■

We are now in position to prove Lemma 3.

Proof of Lemma 3 We will show that for each $1 \leq i < j \leq p$ the probability that $0 < |I_{ij}| \leq 1/(8\theta)$ and there exists a solution to (6) not supported on I_{ij} is bounded from above by $1/p^4$. This will imply the lemma, since by the union bound over all $i < j$,

$$\begin{aligned} &\mathbb{P}(\text{Property P1 does not hold}) \qquad (7)\\ &\leq \sum_{1\leq i<j\leq p} \mathbb{P}(0 < |I_{ij}| \leq 1/(8\theta) \text{ \& there exists a solution } z_* \text{ to (6) not supported on } I_{ij}). \end{aligned}$$

Let us thus fix i, j and let

$$S = \{l \in [p] \colon \exists_{k\in I_{ij}} X_{kl} \neq 0\}.$$

Moreover, for any set $I \subseteq [n]$, define the event

$$\mathcal{A}_I = \{I_{ij} = I\}.$$

By independence of the random variables R_{ij}, χ_{ij}, for each $k \notin \{i, j\}$, if $0 < |I| \leq 1/(8\theta)$, then

$$\mathbb{P}(k \in S|\mathcal{A}_I) \leq 1 - (1-\theta)^{|I|} \leq 1 - e^{-2\theta|I|} \leq 1 - e^{-\frac{1}{4}} < \frac{1}{4},$$

where the second inequality holds if α is sufficiently small.

Thus, by independence of columns of X and Hoeffding's inequality, if $0 < |I| \leq 1/(8\theta)$, then

$$\mathbb{P}\Big(|S \setminus \{i, j\}| \leq \frac{p-2}{4}\Big|\mathcal{A}_I\Big) \geq 1 - 2e^{-cp} \qquad (8)$$

for some universal constant $c > 0$.

Let z_* be any solution of (6) and denote by z_0 its orthogonal projection on $\mathbb{R}^{I_{ij}} = \{x \in \mathbb{R}^n \colon x_k = 0 \text{ for } k \notin I_{ij}\}$. Set also $z_1 = z_* - z_0$ and let X_S, X_{S^c} be the matrices obtained from X by selecting the columns labeled by S and $S^c = [p] \setminus S$ respectively. By the triangle inequality, and the fact that $z_0^T X_{S^c} = 0$, we get

$$\begin{aligned} \|z_*^T X\|_1 &= \|(z_0^T + z_1^T)X_S\|_1 + \|(z_0^T + z_1^T)X_{S^c}\|_1\\ &\geq \|z_0^T X_S\|_1 - \|z_1^T X_S\|_1 + \|z_1^T X\|_1 - \|z_1^T X_S\|_1\\ &= \|z_0^T X\|_1 + (\|z_1^T X\|_1 - 2\|z_1^T X_S\|_1). \qquad (9) \end{aligned}$$

For $J \subseteq [p] \setminus \{i,j\}$ define the events

$$\mathcal{S}_J = \{S \setminus \{i,j\} = J\}.$$

For the moment let us restrict our attention to the event $\mathcal{A}_I \cap \mathcal{S}_J$, for a fixed (but arbitrary) $I \subseteq [n]$, satisfying $0 < |I| \leq 1/(8\theta)$ and $J \subseteq [p] \setminus \{i,j\}$, satisfying $|J| \leq (p-2)/4$.

Denote by X' the $|I^c| \times (p-2)$ matrix obtained by restricting X to the rows from I^c and columns from $[p] \setminus \{i,j\}$. If, slightly abusing the notation, we identify z_1 with a vector from $\mathbb{R}^{|I^c|}$, on the event $\mathcal{A}_I \cap \mathcal{S}_J$ we have

$$\|z_1^T X\|_1 - 2\|z_1^T X_S\|_1 = \|z_1^T X'\|_1 - 2\|z_1^T X'_{S\setminus\{i,j\}}\|_1 = \|z_1^T X'\|_1 - 2\|z_1^T X'_J\|_1, \tag{10}$$

where in the we first equality we used used the fact that $z_1^T X e_i = z_1^T X e_j = 0$ and the second one follows from the definition of the event $\mathcal{S}_J$.

Due to independence and identical distribution of the entries of X, conditionally on the event $\mathcal{A}_I \cap \mathcal{S}_J$ the matrix X' still follows the Bernoulli-Subgaussian model with parameter θ. This matrix is of size $|I^c| \times (p-2)$, therefore if the constant C' is large enough and $p > C' n \log n$, it satisfies the assumptions of Lemma 5 (with $p-2$ instead of p and $|I^c|$ instead of n). Since $|J| \leq (p-2)/4$, a conditional application of Lemma 5 gives

$$\begin{aligned}\mathbb{P}\Big(\text{for all } v \in \mathbb{R}^{|I^c|}: \quad \|v^T X'\|_1 - 2\|v^T X'_J\|_1 \geq \frac{(p-2)\mu}{32}\sqrt{\frac{\theta}{|I^c|}}\|v\|_1 \Big| \mathcal{A}_I \cap \mathcal{S}_J\Big) \\ \geq 1 - (p-2)^{-8} \geq 1 - 2p^{-8},\end{aligned} \tag{11}$$

where the last inequality holds for $p > C'$ and C' sufficiently large.

Note that by the definition of z_0, we have $b_{ij}^T z_0 = b_{ij}^T z_* = 1$, therefore z_0 is a feasible candidate for the solution of the optimization problem (6). Thus, by (9) and (10), we have $\|z_1^T X'\|_1 - 2\|z_1^T X'_J\|_1 \leq 0$ and as a consequence, if $z_1 \neq 0$ then the event of inequality (11) does not hold. Thus, for $0 < |I| \leq 1/(8\theta)$ and $|J| \leq (p-2)/4$, we get

$$\mathbb{P}(\text{for some solution } z_* \text{ to (6)}, z_1 \neq 0 | \mathcal{A}_I \cap \mathcal{S}_J) \leq 2p^{-8}. \tag{12}$$

We are now ready to finish the proof. To shorten the notation, let us denote

$$\mathcal{B} = \{\text{for some solution } z_* \text{ to (6)}, z_1 \neq 0 \text{ and } 0 < |I_{ij}| \leq 1/(8\theta)\}.$$

By (8) we get

$$\begin{aligned}\mathbb{P}\Big(\mathcal{B} \cap \{|S \setminus \{i,j\}| > (p-2)/4\}\Big) &= \sum_{I \subseteq [n]:\ 0<|I|\leq 1/(8\theta)} \mathbb{P}(\mathcal{B} \cap \mathcal{A}_I \cap \{|S'| > (p-2)/4\}) \\ &\leq \sum_{I \subseteq [n]:\ 0<|I|\leq 1/(8\theta)} \mathbb{P}(\mathcal{A}_I \cap \{|S'| > (p-2)/4\}) \\ &= \sum_{I \subseteq [n]:\ 0<|I|\leq 1/(8\theta)} \mathbb{P}(|S'| > (p-2)/4 | \mathcal{A}_I)\mathbb{P}(\mathcal{A}_I) \\ &\leq 2e^{-cp} \sum_{I \subseteq [n]:\ 0<|I|\leq 1/(8\theta)} \mathbb{P}(\mathcal{A}_I) \leq 2e^{-cp},\end{aligned}$$

where the second to last inequality follows from (8) and the last one from the pairwise disjointness of the events $\mathcal{A}_I$.

Similarly,

$$\begin{aligned}\mathbb{P}(\mathcal{B}\cap\{|S\setminus\{i,j\}|\le(p-2)/4\}) &= \sum_{\substack{I\subseteq[n]:\\ 0<|I|\le 1/(8\theta)}}\ \sum_{\substack{J\subseteq[p]\setminus\{i,j\}:\\ |J|\le(p-2)/4}} \mathbb{P}(\mathcal{B}\cap\mathcal{A}_I\cap\mathcal{S}_J)\\ &\le \sum_{\substack{I\subseteq[n]:\\ 0<|I|\le 1/(8\theta)}}\ \sum_{\substack{J\subseteq[p]\setminus\{i,j\}:\\ |J|\le(p-2)/4}} \mathbb{P}(\mathcal{B}|\mathcal{A}_I\cap\mathcal{S}_J)\mathbb{P}(\mathcal{A}_I\cap\mathcal{S}_J)\\ &\le 2p^{-8}\sum_{\substack{I\subseteq[n]:\\ 0<|I|\le 1/(8\theta)}}\ \sum_{\substack{J\subseteq[p]\setminus\{i,j\}:\\ |J|\le(p-2)/4}} \mathbb{P}(\mathcal{A}_I\cap\mathcal{S}_J)\le 2p^{-8},\end{aligned}$$

where we used (12) and disjointness of the events $\mathcal{A}_I\cap\mathcal{S}_J$. Combining the two last inequalities, we get

$$\mathbb{P}(\mathcal{B})\le 2e^{-cp}+2p^{-8}\le p^{-4}$$

for $p>Cn\log n$ with a sufficiently large absolute constant C. By (7) this ends the proof of the lemma. ■

Step 3. If $(\operatorname{supp} Xe_i)\cup(\operatorname{supp} Xe_j)$ is small, then $z_*=\lambda e_k$ where $k=\operatorname{argmax}_{1\le l\le n}|b_{ij}(l)|$.

At this step one proves the following lemma (Lemma 12 in Spielman, Wang, and Wright 2012a). Since no changes with respect to the original argument are required (we do not use Proposition 2 here), we do not reproduce the proof and refer the Reader to (Spielman, Wang, and Wright, 2012a) for details. We remark that although the lemma is formulated therein for symmetric variables, the symmetry assumption is not used in its proof.

Below, by $|b|_1^\downarrow\ge|b|_2^\downarrow\ge\ldots\ge|b|_n^\downarrow$, we denote the nonincreasing rearrangement of the sequence $|b_1|,\ldots,|b_n|$, while for $J\subseteq[n]$, X^J denotes the matrix obtained from X by selecting the rows indexed by the set J.

Lemma 6 *There exist two positive constants c_1,c_2 such that the following holds. For any $\gamma>0$ and $s\in\mathbb{Z}_+$, such that $\theta s<\gamma/8$ and p such that*

$$p\ge\max\Big\{\frac{c_1 s\log n}{\theta\gamma^2},n\Big\},\quad \textit{and}\quad \frac{p}{\log p}\ge\frac{c_2}{\theta\gamma^2},$$

with probability at least $1-4p^{-10}$, the random matrix X has the following property:

(P2) For every $J\subseteq[n]$ with $|J|=s$ and every $b\in\mathbb{R}^s$, satisfying $\frac{|b|_2^\downarrow}{|b|_1^\downarrow}\le 1-\gamma$, the solution to the restricted problem

$$\textit{minimize } \|z^TX^J\|_1 \textit{ subject to } b^Tz=1, \tag{13}$$

is unique, 1-sparse, and is supported on the index of the largest entry of b.

Step 4. Conclusion of the proof.

Set $s = 12\theta n + 1$. Our first goal is to prove that with probability at least $1 - 1/p^2$, for all $k \in [n]$, there exist $i, j \in [p]$, $i \neq j$ such that the vector $b = Xe_i + Xe_j$ satisfies the assumptions of Lemma 6, $|b|_1^{\downarrow} = |b_k|$ and $I_{ij} := (\operatorname{supp} Xe_i) \cup (\operatorname{supp} Xe_j)$ satisfies $0 < |I_{ij}| \leq 1/(8\theta)$, which will allow us to take advantage of Lemma 3. This will already imply that the solution to the problem (6) for such i, j produces a multiple of the k-th row of X.

Note that we have

$$\mathbb{E}R_{ij}^2 \leq 4\int_0^\infty te^{-t^2/2}dt = 4.$$

Since $\mathbb{E}|R_{ij}| = \mu \geq \frac{1}{10}$, by the Paley-Zygmund inequality, see e.g., Corollary 3.3.2. in (de la Peña and Giné, 1999), we have

$$\mathbb{P}(|R_{ij}| \geq \frac{1}{20}) \geq \frac{3}{4}\frac{(\mathbb{E}|R_{ij}|)^2}{\mathbb{E}R_{ij}^2} \geq c_0$$

for some universal constant $c_0 > 0$. In particular $\mathbb{P}(|R_{ij}| = 0) < 1 - \frac{c_0}{2}$. Let q be any $(1 - c_0/(2s))$-quantile of $|R_{ij}|$, i.e., $\mathbb{P}(|R_{ij}| \leq q) \geq (1 - c_0/(2s))$ and $\mathbb{P}(|R_{ij}| \geq q) \geq c_0/(2s)$. In particular, since $s \geq 1$, we get $q > 0$. We have $\mathbb{P}(R_{ij} \geq q) \geq c_0/(4s)$ or $\mathbb{P}(R_{ij} \leq -q) \geq c_0/(4s)$. Let us assume that $\mathbb{P}(R_{ij} \geq q) \geq c_0/(4s)$, the other case is analogous.

Before we proceed with the formal proof, which due to many events under consideration may appear technical, let us provide its informal description. Let us focus on a single value of k (at the end of the argument we will take a union bound over all $k \leq n$). We will first prove that among the first $p/2$ columns of the matrix X there is one (say Xe_i) which has few nonzero entries, the k-th entry exceeds the quantile q and all the other entries are smaller then q in the absolute value. This corresponds to the events $\mathcal{E}_{ki}$ and $\mathcal{A}_k$ considered below. Once we establish that this holds with high probability (equation (14)), we will fix a single column with this property (say with the smallest index) and will prove that conditionally on this event among the $p/2$ last columns of X we can find a column (say Xe_j) with the same properties and such that the only entry which is nonzero in both Xe_i and Xe_j is the k-th one (which corresponds to the event $\mathcal{B}_k$ below and is the content of equation (17)). This will imply that

- the set I_{ij} satisfies the premises of the implication of Lemma 3 (it is nonempty and not too large),
- the k-th entry of $b_{ij} = Xe_i + Xe_j$ exceeds $2q$ while all the other entries are smaller than q in absolute value, which allows to use Lemma 6 with $\gamma = 1/2$.

Combining the two lemmas will allow us to conclude that the solution to (6) produces a nonzero multiple of e_k, i.e., the solution to (5) produces a nonzero multiple of the k-th row of X.

Establishing the aforesaid properties is not difficult and relies just on the independence of entries. In essence it can be reduced to saying that in a sequence of Bernoulli trials with probability of success equal to ρ, it is highly unlikely that we will have to wait much longer than $1/\rho$ for the first success. Specifically, if $\rho > c/n$, then under our assumptions on p, the probability that no success occurs in $p/2$ steps is smaller than $1/p^4$ (see e.g.,

equation (16) below). In the proof the trials correspond to columns of X and success to the conjunction of the properties stated above. Both parts of the proof rely on estimating the probability of success from below (in the second part it is the conditional probability, since the event in question depends on the first part). The main reason behind technical (notational) difficulties is that one needs to explore independence of the variables χ_{ij} and R_{ij} in the right order to be able to take advantage of the already established bounds in consecutive steps.

Define thus the event $\mathcal{E}_{ki}$ as

$$\mathcal{E}_{ki} = \Big\{\chi_{ki} = 1,\ |\{r \in [n] \setminus \{k\} \colon \chi_{ri} = 1\}| \le (s-1)/2,\ R_{ki} \ge q,\ \forall_{r \ne k}\, \chi_{ri} = 1 \implies |R_{ri}| \le q\Big\}$$

(see the description above for the motivation of this and subsequent definitions).

We will assume that $p \ge 2Cn \log n$ for some numerical constant C to be fixed later on. For $k \in [n]$, consider the events

$$\mathcal{A}_k = \bigcup_{1 \le i \le \lfloor p/2 \rfloor} \mathcal{E}_{ki}$$

and

$$\mathcal{B}_k = \bigcup_{1 \le i \le \lfloor p/2 \rfloor} \bigcup_{\lfloor p/2 \rfloor < j \le p} \Big(\mathcal{E}_{ki} \cap \mathcal{E}_{kj} \cap \Big\{\{l \in [n] \colon \chi_{li} = \chi_{lj} = 1\} = \{k\}\Big\}\Big).$$

We will first show that for all $k \in [n]$,

$$\mathbb{P}(\mathcal{A}_k) \ge 1 - \frac{1}{p^4}, \tag{14}$$

which we will use to prove that

$$\mathbb{P}(\mathcal{B}_k) \ge 1 - \frac{1}{p^3}. \tag{15}$$

Let us start with the proof of (14). Set $\mathcal{B}_{ki} = \{|\{r \in [n] \setminus \{k\} \colon \chi_{rk} = 1\}| \le (s-1)/2\}$. By independence we have

$$\begin{aligned}\mathbb{P}(\mathcal{E}_{ki}) &= \mathbb{P}(\chi_{ki} = 1)\mathbb{P}(R_{ki} \ge q)\mathbb{P}(\mathcal{B}_{ki})\mathbb{P}(\forall_{r \ne k}\, \chi_{ri} = 1 \implies |R_{ri}| \le q | \mathcal{B}_{ki}) \\ &\ge \theta \frac{c_0}{4s}\Big(1 - \frac{2\theta(n-1)}{s-1}\Big)\Big(1 - \frac{c_0}{2s}\Big)^{(s-1)/2},\end{aligned}$$

where to estimate $\mathbb{P}(B_{ki})$ we used Markov's inequality. The last factor comes from the definition of q as the $(1 - c_0/(2s))$-quantile of R_{ij}. The right hand side above is bounded from below by c_1/n for some universal constant c_1. Therefore if the constant C is large enough, we obtain

$$\mathbb{P}\Big(\bigcap_{1 \le i \le \lfloor p/2 \rfloor} \mathcal{E}_{ki}^c\Big) \le \Big(1 - \frac{c_1}{n}\Big)^{\lfloor p/2 \rfloor} \le \exp(-c_1 p/(4n)) \le \exp(-4 \log p) = \frac{1}{p^4}, \tag{16}$$

where we used the inequality $p/\log p \ge 16 c_1^{-1} n$ for $p \ge Cn \log n$. We have thus established (14).

Let us now pass to (15). Denote by $\mathcal{F}_1$ the σ-field generated by $\chi_{ki}, R_{ki}, k \in [n], 1 \leq i \leq \lfloor p/2 \rfloor$. Note that $\mathcal{A}_k \in \mathcal{F}_1$.

For $\omega \in \mathcal{A}_k$ define $i_{\min}(\omega) = \min\{1 \leq i \leq \lfloor p/2 \rfloor \colon \omega \in \mathcal{E}_{ki}\}$. Note that on $\mathcal{A}_k$,

$$\mathbb{P}(\mathcal{B}_k|\mathcal{F}_1) \geq \mathbb{P}\Big(\bigcup_{\lfloor p/2 \rfloor < j \leq p} \Big(\mathcal{E}_{kj} \cap \Big\{\{l \in [n] \colon \chi_{li_{\min}} = \chi_{lj} = 1\} = \{k\}\Big\}\Big)\Big|\mathcal{F}_1\Big)$$

Define

$$\mathcal{C}_{kj} = \{|\{r \in [n] \setminus \{k\} \colon \chi_{rj} = 1\}| \leq (s-1)/2\} \cap \Big\{\{l \in [n] \colon \chi_{li_{\min}} = \chi_{lj} = 1\} = \{k\}\Big\}.$$

Similarly as in the argument leading to (14), for fixed j, using the independence of the variables χ_{lm}, R_{lm} and properties of the conditional probability, we obtain on the event $\mathcal{A}_k$,

$$\begin{aligned}
&\mathbb{P}\Big(\mathcal{E}_{kj} \cap \Big\{\{l \in [n] \colon \chi_{li_{\min}} = \chi_{lj} = 1\} = \{k\}\Big\}\Big|\mathcal{F}_1\Big) \\
&= \mathbb{P}(R_{kj} \geq q)\mathbb{E}\Big(\mathbf{1}_{\mathcal{C}_{kj}}\mathbb{P}(\forall_{r \neq k}\, \chi_{rj} = 1 \implies |R_{rj}| \leq q|\mathcal{C}_{kj}, \mathcal{F}_1)|\mathcal{F}_1\Big) \\
&\geq \mathbb{P}(R_{kj} \geq q)\mathbb{E}\Big(\mathbf{1}_{\mathcal{C}_{kj}}\Big(1 - \frac{c_0}{2s}\Big)^{\frac{s-1}{2}}\Big|\mathcal{F}_1\Big) \\
&= \mathbb{P}(R_{kj} \geq q)\Big(1 - \frac{c_0}{2s}\Big)^{\frac{s-1}{2}}\mathbb{P}(\mathcal{C}_{kj}|\mathcal{F}_1) \\
&\geq \frac{c_0}{4s}\Big(1 - \frac{c_0}{2s}\Big)^{\frac{s-1}{2}} \times \\
&\Big(\mathbb{P}\Big(\{l \in [n] \colon \chi_{li_{\min}} = \chi_{lj} = 1\} = \{k\}\Big|\mathcal{F}_1\Big) - \mathbb{P}\Big(\chi_{kj} = 1,\ |\{r \in [n] \setminus \{k\} \colon \chi_{rj} = 1\}| > \frac{s-1}{2}\Big|\mathcal{F}_1\Big)\Big) \\
&\geq \frac{c_0}{4s}\Big(1 - \frac{c_0}{4s}\Big)^{\frac{s-1}{2}}\Big(\theta(1-\theta)^{(s-1)/2} - \theta\frac{2\theta(n-1)}{s-1}\Big),
\end{aligned}$$

where in the last line we again used Markov's inequality.

Now recall that $\theta \leq \frac{\alpha}{\sqrt{n}}$ for some universal constant α. If α is small enough then $1 - \theta \geq e^{-2\theta}$ and

$$(1-\theta)^{(s-1)/2} \geq e^{-\theta(s-1)} = e^{-12\theta^2 n} \geq e^{-12\alpha^2} \geq \frac{1}{3}.$$

Since $\frac{2\theta(n-1)}{s-1} \leq \frac{1}{6}$, this implies that

$$\mathbb{P}\Big(\mathcal{E}_{kj} \cap \Big\{\{l \in [n] \colon \chi_{lj_{\min}} = \chi_{lj} = 1\} = \{k\}\Big\}\Big|\mathcal{F}_1\Big) \geq \frac{c_2}{n}$$

for some positive universal constant c_2. Since the events $\mathcal{E}_{kj} \cap \Big\{\{l \in [n] \colon \chi_{lj_{\min}} = \chi_{lk} = 1\} = \{k\}\Big\}$, $\lfloor p/2 \rfloor < k \leq p$ are conditionally independent, given $\mathcal{F}_1$, we obtain that on $\mathcal{A}_k$,

$$\mathbb{P}(\mathcal{B}_k^c|\mathcal{F}_1) \leq \Big(1 - \frac{c_2}{n}\Big)^{\lfloor p/2 \rfloor} \leq \frac{1}{p^4}, \tag{17}$$

provided C is a sufficiently large universal constant. Now, using (14), we get

$$\mathbb{P}(\mathcal{B}_k) \geq \mathbb{E}\mathbf{1}_{\mathcal{A}_k}\mathbb{P}(\mathcal{B}_k|\mathcal{F}_1) \geq \mathbb{P}(\mathcal{A}_k)\Big(1-\frac{1}{p^4}\Big) \geq \Big(1-\frac{1}{p^4}\Big)^2 \geq 1-\frac{1}{p^3},$$

proving (15).

Taking the union bound over $k \in [n]$, we get

$$\mathbb{P}\Big(\bigcap_{1\leq k\leq n}\mathcal{B}_k\Big) \geq 1-\frac{1}{p^2}.$$

Set $\gamma = 1/2$ and observe that if C is large enough and α small enough, then the assumptions of Lemma 3 and Lemma 6 are satisfied. In particular $s = 12\theta n + 1 \leq \frac{\gamma}{8\theta} \leq \frac{1}{8\theta}$. Recall the properties P1 and P2 considered in the said lemmas. Consider the event $\mathcal{A} = \bigcap_{1\leq k\leq n}\mathcal{B}_k \cap \{\text{properties P1 and P2 hold}\}$ and note that $\mathbb{P}(\mathcal{A}) \geq 1-\frac{1}{p}$. On the event $\mathcal{A}$, for every k, there exist $1 \leq i < j \leq p$, such that

- $1 \leq |I_{ij}| \leq s \leq \frac{\gamma}{8\theta} \leq \frac{1}{8\theta}$,
- the largest entry of b_{ij} (in absolute value) equals $b_{ij}(k) \geq 2q > 0$ whereas the remaining entries do not exceed q,

In particular, by property P1 we obtain that any solution z_* to the problem (6) satisfies $\operatorname{supp} z_* \subseteq I_{ij}$. Therefore for some (any) $J \supseteq I_{ij}$ with $|J| = \lfloor s \rfloor$, we obtain (identifying vectors supported on J with their restrictions to J), that z_* is in fact a solution to the restricted problem (13) with $b = b_{ij}$, which by property P2 implies that $z_* = \lambda e_k$ for some $\lambda \neq 0$.

According to the discussion at the beginning of Step 1, this means that the solution w_* to (5) satisfies $w_*^T Y = \lambda e_k^T X$, i.e., the algorithm, when analyzing the vector b_{ij}, will add a multiple of the k-th row of X to the collection S.

This ends the proof of Theorem 1. ■

3. Proof of Proposition 2

The first tool we will need is the classical Bernstein's inequality, see e.g., Lemma 2.2.11 in (van der Vaart and Wellner, 1996).

Lemma 7 (Bernstein's inequality) *Let $Y_1, \ldots, Y_p$ be independent mean zero random variables such that for some constants M, v and every integer $k \geq 2$, $\mathbb{E}|Y_i|^k \leq k! M^{k-2} v/2$. Then, for every $t > 0$,*

$$\mathbb{P}\Big(\Big|\sum_{i=1}^p Y_i\Big| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{2(pv+Mt)}\Big).$$

As a consequence, for every $q \geq 2$,

$$\Big\|\sum_{i=1}^p Y_i\Big\|_q \leq C(\sqrt{qpv} + qM), \tag{18}$$

where C is a universal constant.

Another (also quite standard) tool we will rely on is the contraction principle for empirical processes due to Talagrand, see Theorem 4.12. in (Ledoux and Talagrand, 1991).

Lemma 8 (Talagrand's contraction principle) *Let $F\colon \mathbb{R}_+ \to \mathbb{R}_+$ be convex and increasing. Let further $\varphi\colon \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz function such that $\varphi(0) = 0$. For every bounded subset T of $\mathbb{R}^n$, if $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. Rademacher variables, then*

$$\mathbb{E}F\Big(\sup_{t\in T}\frac{1}{2}\Big|\sum_{i=1}^n \varphi(t_i)\varepsilon_i\Big|\Big) \leq \mathbb{E}F\Big(\sup_{t\in T}\Big|\sum_{i=1}^n t_i\varepsilon_i\Big|\Big)$$

We can now present the proof of Proposition 2.

Proof of Proposition 2 Let $\varepsilon_1, \ldots, \varepsilon_p$ be i.i.d. Rademacher variables, independent of the sequences (U_i), (χ_i). By the symmetrization inequality, see e.g., (Ledoux and Talagrand, 1991, Lemma 6.3.) or (van der Vaart and Wellner, 1996, Lemma 2.31), we have

$$\mathbb{E}W^q \leq 2^q\mathbb{E}\sup_{x\in B_1^n}\Big|\frac{1}{p}\sum_{i=1}^p \varepsilon_i|x^TZ_i|\Big|^q.$$

Now, since the function $t \mapsto |t|$ is a contraction, an application of Lemma 8 with $F(x) = |x|^q$, conditionally on Z_i, gives

$$\begin{aligned}\mathbb{E}W^q &\leq 2^{2q}\mathbb{E}\sup_{x\in B_1^n}\Big|\frac{1}{p}\sum_{i=1}^p \varepsilon_i x^TZ_i\Big|^q = \frac{2^{2q}}{p^q}\mathbb{E}\sup_{x\in B_1^n}\Big|x^T\sum_{i=1}^p \varepsilon_i Z_i\Big|^q\\ &= \frac{2^{2q}}{p^q}\mathbb{E}\Big\|\sum_{i=1}^p \varepsilon_i Z_i\Big\|_\infty^q = \frac{2^{2q}}{p^q}\mathbb{E}\max_{1\leq j\leq n}\Big|\sum_{i=1}^p \varepsilon_i Z_i(j)\Big|^q\\ &\leq \frac{2^{2q}}{p^q}\sum_{j=1}^n \mathbb{E}\Big|\sum_{i=1}^p \varepsilon_i Z_i(j)\Big|^q. \qquad (19)\end{aligned}$$

Now, for every i, j and every integer $k \geq 2$ we have

$$\mathbb{E}|Z_i(j)|^k = \theta\mathbb{E}|U_i(j)|^k \leq \theta M^k k!\mathbb{E}e^{|U_i(j)|/M} \leq 2k!\theta M^k = k!vM^{k-2}/2$$

with $v = 4\theta M^2$. Thus by the moment version (18) of Bernstein's inequality for some universal constant C we get

$$\mathbb{E}\Big|\sum_{i=1}^p \varepsilon_i X_i(j)\Big|^q \leq C^q\Big(\sqrt{qp\theta}M + qM\Big)^q,$$

which, when combined with (19), yields for $q \geq \log n$,

$$\|W\|_q \leq \frac{4Ce}{p}(\sqrt{p\theta q} + q)M.$$

The first part of the proposition follows by adjusting the constant C. The tail bound is a direct consequence of the Chebyshev inequality for the q-th moment. ■

References

R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *J. Amer. Math. Soc.*, 23(2):535–561, 2010.

J. Błasiok and J. Nelson. An improved analysis of the ER-SpUD dictionary learning algorithm. *43rd International Colloquium on Automata, Languages and Programming*, 2016.

A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Rev.*, 51(1):34–81, 2009.

V. H. de la Peña and E. Giné. *Decoupling*. Probability and its Applications. Springer-Verlag, New York, 1999.

P. Georgiev, F. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks*, 16(4), 2005.

K. Kreutz-Delgado, J. Murray, B. Rao, K. Engan, T. Lee, and T. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15(20):349–396, 2003.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3)*. Springer-Verlag, Berlin, 1991.

K. Luh and V. Vu. Dictionary learning with few samples and matrix concentration. *IEEE Transactions on Information Theory*, 62(3):1516–1527, 2016.

S. Mendelson. On weakly bounded empirical processes. *Math. Ann.*, 340(2):293–314, 2008.

B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6538):607–609, 1996.

R. Rubinstein, A. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

D. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries (long version). Preprint, 2012a. URL `http://www.columbia.edu/~jw2966/papers/SWW12-pp.pdf`.

D. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 23, 2012b. 25th Annual Conference on Learning Theory (COLT).

J. Sun, Q. Qing, and J. Wright. Complete dictionary recovery over the sphere II: recovery by riemannian trust-region method. Preprint, 2015. URL `http://arxiv.org/abs/1511.04777`.

M. Talagrand. *Upper and Lower Bounds for Stochastic Processes*. Springer, Heidelberg, 2014.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE Trans. Image Process.*, 19(11):2861–2873, 2010.

M. Zibulevsky and B. Pearlmutter. Blind source separation by sparse decomposition. *Neural Computation*, 13(4), 2001.