2015 IEEE 56th Annual Symposium on Foundations of Computer Science

Random matrices: l_1 concentration and dictionary learning with few samples

Kyle Luh and Van Vu
Department of Mathematics, Yale University
New Haven, CT
Email: kyle.luh@yale.edu, van.vu@yale.edu

Abstract

Let X be a sparse random matrix of size $n \times p$ ($p \gg n$). We prove that if $p \geq Cn \log^4 n$, then with probability $1 - o(1)$, $\|X^T v\|_1$ is close to its expectation for all vectors $v \in \mathbb{R}^n$ (simultaneously). The bound on p is sharp up to the polylogarithmic factor.

The study of this problem is directly motivated by an application. Let A be an $n \times n$ matrix, X be an $n \times p$ matrix and $Y = AX$. A challenging and important problem in data analysis, motivated by dictionary learning and other practical problems, is to recover both A and X, given Y. Under normal circumstances, it is clear that this problem is underdetermined. However, in the case when X is sparse and random, Spielman, Wang and Wright showed that one can recover both A and X efficiently from Y with high probability, given that p (the number of samples) is sufficiently large. Their method works for $p \geq Cn^2 \log^2 n$ and they conjectured that $p \geq Cn \log n$ suffices. The bound $n \log n$ is sharp for an obvious information theoretical reason. The matrix concentration result verifies the Spielman et. al. conjecture up to a $\log^3 n$ factor.

Our proof of the concentration result is based on two ideas. The first is an economical way to apply the union bound. The second is a refined version of Bernstein's concentration inequality for a sum of independent variables. Both have nothing to do with random matrices and are applicable in general settings.

Keywords

Dictionary learning, matrix concentration

I. Introduction

In this paper, we consider random Bernoulli-subgaussian matrices, defined as follows. Let X be a matrix of size $n \times p$ with iid entries x_{ij}, where

$$x_{ij} := \chi_{ij} \xi_{ij}, \tag{I.1}$$

where χ_{ij} are iid random variables with $\mathbf{P}(\chi_{ij} = 1) = \theta$, $\mathbf{P}(\chi_{ij} = 0) = 1 - \theta$, and ξ_{ij} are iid random variables with mean 0, variance bounded by 1,

$$\mathbf{E}|\xi| \in [1/10, 1],$$

and

$$\mathbf{P}(|\xi| \geq t) \leq 2 \exp(-t^2/2).$$

This model includes many important distributions such as the standard Gaussians and Rademachers. The $1/10$ is introduced for convenience of analysis and not critical to the argument.

For a vector $v \in \mathbb{R}^n$, let $\mu_v := \mathbf{E}\|X^T v\|_1$. Let c be a small positive constant ($c = .1$ suffices) and let $Bad(v)$ be the event that $|\|X^T v\|_1 - \mu_v\| \geq c\mu_v$. We want to have

$$\mathbf{P}(\cup_{v \in \mathbb{R}^n} Bad(v)) = o(1). \tag{I.2}$$

In other words, with high probability, $\|X^T v\|_1$ does not deviate significantly from its mean, simultaneously for all $v \in \mathbb{R}^n$.

We aim to find the smallest value of p which guarantees (I.2). Notice that $\|X^T v\|$ is the sum of p iid random variables $|X_i v|$ where X_i are the rows of X. Thus, intuitively the larger p is, the more $\|X^T v\|$ concentrates. From below, we observe that (I.2) fails if $p \leq n$, since in this case for any matrix X one can find a v such that $X^T v = 0$ and $\mu_v \geq 1$ (we can take v arbitrarily long). Spielman, Wang, and Wright [15] showed that $p \geq Cn^2 \log^2 n$ suffices. Our main contribution is the following

Theorem I.1. *For any constant $c > 0$ there is a constant $C > 0$ such that* (I.2) *holds for any $p \geq Cn \log^4 n$.*

0272-5428/15 $31.00 © 2015 IEEE
DOI 10.1109/FOCS.2015.90

Theorem I.1 is of interest for several reasons. First, while concentration inequalities for random matrices are abundant, most of them concern the spectral or l_2 norm. We have not seen one which addresses the l_1 norm as in this theorem. As sparsity plays crucial role in data analysis, techniques involving l_1 norm (such as l_1 optimization) become more and more important. Second, in the proof we introduce two general ideas, which seem to be applicable in many settings. The first is an economical way to apply the union bound and the second is a refined version of Bernstein's concentration inequality for sums of independent variables. Finally, our study is directly motivated by a matrix recovery problem of fundamental interest in data analysis, which contains dictionary learning as a special case. As an application of Theorem I.1, we are able to prove (up to a logarithmic term) a conjecture of Spielman et. al. concerning the optimal number of random samples one needs to recover a hidden dictionary of size n.

The rest of the paper is organized as follows. In the next section, we discuss the application. In Section III, we present the main ideas behind the proof of Theorem I.1. The details follows next in Section IV. Section V contains the accompanying algorithms and a brief summary of the analysis in [15] that motivated our investigation. Section VI addresses a generalization to rectangular dictionaries. Section VII introduces a new algorithm that achieves the optimal bound in the sparse regime. In Section VIII, we conclude with the results of some numerical experiments of the various algorithms.

Acknowledgement. We would like to thank D. Spielman for bringing the concentration problem to our attention and the anonymous reviewers for their helpful comments.

II. Matrix recovery and dictionary learning

Let A be an $n \times n$ invertible matrix and X be an $n \times p$ matrix; set $Y := AX$. The aim of this paper is to study the following recovery problem:

Given Y, reconstruct A and X.

It is clear that in the equation

$$Y = AX, \tag{II.1}$$

we have $n^2 + np$ unknowns (the entries of A and X), and only np equations (given by the entries of Y). Thus, the problem is underdetermined and one cannot hope for a unique solution. However, in practice, X is frequently a sparse matrix. If X is sparse, the number of unknowns decreases dramatically, as the majority of entries of X are zero. The name of the game here is to find the minimum value of p, the number of observations, which guarantees a unique recovery (e.g. [2] and [6]). Another direction, which we do not address, is to require less stringent sparsity requirements at the cost of larger sample complexity (cf. [16]).

One real-life application that motivates the studies of this problem is dictionary learning. The matrix A can be seen as a hidden dictionary, with its columns being the words. X is a sparse sample matrix. This means that in the columns of Y we observe linear combinations of a few columns of A. From these observations, we would like to recover the dictionary. An archetypal example is facial recognition [19] [10]. A database of observed faces is used to generate the dictionary and once the dictionary is found, the problem of storing and transmitting facial images can be done very efficiently, as all one needs is to store and transmit few coefficients. In fact, such dictionary-learning techniques can be utilized to recognize faces that are partially occluded or corrupted with noise [18]. For more discussion and real-life examples, we refer to [9], [12] and the references therein. Another practical situation in which the recovery problem appears essential is blind source separation and we refer the reader to [21] for more details.

There have been many approaches to efficient recovery beginning with the work of [12]. Let us mention, among others, online dictionary learning by [11], SIV [7], the relative Newton method for source separation by [20], the Method of Optimal Directions by [4], K-SVD in [1], and scalable variants in [11].

While various different approaches have been considered, there have not been many rigorous results concerning performance. The first such result has been obtained by Spielman, Wang and Wright [15] concerning recovery with random samples; in other words, X is a random sparse matrix. Before stating their result, we need to discuss the meaning of *unique*. First, notice that if $Y = AX$, then $Y = (AV)(V^{-1}X)$ for any diagonal matrix V with non-zero diagonal entries. Furthermore, one can freely permute the columns of A and the rows of X accordingly while keeping Y the same. In the rest of the paper, unique recovery will be understood modulo these two operations. Spielman et. al. proved

Theorem II.1. *There are constants $C > 0, C' > 0$ such that the following holds. Let A be an invertible $n \times n$ matrix and X a sparse random $n \times p$ matrix with $2/n \leq \theta \leq C'/\sqrt{n}$ and ξ_{ij} having a symmetric distribution. Then for $p \geq Cn^2 \log^2 n$, one can efficiently find a solution with probability $1 - o(1)$.*

Here and later, efficient means polynomial time. The algorithm designed for this purpose is called ER-SpUD, whose main subroutine is l_1 optimization. We are going to present and discuss this algorithm in Section V. In the dictionary learning problem, p is the number of measurements, and it is important to optimize its value. From below, it is easy to see that we must have $p \geq cn \log n$ for some constant $c > 0$. Indeed, if $\theta = 2/n$ (or c'/n for any constant c') and $p < cn \log n$ for a sufficiently small constant c, then the coupon collector argument shows that with probability $1 - o(1)$, X has an all-zero row. In this case, changing the corresponding column of A will not effect Y, and an unique recovery is hopeless. Spielman et. al. conjecture

Conjecture II.2. *There are constants $C > 0, \alpha > 0$ such that the following holds. Let A be an invertible $n \times n$ matrix and X a sparse random $n \times p$ matrix with $2/n \leq \theta \leq \alpha/\sqrt{n}$. Then for $p \geq Cn \log n$, one can efficiently find a solution with probability $1 - o(1)$.*

As a matter of fact, they believe that ER-SpUD should perform well as long as $p \geq Cn \log n$, for some large constant C. They also proved that if one does not cared about the running time of the algorithm, then $p \geq Cn \log n$ suffices.

The analysis in [15] boils down to the concentration question in the previous section, and Theorem II.1 holds for any p which guarantees (I.2) (see Section V for more details). Using Theorem I.1, we can prove

Theorem II.3. *There are constants $C > 0, C' > 0$ such that the following holds. Let A be an invertible $n \times n$ matrix and X a sparse random $n \times p$ matrix with $2/n \leq \theta \leq C'/\sqrt{n}$. Then for $p \geq Cn \log^4 n$, one can efficiently find a solution with probability $1 - o(1)$.*

Our p is within a $\log^3 n$ factor from the conjectured bound. Furthermore, we can drop the assumption that ξ_{ij} are symmetric from Theorem II.1.

To conclude this section, let us mention two refinements of Theorem I.1. First, combining the proof of Theorem II.3 with a result from random matrix theory, we obtain the following more general result, which handles the case when A is rectangular

Theorem II.4. *There are constants $C, \alpha > 0$ such that the following holds. Let $n > m$ and A be an $n \times m$ matrix of rank m and and X a sparse random $m \times p$ matrix with $2/n \leq \theta \leq \alpha/\sqrt{n}$. Then for $p \geq Cn \log^4 n$, one can efficiently find a solution with probability $1 - o(1)$*

Second, in the sparest case $\theta := \Theta(1/n)$, we can obtain the optimal bound $p = Cn \log n$, using a different algorithm, proving Conjecture II.2 in this regime.

Theorem II.5. *For any $c > 0$ there is a constant $C > 0$ such that the following holds. Let A be an invertible $n \times n$ matrix and X a sparse random $n \times p$ matrix with $\theta = c/n$. Then for $p \geq Cn \log n$, one can efficiently find a solution with probability $1 - o(1)$*

III. The main ideas and lemmas

A. The standard ϵ-net argument

Let us recall our task. For a vector $v \in \mathbb{R}^n$, let $\mu_v := \mathbf{E}\|X^T v\|_1$. Let c be a small positive constant ($c = .1$ suffices) and let $Bad(v)$ be the event that $|\|X^T v\|_1 - \mu_v| \geq c\mu_v$. We want to show that if p is sufficiently large, then

$$\mathbf{P}(\cup_{v \in \mathbb{R}^n} Bad(v)) = o(1). \tag{III.1}$$

For the sake of presentation, let us assume that the random variables ξ_{ij} are Rademacher (taking values ± 1 with probability $1/2$); the entries x_{ij} of X have the form $x_{ij} = \chi_{ij}\xi_{ij}$, where χ_{ij} are iid indicator variables with mean θ. We start by a quick proof of the bound $p \geq Cn^2 \log^2 n$ obtained in [15]. Notice that the union in (I.2) contains infinitely many terms. The standard way to handle this is to use an ϵ-net argument.

Definition III.1. A set $\mathcal{N} \subset \mathbb{R}^n$ is an ϵ-net of a set $D \subset \mathbb{R}^n$ in l_q norm, for some $0 < q \leq \infty$, if for any $x \in D$ there is $y \in \mathcal{N}$ so that $\|x - y\|_q \leq \epsilon$. The unit sphere in l_q norm consists of vectors v where $\|v\|_q = 1$. B denotes the unit sphere in l_1 norm.

Considering the vectors in B is sufficient to prove the result. It is easy to show that for any $v \in B$

$$\mu_{min} := p\sqrt{\theta/n} \leq \mu_v \leq p\theta := \mu_{\max},$$

where the lower bounds attend at $v = \frac{1}{n}\mathbf{1}$ ($\mathbf{1}$ is the all one vector) and the upper bound at $v = (1, 0, \ldots, 0)$. Let $\mathcal{N}_0$ be the set of all vectors in B whose coordinates are integer multiples of n^{-3}. Any vector in B would be of distance at most n^{-2} in l_1 norm from some vector in $\mathcal{N}_0$ (thus $\mathcal{N}_0$ is an n^{-2}-net of B). A short consideration shows that if $u, v \in B$ are within n^{-2} of each other, then

$$|\mu_v - \mu_u| = o(\mu_{min}).$$

Thus, to prove (I.2), it suffices to show that

$$\mathbf{P}(\cup_{v\in\mathcal{N}_0} Bad(v)) = o(1). \tag{III.2}$$

In order to bound $\mathbf{P}(\cup_{v\in\mathcal{N}_0} Bad(v))$, let us first bound $\mathbf{P}(Bad(v))$ for any B. Notice that

$$\|X^T v\|_1 = \sum_{i=1}^{p} |X_i v|,$$

where X_i are the columns of X. The random variables $|X_i v|$ are iid, and one is poised to apply another standard tool, Bernstein's inequality for the sum of independent random variables.

Lemma III.2. *Let $Z_1, \ldots, Z_n$ be independent random variables such that $|Z_i| \leq \tau$ with probability 1. Let $S := \sum_{i=1}^{n} Z_i$. Then for any $T > 0$*

$$\max\{\mathbf{P}(S - \mathbf{E}S \leq -T), \mathbf{P}(S - \mathbf{E}S \geq T)\} \leq \exp(-\frac{T^2}{2(\mathbf{Var}S + T\tau)}) \leq \exp(-\min\{\frac{T^2}{4\mathbf{Var}S}, \frac{T}{4\tau}\}).$$

In our case $Z_i = |X_i v| = \sum_{i=1}^{n} \mathbf{X}_{ij} v_j$. As $|x_{ij} = \chi_{ij}\xi_{ij}| \leq 1$ with probability 1 (we assume that ξ_{ij} are Rademacher)

$$|Z_i| \leq \sum_{i=1}^{n} |v_j| = \|v\|_1 = 1$$

with probability 1. This means we can set $\tau = 1$. Furthermore

$$\mathbf{Var}\sum_{i=1}^{p} Z_i = p\mathbf{Var}Z_i \leq p\mathbf{E}|X_i v|^2 = p\sum_{j=1}^{n} \theta v_j^2 \leq p\theta \sum_{j=1}^{n} |v_j| = p\theta.$$

Finally, one can set $T = c\mu_{min} = cp\sqrt{\theta/n}$. Lemma III.2 implies that

$$\mathbf{P}(Bad(v) \leq 2\exp(-\min\{\frac{c^2 p^2 \theta/n}{4p\theta}, \frac{cp\sqrt{\theta/n}}{4}\}) = 2\exp(-\frac{c^2 p}{4n})$$

since $\sqrt{\theta/n} \geq 1/n$ as $\theta \geq 1/n$.

Using the union bound

$$\mathbf{P}(\cup_{v\in\mathcal{N}_0} Bad(v) \leq \sum_{v\in\mathcal{N}_0} \mathbf{P}(Bad(v)) \tag{III.3}$$

we obtain

$$\mathbf{P}(\cup_{v\in\mathcal{N}_0} Bad(v)) \leq |\mathcal{N}_0| \times 2\exp(-\frac{c^2 p}{4n}).$$

It is easy to check that $\mathcal{N}_0 = \exp(\Omega(n\log n))$. So, in order to make the RHS $o(1)$, we need $p \geq Cn^2 \log n$ for a sufficiently large constant C. For the case when ξ_{ij} are not Bernoulli (but still subgaussian) the calculation in [15] requires an extra logarithm term, which results in the bound $p \geq Cn^2 \log^2 n$.

B. New ingredients

Our first idea is to find a more efficient variant of the union bound

$$\mathbf{P}(\cup_{v\in\mathcal{N}_0} Bad(v)) \le \sum_{v\in\mathcal{N}_0} \mathbf{P}(Bad(v)).$$

Motivated by the inclusion-exclusion formula we try to capture some gain when $\mathbf{P}(Bad(u)\cap Bad(v))$ is large for many pairs u, v. We observe that if we can group the elements v of the net into clusters so that within each cluster, the events $Bad(v)$ (seen as subsets of the underlying probability space) are close to each other. Assume, for a moment, that one can split the net $\mathcal{N}_0$ into m disjoint clusters $\mathcal{C}_i$, $1 \le i \le m$, so that if u and v belong to the same cluster $\mathbf{P}(Bad(u)\backslash Bad(v)) \le p_1$, where p_1 is much smaller than p_0, then

$$\mathbf{P}(\cup_{v\in\mathcal{C}_i} Bad(v)) \le \mathbf{P}(Bad(v^{[i]})) + |\mathcal{C}_i| p_1,$$

where $v^{[i]}$ is a representative point in $\mathcal{C}_i$. Summing over i, one obtains

$$\mathbf{P}(\cup_{v\in\mathcal{N}_0} Bad(v)) \le \sum_{i=1}^{m} \mathbf{P}(\cup_{v\in\mathcal{C}_i} Bad(v)) \le \sum_{i=1}^{m} \mathbf{P}(Bad(v^{[i]})) + |\mathcal{N}_0| p_1 \le m p_0 + |\mathcal{N}_0| p_1. \tag{III.4}$$

We gain significantly if p_1 is much smaller than p_0 and m is much smaller than $|\mathcal{N}_0|$. Next, viewing the set of representatives $v^{[i]}$ as a new net $\mathcal{N}_1$, we can iterate the argument, obtaining the following lemma.

Lemma III.3. *Let $\mathcal{P}$ be a probability space. Let $\mathcal{N} = \mathcal{N}_0$ be a finite set, where to each element $v \in \mathcal{N}_0$ we associate a set $Bad_0(v) \subset \mathcal{P}$. Assume that we can construct a sequence of sets*

$$\mathcal{N}_L, \mathcal{N}_{L-1}, \dots, \mathcal{N}_0,$$

and for each $u \in \mathcal{N}_l, 1 \le l \le L$ an event $Bad_l(u)$ such that the following holds. For each $v \in \mathcal{N}_{l-1}$, there is $u \in \mathcal{N}_l$ such that $\mathbf{P}(Bad_{l-1}(v)\backslash Bad_l(u)) \le p_l$ and for each $u \in \mathcal{N}_L$, $\mathbf{P}(Bad_L(u)) \le p_0$. Then

$$\mathbf{P}(\cup_{v\in\mathcal{N}_0} Bad_0(v)) \le |\mathcal{N}_L| p_0 + \sum_{l=1}^{L} |\mathcal{N}_{l-1}| p_l. \tag{III.5}$$

The construction of $\mathcal{N}_l$ are of critical importance, and we are going to construct them using the l_∞ distance, rather than the obvious choice of l_1. *This is the key point of our method.*

The next main technical ingredient is a more efficient way of using Bernstein's inequality, Lemma III.2. Recall the bound

$$\mathbf{P}(|S - \mathbf{E}S| \ge T) \le 2\exp(-\frac{T^2}{2(\mathbf{Var}S + T\tau)}) \le 2\exp(-\min\{\frac{T^2}{4\mathbf{Var}S}, \frac{T}{4\tau}\}). \tag{III.6}$$

The first term $\frac{T^2}{4\mathbf{Var}S}$ on the right most formula is usually optimal. However, we need to improve the second term. The idea is to replace τ with a smaller quantity τ' such that the probability that $|Z_i| \le \tau'$ is close to 1. Let us illustrate this idea with the upper tail. Set $\mu := \mathbf{E}S$, we consider

$$\mathbf{P}(S \ge \mu + T).$$

Write

$$Z_i := Z_i\mathbf{J}_i + Z_i\mathbf{I}_i$$

where $\mathbf{J}_i$ is the indicator of the event $|Z_i| \le \tau'$ and $\mathbf{I}_i = 1 - \mathbf{J}_i$. Thus

$$S := \sum_i Z_i\mathbf{I}_i + \sum_i Z_i\mathbf{J}_i = Q + S(1).$$

Let μ_j be the expectation of $S(j)$. Then

$$\mathbf{P}(S \ge \mu + T) \le \mathbf{P}(Q \ge \mu_0 + T/2) + \mathbf{P}(S(1) \ge \mu_1 + T/2).$$

We can use Lemma III.2 to bound $\mathbf{P}(Q \geq \mu_0 + T/2)$, which provides a bound better than (III.6) as now $\tau' < \tau$. On the other hand, if the probability that $|Z_i| \geq \tau'$ is small, then we can bound $\mathbf{P}(S(1) \geq \mu_1 + T/2)$ in a different way, exploiting the fact that there will be very few non-zero summands in $S(1)$.

We can (and have to) further refine this idea by considering a sequence of τ', breaking S into the sum of Q and $S(k), 1 \leq k \leq M$, for a properly chosen M. This will be our leading idea to bound the difference probability p_l in the next section.

On the abstract level, our method bears a similarity to the chaining argument from the theory of Banach spaces. We are going to discuss this point in Section IV-G.

IV. Proof of Theorem I.1

For the sake of presentation, we assume that $x_{ij} = \chi_{ij}\xi_{ij}$ where χ_{ij} are iid Bernoulli random variables with mean θ and ξ_{ij} are iid Rademachers random variables. In fact, $p \geq Cn\log^3 n$ is sufficient for the Rademacher case. The proof can be easily modified for ξ_{ij} being general sub-gaussian at the cost of a $\sqrt{\log n}$ factor in the bound for p (See Section IV-F). We recall the notation $\mu_{min} = p\sqrt{\theta/n}, \mu_{max} = p\theta$; $\mu_v := \mathbf{E}\|X^T v\|_1$. B is the set of all vectors of unit l_1 norm.

We set $p = Cn\log^3 n$, for a sufficiently large constant C. Let $T := \frac{c_0\mu_{min}}{\log n}$ for a small constant $c_0 > 0$ and $K := \lceil \frac{6\mu_{max}}{T} \rceil$.

A. α-nets in l_∞ norm

Lemma IV.1. *For any $1 \geq \alpha \geq 2/n$, B admits an α-net in l_∞ norm of size at most* $\exp(2\alpha^{-1}\log n)$.

Proof: Let $\mathcal{N}$ be the collection of all vectors $v \in B$, whose coordinates are integer multiples of α. Obviously, $\mathcal{N}$ is an α-net of B in l_∞ norm. Furthermore, any $v \in \mathcal{N}$ satisfies $\|v\|_1 \leq 1$, so it has at most $k := \alpha^{-1}$ non-zero coordinates. If a coordinate is non-zero, it can take at most $2\alpha^{-1} + 1 \leq 3k$ values. Therefore,

$$|\mathcal{N}| \leq \sum_{i=0}^{k} \binom{n}{i} (3k)^k.$$

As $\alpha \geq 2/n$, the RHS is at most

$$n\binom{n}{k}(3k)^k \leq n(\frac{en}{k} \times 3k)^k = n(2en)^k \leq \exp(2\alpha^{-1}\log n).$$

■

The key here is that we consider an α-net in l_∞ norm, rather than in l_1 norm, which appears to be a natural choice.

B. Building a nested sequence

Recall that $\mathcal{N}_0$ is the set of vectors v in B whose coordinates are integer multiples of n^{-3}. We have

$$|\mathcal{N}_0| \leq (2n^3 + 1)^n \leq \exp(4n\log n). \tag{IV.1}$$

Consider the sequence $\alpha_0 = 2/n; \alpha_l = 2\alpha_{l-1}$ for $l = 1, \ldots, L$, where $L \leq \log_2 n$ is the first index such that $\alpha_L > 1/2$. Let $\mathcal{N}_l'$ be an α_l-net of B in the l_∞ norm. By Lemma IV.1, we can choose $\mathcal{N}_l'$ such that

$$|\mathcal{N}_l'| \leq \exp(2\alpha_l^{-1}\log n). \tag{IV.2}$$

We now build a nested sequence $\mathcal{N}_L \subset \mathcal{N}_{L-1} \subset \cdots \subset \mathcal{N}_1 \subset \mathcal{N}_0$ as follows. Assume that $\mathcal{N}_{l-1}$ has been built. Use the points in $\mathcal{N}_l'$ as centers to construct a Voronoi partition of the points of $\mathcal{N}_{l-1}$ with respect to the l_∞ norm (ties are broken arbitrarily). For each point $u \in \mathcal{N}_l'$, let C_u be the subset of $\mathcal{N}_{l-1}$ corresponds to u. By definition, $\|u - v\|_\infty \leq \alpha_l$ for any $v \in C_u$,

Partition the interval $[\mu_{min}, \mu_{max}] = [p\sqrt{\theta/n}, p\theta]$ into K intervals $I_1, \ldots, I_K$ of equal lengths. We partition C_u further into K subsets $C_{u,j}, 1 \leq j \leq K$, where $v \in C_{u,j}$ if $\mathbf{E}\|Xv\|_1 \in I_j$. By this construction, if v, w belong to the same $C_{u,j}$, then by the definition of K, we have the key relations

$$\|v - w\|_\infty \leq 2\alpha_l \text{ and } |\mathbf{E}\|Xv\|_1 - \mathbf{E}\|Xw\|_1| \leq p\theta/K \leq T/6. \tag{IV.3}$$

From each set $C_{u,j}$ choose an arbitrary element v. Thus, each $u \in \mathcal{N}'_l$ gives rise to a set R_u of K elements (R stands for representative). Define

$$\mathcal{N}_l := \cup_{u \in \mathcal{N}'_l} R_u.$$

It is clear that $\mathcal{N}_l \subset \mathcal{N}_{l-1}$ and

$$|\mathcal{N}_l| \le K|\mathcal{N}'_l| \le K \exp(2\alpha_l^{-1} \log n). \tag{IV.4}$$

C. Bounding the differences

Consider the construction of $\mathcal{N}_l$, $1 \le l \le L$, from Section IV-A. Let $v \in \mathcal{N}_l$. Thus, $v \in C_{u,j}$ for some $u \in \mathcal{N}'_l$ and $1 \le j \le K$. Consider another point $w \in \mathcal{N}_{u,j}$. Our main task is to show

Lemma IV.2. *For all pairs v, w as above*

$$\rho(v,w) := \mathbf{P}(|\|X^T v\|_1 - \|X^T w\|_1| \ge T) \le \exp(-5\alpha_l^{-1} \log n). \tag{IV.5}$$

The rest of this section is devoted to the proof of this lemma. By (IV.3), we have

$$\|v - w\|_\infty \le 2\alpha_l \ \text{ and } \ |\mathbf{E}\|X^T v\|_1 - \mathbf{E}\|X^T w\|_1| \le p\theta/K \le T/6. \tag{IV.6}$$

Define $Z_i = |X_i v| - |X_i w|$, where X_i is the ith row of X^T; we have

$$\|X^T v\|_1 - \|X^T w\|_1 = \sum_{i=1}^p (|X_i v| - |X_i w|) = \sum_{i=1}^p Z_i.$$

Set $S := \sum_{i=1}^p Z_i$; by symmetry, it suffices to bound

$$\mathbf{P}(Z_1 + \cdots + Z_p \ge T) := \mathbf{P}(S \ge T).$$

Notice that by the triangle inequality

$$|Z_i| = \Big| |X_i v| - |X_i w| \Big| \le |X_i (v - w)|.$$

Therefore,

$$\mathbf{Var} Z_i \le \mathbf{E} Z_i^2 \le \mathbf{E}|X_i(v-w)|^2 = \theta \sum_{j=1}^n (v_j - w_j)^2.$$

Recall that $\|v\|, \|w\| \le 1$ and $\|v - w\|_\infty \le \alpha_l$. Therefore

$$\sum_{j=1}^n (v_j - w_j)^2 \le \alpha_l \sum_{j=1}^n |v_j| + |w_j| = 2\alpha_l.$$

This implies

$$\mathbf{Var} Z_i \le \mathbf{E} Z_i^2 \le 2\alpha_l \theta. \tag{IV.7}$$

We denote by $\mathbf{I}_{i,k}$ the event that $\tau_k < Z_i \le \tau_{k-1}$ for $k = 1, \dots, M$ and J_i the event that $|Z_i| \le \tau_M$, for a sequence τ_k, $k = 0, \dots, M$, where $\tau_0 = 2$; $\tau_i = 2^{-i}\tau_0$ and M is the first index so that

$$\min\{\frac{\tau_M^2}{8\alpha_l \theta}, \frac{\tau_M}{4\alpha_l}\} \ge 8 \log n. \tag{IV.8}$$

Note that if $\alpha_l \le \frac{1}{32} \log^{-1} n$ then such an index $M \ge 1$ exists. We will proceed with this assumption and cover the remaining cases at the end of the proof. Apparently,

$$Z_i \le \sum_{i=1}^k Z_i \mathbf{I}_{i,k} + Z_i J_i.$$

Set $S(k) = \sum_{i=1}^{p} Z_i \mathbf{I}_{i,k}$ for $k = 1, \dots, M$ and $Q = \sum_{i=1}^{p} Z_i \mathbf{J}_i$. We have

$$\mathbf{P}(S \geq T) \leq \mathbf{P}(Q \geq T/2) + \sum_{k=1}^{M} \mathbf{P}(S(k) \geq \frac{T}{2M}).$$

To bound $\mathbf{P}(Q \geq T/2)$, we notice that (see (IV.11)) the choice of τ_M guarantees that $\mathbf{P}(\mathbf{J}_i) \geq 1 - 2n^{-8}$ for all $i = 1, \dots, p$. As $|Z_i| \leq 2$ with probability 1, it follows that

$$|\mathbf{E} Z_i \mathbf{J}_i - \mathbf{E} Z_i| \leq 4n^{-8}$$

and so

$$|\mathbf{E}Q - \mathbf{E}S| \leq 4pn^{-8} = o(n^{-6}),$$

as $p = o(n^2)$. On the other hand, by (IV.6), $T \geq 5(\mathbf{E}S + n^{-6})$. Thus

$$\mathbf{P}(Q \geq T/2) \leq \mathbf{P}(Q \geq \mathbf{E}Q + T/4).$$

By definition, Q is sum of p iid random variables, each is bounded by τ_M in absolute value with probability 1. Furthermore, by (IV.7)

$$\mathbf{Var} Q = p \mathbf{Var} Z_1 \mathbf{J}_1 \leq p \mathbf{E} Z_1^2 \leq 2\alpha_l \theta p.$$

By Lemma III.2, we have

$$\mathbf{P}(Q \geq \mathbf{E}Q + T/4) \leq 2(\exp(-\min\{\frac{(T/4)^2}{8\alpha_l \theta p}, \frac{T/4}{4\tau_M}\}) = 2\exp(-\min\{\frac{T}{128\alpha_l \theta p}, \frac{T}{16\tau_M}\}). \tag{IV.9}$$

Now we bound $\mathbf{P}(S(k) \geq \frac{T}{2M})$, for $k = 1, \dots, M$. Recall that $S(k) := \sum_{i=1}^{p} Z_i \mathbf{I}_{i,k}$ is a sum of iid non-negative random variables, each is either 0 or in (τ_k and τ_{k-1}]. Thus, if $S(k) \geq T/2M$ there must be at least $p_k := \frac{T/2M}{\tau_{k-1}}$ indices i such that $Z_i > \tau_k$. Let ρ_k be the probability that $Z_1 > \tau_k$. Then by the union bound and the fact that $p = o(n^2)$,

$$\mathbf{P}(S(k) \geq \frac{T}{2M}) \leq \binom{p}{p_k} \rho_k^{p_k} \leq (\frac{ep}{p_k}\rho_k)^{p_k} \leq (\frac{n^2}{2}\rho_k)^{p_k}. \tag{IV.10}$$

To complete the analysis, we need to estimate ρ_k. By definition

$$\rho_k := \mathbf{P}(|X_1 v| - |X_1 w| > \tau_k) \leq \mathbf{P}(|X_1(v-w)| \geq \tau_k).$$

The random variable $\tilde{Z}_1 := X_1(v-w) = \sum_{j=1}^{n} \xi_j (v_j - w_j)$ has mean 0. Furthermore, by (IV.7), $\mathbf{Var}\tilde{Z}_1 \leq \tilde{Z}_1^2 \leq 2\alpha_l \theta$. Finally, each term $\xi_j(v_j - w_j)$ is at most α_l in absolute value. Thus Lemma III.2 implies

$$\rho_k \leq \mathbf{P}(|\tilde{Z}_1| \geq \tau_k) \leq 2(\exp(-\min\{\frac{\tau_k^2}{8\alpha_l\theta}, \frac{\tau_k}{4\alpha_l}\}). \tag{IV.11}$$

This and (IV.10) yield

$$\mathbf{P}(S(k) \geq \frac{T}{2M}) \leq 2\exp(-\Big(\min\{\frac{\tau_k^2}{8\alpha_l\theta}, \frac{\tau_k}{4\alpha_l}\} + 2\log n\Big)p_k). \tag{IV.12}$$

By (IV.8),

$$\min\{\frac{\tau_k^2}{8\alpha_l\theta}, \frac{\tau_k}{4\alpha_l}\} \geq 8\log n,$$

so

$$\Big(\min\{\frac{\tau_k^2}{8\alpha_l\theta}, \frac{\tau_k}{4\alpha_l}\} + 2\log n\Big)p_k \geq \frac{1}{2}\min\{\frac{\tau_k^2}{8\alpha_l\theta}p_k, \frac{\tau_k}{4\alpha_l}p_k\}.$$

By definition $p_k = \frac{T/2M}{\tau_{k-1}} = \frac{T/4M}{\tau_k}$, as $\tau_{k-1} = 2\tau_k$. Therefore,

$$\frac{1}{2}\frac{\tau_k^2}{8\alpha_l\theta}p_k = \frac{\tau_k T}{64M\alpha_l\theta}$$

and

$$\frac{1}{2}\frac{\tau_k}{4\alpha_l}p_k = \frac{T}{32M\alpha_l}.$$

By (IV.9) and (IV.12), we conclude that

$$\mathbf{P}(S \geq T) \leq 2\exp(-\min\{\frac{T^2}{128\alpha_l\theta p}, \frac{T}{16\tau_M}\}) + \sum_{k=1}^{M} 2\exp(-\min\{\frac{\tau_k T}{64M\alpha_l\theta}, \frac{T}{32M\alpha_l}\}). \tag{IV.13}$$

A routine verification (see Section IV-E) shows that once $p \geq Cn\log^3 n$ for a sufficient large constant C, then the RHS in (IV.13) is at most $\exp(-5\alpha_l^{-1}\log n)$, completing the proof for the case $\alpha_l \leq \frac{1}{32}\log^{-1} n$.

To complete the proof, we now treat the remaining case when $\alpha_l \geq \frac{1}{32}\log^{-1} n$.. In this case, we do not need to split Z_i. Recall $S = Z_1 + \cdots + Z_p$ where $|Z_i| \leq 2$ with probability 1, $\mathbf{E}S \leq T/6$ and $\mathbf{Var}S \leq 2p\theta\alpha_l$. By Lemma III.2, we have

$$\mathbf{P}(S \geq T) \leq \mathbf{P}(S \geq \mathbf{E}S + T/2) \leq \exp(-\min\{\frac{T^2}{8p\theta\alpha_l}, \frac{T}{8}\}).$$

By the analysis of (IV.13), we already know that $\frac{T^2}{8p\theta\alpha_l} \geq 5\alpha_l^{-1}\log n$. On the other hand, as $\alpha_l \geq \frac{1}{32}\log^{-1} n$

$$\frac{T}{8} = \frac{c_0 p\sqrt{\theta/n}}{8\log n} = \frac{c_0 C}{8}\sqrt{\theta n}\log^2 n \geq 5\alpha_l^{-1}\log n,$$

given that $c_0 C$ is sufficiently large. This completes the proof.

D. Proof of the Concentration lemma

For $v \in \mathcal{N}_l, 0 \leq l \leq L$, let $Bad_l(v)$ be the event that $|\|Xv\|_1 - \mu_v| \geq 2(L+1-l)T$. For $l = 0$, $2(L+1-l_T = 2(L+1)T \leq \frac{2c_0(\log_2 n+1)\mu_{min}}{\log n} \leq 4c_0\mu_{min}$. Thus,

$$\mathbf{P}(\cup_{v\in\mathcal{N}_0}|\|X^T v\|_1 - \mu_v| \geq 4c_0\mu_{min}) \leq \mathbf{P}(\cup_{v\in\mathcal{N}_0} Bad_0(v)).$$

Assume that there is a number p_0 such that $\mathbf{P}(Bad_0(v)) \leq p_0$ for all $v \in \mathcal{N}_0$. Assume furthermore that for any $1 \leq l \leq L$, there is a number p_l such that for $v \in \mathcal{N}_l$ and $w \in \mathcal{N}_{l-1}$ where v is the representative of the set $C_{(u,k)}$ that contains w (see the construction in Section IV-B).

$$\mathbf{P}(Bad_l(w)\backslash Bad_{l-1}(v)) \leq p_l.$$

Then by Lemma III.3

$$\mathbf{P}(\cup_{v\in\mathcal{N}_0}) \leq |\mathcal{N}_L|p_0 + \sum_{l=1}^{L}|\mathcal{N}_{l-1}|p_l.$$

To find p_l, notice that if $Bad_{l-1}(w)$ holds and $Bad_l(v)$ does not, then $|\|X^T w\|_1 - \mu_w| \geq 2(L+2-l)T$ and $|\|X^T v\|_1 - \mu_v| \leq 2(L+1-l)T$. By (IV.3), $|\mu_v - \mu_w| \leq T$. It thus follows that

$$|\|X^T w\|_1 - \|X^T v\|_1| \geq T.$$

By the main lemma of Section IV-C, we know that the probability of this event is at most $p_l := \exp(-5\alpha_l^{-1}\log n)$, for all l. Recall from Section IV-B that

$$|\mathcal{N}_l| \leq K\exp(2\alpha_l^{-1}\log n) = K\exp(4\alpha_l^{-2}\log n),$$

we have

$$\sum_{l=1}^{L} |\mathcal{N}_{l-1}| p_l \leq \sum_{l=1}^{L} \exp(-4\alpha_l^{-1} \log n) \times K \exp(4\alpha_l^{-1} \log n).$$

Since $K = O(n^{1/2})$ and $\alpha_l^{-1} \log n \geq \log n$, the RHS is at most

$$\sum_{l=1}^{L'} \exp(-.5\alpha_l^{-1} \log n) = o(1).$$

To conclude, notice that by Lemma III.2, we can set $p_0 := 2\exp(-\min\{\frac{T^2}{8p\theta}, \frac{T}{8}\})$. As $|\mathcal{N}_L| \leq \exp(-2\alpha_L^{-1} \log n) \leq \exp(4 \log n)$ since $\alpha_L \geq 1/2$, we have

$$p_0 |\mathcal{N}_0| = o(1),$$

as long as $\min\{\frac{T^2}{8p\theta}, \frac{T}{8}\} \geq 5 \log n$. This condition holds if $p \geq Cn \log^3 n$ for a sufficiently large constant C. This implies that

$$\mathbf{P}(\cup_{v \in \mathcal{N}_0} \{\|X^T v - \mu_v\| \geq 4c_0 \mu_{min}) = o(1),$$

and we are done by (III.2).

E. The magnitude of p

We present the routine verification concerning the exponents in (IV.13). This is the only place where the magnitude of p matters. Recall that $T = \frac{c_0 \mu_{min}}{\log n} = \frac{c_0 p \sqrt{\theta/n}}{\log n}$ and $p = Cn \log^3 n$ (since for the sake of exposition we are only considering the Rademacher case). We have

$$\frac{T^2}{128\alpha_l \theta p} = \frac{c_0^2 p^2 \theta / n}{128 \theta p \log^2 n} \alpha_l^{-1} = \frac{c_0^2 p}{n} \alpha_l^{-1} = c_0^2 C \alpha_l^{-1} \log n \geq 4.1 \alpha_l^{-1} \log n,$$

provided that $c_0^2 C \geq 4.1$.

By the definition of M in (IV.8), we have

$$32 \log n \geq \min\{\frac{\tau_M^2}{8\alpha_l \theta}, \frac{\tau_M}{4\alpha_l}\} \geq 8 \log n.$$

This implies that

$$\tau_M \leq \max\{16\sqrt{\alpha_l \theta \log n}, 128 \alpha_l \log n\}.$$

It follows that

$$\frac{T}{16\tau_M} \geq \min\{\frac{T}{256\sqrt{\alpha_\theta \log n}}, \frac{T}{2048\sqrt{2}\alpha_l \log n}\}.$$

By the definition of p and T

$$\frac{T}{256\sqrt{\alpha_\theta \log n}} = \frac{c_0 p}{256\sqrt{\alpha_l n \log^3 n}} = \alpha_l^{-1} c_0 C \log n \sqrt{n\alpha_l} \geq 4.1 \alpha_l^{-1} \log n,$$

since $c_0 C \geq 4.1$ and $n\alpha_l \geq n\alpha_0 \geq n\frac{2}{n} > 1$. Furthermore,

$$\frac{T}{2048\sqrt{2}\alpha_l \log n} = \alpha_l^{-1} \frac{c_0 C n \log^3 n \sqrt{\theta/n}}{2048\sqrt{2} \log n} = \omega(\alpha_l^{-1} \log n).$$

Next, we bound the exponent $\frac{T}{32M\alpha_l}$. As $M \leq \log n$, we have

$$\frac{T}{32M\alpha_l} \geq \frac{c_0 C n \log^2 n \sqrt{\theta/n}}{32\alpha_l \log n} = \alpha_l^{-1} \frac{c_0 C}{32} \sqrt{\theta n} \log n \geq 4.1 \alpha_l^{-1} \log n,$$

provided that $c_0 C/32 \geq 4.1$, since $\theta n \geq 1$.

Finally, we bound the exponent $\frac{\tau_k T}{64M\alpha_l\theta}$. By definition $\frac{\tau_k^2}{8\alpha_l\theta} \geq 8\log n$ and $M \leq \log n$ thus

$$\frac{\tau_k T}{64M\alpha_l\theta} \geq \frac{8\sqrt{\alpha_l\theta\log n}T}{64\log n\alpha_l\theta} = \alpha_l^{-1}\frac{c_0 C}{8}\sqrt{n\alpha_l}\log^{3/2} n = \omega(\alpha_l^{-1}\log n),$$

concluding the proof.

F. Extension from Rademacher to general sub-gaussian variables

We introduce the truncation operator $T_\tau : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ as

$$(T_\tau[M])_{ij} = \begin{cases} M_{ij} & |M_{ij}| \leq \tau \\ 0 & else \end{cases}$$

Let $\tau = \sqrt{C\log n}$ and let

$$X' = T_\tau[X].$$

For C sufficiently large, the probability that $X' = X$ is $1 - o(1)$. This allows us to work with random matrix whose entries are bounded by τ (instead of 1 as in the Rademacher case). The same proof will go through if we increase p by $C_1\tau$, for a sufficiently large constant C_1. This means $p = O(n\log^{3.5} n)$ suffices. We round 3.5 up to 4 for a cosmetic reason.

G. Concluding remarks

There is a connection between the method of our proof and Fernique's chaining argument [5] (see [17] for a survey). The goal of the chaining method is to bound the supremum $\sup_{t\in B} X_t$ where B is a domain in a metrics space and X_t is a Gaussian process. In this case, the bad event $Bad(v)$ can roughly be defined as $X_v \geq M_v$, for some candidate value M_v. One then considers a chain of sets in order to bound $\mathbf{P}(\cup_{v\in B} Bad(v))$. This, in spirit, is similar to the purpose of Lemma III.3.

After this, the arguments become different in all aspects. First, in our setting, the bad event $Bad(v)$ can have any nature. Next, in the chaining argument, the sets $\mathcal{N}_j$ are defined using the metrics of B, while in our case, it is crucial to use a different metrics. We construct $\mathcal{N}_j$ using the l_∞ norm, rather than the natural l_1 norm used to define the domain B. Finally, in the chaining case it is easy to bound $\mathbf{P}(Bad(u)\backslash Bad(v))$, using the fact that $\mathbf{P}(|X_u - X_v| \geq t) \leq 2\exp(-\frac{t^2}{dist(u,v)^2})$, which is the basic property of a Gaussian process. In our case, bounding $\mathbf{P}(Bad(u)\backslash Bad(v))$ is an essential step (Lemma IV.2), which requires the development of the refined Bernstein's inequality.

Concentration results similar to that of Theorem I.1 appear often in the context of l_p subspace embeddings and randomized numerical linear algebra (cf. [14] and the refrences therein). Our techniques readily generalize to other norms besides the l_1 norm and so may be of use in these contexts as well.

V. The algorithm and concentration of random matrices

As the algorithm and analysis are discussed extensively in [15], we will be brief and the readers can consult [15] for more details. [15] introduces the dictionary learning algorithm ER-SpUD. The key insight in the design of ER-SpUD is that the rows of $\mathbf{X}$ are likely to be the sparsest vectors in the row space of $\mathbf{Y}$. (This observation also appeared [21] and [11].) [15] proposed to find these vectors by considering the following optimization problems.

$$\text{minimize } \|w^T\mathbf{Y}\|_1 \text{ subject to } r^T w = 1$$

where r is a row of two columns of $\mathbf{Y}$.

Using l_1 optimization for finding sparse vectors is a natural idea, and the authors of [15] pointed out that such an approach was already proposed in [13] and [8]. The difference is the new constraint $r^T w = 1$. (Earlier works used different constraints.)

By a change of variables $z = A^T w$, $b = A^{-1} r$, we can consider the equivalent problem

$$\text{minimize } \|z^T\mathbf{X}\|_1 \text{ subject to } b^T z = 1. \tag{V.1}$$

The algorithm presented in [15] is outlined below (for those familiar with [15], note that we are presenting the two-column version of ER-SpUD):

Algorithm 1 ER-SpUD

Randomly pair the columns of Y into $p/2$ groups $g_j = \{Ye_{j_1}, Ye_{j_2}\}$
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 $min_w \|w^T \mathbf{Y}\|_1$ subject to $(\mathbf{Y}r_j)^T w = 1$, and set $s_j = w^T \mathbf{Y}$.
Use Greedy algorithm to reconstruct $\mathbf{X}$ and A.

Algorithm 2 Greedy

Require: $S = \{s_1, \ldots, s_T\} \subset \mathbb{R}^p$
For $i = 1 \ldots n$
  REPEAT
    $l \leftarrow arg\ min_{s_l \in S} \|s_l\|_0$, breaking ties arbitrarily
    $x_i = s_l$
    $S = S \backslash \{s_l\}$
  UNTIL $rank([x_1, \ldots, x_i]) = i$
Set $\mathbf{X} = [x_1, \ldots, x_i]^T$, and $A = \mathbf{Y}\mathbf{Y}^T(\mathbf{X}\mathbf{Y}^T)^{-1}$

A key technical step in analyzing ER-SpUD is the following lemma, which asserts that if p is sufficiently large, then with high probability $\|X^T v\|_1$ is close to its mean, simultaneously for all unit vectors $v \in \mathbb{R}^n$.

Lemma V.1. *For every constant $1 \geq \delta > 0$ there is a constant $C_0 > 0$ such that the following holds. If $\theta \geq \frac{1}{n}$ and $p \geq C_0 n^2 \log^2 n$, then with probability $1 - o(1)$, for all $v \in \mathbb{R}^n$*

$$|\|X^T v\|_1 - \mathbf{E}\|X^T v\|_1| \leq \delta \mathbf{E}\|X^T v\|_1. \tag{V.2}$$

This lemma appears implicitly in [15]. Dan Spielman pointed out to us that this would imply the critical [15, Lemma 17]. The bound $p \geq Cn^2 \log^2 n$ is of importance in the proof of this lemma.

Our Theorem I.1, which pushes p to $Cn \log^4 n$, is an improved version of Lemma V.1.

With Theorem I.1 in hand, let us now sketch the proof of Theorem II.3, following the analysis in [15].

Notice that if the solution of the l_1 optimization problem, z_*, is 1-sparse, then the algorithm will recover a row of X. The proof of the theorem relies on showing that z_*, is supported on the non-zero indices of b and that with high-probability, z_* is in fact 1-sparse. The first goal allows us to focus our attention on a submatrix of $\mathbf{X}$ which will be convenient for technical reasons. To address this first issue, we prove the following.

Lemma V.2. *Suppose that $\mathbf{X}$ satisfies the Bernoulli-Subgaussian model. There exists a numerical constant $C > 0$ such that if $\theta n \geq 2$ and*

$$p > Cn \log^4 n$$

then the random matrix $\mathbf{X}$ has the following property with probability at least $1 - o(1)$.

***(P1)** For every b satisfying $\|b\|_0 \leq 1/8\theta$, any solution z_* to the optimization problem V.1 has $supp(z_*) \subseteq supp(b)$.*

Sketch of the Proof of Lemma V.2. We let J be the indices of the s non-zero entries of b. Let S be the indices of the nonzero columns in $\mathbf{X}_J$, and let $z_0 = P_J z_*$ (the restriction to those coordinates indexed by J). Define $z_1 = z_* - z_0$. We demonstrate that z_0 has at least as low an objective as z_* so z_1 must be zero. One can show using the triangle inequality that

$$\|z_*^T \mathbf{X}\|_1 \geq \|z_0^T \mathbf{X}\|_1 - 2\|z_1^T \mathbf{X}^S\|_1 + \|z_1^T \mathbf{X}\|_1.$$

Thus, if $\|z_1^T \mathbf{X}\|_1 - 2\|z_1^T \mathbf{X}^S\|_1 > 0$, then z_0 has a lower objective value. We need this inequality to hold for all z with high probability. Notice that

$$\mathbf{E}[\|z^T \mathbf{X}\|_1 - 2\|z^T \mathbf{X}^S\|_1] = (p - 2|S|)\mathbf{E}|z^T \mathbf{X}_1|$$

It is easy to show that $|S| < p/4$ with high probability so $(p - 2|S|) > 0$ with high probability. Therefore, if we can show that $\|z^T \mathbf{X}\|_1 - 2\|z^T \mathbf{X}^S\|_1$ is concentrated near its positive expectation we are done.

We see that it suffices to show the result for the worst case $|S| = p/4$. Now we make critical use of Theorem I.1, which asserts that with high probability,

$$\|z^T\mathbf{X}\|_1 \geq \frac{5}{8}\mathbf{E}\|z^T\mathbf{X}\|_1 = \frac{5p}{8}\mathbf{E}|z^T\mathbf{X}_1|.$$

and

$$\|z^T\mathbf{X}^S\|_1 \leq \frac{1}{2}\mathbf{E}\|z^T\mathbf{X}^S\|_1 = \frac{p}{8}\mathbf{E}|z^T\mathbf{X}_1|.$$

so

$$\|z^T\mathbf{X}\|_1 - 2\|z^T\mathbf{X}^S\|_1 \geq \frac{p}{2}\mathbf{E}|z^T\mathbf{X}_1| > 0.$$

Having proved Lemma V.2, the rest of the proof is relatively simple and follows [15] exactly. The success of the algorithm now depends on the existence of a sufficient gap between the largest and second largest entry in b. The intuition is that if $\mathbf{X}$ preserved the l_1 norm exactly, i.e. $\|z^T\mathbf{X}\|_1 = c\|z\|_1$, then the minimization procedure will output the vector z of smallest l_1 norm such that $b^Tz = 1$, which is just e_{j_*}/b_{j_*}, where j_* is the index of the element of b with the largest magnitude. However, $\mathbf{X}$ only preserves the l_1 norm in an approximate sense. Yet, the algorithm will still extract a column of $\mathbf{X}$ if there is a significant gap between the largest element of b and the second largest.

VI. Rectangular dictionaries and Theorem II.4

We now present a generalization of ER-SpUD, which enables us to deal with a rectangular dictionary. Consider a full rank matrix A of size $n > m$, such that $n > m$, and the equation $AX = Y$. To deal with this setting, we first augment A to be a square, $n \times n$, invertible matrix. Of course, the issue is that one does not know A, and also need to figure out how the augmentation changes the product Y.

We can solve this issue using a random augmentation. For instance, we can use $n \times (n-m)$ gaussian matrix B to augment A to a square matrix A' (the entries in B are iid standard gaussian). It is trivial that the augmented matrix has full rank with probability 1, since the probability that a gaussian vector belongs to any fixed hyperplane is zero. We can also augment $\mathbf{X}$ from an $m \times p$ matrix to a $n \times p$ matrix, $\mathbf{X}'$ by an $(n-m) \times p$ random matrix Z with entries iid to those of $\mathbf{X}$. This augmentation process yields a matrix equation

$$Y' = A'X'$$

where $\mathbf{Y}' = \mathbf{Y} + E$ where $E = BZ$ (Figure 1). In practice, we can first generate B, Z, then compute $E := BZ$ and construct $\mathbf{Y}' := Y + E'$. Next then apply the ER-SpUD algorithm to the equation $Y' = A'X'$ to recover A' and $\mathbf{X}'$ with high probability. From these two matrices, we can then deduce A and $\mathbf{X}$.

Using a gaussian (or any continuous) augmentation is convenient, as the resulting matrix is obviously full rank. However, it is, in some way, a cheat. Apparently, a gaussian number does not have any finite representation, thus it takes forever to read the input, let alone process it. A common practice is to truncate (as a matter of fact, the computer only generates a finite approximation of the gaussian numbers anyway), and hope that the truncation is fine for our purpose. But then we face a non-trivial theoretical question to analyze this approximation. How many decimal places are enough ? Even if we can prove a guarantee here, using it in practice would require computing with a matrix with many long entries, which significantly increases the running time.

We can avoid this problem by using random matrices with discrete distributions, such as ± 1. The technical issue now is to prove the full rank property. This is a highly non-trivial problem,but luckily was taken care of in the following result of Bourgain, Vu, and Wood [3].

Theorem VI.1. *For every $\epsilon > 0$ there exists $\delta > 0$ such that the following holds. Let $N_{f,n}$ be an n by n complex matrix in which f rows contain fixed, non-random entries and where the other rows contain entries that are independent discrete random variables. If the fixed rows have co-rank k and if for every random entry α, we have $max_x\mathbf{P}(\alpha = x) \leq 1 - \epsilon$, then for all sufficiently large n*

$$\mathbf{P}(N_{f,n} \textit{ has co-rank} > k) \leq (1-\delta)^{n-f}.$$

Letting, $k = 0$ and $f = m$, the result shows that if we augment A by $n \times (m-n)$ random Bernoulli matrix, this new matrix, A', will be nonsingular with high probability, given that $n - m = \omega(1)$.

We summarize our reasoning in the following algorithm.

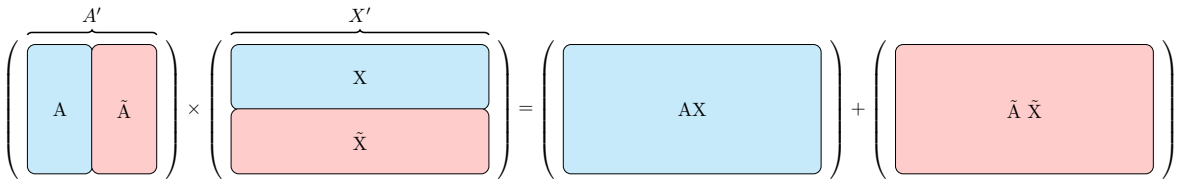

Figure 1. Rectangular A with $n > m$

Algorithm 3 Rectangular Algorithm

1: Generate a $(n-m) \times p$ matrix Z with iid random variables that agree with the model for X.
2: Generate a $n \times (n-m)$ matrix B with iid entries (either Gaussian or Rademacher).
3: Run ER-SpUD on $Y' = Y + BZ$
4: Remove the rows of A' and the columns of X' from the output of ER-SpUD.

VII. Optimal bound for very sparse random matrices

In this section, we discuss Theorem II.5. We present a simple algorithm (see below) and use this algorithm to prove Theorem II.5, obtaining the optimal bound $p = Cn \log n$.

Algorithm 4 Very-sparse Algorithm

1: Partition the columns of Y into a minimum number of groups G_i whose members are multiples of each other.
2: Choose representatives of those G_i with more than two members to be the columns of A up to scaling.

Proof of Theorem II.5. Since A is nonsingular, any two columns of Y that are multiples of each other must be linear combinations of the same columns of A. For a group G_i to have more than two members would require that there be more than two columns in X with their non-zero entries in the same rows.

Definition VII.1. We say that a set of columns are aligned if they each have more than one nonzero entry and their non-zero entries occur in the same positions.

Lemma VII.2. *The probability that X has more than two aligned columns is $o(1)$.*

Thus, the algorithm is likely to yield only columns of A. We now need to show that all the columns of A will be outputted with high probability.

Definition VII.3. We say the column $\mathbf{a}$ of A is k-represented if some group G_i consists of multiples of $\mathbf{a}$ and $|G_i| = k$. In particular, if no multiple of the jth column, $\mathbf{a_j}$, shows up in the columns of Y then $\mathbf{a_j}$ is 0-represented. A column is well represented if it is k-represented for $k > 2$.

Notice that the algorithm will output a multiple of every column that is well represented.
The following lemma finishes the proof of Theorem II.5.

Lemma VII.4. *The probability that every column $\mathbf{a_i}$ is well represented is $1 - o(1)$.*

A. Proofs of Sparse Algorithm

Proof of Lemma VII.2. Given the choice of θ, we know that the number of nonzero entries in any column of X will converge to the Poisson distribution. We ignore the $o(1/n)$ error terms from this approximation in later calculations to alleviate clutter. To calculate the probability, we condition on the number of nonzero entries, and then we bound the probability that three specific columns have the required property, and finally we use the union bound. This yields an upper bound of

$$\binom{n}{3} \sum_{k \geq 2} \frac{e^{-3c}}{(k!)^3} \frac{1}{\binom{n}{k}^2} = o(1)$$

■

Proof of Lemma VII.4. By the union bound,

$$\mathbf{P}(\exists i \text{ such that } \mathbf{a_i} \text{ is not well represented}) \leq n\mathbf{P}(\mathbf{a_1} \text{ is not well represented})$$

Partitioning into disjoint events yields

$$\mathbf{P}(\mathbf{a_1} \text{ is not well represented}) = \sum_{j=0}^{2} \mathbf{P}(\mathbf{a_1} \text{ is } j\text{-represented})$$

Notice that a multiple of $\mathbf{a_1}$, say $a * \mathbf{a_1}$, appears as a column of Y if and only if $a * \mathbf{e_1} = (a, 0, 0, \ldots, 0)^T$, with $a \neq 0$, is X^j, the jth column of X, for some j. Now, using the Poisson approximation we can bound each term in the summand. For example, for the probability of being 0-represented, we can divide into the case that X^i does not have exactly one non-zero element and the case that X^i has exactly one non-zero term but not in the first row. We use C to indicate an absolute constant which may change with each appearance.

$$\mathbf{P}(\mathbf{a_1} \text{ is 0-represented}) \leq \left((1 - ce^{-c}) + e^{-c}\frac{n-1}{n}\right)^p \leq C\exp(-Cp/n)$$

Similarly,

$$\mathbf{P}(\mathbf{a_1} \text{ is 1-represented}) \leq n\left(\frac{ce^{-c}}{n}\right)\left((1 - ce^{-c}) + e^{-c}\frac{n-1}{n}\right)^{p-1} \leq C\exp(-Cp/n)$$

and

$$\mathbf{P}(\mathbf{a_1} \text{ is 2-represented}) \leq \binom{n}{2}\left(\frac{ce^{-c}}{n}\right)^2\left((1 - ce^{-c}) + e^{-c}\frac{n-1}{n}\right)^{p-2} \leq C\exp(-Cp/n)$$

Thus,

$$\mathbf{P}(\mathbf{a_1} \text{ is not well represented}) \leq C\exp(\log n - Cp/n) = o(1)$$

for $p = C'n\log n$ for a large enough C'. ■

VIII. Numerical Simulations

We demonstrate that the efficiency of the ER-SpUD algorithm is not improved with larger p values beyond the threshold conjectured. In Figure 2, we have chosen A to be an $n \times n$ matrix of independent $N(0,1)$ random variables. The $n \times p$ matrix X has k randomly chosen non-zero entries which are Rademacher. The graph on the left of Figure 2 is generated with $p = 5n\log n$ and the one on the right with $p = 5n^2\log^2 n$. For both graphs, n varies from 10 to 60 and k from 1 to 10. Accuracy is measured in terms of relative error:

$$re(A', A) = min_{\Pi,\Lambda}\|A'\Lambda\Pi - A\|_F/\|A\|_F$$

The average relative error over ten trials is reported.

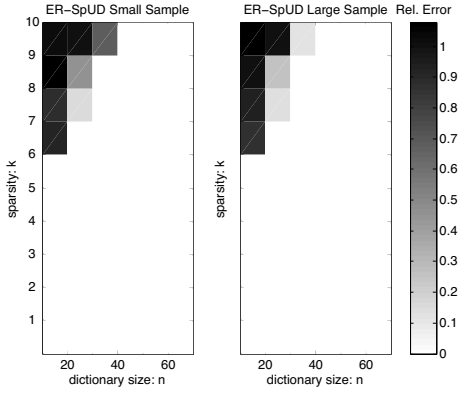

Figure 2. Mean relative errors of ER-SpUD with $p = 5n\log n$ versus $p = 5n^2\log^2 n$

We then ran our Algorithm VII in a sparse regime to compare its performance with that of ER-SpUD (see Figure 3. A was as before, but since our algorithm relies on the appearance of 1-sparse columns in X, we cannot fix sparsity as in our first experiments. Rather, we vary the Bernoulli parameter θ from 0.02 to 0.18, and the χ_{ij} are Rademacher. One can see the expected phase transition at which point the matrix X is no longer sparse enough for our algorithm. In the regime for which the algorithm was designed, the relative error of our output is on the same order as that of ER-SpUD. Furthermore, our algorithm runs much quicker and has no trouble with inputs of size up to $n = 500$. (The numerical experiments were completed on a Macbook Pro.)

Finally, we compare the outcome of our optimal p value with that of a much larger sample size ($p = O(n^2 \log^2 n)$). We let n range from 10 to 200 and θ from 0.01 to 0.08. Figure 4 shows that the efficacy of the algorithm is not much improved despite the dramatic increase in p. The threshold for failure is identical.

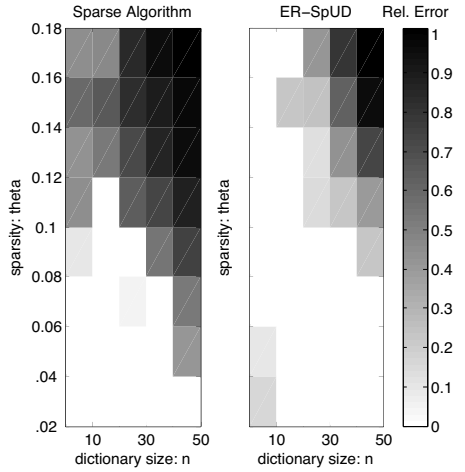

Figure 3. Mean relative errors with varying sparsity θ. Here, $p = 5n \log n$.

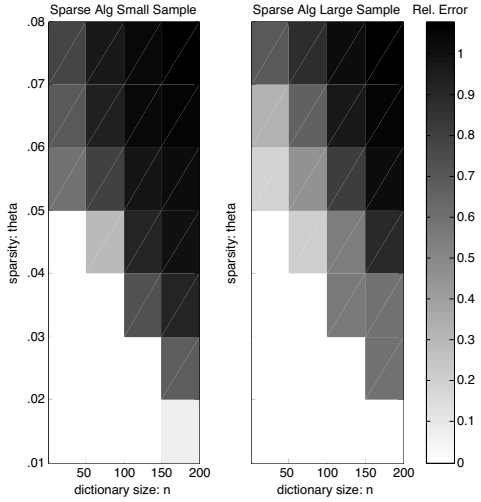

Figure 4. Mean relative errors of Algorithm VII with $p = 5n \log n$ versus $p = 5n^2 \log^2 n$

REFERENCES

[1] Michal Aharon, Michael Elad, and Alfred Bruckstein. The k-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54(11):4311–4322, 2006.

[2] Michal Aharon, Michael Elad, and Alfred M Bruckstein. On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them. *Linear algebra and its applications*, 416(1):48–67, 2006.

[3] Jean Bourgain, Van H Vu, and Philip Matchett Wood. On the singularity probability of discrete random matrices. *Journal of Functional Analysis*, 258(2):559–603, 2010.

[4] Kjersti Engan, Sven Ole Aase, and J Hakon Husoy. Method of optimal directions for frame design. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 5, pages 2443–2446. IEEE, 1999.

[5] X Fernique. Regularite de processus gaussien. In *Invent Math.*, pages 304–321. 1971.

[6] Pando Georgiev, Fabian Theis, and Andrzej Cichocki. Blind source separation and sparse component analysis of overcomplete mixtures. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 5, pages V–493. IEEE, 2004.

[7] Lee-Ad Gottlieb and Tyler Neylon. Matrix sparsification and the sparse null space problem. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 205–218. Springer, 2010.

[8] Florent Jaillet, Rémi Gribonval, Mark D Plumbley, and Hadi Zayyani. An l1 criterion for dictionary learning by subspace identification. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5482–5485. IEEE, 2010.

[9] Kenneth Kreutz-Delgado, Joseph F Murray, Bhaskar D Rao, Kjersti Engan, Te-Won Lee, and Terrence J Sejnowski. Dictionary learning algorithms for sparse representation. *Neural computation*, 15(2):349–396, 2003.

[10] Liangyue Li, Sheng Li, and Yun Fu. Discriminative dictionary learning with low-rank regularization for face recognition. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[11] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.

[12] Bruno A Olshausen et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[13] Mark D Plumbley. Dictionary learning for l1-exact sparse coding. In *Independent Component Analysis and Signal Separation*, pages 406–413. Springer, 2007.

[14] Christian Sohler and David P Woodruff. Subspace embeddings for the l 1-norm with applications. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 755–764. ACM, 2011.

[15] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 3087–3090. AAAI Press, 2013.

[16] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere. *arXiv preprint arXiv:1504.06785*, 2015.

[17] Michel Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, pages 1049–1103, 1996.

[18] John Wright, Allen Y Yang, Arvind Ganesh, Shankar S Sastry, and Yi Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[19] Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.

[20] Michael Zibulevsky. Blind source separation with relative newton method. In *Proc. ICA*, volume 2003, pages 897–902, 2003.

[21] Michael Zibulevsky and Barak A Pearlmutter. Blind source separation by sparse decomposition. In *AeroSense 2000*, pages 165–174. International Society for Optics and Photonics, 2000.