An improved analysis of the ER-SpUD dictionary learning algorithm

Jarosław Błasiok* Jelani Nelson†

February 19, 2016

Abstract

In *dictionary learning* we observe $Y = AX + E$ for some $Y \in \mathbb{R}^{n\times p}$, $A \in \mathbb{R}^{m\times n}$, and $X \in \mathbb{R}^{m\times p}$, where $p \geq \max\{n, m\}$, and typically $m \geq n$. The matrix Y is observed, and A, X, E are unknown. Here E is a "noise" matrix of small norm, and X is column-wise sparse. The matrix A is referred to as a *dictionary*, and its columns as *atoms*. Then, given some small number p of samples, i.e. columns of Y, the goal is to learn the dictionary A up to small error, as well as the coefficient matrix X. In applications one could for example think of each column of Y as a distinct image in a database. The motivation is that in many applications data is expected to sparse when represented by atoms in the "right" dictionary A (e.g. images in the Haar wavelet basis), and the goal is to learn A from the data to then use it for other applications.

Recently, the work of [SWW12] proposed the dictionary learning algorithm **ER-SpUD** with provable guarantees when $E = 0$ and $m = n$. That work showed that if X has independent entries with an expected θn non-zeroes per column for $1/n \lesssim \theta \lesssim 1/\sqrt{n}$, and with non-zero entries being subgaussian, then for $p \gtrsim n^2 \log^2 n$ with high probability **ER-SpUD** outputs matrices A', X' which equal A, X up to permuting and scaling columns (resp. rows) of A (resp. X). They conjectured that $p \gtrsim n \log n$ suffices, which they showed was information theoretically necessary for *any* algorithm to succeed when $\theta \simeq 1/n$. Significant progress toward showing that $p \gtrsim n \log^4 n$ might suffice was later obtained in [LV15].

In this work, we show that for a slight variant of **ER-SpUD**, $p \gtrsim n \log(n/\delta)$ samples suffice for successful recovery with probability $1 - \delta$. We also show that without our slight variation made to **ER-SpUD**, $p \gtrsim n^{1.99}$ samples are required even to learn A, X with a small success probability of $1/\mathit{poly}(n)$. This resolves the main conjecture of [SWW12], and contradicts a result of [LV15], which claimed that $p \gtrsim n \log^4 n$ guarantees high probability of success for the original **ER-SpUD** algorithm.

1 Introduction

The *dictionary learning* or *sparse coding* problem is defined as follows. There is a hidden set of vectors $a_1, a_2, \ldots a_m \in \mathbb{R}^n$ (called a "dictionary"), with $\mathit{span}\{a_1, \ldots a_m\} = \mathbb{R}^n$. We are given a sequence of samples $y_i = Ax_i + \epsilon_i$, where each x_i is a sparse vector and ϵ_i is noise. In other words

*Harvard University, Cambridge, MA. jblasiok@g.harvard.edu. Supported by NSF grant IIS-1447471.

†Harvard University, Cambridge, MA. minilek@seas.harvard.edu. Supported by NSF grant IIS-1447471 and CAREER award CCF-1350670, ONR grant N00014-14-1-0632 and Young Investigator award N00014-15-1-2388, and a Google Faculty Research Award.

ref	sample complexity	noise	overcomplete	sparsity	arbitrary dict.
[SWW12]	$\mathcal{O}(n^2 \log^2 n)$	No	No	$\mathcal{O}(\sqrt{n})$	Yes
[AAJ+14]	$\mathcal{O}(m^2)$	No	Yes	$\mathcal{O}(n^{1/4})$	No
[AGM14]	$\mathcal{O}(m^2 s^{-2} + s^2 m)$	Yes	Yes	$\mathcal{O}(\min(m^{2/5}, \frac{\sqrt{n}}{\log n}))$	No
[AGM14]	$\mathcal{O}(\text{poly}(m))$	Yes	Yes	$\mathcal{O}(n^{1/2-\epsilon})$	No
[ABGM14]*	$\mathcal{O}(\text{poly}(m))$	No	Yes	$\mathcal{O}(n/\text{polylog}(n))$	No
[BKS15]	$\mathcal{O}(\text{poly}(m))$	Yes	Yes	$\mathcal{O}(n^{1-\epsilon})$	Yes
[BKS15]*	$\mathcal{O}(\text{poly}(m))$	Yes	Yes	$\mathcal{O}(n)$	Yes
[SQW15]	$\mathcal{O}(\text{poly}(m, \kappa(A))))$	No	No	$\mathcal{O}(n)$	Yes
[VX15]	$\mathcal{O}(\text{poly}(n))$	Yes	No	$\mathcal{O}(n)$	Yes
[LV15][1]	$\mathcal{O}(n \log^4 n)$	No	No	$\mathcal{O}(\sqrt{n})$	Yes
This work	$\mathcal{O}(n \log n)$	No	No	$\mathcal{O}(\sqrt{n})$	Yes

Figure 1: Comparison of algorithms with proven guarantees for dictionary learning. Last column indicates whether the dictionary can be arbitrary, or if additional structure is assumed in order to guarantee recovery. Algorithms marked with star require quasi-polynomial running time. $\kappa(A)$ denotes condition number.

each y_i is close to a linear combination of few vectors a_k. The goal is to recover both matrix A and the sparse representations x_i. We can write it as a matrix equation

$$Y = AX + E$$

where the vectors y_i are the columns of Y, and x_i are columns of X. Let $A \in \mathbb{R}^{n\times m}$ and $X \in \mathbb{R}^{m\times p}$. Traditionally, and as motivated by applications, the interesting regime of parameters is when A is of full row rank (in particular $n \leq m$) [AAJ+14].

The dictionary learning problem is motivated by the intuition that the dictionary A is in some sense the "right" spanning set for representing vectors y_i since it allows sparse representation. In some domains this correct basis is known thanks to a deep understanding of the domain in question: for example the Fourier basis for audio processing, or Haar wavelets for images. Here we want to infer analogous "nice" representations of the data from the data itself. As it turns out, even in situations such as audio and image processing in which traditional transforms are useful, replacing them with dictionaries learned directly from data turned out to improve quality of the solution (see for example [EA06], which applied a dictionary learning algorithm for image denoising).

This problem has found a tremendous number of applications in various areas, such as image and video processing (e.g. [MBP+09, BE08, EA06]; see [MBP14] for more references), image classification [RBL+07, MBP+08] as well as neurobiology [LYB+14]. Given its huge practical importance, a number of effective heuristics for dictionary learning were proposed [AEB06, MBPS10] — those are based on iterative methods for solving the (non-convex) optimization problem of minimizing the sparsity of X' subject to Y being close to $A'X'$. Some of these algorithms work well in practice but without provable guarantees.

1.1 Prior work

[1]As written, their work has certain errors which we discuss later in detail. Nevertheless, using some of our

Until recently there was little theoretical understanding of the dictionary learning problem. Spielman, Wang and Wright in [SWW12] proposed the first algorithm that provably solves this problem in some regime of parameters. More concretely, they assumed no presence of noise (i.e. $E = 0$), and that A is a basis (that is $n = m$), potentially adversarially chosen. The vectors x_i are sampled independently at random from some distribution — specifically, each entry $x_{i,j}$ is nonzero with probability $1 - \theta$, and once it is nonzero, it is a symmetric subgaussian random variable (i.e. with tails decaying at least as fast as a gaussian), independent from every other entry. Henceforth we say that a matrix $X \in \mathbb{R}^{n\times p}$ follows the *Bernoulli-subgaussian model* with parameter θ, if the entries $X_{i,j}$ are i.i.d. with $X_{i,j} = \chi_{i,j} g_{i,j}$, where $\chi_{i,j} \in \{0, 1\}$ are Bernoulli random variables with $\mathbb{E}\,\chi_{i,j} = \theta$, and $g_{i,j}$ are symmetric subgaussian random variables. We also say that X follows the Bernoulli-Rademacher model if $g_{i,j}$ in the above definition are independent Rademachers (i.e. uniform ± 1).

Under the Bernoulli-subgaussian model for X, [SWW12] proved that once the number of samples p is $\Omega(n \log n)$ and the sparsity $s = \theta n$ (i.e. expected number of nonzero entries in each column of X) is at least constant and at most $\mathcal{O}(n)$, the matrix Y with high probability has a unique decomposition as a product $Y = AX$, up to permuting and rescaling rows of X. Moreover, the number of samples $p = \Omega(n \log n)$ was proven to be optimal in the constant sparsity regime $s = \Theta(1)$. In particular, it is possible in principle to find such a decomposition information-theoretically, but unfortunately not necessarily with an efficient algorithm.

In addition to the above, they proposed an efficient algorithm **ER-SpUD** (*Efficient Recovery of Sparsely Used Dictionaries*) to find this unique decomposition, in a more restricted regime of parameters. Namely, they proposed an algorithm and proved that it finds correctly the unique decomposition $Y = AX$, with high probability over X, as long as the sparsity s is at least constant and at most $\mathcal{O}(\sqrt{n})$, and the number of samples p is at least $\Omega(n^2 \log^2 n)$. The low sparsity constraint was inherent to their solution: according to the proof in the same paper, if $s = \Omega(\sqrt{n \log n})$ the algorithm with high probability fails to find the correct decomposition. They conjectured however, that with the number of samples p as small as $\mathcal{O}(n \log n)$, **ER-SpUD** should return the correct decomposition with high probability, matching the sample lower bound for when $s = O(1)$.

Since then, much more theoretical work has been dedicated to the dictionary learning problem; see Figure 1. In the work of Agarwal et al. [AAJ$^+$14], and independently Arora et al. [AGM14], an algorithm was proposed that works for overcomplete dictionaries A (i.e. when $m > n$), under additional structural assumptions on A — namely that A is *incoherent*, i.e. the projection of any standard basis vector onto the column space of A has small norm. The algorithm presented in [AAJ$^+$14] require $p = \tilde{\mathcal{O}}(m^2)$ samples, where $\tilde{\mathcal{O}}(f) = \mathcal{O}(f \cdot \log^{O(1)}(f))$. More detailed analysis of the dependence between sparsity and number of samples was provided in the work [AGM14] for their algorithm — for $s = \mathcal{O}(\min(\frac{\sqrt{n}}{\log n}, m^{2/5}))$, they require $\tilde{\Omega}(m^2 s^{-2} + ms^2)$ samples; if s is larger than $m^{2/5}$, but smaller than $\mathcal{O}(\min(m^{1/2-\varepsilon}, \frac{\sqrt{n}}{\log n}))$ the algorithm require $\mathcal{O}(m^C)$ samples, where C is a large constant depending on ε. In the lowest sparsity regime, i.e. $s = \mathcal{O}(\mathrm{polylog}(n))$, the sample complexity stated in their analysis simplifies to $\tilde{\Omega}(m^2)$, for comparison in the most favorable sparsity regime $s = \Theta(m^{1/4})$, the number of samples necessary for correct recovery is $\Omega(m^{3/2})$. The work [AGM14] also proves correct recovery by this algorithm in the presence of noise. Later Arora et al. [ABGM14] gave a quasipolynomial time algorithm working for sparsity up to $\mathcal{O}(n/\mathrm{polylog}(n))$, but

approaches we believe it should be possible to salvage their sample complexity bound in the Bernoulli-gaussian model for X, but not in the more general Bernoulli-subgaussian model (since in particular, $p \gtrsim n^{1.99}$ samples are required for that algorithm even to succeed with polynomially small *success* probability; see Section A.

under much stronger assumptions on the structure of A. Those assumptions include in particular, that the dictionary A itself is assumed to be sparse, which is violated in many natural examples, e.g. the discrete Fourier basis. They prove that their algorithm correctly recovers the hidden dictionary given access to $p = \mathcal{O}(m^C)$ samples, for some unspecified constant C.

Barak et al. [BKS15] proposed an algorithm fitting in the Sum-of-Squares framework, which works in polynomial time for sparsity $\mathcal{O}(n^{1-\epsilon})$ and in quasipolynomial time for sparsity as large as $\mathcal{O}(n)$, again given access to $\mathcal{O}(m^C)$ samples for some unspecified constant C. Moreover, this algorithm works under the presence of noise and a more general model of X. In particular, coordinates within a single column are not required to be fully independent. Recently, Sun et al. [SQW15] proposed a polynomial time algorithm for the case when $n = m$ and sparsity is as large as $\mathcal{O}(n)$. Their result works in the similar model as in [SWW12], without any additional assumptions on the matrix A, and with matrix X having independent entries that are product of Bernoulli and gaussian random variables (as opposed to the weaker subgaussian assumption in [SWW12]). The sample complexity depends polynomially on n and the condition number of the dictionary matrix A. In particular, in the low sparsity regime ($s = \Theta(\mathrm{polylog}(n))$), this sample complexity is as large as $\tilde{\Omega}(n^9)$ even if the matrix A is well conditioned.

Work on Independent Component Analysis (ICA) [FJK96, NR09, BRV13, AGMS15, GVX14, VX15] is also relevant to the dictionary learning problem. In this problem, again one is given $Y = AX + E$ for square A, with the assumption that the entries of X are i.i.d. (and X need not necessarily be sparse). The works in ICA then say that A, X can be efficiently recovered using few samples, but where the sample complexity depends on the distribution of entries of X. For example in the case of Bernoulli-Rademacher entries with $\theta = 1/n$ (constant sparsity per column of X), these works require large polynomial sample complexity. For example, [VX15, Theorem 1] implies a sufficient sample complexity in this setting of $p \gg n^{12}$.

From Figure 1, one can see that the "holy grail" of dictionary learning is to achieve the following features simultaneously: (1) low sample complexity, i.e. nearly-linear in the dimension n and number of atoms m, (2) the ability to handle noise (the more noise handled the better), (3) handling overcomplete dictionaries, (4) handling a larger range of sparsity, with $s = O(n)$ being the best, (5) making no assumptions on the dictionary A, (6) a fast algorithm to actually learn the dictionary from samples, and (7) making few assumptions on the matrix X.

Most of the aforementioned results focus on weakening the sparsity constraint under which it is possible to perform efficient learning, or handling overcomplete dictionaries or noise. These all, however, come at an expense: the number of samples necessary for those algorithms to provably work is quite large, often of order n^C for some large constant C. Some of the algorithms also make strong assumptions on A, and/or have quasi-polynomial running time.

Recently, Luh and Vu in [LV15] made significant progress toward showing that the **ER-SpUD** algorithm proposed in [SWW12] actually solves the dictionary learning problem already with $p = \mathcal{O}(n \log^4 n)$ samples. They claimed to prove that this p in fact suffices for dictionary learning. In fact however, several probabilistic events were analyzed in [SWW12], and if they all occurred then **ER-SpUD** performed correct recovery. The work [LV15] analyzed arguably the most complex of these events more efficiently, showing a certain crucial inequality held with good probability when $p \gtrsim n \log^4 n$. Unfortunately there is a gap: [SWW12] required this inequality to hold for exponentially many settings of variables, and thus one wants the inequality to hold for any fixed instantiation with very high probability to then union bound, and [LV15] does not provide such a probabilistic analysis (see Remark 5). More seriously, there are other events defined in [SWW12] which *require* $p \gtrsim n^2$ to hold whp in the Bernoulli-subgaussian model (except in the case the subgaussians are actual gaussians), and [LV15] did not discuss these events at all (see for example Remark 7). In fact, in Section A we prove that in the Bernoulli-Rademacher model the **ER-SpUD** algorithm of [SWW12] actually *requires* $p \gtrsim n^{1.99}$ to succeed with probability even polynomially small in n, contradicting the main result of [LV15] which claimed $1 - o(1)$ successful learning for p nearly linear in n.

Our contribution: We very slightly modify the algorithm **ER-SpUD** to obtain another polynomial-time dictionary learning algorithm "**ER-SpUD(DCv2)**", which circumvents our $p \gtrsim n^{1.99}$ lower bound for **ER-SpUD** in the Bernoulli-subgaussian model. We then show that **ER-SpUD(DCv2)** provides correct dictionary learning with probability $1 - \delta$ with sparsity $s = \mathcal{O}(\sqrt{n})$ as long as $p \gtrsim n \log(n/\delta)$. In particular our result shows that a slight modification of **ER-SpUD** provides correct dictionary learning for complete dictionaries with no noise, which provably works with high probability using $p \gtrsim n \log n$ samples. This resolves the main open problem of [SWW12].

Furthermore, the work of [LV15] observed that the method of their proof is connected to generic chaining, but that after a certain point the methods "become different in all aspects" [LV15, Section G]. They also advertised and proved a new "refined version of Bernstein's concentration inequality for a sum of independent variables". Unlike their work, our analysis has the benefit of using standard off-the-shelf concentration and chaining results, thus making the proof simpler and more easily accessible since it is less ad-hoc.

1.2 Approach overview

In Figure 2 we give the algorithm **ER-SpUD(DCv2)** analyzed in this work, a slight modification of **ER-SpUD(DC)** from [SWW12]. The only difference between DCv2 and the original DC variant in [SWW12] is that we try all $\binom{p}{2}$ pairings of columns, whereas DC tried a random pairing of the p columns into $p/2$ pairs. As we will see soon, one of the several conditions in [SWW12] necessary for their proof of successful recovery of (A, X) from Y actually requires $p = \Omega(n^2)$ if using the DC variant (see Remark 7), and hence our switch to DCv2 allows p to be reduced to $\mathcal{O}(n \log n)$. In any case, this issue is easily circumvented by switching to DCv2 as we shall soon justify.

Henceforth when we refer to **ER-SpUD**, we are referring to **ER-SpUD(DCv2)** unless we state otherwise.

The main insight in the recovery analysis of [SWW12] is that the last line of the **ER-SpUD** pseudocode in Figure 2 can be rewritten (only in the analysis, since A, X are unknown) as $\min_w \|w^T AX\|_1$ subject to $(A(Xe_{j_1} + Xe_{j_2}))^T w = 1$. Then writing $z = A^T w$, this linear program (LP) is equivalent to the secondary LP $\min_z \|z^T X\|_1$ subject to $b_j^T z = 1$, since we could recover $w = (A^T)^{-1} z$ since A is invertible. Here b_j denotes $Xe_{j_1} + Xe_{j_2}$. The ideal case then is that the only optimal solution to the second LP will be a vector z_* that is 1-sparse. In this case, the solution to the LP that we *actually* solve is equal to $w_* = (A^T)^{-1} z_* = (z_*^T A^{-1})^T$ and thus a scaled row of A^{-1}, implying $w_*^T Y$ is a scaled row of X. Thus, if z_* is 1-sparse in the second LP, then the solution to the first LP allows us to recover a scaled row of X.

The work [SWW12] then outlines certain conditions for X that, if they hold, guarantee correct recovery of (A, X). We now state these deterministic conditions, as per [SWW12], which imply correct recovery of (A, X) via **ER-SpUD** when they all simultaneously hold.

ER-SpUD(DCv2): `Exact Recovery of Sparsely-Used Dictionaries using the sum of two columns of` Y `as constraint vectors.`

1. Create all $T = \binom{p}{2}$ pairings of columns of $\boldsymbol{Y}$ and for $j \in [T]$ write $g_j = \{\boldsymbol{Y}\boldsymbol{e}_{j_1}, \boldsymbol{Y}\boldsymbol{e}_{j_2}\}$.

2. For $j = 1 \dots T$

   Let $\boldsymbol{r}_j = \boldsymbol{Y}\boldsymbol{e}_{j_1} + \boldsymbol{Y}\boldsymbol{e}_{j_2}$, where $g_j = \{\boldsymbol{Y}\boldsymbol{e}_{j_1}, \boldsymbol{Y}\boldsymbol{e}_{j_2}\}$.
   Solve $\min_{\boldsymbol{w}} \ \|\boldsymbol{w}^T\boldsymbol{Y}\|_1$ subject to $\boldsymbol{r}_j^T\boldsymbol{w} = 1$, and set $\boldsymbol{s}_j = \boldsymbol{w}^T\boldsymbol{Y}$.

Greedy: `A Greedy Algorithm to Reconstruct` $\boldsymbol{X}$ `and` $\boldsymbol{A}$`.`

1. **REQUIRE:** $\mathcal{S} = \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_T\} \subset \mathbb{R}^p$.

2. For $i = 1 \dots n$

   REPEAT

   $l \leftarrow \arg\min_{\boldsymbol{s}_l \in \mathcal{S}} \|\boldsymbol{s}_l\|_0$, breaking ties arbitrarily
   $\boldsymbol{x}_i = \boldsymbol{s}_l$
   $\mathcal{S} = \mathcal{S} \backslash \{\boldsymbol{s}_l\}$

   UNTIL `rank(`$[\boldsymbol{x}_1, \dots, \boldsymbol{x}_i]$`)`$= i$

3. Set $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^T$, and $\boldsymbol{A} = \boldsymbol{Y}\boldsymbol{Y}^T(\boldsymbol{X}\boldsymbol{Y}^T)^{-1}$.

Figure 2: **ER-SpUD** recovery algorithm.

(P0) Every row of X has positive support size at most $(10/9)\theta p$. Furthermore, every linear combination of rows of X in which at least two of the coefficients in the linear combination are non-zero has support size at least $(11/9)\theta p$.

(P1) For every b satisfying $\|b\|_0 \le 1/(8\theta)$, any solution z_* to the optimization problem

$$\min \|z^T X\|_1 \text{ subject to } b^T z = 1 \tag{1}$$

has $\text{support}(z_*) \subseteq \text{support}(b)$.

(P2) Let q be $\frac{1}{8\theta}$. For every $J \in \binom{[n]}{q}$ and every $b \in \mathbb{R}^n$ satisfying $|b|_{(2)}/|b|_{(1)} \le 1/2$, the solution to the restricted problem

$$\|z^T X_{J,*}\|_1 \text{ subject to } b^T z = 1 \tag{2}$$

is unique, 1-sparse, and is supported on the index of the largest entry of b. Here $|b|$ is the vector whose ith entry is $|b_i|$, and $|b|_{(j)}$ is the jth largest entry of $|b|$. Also, $X_{J,*}$ denotes the submatrix of X with rows in J.

(P3) For every $i \in [n]$ there exist a pair of columns Xe_{j_1} and Xe_{j_2} in X such that for $b = Xe_{j_1} + Xe_{j_2}$ with support J, we have that $0 < |J| \le 1/(8\theta)$, $|b|_{(2)}/|b|_{(1)} \le 1/2$, and the unique largest entry of $|b|$ has index i.

The main result of [SWW12] is then obtained by proving the following theorem, and then by showing that **(P0)**–**(P3)** all hold whp for $p \gtrsim n^2 \log^2 n$.

Theorem 1 ([SWW12])**.** *Suppose conditions* **(P0)**–**(P3)** *all hold. Then* ***ER-SpUD*** *and* ***Greedy*** *from Figure 2 recover (A', X') such that $X' = \Pi D X$ and $A = AD^{-1}\Pi^{-1}$ for some diagonal scaling matrix D and permutation matrix Π. That is, the recovered (A', X') are correct up to scaling and permuting rows (resp. columns) of X (resp. A).*

It was implicit in [SWW12], and made explicit in [LV15], that to analyze the probability **(P1)** holding as a function of p, it suffices to prove some upper bound on some stochastic process. Namely, [LV15] proves that for Π a Bernoulli-subgaussian matrix with p rows, for $p = \Omega(n \log^4 n)$

$$\mathbb{P}\left(\sup_{\|v\|_1=1} |\|\Pi v\|_1 - \mathbb{E}\,\|\Pi v\|_1| < c_0 \mu_{min}\right) > 1 - o(1) \tag{3}$$

for some constant $c_0 < 1$, and $\mu_{min} := \inf_{\|v\|_1=1} \mathbb{E}\,\|X^T v\|_1$. Both [SWW12, LV15] though required the stochastic process of Eq. (3) to be bounded for roughly $\binom{n}{1/(8\theta)}$ choices of Π, formed by taking various submatrices of X^T. The naive approach is to then argue that the inequality holds with failure probability $\ll 1/\binom{n}{1/(8\theta)}$ for a fixed Π so then union bound over all such submatrices. Unfortunately the failure probability in [LV15] was not made explicit and was only given as $o(1)$. In fact, it is likely that making the failure probability explicit would force $p \gg n^{3/2}$ for some sparsity settings (see Remark 5).

We show that, first of all, **(P1)** can be relaxed to some **(P1')** such that it suffices to only show Eq. (3) holds for polynomially many submatrices of X; showing **(P1')** suffices requires only a very minor change in the previous analysis of [SWW12]. Next, more importantly, show that $p \gtrsim n \log(n/\delta)$ suffices for Eq. (3) to hold with probability $1 - \delta$. This is one of our main technical contributions, and is established using a generic chaining argument [Tal14]. It is worth pointing out that simpler chaining inequalities, such as Dudley's inequality, would yield suboptimal results in our setting by logarithmic factors.

Next, we also show that **(P2)** can be weakened to some other event **(P2')** that holds whp as long as $p \gtrsim \theta^{-1}\log(n/\delta)$. Establishing this only requires a minor change in the analysis of [SWW12].

Finally, in Lemma 6 we show that event **(P3)** holds whp for $p \gtrsim n\log(n/\delta)$. This is the part where the modification of the algorithm was necessary, so that pairs of columns Xe_{j_1} and Xe_{j_2} mentioned in this condition refers to all $\binom{p}{2}$ pairs of columns, as opposed to a fixed pairing (with $\lfloor \frac{p}{2} \rfloor$ pairs). Note that this condition actually fails to hold for the unmodified version of the algorithm with $p \ll n^2$, for example when the matrix X is drawn from the Bernoulli-Rademacher model, which is the main reason the unmodified algorithm fails to perform recovery (see Section A).

1.3 Recent and independent work

In a recent and independent work, Adamczak showed a main result similar to ours [Ada16]. In particular, he showed that by making the same modification to **ER-SpUD** that we have made (**ER-SpUD(DCv2)**), $p \gtrsim n\log n$ suffices for successful dictionary learning with probability $1 - 1/(n\log n)$. Unlike our analysis which is based on Bernstein's inequality and generic chaining, the proof in [Ada16] combines Bernstein's inequality with Talagrand's contraction principle, which leads to an overall simpler proof than ours. The main differences in the results themselves are that attention in [Ada16] was not given to dependence of p on the failure probability δ, and the analysis of our Section A that **ER-SpUD(DC)** fails for $p \ll n^2$ also does not appear there, so that our stated results are slightly stronger in these regards.

2 Sufficient conditions for successful recovery

We first explain why all conditions **(P0)**–**(P3)** holding simultaneously implies **ER-SpUD** correctly recovers (A, X). This argument appears in [SWW12], but since it is quite short we repeat it here for the benefit of the reader. Afterward, we slightly change **(P1)** and **(P2)** to similar conditions **(P1')**, **(P2')** which still suffice for correct operation of **ER-SpUD**, and we show that all conditions above (with **(P1)** replaced by **(P1')** and **(P2)** by **(P2')**) hold simultaneously with probability $1-\delta$ as long as $p \gtrsim n\log(n/\delta)$. For all the conditions except **(P0)**, the original analysis of [SWW12] required $p \gg n^2$, which we cannot afford here, and hence we provide more efficient analyses here.

Proof (of Theorem 1). We first show that for every row $X_{i,*}$ of X, there is some j so that s_j from the output of **ER-SpUD** is some scaling of $X_{i,*}$. By **(P3)** there is some pair of columns Xe_{j_1}, Xe_{j_2} so that their sum b has support J with $0 < |J| \le 1/(8\theta)$, and $|b|_{(2)}/|b|_{(1)} \le 1 - \gamma_0$, and furthermore the unique largest entry of $|b|$ is at index i. Since $|J| \le 1/(8\theta)$, **(P1)** implies any solution z_* to (1) has support contained in J. Therefore (1) has the same set of optimal solutions as (2). By **(P2)** we thus know that the optimal solution is some z_* which is 1-sparse, supported only on index i. Therefore the corresponding w_* obtained from **ER-SpUD** is some scaling of $X_{i,*}$.

The above only shows all rows of X appear as some s_j (possibly scaled). However, many s_j found may not be any scaled row of X at all. We now complete the proof. First, observe **(P0)** implies X has rank n (if not, then either some row of X is zero, which **(P0)** forbids, or some linear combination of at least two rows is zero, but the zero vector has sparsity $0 < (11/9)\theta p$, and thus this also cannot happen). Therefore, the n rows of X are exactly the n sparsest vectors in the rowspace of X (up to scaling). Since they all appear as outputs of **ER-SpUD**, scaled, they are then exactly the n rows returned by **Greedy** in some order. Thus **Greedy** returns $X' = \Pi D X$ as desired. Noting $Y = AX$, we see $A = YX^T(XX^T)^{-1}$. Meanwhile, **Greedy** returns

$$A' = YX'^T(X'X'^T)^{-1} = YX^T D\Pi^T(\Pi^T)^{-1}D^{-1}(XX^T)^{-1}D^{-1}\Pi^{-1} = AD^{-1}\Pi^{-1}.$$

■

Remark 2. It is worth noting that the proof of Theorem 1 implies that **Greedy** could be replaced by the following simpler algorithm and still maintain correctness under **(P0)**–**(P3)**: for each s_j in order, remove any other $s_{j'}$ which are scaled copies of s_j, then return the n sparsest s_j remaining to be the rows of X.

In the proof of Theorem 1, observe that **(P1)** is not invoked for *every* one of the possible sparsity patterns for b (of which there are at least $\binom{n}{q}$ where $s = 1/(8\theta)$), and **(P2)** is not invoked for all possible choices of J. Rather, in the proof, the effects of **(P1)** and **(P2)** are only needed for the at most $\binom{p}{2}$ vectors b that are non-zero, at most $1/(8\theta)$-sparse, and expressible as the sum of two columns of X. We now define **(P1')**, **(P2')** as follows.

(P1') For every b that can be expressed as the sum of two columns of X,

$$\forall v \in \mathbb{R}^{|\bar{J}|},\ \|v^T X_{\bar{J},*}\|_1 - 2\|v^T X_{\bar{J},S}\|_1 > Cp\sqrt{\frac{\theta}{|\bar{J}|}}\|v\|_1 \tag{4}$$

and

$$|S| < p/4 \tag{5}$$

where $C > 0$ is some fixed constant, $J = \text{support}(b)$, $\bar{J} = [n]\backslash J$, and $S \subseteq [p]$ is the set of columns of X with support intersecting J.

(P2') Let q be $\frac{1}{8\theta}$. For every b equaling the sum of two columns of X and with $J \subset [n]$ its support, let $b' \in \mathbb{R}^{|J|}$ be the projection of b onto its support. If $0 < |J| \le q = 1/(8\theta)$ and $|b|_{(2)}/|b|_{(1)} \le 1/2$, then the solution to the restricted problem

$$\|z^T X_{J,*}\|_1 \text{ subject to } (b')^T z = 1 \tag{6}$$

is unique, 1-sparse, and is supported on the index of the largest entry of b'. Here $|b'|$ is the vector whose ith entry is $|b'_i|$, and $|b'|_{(j)}$ is the jth largest entry of $|b'|$. Also, $X_{J,*}$ denotes the submatrix of X with rows in J.

The following corollary then is immediate from the proof of Theorem 1 and the fact that **(P1')** implies that, for any $b \ne 0$ with $|\bar{J}| = \Omega(n)$ (which holds for $0 < |J| < 1/(8\theta) = O(\sqrt{n})$ as per **(P3)**), it holds that the optimal solution z_* to min $\|z^T X\|_1$ subject to $b^T z = 1$ has support$(z_*) \subseteq$ support(b) (see the proofs of [SWW12, Lemma 11] and [LV15, Lemma V.2]).

Corollary 3. *Suppose conditions* **(P0)**, **(P1')**, **(P2')**, *and* **(P3)** *all hold. Then* ***ER-SpUD*** *and* ***Greedy*** *from Figure 2 recover (A', X') such that $X' = \Pi D X$ and $A = AD^{-1}\Pi^{-1}$ for some diagonal scaling matrix D and permutation matrix Π. That is, the recovered (A', X') are correct up to scaling and permuting rows (resp. columns) of X (resp. A).*

We now show **(P0)**, **(P1')**, **(P2')**, and **(P3)** all simultaneously hold with probability $1-\delta$ as long as $p \gtrsim n\log(n/\delta)$ and $1/n \lesssim \theta \lesssim 1/\sqrt{n}$, which when combined with Corollary 3 implies that **ER-SpUD** has the desired correctness guarantee under this same regime for p, θ.

Theorem 4. *For $p \gtrsim n\log(n/\delta)$ and $1/n \lesssim \theta \lesssim 1/\sqrt{n}$,*

$$\mathbb{P}(\neg(\mathbf{P0}) \vee \neg(\mathbf{P1'}) \vee \neg(\mathbf{P2'}) \vee \neg(\mathbf{P3})) < \delta \tag{7}$$

Proof. We will show the right hand side of (7) is at most $C\delta$ for some C, then the theorem follows by rescaling δ. We use the union bound.

First, $\mathbb{P}(\neg(\mathbf{P0})) < \delta$ was already shown, even under the weaker conditions $p \gtrsim n\log n + \theta^{-1}\log(n/\delta)$ and $1/n \lesssim \theta \le 1$, in [SWW12, Theorem 3]. We thus do not provide an analysis here.

For **(P1)**–**(P3)**, the analyses in [SWW12] required $p \gg n^2$ for any non-trivially small failure probability. We thus now provide our analyses for **(P1')**, **(P2')**, and **(P3)**. Relaxing the requirement on p for **(P3)** to hold with high probability required us to switch from **ER-SpUD(DC)** to **ER-SpUD(DCv2)**.

For **(P1')**, the analysis is almost identical to the proofs of [SWW12, Lemma 11] and [LV15, Lemma V.2] regarding **(P1)**. We repeat the slightly modified argument here for **(P1')**. Let b be a particular sum of two columns of X. We will show that the condition of **(P1')** fails to hold for b with probability at most δ/p^2, which implies $\mathbb{P}(\neg\textbf{(P1')}) \le \delta$ by a union bound over all $\binom{p}{2}$ such b. Let J, S be as in the definition of **(P1')** above. Define the event $\mathcal{E}_S$ as the event that $|S| < p/4$. Since $\theta n \le c\sqrt{n}$ for some small $c > 0$, if $b = X_{*,j_1} + X_{*,j_2}$, it follows that any column index $j \notin \{j_1, j_2\}$ has support intersecting J with probability at most $1/10$ (by making c sufficiently small). Thus $\mathbb{E}\,|S| < p/10$, implying $\mathbb{P}(\neg\mathcal{E}_S) = \mathbb{P}(|S| \ge p/4)$ is at most $\exp(-\Omega(p)) \le \delta/p^2$ by the Chernoff bound and fact that $p \gtrsim \log(p^2/\delta)$.

The definition of $\mathcal{E}_N$ is the following event:

$$\forall v \in \mathbb{R}^{|\bar{J}|},\ \|v^T X_{\bar{J},*}\|_1 - 2\|v^T X_{\bar{J},S}\|_1 > Cp\sqrt{\frac{\theta}{|\bar{J}|}}\|v\|_1 \tag{8}$$

for some constant C, where $\bar{J}$ denotes $[n]\backslash J$. Note though that $X_{\bar{J},*}$ is itself a matrix of i.i.d. Bernoulli-subgaussian entries (except for the two columns j_1, j_2, which are both zero). Thus setting $\Pi = X_{\bar{J},*}^T$ and applying Theorem 23 with our choice of p, with probability at least $1-\delta/p^2$, for all $v \in B_1$,

$$\|v^T X_{\bar{J},*}\|_1 \ge \frac{7}{8}\,\mathbb{E}\,\|v^T X_{\bar{J},*}\|_1 = \frac{7p}{8}\,\mathbb{E}\,|v^T (X_{\bar{J},*})_{*,1}| \stackrel{\text{def}}{=} \frac{7p}{8}\alpha(v), \tag{9}$$

where $(X_{\bar{J},*})_{*,1}$ clumsily denotes the first column of the matrix $X_{\bar{J},*}$. The last inequality follows from [SWW12, Lemma 16]. Also, conditioned on $\mathcal{E}_S$, $|S| < p/4$. Let X' be the matrix $X_{\bar{J},S}$ padded with $p/4 - |S|$ additional columns, each independent of but identically distributed to the columns of X. Then, even conditioned on $\mathcal{E}_S$, X' is a $|\bar{J}| \times p/4$ matrix of i.i.d. Bernoulli-gaussian entries (except for two columns which are both identically zero, corresponding to j_1, j_2). Thus applying Theorem 23 to $\Pi = (X')^T$, with probability at least $1-\delta/p^2$, for all $v \in B_1$,

$$\|v^T X'\|_1 \le \frac{3}{2}\,\mathbb{E}\,\|v^T X'\|_1 = \frac{3p}{8}\,\mathbb{E}\,|v^T X'_{*,1}| = \frac{3p}{8}\alpha(v). \tag{10}$$

Then by combining (9), (10) and scaling by $\|v\|_1$, we see that the left hand side of (8) is at least

$$\frac{p}{8}\alpha(v) \gtrsim p\sqrt{\frac{\theta}{|\bar{J}|}}\|v\|_1, \tag{11}$$

with the inequality following from [SWW12, Lemma 16].

We finally now analyze the probability that **(P2')** holds. It is implied by the proof of [SWW12, Lemma 12] that for **(P2')** to hold, it suffices for the following three equations to hold, where $\mathcal{B} = \{Xe_{j_1} + Xe_{j_2} : 1 \le j_1 < j_2 \le p\}$ is the set of all sums of pairs of columns in X:

$$\|X\|_{\ell_\infty \to \ell_\infty} \le (1+\varepsilon)\mu\theta p \tag{12}$$

$$\forall j \in [n],\ \|X_{[n]\setminus\{j\}, \bar{T}_j}\|_{\ell_\infty \to \ell_\infty} \le \alpha\mu\theta p \tag{13}$$

$$\forall J \subseteq [n] \text{ s.t. } \exists b \in \mathcal{B},\ J = \mathrm{support}(b),\ \|X_{j,\Omega_{J,j}}\|_1 \ge \beta\mu\theta p. \tag{14}$$

for some particular constants $\varepsilon, \alpha, \beta > 0$ (specifically [SWW12] pick $\beta = 7/8$, $\alpha = \varepsilon = 1/8$). Also, for $X_{i,j} = \chi_{i,j} g_{i,j}$, μ denotes the constant $\mathbb{E}|g_{i,j}|$. Here $\|M\|_{\ell_p \to \ell_p}$ is the ℓ_p to ℓ_p operator norm, which in the case of $p = \infty$ is simply the largest ℓ_1 norm of any row of M. Furthermore, $T_j = \{i : X_{j,i} = 0\}$, and

$$\Omega_{J,j} = \{\ell : X_{j,\ell} \ne 0 \text{ and } X_{j',\ell} = 0,\ \forall\, j' \in J\setminus\{j\}\}.$$

It is already shown in [SWW12, Lemma 18] that (12) and (13) fail to hold with probability at most $Cn\exp(-c\theta p) < \delta$, by choice of p, where the constant c depends on α, ε. For condition (14), the proof of [SWW12, Lemma 12] shows that for any fixed J as in (14) and $j \in J$ (see Eqn. (56) of [SWW12]),

$$\mathbb{P}(\|X_{j,\Omega_{J,j}}\|_1 \le \beta\mu\theta p) \le 4\exp\left(-\frac{c\theta p}{256}\right)$$

Thus by a union bound over all $b \in \mathcal{B}$, (14) holds with probability at least $1 - \delta$.

We upper bound $\mathbb{P}(\neg\mathbf{(P3)})$ separately in Lemma 6. □

Remark 5. The work [LV15] showed a weaker version of Theorem 23 in which p was required to be $\Omega(n \log^4 n)$, and where the failure probability was shown to be some non-explicit value $\delta = o(1)$. [LV15] then claimed that this was sufficient to show that **ER-SpUD(DC)** was correct with probability $1 - o(1)$. Unfortunately, it appears there were a few gaps in their analysis. First, [LV15] relied on conditions **(P2)** and **(P3)** from [SWW12] (as rewritten above) both holding, but the only known probabilistic analyses of these conditions, given in [SWW12], required $p \gg n^2$. Secondly, the proof sketch of [LV15, Lemma V.2] showing that the condition of (22) suffices to imply **(P1)** actually invoked the inequality

$$\sup_{v \in B_1} |\|\Pi v\|_1 - \mathbb{E}\|\Pi v\|_1| \le \varepsilon \cdot \mathbb{E}\|\Pi v\|_1$$

for at least $\binom{n}{q}$ choices of Π where $s = 1/(8\theta)$ (specifically $\Pi = X_{*,S}$ for $\binom{n}{q}$ choices of S). Thus to apply a probabilistic inequality of the form (22) to imply **(P1)**, one actually needs a specific δ and not just $\delta = o(1)$; in particular one needs $\delta \ll 1/\binom{n}{q}$ to union bound over all S, which for the largest value of $\theta \simeq 1/\sqrt{n}$, means one needs $\delta \ll \exp(-C\sqrt{n}\log n)$. Thus even if (22) held for $p \gtrsim n\log(1/\delta)$, one would still need $p \gtrsim n^{3/2}\log n$ for the analysis there to imply that **(P1)** holds with positive probability.

We are now going to show that with probability $1 - \delta$ condition **(P3)** holds. Let us present an intuition behind the proof before we delve into technical details.

Consider the special case that $X_{ij} = b_{ij} g_{ij}$ with Bernoulli random variable b_{ij} and independent *continuous* subgaussian random variable g_{ij}. In such a case there would exist some fixed threshold

t_0, such that $\mathbb{P}(|X_{ij}| > t_0) = \frac{1}{n}$ — it would mean that a constant fraction of columns would have *unique* entry larger than this threshold. For a single index $i \in [n]$ we would expect that at least $C\frac{p}{n} > \log \frac{n}{\delta}$ columns have a unique entry larger than t_0 and such that this entry has index i. Let us focus on this set of columns. If supports of any two such columns had common intersection exactly equal to $\{i\}$ — and if the sign on this i-th coordinate were matching, then in fact sum of those two columns would exhibit a factor two gap between the largest and the second largest entry, with largest entry being on the i-th position — indeed, entry on position i would have magnitude larger than $2t_0$, whereas all other entries are at most t_0 in absolute value. We can expect to find such a pair with probability $1 - \frac{\delta}{n}$, as all columns are expected to be $\mathcal{O}(\sqrt{n})$ sparse — therefore for a fixed pair containing $\{i\}$, their supports would intersect on exactly $\{i\}$ with constant probability. We then prove that there exist such a pair with probability at least $\frac{\delta}{n}$ for every fixed i, and hence by union bound property **(P3)** holds with probability δ.

In the actual proof we do not assume that g_{ij} is continuous, and hence a threshold t_0 for which $\mathbb{P}(|X_{ij}| > t_0) = \frac{1}{n}$ might not exist, and the proof is slightly more complicated, but it follows the same general intuition.

Lemma 6. *Let $X \in \mathbb{R}^{n \times p}$ be a Bernoulli-Subgaussian matrix with $\theta = \mathcal{O}(\frac{1}{\sqrt{n}})$. If $p = \Omega(n \log \frac{n}{\delta})$, then with probability at least $1 - \delta$ condition* **(P3)** *holds.*

Proof. Take $t_0 := \inf\{t : \mathbb{P}(|X_{ij}| > t) < \frac{1}{n}\}$. Observe that $\mathbb{P}(|X_{ij}| > t_0) \leq \frac{1}{n} \leq \mathbb{P}(|X_{ij}| \geq t_0)$.

Let us define y_k to be the k-th column of X. Let l_k be the number of coordinates of y_k strictly larger than t_0, and s_k be the number of coordinates of y_k of size larger or equal to t_0. We will say that column k is well-separated if $l_k \leq 1 \leq s_k$.

We claim that for fixed k, the probability that y_k is well-separated is bounded away from zero. Clearly probability that a column is not well-separated is equal to $\mathbb{P}(l_k > 1) + \mathbb{P}(s_k = 0)$. Let us consider two cases. If $\mathbb{P}(|X_{ij}| > t_0) < \frac{1}{5n}$, then $\mathbb{E}\, l_k \leq \frac{1}{5}$ and $\mathbb{P}(l_k > 1) \leq \frac{1}{5}$ by Markov inequality. Moreover

$$\mathbb{P}(s_k = 0) = 1 - \prod_j (1 - \mathbb{P}(X_{j,k} \geq t_0)) \leq 1 - \left(1 - \frac{1}{n}\right)^n \leq 1 - e^{-1}\left(1 - \frac{1}{n}\right)$$

And finally probability that a single column is not well-separated is at most $1 - e^{-1}(1 - \frac{1}{n}) + \frac{1}{5}$, which is at most 0.9 for sufficiently large n.

On the other hand, if $\mathbb{P}(|X_{ij}| > t_0) \geq \frac{1}{5n}$, we will prove that with probability bounded away from zero, there is exactly one entry of a column y that is greater than t_0.

Indeed, if $\alpha := \mathbb{P}(|X_{ij}| > t_0)n$, we have

$$\begin{aligned}
\mathbb{P}(l_k = 1) &= n\frac{\alpha}{n}(1 - \frac{\alpha}{n})^{n-1} \\
&= (1 - \frac{\alpha}{n})^{-1}\alpha(1 - \frac{\alpha}{n})^n \\
&\geq \alpha\left(1 - \frac{\alpha}{n}\right)^{-1}\left(1 - \frac{\alpha^2}{n}\right)e^{-\alpha}
\end{aligned}$$

We know that $\alpha \in [\frac{1}{5}, 1]$, and therefore the last expression is bounded away from zero, for large enough n.

Let $C \in (0, 1)$ be a constant, such that the probability for a column to be well-separated is at least C. We will prove now that with $\theta = \mathcal{O}(1/\sqrt{n})$ (where constant hidden in $\mathcal{O}$ notation depends on C), the probability that a column has support of size larger than $\frac{\sqrt{n}}{4}$ is at most $\frac{C}{2}$.

Indeed, let $S_k \subset [n]$ be the support of the k-th column of X. We can assume that θ is such that $n\theta < \frac{C\sqrt{n}}{8}$, so that $\mathbb{E}|S_k| = n\theta < \frac{C\sqrt{n}}{8}$. Now by Markov inequality $\mathbb{P}(|S_k| > \frac{\sqrt{n}}{4}) = \mathbb{P}(|S_k| > \frac{\mathbb{E}|S_k|}{C/2}) < C/2$.

Now, by union bound, any fixed column k of matrix X simultaneously is well-separated and have support of size at most $\frac{\sqrt{n}}{4}$ with probability at least $\frac{C}{2}$. Let us define a set $W \subset [p]$, such that $k \in W$ if and only if column $X_{*,k}$ is well-separated and $|\operatorname{support}(X_{*,k})| < \frac{\sqrt{n}}{4}$.

Fix some index $i \in [n]$. We wish to prove that with probability at least $1 - \frac{1}{n\delta}$ there exist a pair of columns j, k such that for $b = X_{*,j} + X_{*,k}$, we have $b_i > 2b_l$ for all $l \neq i$, and moreover $|\operatorname{support}(b)| < \frac{\sqrt{n}}{2}$. Indeed, let $W_i \subset W$, be a set of those indices $k \in W$, such that $X_{i,k} \geq t_0$. Observe that for a fixed $k \in [p]$, we have $\mathbb{P}(k \in W_i) = \mathbb{P}(k \in W)\,\mathbb{P}(k \in W_i | k \in W) \geq \frac{C}{2} \cdot \frac{1}{n}$. That is $\mathbb{E}|W_i| \geq \frac{Cp}{2n}$, and by Chernoff bound $\mathbb{P}(W_i < \mathbb{E}|W_i|/2) < \exp(-\mathcal{O}(\frac{p}{n})) = \frac{\delta}{2n}$ for $p = \Omega(n \log \frac{n}{\delta})$.

Let us condition on the event that $|W_i| > \frac{Cp}{4n}$. We take $W_i^+ = \{k \in W_i : X_{i,k} > 0\}$, and let us assume without loss of generality that $|W_i^+| \geq \frac{1}{2}|W_i|$ (otherwise we can use a symmetric argument on the complement — that is, on the set of columns for which $X_{i,k}$ is negative). We wish to prove that with probability at least $1 - \frac{\delta}{2n}$ there exist a pair of indices $j_1, j_2 \in W_i^+$, such that $S_{j_1} \cap S_{j_2} = \{i\}$.

Observe that for $k \in [p]$, $S_k \setminus \{i\}$ conditioned on the fact that $k \in W_i^+$ and $|W_i| \geq \frac{Cp}{4n}$ has distribution supported on sets of size at most $\frac{\sqrt{n}}{4}$ and permutationally invariant. Therefore a pair of two independent such sets is disjoint with probability bounded at least $1 - n\left(\frac{\sqrt{n}}{4}\right)^2 = \frac{15}{16}$. Consider an arbitrary pairing of elements in W_i^+. We have $\lfloor \frac{|W_i|}{2} \rfloor$ pairs of indices (k_1^j, k_2^j), which is at least $\frac{Cp}{8n} = \Omega(\log \frac{n}{\delta})$, and therefore with probability at least $1 - \frac{\delta}{2n}$ there is a pair such that $S_{k_1^j} \cap S_{k_2^j} = \{i\}$. Take $b := X_{*,k_1^j} + X_{*,k_2^j}$; we have $b_i \geq 2t_0$ because both $k_1^j, k_2^j \in W_i^+$. On the other hand for $i' \neq i$ only one of $X_{i',k_1^j}, X_{i',k_2^j}$ is nonzero, and if it is — it is at most t_0 (again by definition of W_i). Hence $|b_{i'}| \leq t_0$ as expected. Moreover, the size of support of b is at most $\frac{\sqrt{n}}{2}$, and by assumption that $\theta = \mathcal{O}(\frac{1}{\sqrt{n}})$, this value is at most $\frac{1}{8\theta}$

Finally, by union bound over all $i \in [n]$, the statement of the lemma holds — for fixed i with probability at most $\frac{\delta}{2n}$ set W_i fails to be large enough, and conditioned on this set being large, with probability at most $\frac{\delta}{2n}$ it fails to contain two elements of interests.

□

Remark 7. If one uses **ER-SpUD(DC)** and not **ER-SpUD(DCv2)**, then condition **(P3)** actually *requires* $p \gg n^2$ to hold with non-negligible probability. To see this, we first describe the difference between **ER-SpUD(DCv2)** and the **ER-SpUD(DC)** algorithm of [SWW12]. In **ER-SpUD(DC)**, rather than try all $T = \binom{p}{2}$ pairings of columns as in **ER-SpUD(DCv2)**, it only tries $p/2$ column pairs formed by randomly pairing the p columns with each other.

Now, consider the case of sparsity $s = 1$, i.e. $\theta = 1/n$, with $p \gtrsim n \log n$. Then for $i \in [n]$ if we let q_i be the number of columns of X with support containing i, then by the Chernoff bound whp for all i we have $q_i = \Theta(p/n)$. Fix an i and consider the q_i columns $j_1, \ldots, j_{q_i}$ containing i in their support. Note that the expected number of these q_i columns that are randomly paired with another one of the same q_i columns by **ER-SpUD(DC)** is $q_i(q_i-1)/(p-1) = \Theta(p/n^2)$, which is $o(1)$ for $p = o(n^2)$ so that **(P3)** is likely to fail. In fact, essentially this same argument shows that unless $p = \Omega(n^2 \log n)$, it is likely that there will be *some* $i \in [n]$ such that none of the q_i columns containing i in its support will be paired with each other. In Section A we show that not only does **(P3)** fail whp for $p = o(n^2)$, but in fact **ER-SpUD** itself fails for $p \ll n^2$.

3 Concentration and chaining background

We now provide some preliminary definitions and results we will need to prove our chaining theorem, Theorem 23. As per Corollary 3 and the proof of Theorem 4, Theorem 23 fits in to show that **ER-SpUD** achieves correct recovery with probability $1-\delta$ for $p \gtrsim n\log(n/\delta)$ and $1/n \lesssim \theta \lesssim 1/\sqrt{n}$.

3.1 Tail bounds

For a random variable Z, we make the standard definition $\psi_Z(\lambda) = \ln \mathbb{E}\, e^{\lambda Z}$ (e.g. [BLM13, Section 2.2]). The following lemma is then immediate.

Lemma 8. *If Z, Z' are independent, then $\psi_{Z+Z'}(\lambda) = \psi_Z(\lambda) + \psi_{Z'}(\lambda)$.*

The following definition and facts concerning subgamma random variables are standard [BLM13, Section 2.4].

Definition 9. *A random variable Z is said to be (σ, B)-*subgamma *if $\mathbb{E}\, Z = 0$ and $\psi_Z(\lambda) \le \lambda^2\sigma^2/(2(1 - B\lambda))$ for all $|\lambda| < 1/|B|$.*

Lemma 10 (Basic properties of subgamma random variables)**.** *It holds that*

1. *If Z is (σ, B)-subgamma and $\alpha > 0$, then αZ is $(\alpha\sigma, \alpha B)$-subgamma.*

2. *If $Z_1, \ldots, Z_n$ are independent and each Z_i is (σ_i, B_i)-subgamma, then $\sum_{i=1}^n Z_i$ is $(\sqrt{\sum_i \sigma_i^2}, \min_i B_i)$-subgamma.*

3. *If Z is (σ, B)-subgamma, then*

$$\mathbb{P}(|Z| > \lambda) \lesssim \exp\left(-\frac{\lambda^2}{2\sigma^2}\right) + \exp\left(-\frac{\lambda}{2B}\right)$$

4. *If Y is a symmetric (σ, B)-subgamma random variable, and Z is symmetric such that $|Z| \le |Y|$ with probability 1, then Z is (σ, B)-subgamma.*

Remark 11. The work [LV15], in addition to providing a bound on p sufficient for (22) to hold for some $\delta = o(1)$, also advertised a new "refined version of Bernstein's concentration inequality for a sum of independent variables". In fact though, their concentration inequality is equivalent to the statement that a sum of subgamma random variables is subgamma with new parameters as per above, which is a known fact (the reader is encouraged to read the excellent treatment of sums of independent random variables in [BLM13, Section 2], from which the above definitions and lemmas are taken).

Definition 12. *A random variable Z is said to be σ-*subgaussian *if $\mathbb{E}\, Z = 0$ and $\psi_Z(\lambda) \le \lambda^2\sigma^2/2$. It is called simply subgaussian if it is 1-subgaussian.*

Fact 13. *If Z is a subgaussian random variable, and $D \in \{0, 1\}$ is Bernoulli random variable, independent from Z with $\mathbb{E}\, D = \theta$, then ZD is $(\sqrt{2\theta}, 1)$-subgamma.*

Proof. Take $\lambda < 1$. We have

$$\exp(\psi_{DZ}(\lambda)) = \mathbb{E}\exp(DZ\lambda) = 1 - \theta + \theta\,\mathbb{E}\exp(Z\lambda) \le 1 - \theta + \theta\exp(\lambda^2/2)$$

We can expand $\exp(\lambda^2/2)$, to get

$$\begin{aligned}
\exp(\psi_{DZ}(\lambda)) &\le 1 - \theta + \theta\exp(\lambda^2/2) \\
&= 1 - \theta + \theta\left(\sum_{k=0}^{\infty} \frac{\lambda^{2k}}{2^k k!}\right) \\
&= 1 + \frac{\theta\lambda^2}{2} + \frac{\theta\lambda^2}{4}\left(\sum_{k=2}^{\infty} \frac{\lambda^{2k-2}}{2^{k-2}k!}\right) \\
&\le 1 + \frac{\lambda^2\theta}{2} + \frac{\lambda^2\theta}{4}\left(\sum_{k=0}^{\infty} 2^{-k}\right) \\
&= 1 + \lambda^2\theta \\
&\le \exp(\lambda^2\theta) \\
&\le \exp(\frac{\lambda^2\theta}{1-\lambda})
\end{aligned}$$

Taking logarithms on both sides proves the result. □

Lemma 14 (Moment and tail bounds equivalence, Lemma 4.10 [LT91])**.** *Let Z be a nonnegative random variable.*

1. *If there exists some constants m, s, b, such that for every λ_1, λ_2*

$$\mathbb{P}(Z > m + \lambda_1 s + \lambda_2 b) < \exp(-\lambda_1^2) + \exp(-\lambda_2), \tag{15}$$

   then for all $p \ge 1$ it holds that $\|Z\|_p \le C(m + \sqrt{p}s + pb)$, where C is a universal constant.

2. *If for some constants s, b and for all $p \ge 1$ it holds that $\|Z\|_p \le m + \sqrt{p}s + pb$, then*

$$\mathbb{P}(Z > C(m + \lambda_1 s + \lambda_2 b)) < \exp(-\lambda_1^2) + \exp(-\lambda_2) \tag{16}$$

   where C is a universal constant.

3.2 Chaining

In this subsection we provide relevant definitions for a technique called *generic chaining*, as well as statements of some of the results in the area. Those tools have been designed to provide answers about the supremum of the fluctuations from mean for a large collection of random variables, when the reasonable bounds for covariances in terms of the geometry of the set of indices are at hand. In particular, those methods reduce questions about such fluctuations to questions about purely geometric quantities of the set of indices, and they proved to be extremely useful in a number of applications. The generic chaining method will be a core of the proof of Theorem 23. For a more detailed exposition of this technique, we refer the reader to an excellent book on that topic [Tal14].

Definition 15 (Admissible sequence)**.** *For an arbitrary set T, we say that a sequence of its subsets $(T_k)_{k=0}^{\infty}$ is admissible if for every number k it is true that $T_k \subset T_{k+1}$ and $|T_k| \leq 2^{2^k}$ for $k \geq 1$ and $|T_0| = 1$.*

Definition 16 (Gamma functionals)**.** *For a metric space (T, d) we define*

$$\gamma_\alpha(T, d) := \inf_{(T_k)} \sup_{x \in T} \sum_{k=0}^{\infty} 2^{k/\alpha} d(x, T_k) \tag{17}$$

where the infimum is taken over all admissible sequences T_k. In the above formula we define as usual $d(x, T_k) := \inf_{t \in T_k} d(x, t)$.

Fact 17. *If d and d' are two metrics such that for $d(t_1, t_2) = Cd'(t_1, t_2)$ for every pair of points t_1, t_2, then $\gamma_\alpha(T, d) = C\gamma_\alpha(T, d')$*

Theorem 18 (Generic chaining [Tal14], Theorem 2.2.23)**.** *Let T be an arbitrary set of indices, and $d_1, d_2 : T \times T \to \mathbb{R}_{\geq 0}$ two metrics on T. Suppose that with any point $t \in T$ we have associated random variable X_t, with $\mathbb{E} X_t = 0$. Suppose moreover, that for any two points $u, w \in T$ we have a tail bound:*

$$\mathbb{P}\left(|X_u - X_v| > \lambda\right) \lesssim \exp\left(-\frac{\lambda^2}{d_1(u, v)^2}\right) + \exp\left(-\frac{\lambda}{d_2(u, v)}\right) \tag{18}$$

Then

$$\mathbb{E} \sup_{u \in T} |X_u| \lesssim \gamma_2(T, d_1) + \gamma_1(T, d_2) \tag{19}$$

Theorem 19 (Dirksen, [Dir15])**.** *Let T be an arbitrary set of indices and $d_1, d_2 : T \times T \to \mathbb{R}_{\geq 0}$ two metrics on T. Suppose that with any point $t \in T$ we have associated random variable X_t, such that $\mathbb{E} X_t = 0$. Suppose moreover that for any two points $u, w \in T$, we have a tail bound*

$$\mathbb{P}(|X_u - X_v| > \lambda) \lesssim \exp\left(-\frac{\lambda^2}{d_1(u, v)^2}\right) + \exp\left(-\frac{\lambda}{d_2(u, v)}\right)$$

Then for there exists an universal constant C, such that for any $u > 0$

$$\mathbb{P}\left(\sup_{u \in T} |X_u| > C(\gamma_2(T, d_1) + \gamma_1(T, d_2) + \sqrt{u}\Delta(T, d_1) + u\Delta(T, d_2))\right) < e^{-u}$$

where $\Delta(T, d) := \sup_{u,v \in T} d(u, v)$.

Remark 20. The work [LV15] observed that the method of their proof is connected to generic chaining, but that after a certain point the methods "become different in all aspects" [LV15, Section G]. As we will see soon in the proof of Theorem 23 in Section 4, our analysis in fact simply uses the generic chaining results above, black box, without any ad hoc adjustments. Thus, in addition to improving the bounds in [LV15], our proof also has the benefit of using standard chaining results, perhaps thus also making the proof more accessible.

In some special cases it is known that bounds obtained via generic chaining are optimal up to a constant factor. We will use two of such results. Strictly speaking these results are not crucial in our analysis (one could also proceed by constructing near-optimal admissible sequences for the two different sets T that arise in our proof), but invoking these results shrinks the length of our final proof significantly.

Theorem 21 (Majorizing measures [Tal14], Theorem 2.4.1)**.** *Let $T \subset \mathbb{R}^n$, and assume that $g = (g_1, \dots g_n)$ is a vector of i.i.d. standard normal random variables. Then*

$$\mathbb{E} \sup_{t \in T} \langle g, t \rangle \simeq \gamma_2(T, d_2) \tag{20}$$

Where d_p is the metric induced by the ℓ_p norm.

Theorem 22 ([Tal14], Theorem 10.2.8)**.** *Let $T \subset \mathbb{R}^n$, and assume that $x = (x_1, \dots, x_n)$ is a vector of i.i.d. standard exponential random variables. Then*

$$\mathbb{E} \sup_{t \in T} \langle t, x \rangle \simeq \gamma_2(T, d_2) + \gamma_1(T, d_\infty) \tag{21}$$

4 Proof of the stochastic process bound

In this section we will prove the following theorem, which provides a stronger form of Eq. (3).

Theorem 23. *Let $\Pi \in \mathbb{R}^{m \times n}$ be a random matrix with i.i.d. random entries $\pi_{ij} = \chi_{ij} g_{ij}$, where $\chi_{ij} \in \{0, 1\}$ is a Bernoulli random variable with $\mathbb{E} \chi_{ij} = \theta$, and g_{ij} symmetric subgaussian random variable. Moreover, assume that $\frac{1}{n} \leq \theta$. When $m = \Omega(\varepsilon^{-2} n \log \frac{n}{\delta})$,*

$$\mathbb{P}_{\Pi} \left(\sup_{v \in B_1} |\|\Pi v\|_1 - \mathbb{E} \|\Pi v\|_1| > \varepsilon \cdot \mathbb{E} \|\Pi v\|_1 \right) < \delta \tag{22}$$

We now prove the theorem. Define $B_1 := \{t \in \mathbb{R}^n : \|t\|_1 \leq 1\}$. For each $v \in B_1$, consider

$$\tilde{X}_v := \|\Pi v\|_1 - \mathbb{E} \|\Pi v\|_1 \tag{23}$$

We wish to prove that with high probability over Π we have

$$\sup_{v \in B_1} |\tilde{X}_v| \leq \varepsilon \mu_{min} \tag{24}$$

where $\mu_{min} := m\sqrt{\frac{\theta}{n}}$ is such that for every $v \in B_1$ we have $\mu_{min} \leq \mathbb{E} \|\Pi v\|_1$ — as it was shown in [SWW12, Lemma 16].

Let $\pi_1, \dots \pi_m$ be rows of matrix Π. With each $v \in B_1$ we associate another random variable

$$X_v := \sum_{i=1}^{m} \sigma_i |\langle \pi_i, v \rangle| \tag{25}$$

Lemma 24. *For every integer p we have*

$$\| \sup_{v \in B_1} |\tilde{X}_v| \|_p \lesssim \| \sup_{v \in B_1} |X_u| \|_p \tag{26}$$

Proof. Without loss of generality consider an even p, so that $|X_u|^p = X_u^p$. Let $\tilde{\Pi}$ be a random matrix, independent and identically distributed as Π. By Jensen inequality we have

$$\begin{aligned}
\| \sup_{v\in B_1} |\tilde{X}_v| \|_p &= \left\| \sup_{v\in B_1} \|\Pi v\|_1 - \mathop{\mathbb{E}}_{\tilde{\Pi}} \|\tilde{\Pi} v\|_1 \right\|_p \\
&\leq \left\| \sup_{v\in B_1} \|\Pi v\|_1 - \|\tilde{\Pi} v\|_1 \right\|_p \\
&= \left\| \sup_{v\in B_1} \sum_{i=1}^m |\langle \pi_i, v\rangle| - |\langle \tilde{\pi}_i, v\rangle| \right\|_p
\end{aligned}$$

Now each summand $|\langle \pi_i, v\rangle| - |\langle \pi_i, v\rangle|$ is symmetric random variable. We can introduce independent random signs σ_i, and the distribution of a summands is unaffected

$$\begin{aligned}
\| \sup_{v\in B_1} |\tilde{X}_v| \|_p &\lesssim \left\| \sup_{v\in B_1} \sum_{i=1}^m \sigma_i(|\langle \pi_i, v\rangle| - |\langle \tilde{\pi}_i, v\rangle|) \right\|_p \\
&\leq \left\| \sup_{v\in B_1} \sum_{i=1}^m \sigma_i |\langle \pi_i, v\rangle| \right\|_p + \left\| \sup_{v\in B_1} \sum_{i=1}^p (-\sigma_i) |\langle \tilde{\pi}_i, v\rangle| \right\|_p \\
&= 2 \left\| \sup_{v\in B_1} \sum_{i=1}^m \sigma_i |\langle \pi_i, v\rangle| \right\|_p \\
&\lesssim \left\| \sup_{v\in B_1} |X_v| \right\|_p
\end{aligned}$$

□

We will first analyze tail behavior of the random variable $\sup_{v\in B_1} |X_v|$, and then use Lemma 24 together with Lemma 14 to obtain tail bounds for the random variable of original interest ($\sup_{v\in B_1} |\tilde{X}_v|$).

In order to use Theorem 19 to obtain tail bounds for supremum of X_u, we need to bound tails of random variables $X_u - X_v$ for $u, v \in B_1$.

Lemma 25. *For every pair of points $u, v \in B_1$, we have*

$$\mathbb{P}(|X_u - X_v| > \lambda) \lesssim \exp\left(-\frac{\lambda^2}{2m\theta \|u-v\|_2^2}\right) + \exp\left(-\frac{\lambda}{\|u-v\|_\infty}\right) \tag{27}$$

Proof. We can write

$$X_u - X_v = \sum_{i=1}^m \sigma_i(|\langle \pi_i, u\rangle| - |\langle \pi_i, v\rangle|) \tag{28}$$

Define $Q_i := \sigma_i(|\langle \pi_i, u\rangle| - |\langle \pi_i, v\rangle|)$. We have $X_u - X_v = \sum_{i=1}^m Q_i$, where all Q_i are symmetric and identically distributed.

Moreover, we have $|Q_i| = ||\langle \pi_i, u\rangle| - |\langle \pi_i, v\rangle|| \leq |\langle \pi_i, u-v\rangle|$. Observe that each π_{ij} is $(\sqrt{2\theta}, 1)$-subgamma. Therefore, by basic properties of subgamma random variables (Lemma 10) we know that $\langle \pi_i, u-v\rangle$ is $(\sqrt{2\theta}\|u-v\|_2, \|u-v\|_\infty)$-subgamma.

Now, as both Q_i and $\langle \pi_i, u-v\rangle$ are symmetric, and always $|Q_i| \leq |\langle \pi_i, u-v\rangle|$, we deduce that each Q_i is also $(\sqrt{2\theta}\|u-v\|_2, \|u-v\|_\infty)$-subgamma.

Finally, $X_u - X_v$, as a sum of independent subgamma random variables is $(\sqrt{2m\theta}\|u-v\|_2^2, \|u-v\|_\infty)$-subgamma. This, together with Lemma 10 implies tail bound

$$\mathbb{P}\left(\left|\sum_{i=1}^m Q_i\right| > \lambda\right) \lesssim \exp\left(\frac{\lambda^2}{2m\theta\|u-v\|_2^2}\right) + \exp\left(\frac{\lambda}{\|u-v\|_\infty}\right) \tag{29}$$

□

With this lemma in hand, we can use Theorem 19, to deduce the tail bound for supremum of $|X_v|$.

$$\mathbb{P}\left(\sup_{v\in B_1} |X_u| > M + \sqrt{u}D_1 + uD_2\right) < e^{-u} \tag{30}$$

Where

$$\begin{aligned} M &:= C_1(\gamma_2(B_1, \sqrt{2m\theta}d_2) + \gamma_1(B_1, d_\infty)) \\ D_1 &:= C_2\Delta(B_1, \sqrt{2m\theta}d_2) \\ D_2 &:= C_3\Delta(B_1, d_\infty) \end{aligned}$$

with d_2, d_∞ — metrics on $\mathbb{R}^n$ induced by norms ℓ_2, ℓ_∞ respectively, and C_1, C_2, C_3 are universal constants.

We claim that, we can deduce similar tail bounds for $\sup_{v\in B_1} |\tilde{X}_u|$. Namely

$$\mathbb{P}\left(\sup_{v\in B_1} |\tilde{X}_u| > L(M + \sqrt{u}D_1 + uD_2)\right) < e^{-u} \tag{31}$$

for some universal constant L.

Indeed by Lemma 14, tail bound of form Eq. (30) implies moment bounds of form $\|\sup_{v\in B_1} |X_v|\|_p \lesssim M + \sqrt{p}D_1 + pD_2$. By Lemma 24, the same (up to a constant) moment bounds are true for $\sup_{v\in B_1} |\tilde{X}_v|$. Finally, applying the other direction of Lemma 14 we deduce similar tail behavior of random variable $\sup_{v\in B_1} \tilde{X}_v$, as in Eq. (31).

If we set $u := \log \frac{1}{\delta}$ in Eq. (31), we will get an upper bound for $\sup_{v\in B_1} \tilde{X}_v$ which is satisfied with probability at least $1-\delta$. We need to understand the values of M, $\sqrt{u}D_1$ and uD_2, for this setting of u, and we will show how to pick m such that sum of those values is smaller than $\varepsilon\mu_{min}$.

Let us focus now on bounding M. We have $\gamma_2(B_1, \sqrt{2m\theta}d_2) = \sqrt{2m\theta}\gamma_2(B_1, d_2)$. We need an upper bound for $\gamma_2(B_1, d_2)$ and $\gamma_1(B_1, d_\infty)$.

Fact 26. $\gamma_2(B_1, d_2) \lesssim \sqrt{\log n}$

Proof. By Theorem 21, we have

$$\gamma_2(B_1, d_2) \lesssim \mathop{\mathbb{E}}_g \sup_{t\in B_1} \langle t, g\rangle \tag{32}$$

where g is a Gaussian vector. By the duality of ℓ_1 and ℓ_∞ norms, for any vector $w \in \mathbb{R}^n$ we have $\sup_{t\in B_1} \langle t, v\rangle = \|v\|_\infty$, so in particular $\mathbb{E}\sup_{t\in B_1} \langle t, g\rangle = \mathbb{E}\,\|g\|_\infty \simeq \sqrt{\log n}$. □

Fact 27. $\gamma_1(B_1, d_\infty) \lesssim \log n$

Proof. By Theorem 22, we have

$$\gamma_1(B_1, d_\infty) \lesssim \mathop{\mathbb{E}}_{x} \sup_{t \in B_1} \langle t, x \rangle \tag{33}$$

where x is a vector of independent standard exponentially distributed random variables. Again $\sup_{t \in B_1} \langle t, x \rangle = \|x\|_\infty$. It is standard fact that $\mathbb{E}\, \|x\|_\infty \simeq \log n$. ☐

Those two facts together with previous discussion yield an upper bound $M \lesssim \sqrt{m\theta \log n} + \log n$.

Moreover, as $d_2(u, v) \leq d_1(u, v)$ for any $u, v \in \mathbb{R}^n$, where d_1 is metric induced by ℓ_1 norm, we can easily upper bound diameter of B_1 in d_2 by diameter of B_1 and d_1 and therefore obtain an upper bound for D_1

$$D_1 = C_1 \Delta(B_1, \sqrt{em\theta} d_2) = C_1 \sqrt{em\theta} \Delta(B_1, d_2) \leq C_1 2 \sqrt{em\theta}$$

and similarly $\Delta(B_1, d_\infty) = 2$. Altogether, we have following inequalities

$$\begin{aligned} M &\lesssim \sqrt{m\theta \log n} + \log n \\ D_1 &\lesssim \sqrt{m\theta} \\ D_2 &\lesssim 1 \end{aligned}$$

Plugging this back to Eq. (31), we have

$$\mathbb{P}\left(sup_{v \in B_1} |\tilde{X}_v| < L_2 \left(\sqrt{m\theta(\log n + \log \frac{1}{\delta}}) + \log n + \log \frac{1}{\delta} \right) \right) < \delta \tag{34}$$

where again L_2 is some constant.

The following inequalities are equivalent:

$$\begin{aligned} L_2 \sqrt{m\theta \log \frac{n}{\delta}} &\leq \frac{1}{2} \varepsilon \mu_{min} \\ L_2 \sqrt{m\theta \log \frac{n}{\delta}} &\leq \frac{1}{2} \varepsilon \sqrt{\frac{\theta}{n}} m \\ \frac{4L_2^2}{\varepsilon^2} n \log \frac{n}{\delta} &\leq m \end{aligned}$$

Similarly, assumption $\theta \geq \frac{1}{n}$, implies that if $m > \frac{2L_2}{\varepsilon} n \log \frac{n}{\delta}$, then also $L_2 \log \frac{n}{\delta} \leq \frac{1}{2} \varepsilon \mu_{min}$, so once m is larger than both those values, Eq. (34) implies

$$\mathbb{P}\left(\sup_{v \in B_1} |\tilde{X}_v| > \varepsilon \mu_{min} \right) < \delta \tag{35}$$

as desired.

Acknowledgments

We thank John Wright for pointing out to us the recent independent work [Ada16].

References

[AAJ$^+$14] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, pages 123–137, 2014.

[ABGM14] Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *CoRR*, abs/1401.0579, 2014.

[Ada16] Radosław Adamczak. A note on the sample complexity of the Er-SpUD algorithm by Spielman, Wang and Wright for exact recovery of sparsely used dictionaries. *CoRR*, abs/1601.02049, 2016.

[AEB06] M. Aharon, M. Elad, and A. Bruckstein. SVDD: An algorithm for designing overcomplete dictionaries for sparse representation. *Trans. Sig. Proc.*, 54(11):4311–4322, November 2006.

[AGM14] Sanjeev Arora, Rong Ge, and Ankur Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Proceedings of The 27th Conference on Learning Theory (COLT)*, pages 779–806, 2014.

[AGMS15] Sanjeev Arora, Rong Ge, Ankur Moitra, and Sushant Sachdeva. Provable ICA with unknown gaussian noise, and implications for gaussian mixtures and autoencoders. *Algorithmica*, 72(1):215–236, 2015.

[BE08] Ori Bryt and Michael Elad. Compression of facial images using the K-SVD algorithm. *J. Visual Communication and Image Representation*, 19(4):270–282, 2008.

[BKS15] Boaz Barak, Jonathan A. Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the 47th Annual ACM on Symposium on Theory of Computing (STOC)*, pages 143–151, 2015.

[BLM13] Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BRV13] Mikhail Belkin, Luis Rademacher, and James R. Voss. Blind signal separation in the presence of gaussian noise. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 270–287, 2013.

[Dir15] Sjoerd Dirksen. Tail bounds via generic chaining. *Electron. J. Probab.*, 20(53):1–29, 2015.

[EA06] Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.

[FJK96] Alan M. Frieze, Mark Jerrum, and Ravi Kannan. Learning linear transformations. In *Proceedings of the 37th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 359–368, 1996.

[GVX14] Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier PCA and robust tensor decomposition. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 584–593, 2014.

[LT91] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, 1991.

[LV15] Kyle Luh and Van Vu. Random matrices: l_1 concentration and dictionary learning with few samples. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 1409–1425, 2015.

[LYB+14] Yuanqing Li, Zhu Liang Yu, Ning Bi, Yong Xu, Zhenghui Gu, and S.-I. Amari. Sparse representation for brain signal processing: A tutorial on methods and applications. *Signal Processing Magazine, IEEE*, 31(3):96–106, May 2014.

[MBP+08] Julien Mairal, Francis R. Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Supervised dictionary learning. In *Proceedings of the 22nd Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 1033–1040, 2008.

[MBP+09] Julien Mairal, Francis R. Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *IEEE 12th International Conference on Computer Vision (ICCV)*, pages 2272–2279, 2009.

[MBP14] Julien Mairal, Francis R. Bach, and Jean Ponce. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

[MBPS10] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[NR09] Phong Q. Nguyen and Oded Regev. Learning a parallelepiped: Cryptanalysis of GGH and NTRU signatures. *J. Cryptology*, 22(2):139–160, 2009.

[RBL+07] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the Twenty-Fourth International Conference on Machine Learning (ICML)*, pages 759–766, 2007.

[SQW15] Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere. *CoRR*, abs/1504.06785, 2015.

[SWW12] Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *The 25th Annual Conference on Learning Theory (COLT)*, pages 37.1–37.18, 2012. Full version: `http://arxiv.org/abs/1206.5882v1`.

[Tal14] Michel Talagrand. *Upper and lower bounds for stochastic processes: modern methods and classical problems*. Springer, 2014.

[VX15] Santosh Vempala and Ying Xiao. Max vs Min: Tensor decomposition and ICA with nearly linear sample complexity. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, pages 1710–1723, 2015.

Appendix

A Lower bounds for sample complexity of ER-SpUD(DC) algorithm in the Bernoulli-Rademacher model

In this section we prove that the modification introduced in this paper to **ER-SpUD(DC)** algorithm is necessary in order to guarantee the correctness for arbitrary Bernoulli-subgaussian X with strictly subquadratic number of samples p. More concretely, we prove that if X follows the Bernoulli-Rademacher model, and $p = \mathcal{O}(n^{2-\varepsilon} \log n)$, then the **ER-SpUD** algorithm actually fails to recover A and X with probability at least $1 - \tilde{\mathcal{O}}(\frac{1}{n^\varepsilon})$. The proof of this theorem relies on few technical lemmas, they are presented later in this section.

Theorem 28. *For every constant C and $\varepsilon \leq 1$, there exist C', such that for sufficiently large n if $2n \leq p \leq Cn^{2-\varepsilon} \log n$ and $X \in \mathbb{R}^{n \times p}$ follows the Bernoulli-Rademacher model with sparsity parameter $\theta := C' \frac{\log n}{n}$, then the **ER-SpUD(DC)** algorithm fails to recover X with probability at least $1 - \mathcal{O}(\frac{\log^5 n}{n^\varepsilon})$.*

Proof. We shall first prove that once following events happens simultaneously, the **ER-SpUD(DC)** algorithm fails to recover X. Later on we will prove that each of those events fails with probability at most $\mathcal{O}(\frac{\log^5 n}{n^\varepsilon})$ — that will be enough to conclude the statement of the theorem.

In what follows, let X_i be the i-th column of X, $j_* \in [n]$ be the index of the row of X with largest number of non-zero entries (for concreteness, the smallest such index), and K be some universal constant the same for fourth and fifth event, and will be specified later. Consider the following events

1. Matrix X is of full rank.
2. For every $i \in [p-1]$ it holds that $|\operatorname{support}(X_i) \cap \operatorname{support}(X_{i+1})| < 2$.
3. For every $i \in [p-1]$ it holds that $j_* \notin (\operatorname{support}(X_i) \cap \operatorname{support}(X_{i+1}))$.
4. Every column of X has at least $K \log n$ nonzero entries.
5. The number of rows of X with largest support size is smaller than $K \log n$.

Let us call those events $\mathcal{E}_1, \ldots, \mathcal{E}_5$ respectively. We claim that under $\mathcal{E}_1, \ldots, \mathcal{E}_5$, the j_*-th row of X would not be recovered by the **ER-SpUD(DC)** algorithm. Indeed, assume for the proof by contradiction, that solving the optimization problem $\min_w \|w^T Y\|_1$ subject to $r_*^T w = 1$, yields a solution such that $w^T Y$ is proportional to the j_*-th row of X, for some r_* which is sum of two consecutive columns of Y. By condition $\mathcal{E}_1$, it means that the solution to the equivalent problem of $\min_z \|z^T X\|_1$ subject to $b_*^T z = 1$ is $z = \pm e_{j_*}$, where b_* is a sum of corresponding columns of X.

Observe that, because the matrix X has entries in $\{-1, 0, 1\}$, the ℓ_1 norm and sparsity of each row is equal, that is for every k we have $\|e_k^T X\|_1 = \|e_k^T X\|_0$.

Now, by condition $\mathcal{E}_2$, at most one coordinate of b_* is of absolute value 2, and all other are either ± 1 or 0. Moreover, if the condition $\mathcal{E}_3$ holds, the entry with absolute value 2 is not j_*-th. If for some k, we had $(b_*)_k = 2$, then taking $z := \frac{1}{2} e_k$, would yield a feasible solution to the optimization problem mentioned above, and one with smaller value of objective function — $\frac{1}{2}\|e_k^T X\|_0$ as opposed to $\|e_{j_*}^T X\|_0$; similarly for $(b_*)_k = -2$. Therefore, all non-zero entries of b_* must have absolute value 1.

Moreover b_* has support of size at least $K \log n$ (as $\mathcal{E}_4$ holds and b_* is a sum of two columns of X with almost disjoint support by $\mathcal{E}_2$), hence $|\operatorname{support} b_*|$ is strictly larger than the number of rows of X with largest support — this number of rows is less than $K \log n$ by $\mathcal{E}_5$. In particular, there is some $k \in \operatorname{support}(b_*)$, such that the k-th row of X has strictly smaller support size than j_*-th row. Again, if this is the case $z = e_{j_*}$ is not a solution to the optimization problem $\min_z \|z^T X\|_1$ subject to $b_*^T z = 1$ — as $z := e_k$ (or $-e_k$) is feasible and with strictly smaller objective value. From this contradiction we conclude, that once all the events in preceding list hold simultaneously, the **ER-SpUD(DC)** algorithm fails in recovering X, and therefore fails to recover the hidden decomposition.

It is now enough to show that each of those events $\mathcal{E}_1 \dots \mathcal{E}_5$ fails with probability at most $\mathcal{O}(\frac{\log^5 n}{n^\varepsilon})$.

Event $\mathcal{E}_1$ fails with probability at most $n(1 - \frac{C' \log n}{n})^n \lesssim \frac{1}{n^{C'-1}}$ by Lemma 30 and assumption that $p > 2n$. For $C' \geq 1$ and large enough n this quantity is smaller than $\frac{\log^5 n}{n^\varepsilon}$.

For event $\mathcal{E}_2$, it holds with probability $1 - \mathcal{O}(\frac{\log^5}{n^\varepsilon})$ simply by union bound — for every fixed i, we have $\mathbb{P}(\operatorname{support}(X_i) \cap \operatorname{support}(X_{i+1}) \geq 2) \leq \binom{n}{2}\theta^4 \lesssim \frac{\log^4 n}{n^2}$ — and we need a union bound over $p \leq C n^{2-\varepsilon} \log n$ such events.

To bound the probability of event $\mathcal{E}_3$, let random set $S \subset [p]$ be the support of j_*-th row of X. In what follows we will condition implicitly on $S \neq \emptyset$ as S is empty with exponentially small probability.

Expected support size of any single row is $p\theta > 2 \log n$, therefore by Chernoff and union bound, we deduce that except with probability smaller than $\frac{1}{n}$ all rows of X has support size smaller than $C_2 p\theta$ for some universal constant C_2. Conditioned on $|S| < C_2 p\theta$, the distribution of S is invariant under permutations of $[n]$, in a sense that for fixed set $S_0 \subset [n]$ probability $\mathbb{P}(S = S_0 | |S| < C_2 p\theta)$ depends only on the size of S_0. In such a case, and because of additional conditioning on S being nonempty, by Lemma 31 for every $i \neq j$ we have

$$\mathbb{P}(i \in S | j \in S \wedge |S| < C_2 p\theta) \leq \mathbb{P}(i \in S | |S| < C_2 p\theta)$$

In particular, for fixed $i \in [p-1]$, we have

$$\begin{aligned}
\mathbb{P}\left(i \in S \wedge (i+1) \in S \,|\, |S| < C_2 p\theta\right) &\leq \mathbb{P}\left(i \in S | |S| < C_2 p\theta\right) \mathbb{P}\left(i+1 \in S | |S| < C_2 p\theta\right) \\
&\leq \frac{C_2^2 p^2 \theta^2}{p^2} \\
&\lesssim \frac{\log^2 n}{n^2}
\end{aligned}$$

Hence, by union bound over all $i \in [p-1]$, it follows that

$$\mathbb{P}(\neg \mathcal{E}_3) \leq \mathbb{P}(|S| \geq C_2 p\theta) + \mathbb{P}(\neg \mathcal{E}_3 | |S| < C_2 p\theta) \lesssim \frac{\log^3 n}{n^\varepsilon}$$

For $\mathcal{E}_4$, we know that the expected number of non-zero entries in a column of X is $\theta n = C' \log n$ — by the Chernoff bound, probability that any such column has sparsity smaller than $K \log n$ is much smaller than $\frac{1}{n^4}$ if we set C' large enough depending on K. Therefore by union bound, they all have sparsity at least $K \log n$ simultaneously with probability at least $1 - \frac{1}{n^2}$.

In order to bound probability of failure for the event $\mathcal{E}_5$, let $s_i \in \mathbb{N}$ for $i \in [n]$ be the size of the support of the i-th row of X. Clearly s_i are Binomial random variable with parameters (p, θ). Take $\gamma := \frac{K_1 \log n}{n}$ (with some constant K_1 that will be specified later) and let $T_0 \in [p]$ be the largest number such that $\mathbb{P}(s_i \geq T_0) \geq \gamma$. We want to apply Lemma 29 for all random variables s_i. Observe that in this setting $\mathbb{E}\, s_i = p\theta \ll \frac{p}{8}$, and on the other hand $\mathbb{E}\, s_i \geq 2 \log n$.

Moreover, observe that $T_0 \leq 4\, \mathbb{E}\, s_i$ — it is enough to show that $\mathbb{P}(s_i \geq 4\, \mathbb{E}\, s_i) \leq \gamma$, and this fact follows from Chernoff bound if K_1 is large enough constant. Therefore, we can apply Lemma 29 to conclude that

$$\mathbb{P}(S_i \geq T_0) \leq K_2 \gamma \max(1, \frac{T_0}{\mathbb{E}\, S_i}) \leq K_3 \gamma$$

where K_2 and K_3 are some universal constants.

Now we want to show that with probability at least $1 - \frac{1}{n}$, number of s_i that are not smaller than T_0, is between 1 and $4K_3\gamma n = 4K_3K_1 \log n =: K_4 \log n$. If we consider indicator random variables $M_i \in \{0, 1\}$, such that $M_i = 1$ if and only if $S_i \geq T_0$, by previous discussion we know that $\gamma \leq \mathbb{P}(M_i = 0) \leq K_3\gamma$, and all M_i are independent. We can now apply the Chernoff bound to bound the probability of $\mathbb{P}(\sum M_i < 1)$ and $\mathbb{P}(\sum M_i \geq 4K_3\gamma n)$ — again, if K_1 is large enough, each of those is much smaller than $\frac{1}{2n}$ — we can now fix constant K_1 large enough so that all three Chernoff bounds yield desired inequalities.

Finally, if the number of rows with support larger than T_0 is between one and $K_4 \log n$, then clearly at most $K_4 \log n$ has the largest support — and therefore the event $\mathcal{E}_5$ holds with $K := K_4$. $\square$

We now prove certain technical lemmas that were used in the proof of Theorem 28.

Lemma 29. *For $\theta \in (0, 1)$ and $p \in \mathbb{N}$, let Q be a Binomial random variable with parameters (p, θ) and let $\gamma \in (0, 1)$ be some fixed threshold. Moreover, let T_0 be the largest natural number such that*

$$\mathbb{P}(Q \geq T_0) \geq \gamma$$

and assume that $T_0 \leq \frac{p}{2}$. Then

$$\mathbb{P}(Q \geq T_0) \leq K\gamma \max(1, \frac{T_0}{\mathbb{E}\, Q})$$

for some universal constant K.

Proof. We start with bounding the ratio

$$\begin{aligned}
\frac{\mathbb{P}(Q = T_0)}{\mathbb{P}(Q = T_0 + 1)} &= \frac{\binom{p}{T_0}\theta^{T_0}(1-\theta)^{p-T_0}}{\binom{p}{T_0+1}\theta^{T_0+1}(1-\theta)^{p-T_0-1}} \\
&= \frac{T_0 + 1}{p - T_0} \cdot \frac{1-\theta}{\theta} \\
&\leq K_1 \frac{T_0}{p\theta} \\
&= K_1 \frac{T_0}{\mathbb{E}\, Q}
\end{aligned}$$

We can rephrase it as $\mathbb{P}(Q = T_0) \leq K_1 \frac{T_0}{\mathbb{E}\, Q} \mathbb{P}(Q = T_0 + 1)$, and clearly $\mathbb{P}(Q = T_0 + 1) \leq \mathbb{P}(Q \geq T_0 + 1)$. Therefore

$$\mathbb{P}(Q = T_0) \leq K_1 \frac{T_0}{\mathbb{E}\, Q} \mathbb{P}(Q \geq T_0 + 1) \tag{36}$$

We can now directly bound the desired probability as follows

$$\begin{aligned}
\mathbb{P}(Q \geq T_0) &= \mathbb{P}(Q = T_0) + \mathbb{P}(Q \geq T_0 + 1) \\
&\leq \left(K_1 \frac{T_0}{\mathbb{E}\, Q} + 1 \right) \mathbb{P}(Q \geq T_0 + 1) \\
&\leq \left(K_1 \frac{T_0}{\mathbb{E}\, Q} + 1 \right) \gamma
\end{aligned}$$

Where the last inequality follows from the assumption that T_0 were largest such that $\mathbb{P}(Q \geq T_0) \geq \gamma$. □

Lemma 30. *Let $X \in \mathbb{R}^{n \times p}$ follow the Bernoulli-Rademacher model with sparsity parameter θ and $p > n$. Then matrix X is of rank n with probability at least $1 - n(1-\theta)^{p-n}$.*

Proof. Let $W_i \subset \mathbb{R}^p$ be a subspace of $\mathbb{R}^p$ spanned by first $i-1$ rows of X. We wish to prove that i-th row of X lies in W_i with probability at most $(1-\theta)^{p-n}$ — if we do this, the claim will follow by the union bound. Fix some i, and let v be the i-th row of X; moreover, let $W^\perp$ be the orthogonal complement of W_i. Clearly $\dim W^\perp \geq p - n$. We will show that for any fixed $W^\perp$

$$\mathbb{P}(v \perp W^\perp) \leq (1-\theta)^{\dim W^\perp}$$

Indeed, let $q := \dim W^\perp$, and consider sequence of indices $i_1, \dots i_q$ together with a basis $u_1, \dots u_q$ of $W^\perp$ such that for every r we have $(u_r)_{i_r} \neq 0$, and for every pair $s < r$ we have $(u_s)_{i_r} = 0$ — such a basis and sequence of indices exists by Gaussian elimination. Now, by the chain rule, we have

$$\mathbb{P}\left(v \perp W^\perp\right) = \prod_{r=1}^{q} \mathbb{P}\left(\langle v, u_r \rangle = 0 | \forall_{s<r} \langle v, u_s \rangle = 0\right) \tag{37}$$

Let us fix some r now. We want to show that $\mathbb{P}\left(\langle v, u_r \rangle = 0 | \forall_{s<r} \langle v, u_r \rangle = 0\right) < (1-\theta)$. Observe that the event $\forall_{s<r} \langle v, u_s \rangle = 0$ is independent of v_{i_r}; moreover, if we fix all values of v_j for $j \neq i_r$, probability of $\langle v, u_r \rangle = 0$ is at most $(1-\theta)$ — there is at most one value of v_{i_r} that would make this inner product equal zero, and v_{i_r} assumes every value with probability at most $(1-\theta)$. Therefore

$$\begin{aligned}
\mathbb{P}\left(\langle v, u_r \rangle = 0 | \forall_{s<r} \langle v, u_r \rangle = 0\right) &= \mathbb{E}\left(\mathbb{P}\left(\langle v, u_r \rangle = 0 | v_1, \dots \hat{v_{i_r}} \dots v_p\right) | \forall_{s<r} \langle v, u_r \rangle = 0\right) \\
&\leq \mathbb{E}\left(1 - \theta | \forall_{s<r} \langle v, u_r \rangle = 0\right) \\
&= 1 - \theta
\end{aligned}$$

Where $v_1, \dots \hat{v_{i_r}} \dots v_d$ denotes omitting the i_r-th index in this sequence.

We can plug this back to Eq. (37) to conclude that $\mathbb{P}\left(v \perp W^\perp\right) \leq (1-\theta)^{\dim W^\perp}$ and the statement of the lemma follows. □

Lemma 31. *Let $S \subset [n]$ be a random set, with permutationally invariant distribution, i.e. such that for fixed S_0, $\mathbb{P}(S = S_0)$ depends only on the size of S_0. Assume moreover, that S is nonempty almost surely. Then for any $i \neq j \in [n]$, we have $\mathbb{P}(i \in S | j \in S) \leq \mathbb{P}(i \in S)$.*

Proof. For $k \in \{0, \dots n\}$, let $p_k := \mathbb{P}(|S| = k)$. Observe that for fixed $i \in S$

$$\mathbb{P}(i \in S) = \sum_{k=1}^{n} p_k \, \mathbb{P}(i \in S \| S| = k) = \sum_{k=1}^{n} p_k \frac{k}{n} \tag{38}$$

On the other hand

$$\begin{aligned}
\mathbb{P}(i \in S | j \in S) &= \frac{1}{1 - p_0} \left(\sum_{k=1}^{n} p_k \, \mathbb{P}(i \in S | j \in S, |S| = k) \right) \\
&= \frac{1}{1 - p_0} \left(\sum_{k=1}^{n} p_k \frac{k-1}{n-1} \right) \\
&= \sum_{k=1}^{n} p_k \frac{k-1}{n-1}
\end{aligned}$$

where the last equality follows from the assumption $\mathbb{P}(S = \emptyset) = 0$.

Then the statement of the lemma follows by explicitly comparing two expressions for $\mathbb{P}(i \in S)$ and $\mathbb{P}(i \in S | j \in S)$, and using inequality $\frac{k-1}{n-1} \leq \frac{k}{n}$. $\square$