An Improved Analysis of the ER-SpUD Dictionary Learning Algorithm*

Jarosław Błasiok[†1] **and Jelani Nelson**[‡2]

1 **Harvard University, Cambridge, MA, USA**
jblasiok@g.harvard.edu
2 **Harvard University, Cambridge, MA, USA**
minilek@g.harvard.edu

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 [24] 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 [17].

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 [24], and contradicts a result of [17], which claimed that $p \gtrsim n \log^4 n$ guarantees high probability of success for the original **ER-SpUD** algorithm.

1998 ACM Subject Classification I.2.6 Learning

Keywords and phrases dictionary learning, stochastic processes, generic chaining

Digital Object Identifier 10.4230/LIPIcs.ICALP.2016.44

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

* A full version of the paper is available at http://arxiv.org/abs/1602.05719.
† JB was supported by NSF grant IIS-1447471.
‡ JN was supported by NSF grant IIS-1447471 and CAREER CCF-1350670, ONR grant N00014-14-1-0632 and Young Investigator N00014-15-1-2388, and a Google Faculty Research Award.

43rd International Colloquium on Automata, Languages, and Programming (ICALP 2016).
Editors: Ioannis Chatzigiannakis, Michael Mitzenmacher, Yuval Rabani, and Davide Sangiorgi;
Article No. 44; pp. 44:1–44:14


Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

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 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 $n \leq m$) [2].

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 [12], 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. [21, 10, 12]; see [19] for more references), image classification [23, 20] as well as neurobiology [16]. Given its huge practical importance, a number of effective heuristics for dictionary learning were proposed [3, 18] – 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

Until recently there was little theoretical understanding of the dictionary learning problem. Spielman, Wang and Wright in [24] 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, [24] 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 and columns of A. 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.

[1] As written, their work has certain errors which we discuss later in detail. Nevertheless, using some of our approaches we believe it should be possible to salvage their sample complexity bound in the

ref	sample complexity	noise	overcomplete	sparsity	arbitrary dict.
[24]	$\mathcal{O}(n^2 \log^2 n)$	No	No	$\mathcal{O}(\sqrt{n})$	Yes
[2]	$\mathcal{O}(m^2)$	No	Yes	$\mathcal{O}(n^{1/4})$	No
[5]	$\mathcal{O}(m^2 s^{-2} + s^2 m)$	Yes	Yes	$\mathcal{O}(\min(m^{2/5}, \frac{\sqrt{n}}{\log n}))$	No
[5]	$\mathcal{O}(\mathrm{poly}(m))$	Yes	Yes	$\mathcal{O}(n^{1/2-\epsilon})$	No
[4]*	$\mathcal{O}(\mathrm{poly}(m))$	No	Yes	$\mathcal{O}(n/\mathrm{polylog}(n))$	No
[7]	$\mathcal{O}(\mathrm{poly}(m))$	Yes	Yes	$\mathcal{O}(n^{1-\epsilon})$	Yes
[7]*	$\mathcal{O}(\mathrm{poly}(m))$	Yes	Yes	$\mathcal{O}(n)$	Yes
[25]	$\mathcal{O}(\mathrm{poly}(m, \kappa(A)))$	No	No	$\mathcal{O}(n)$	Yes
[27]	$\mathcal{O}(\mathrm{poly}(n))$	Yes	No	$\mathcal{O}(n)$	Yes
[17][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.

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. [2], and independently Arora et al. [5], 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 [2] requires $p = \tilde{\mathcal{O}}(m^2)$ samples, where $\tilde{\mathcal{O}}(f) = \mathcal{O}(f \cdot \log^{O(1)}(f))$. A more detailed analysis of the dependence between sparsity and number of samples was provided in the work [5] for their algorithm – for $s = \mathcal{O}(\min(\frac{\sqrt{n}}{\log n}, m^{2/5}))$, they require $\tilde{\Omega}(m^2 s^{-2} + m s^2)$ 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 requires $\mathcal{O}(m^C)$ 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 [5] also proves correct recovery by this algorithm in the presence of noise. Later Arora et al. [4] gave a quasipolynomial time algorithm working for sparsity up to $\mathcal{O}(n/\mathrm{polylog}(n))$, but under much stronger assumptions

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 the full version for details.

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. [7] proposed an algorithm fitting in the Sum-of-Squares framework, which works in polynomial time for sparsity $\mathcal{O}(n^{1-\epsilon})$ for any constant $\epsilon > 0$ 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. [25] proposed a polynomial time algorithm for the case when $n = m$ and sparsity is as large as $\mathcal{O}(n)$. Their result works in a similar model as in [24], 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 [24]). The sample complexity depends polynomially on n and the condition number of the dictionary matrix A. In particular, in the low sparsity regime ($s = \Theta(\text{polylog}(n))$), 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) [13, 22, 8, 6, 14, 27] 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, [27, 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 (i.e. dictionaries for which m may be larger than n), (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 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 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 [17] made significant progress toward showing that the **ER-SpUD** algorithm proposed in [24] 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 [24], and if they all occurred then **ER-SpUD** performed correct recovery. The work [17] analyzed arguably the most complex of these events more efficiently, showing a certain crucial inequality held with good probability when $p \gtrsim n \log^4 n$ ($x \gtrsim y$ means that $x > Ky$ for some universal constant $K > 0$). Unfortunately there is a gap: [24] 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 [17] does not provide such a probabilistic analysis. More seriously, there are other events defined in [24]

which *require* $p \gtrsim n^2$ to hold whp in the Bernoulli-subgaussian model (except in the case the subgaussians are actual gaussians), and [17] did not discuss these events at all. In fact, in the full version we prove that in the Bernoulli-Rademacher model the **ER-SpUD** algorithm of [24] actually *requires* $p \gtrsim n^{1.99}$ to succeed with probability even polynomially small in n, contradicting the main result of [17] 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)**" for the noiseless case with $m = n$, 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 [24].

Furthermore, the work of [17] 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" [17, 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 [24]. The only difference between DCv2 and the original DC variant in [24] 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 show in the full version, one of the several conditions in [24] necessary for their proof of successful recovery of (A, X) from Y actually requires $p = \Omega(n^2)$ if using the DC variant, and hence our switch to DCv2 allows p to be reduced to $\mathcal{O}(n \log n)$.

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 [24] 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 [24] then outlines certain conditions for X that, if they hold, guarantee correct recovery of (A, X). We now state these deterministic conditions, as per [24], which imply correct recovery of (A, X) via **ER-SpUD** when they all simultaneously hold.

(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 $(11/9)\theta p$.

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

1. For all pairs $j_1 < j_2 \in \{1, \dots, p\}$
   Let $\boldsymbol{r}_j = \boldsymbol{Y}\boldsymbol{e}_{j_1} + \boldsymbol{Y}\boldsymbol{e}_{j_2}$
   Solve $\min_{\boldsymbol{w}} \|\boldsymbol{w}^T\boldsymbol{Y}\|_1$ subject to $\boldsymbol{r}_j^T\boldsymbol{w} = 1$, and set $\boldsymbol{s}_j = \boldsymbol{w}^T\boldsymbol{Y}$.
   $j \leftarrow j + 1$

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.

(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 support$(z_*) \subseteq$ 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 [24] 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** ([24]). *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 [24], and made explicit in [17], that to analyze the probability **(P1)** holding as a function of p, it suffices to prove some upper bound on some stochastic process. Namely, [17] 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 [24, 17] 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 [17] was not made explicit and was only given as $o(1)$, so it does not clearly allow for this union bound.

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 [24]. Next, more importantly, we 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 [26]. 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 $p \gtrsim \theta^{-1} \log(n/\delta)$ – this requires only a minor change in the analysis of [24].

Finally, in Lemma 4 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 the full version).

1.3 Recent and independent work

In a recent and independent work, Adamczak showed a main result similar to ours [1]. 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 $1 - 1/p$. Unlike our analysis which is based on Bernstein's inequality and generic chaining, the proof in [1] 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 [1] was not given to dependence of p on the failure probability δ, and the analysis in our full version 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 define **(P1')**, **(P2')** as follows.

(P1') For every b that can be expressed as the sum of two columns of X, $|S| < p/4$ and

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

where $C > 0$ is some fixed constant, $J = \mathrm{support}(b)$, $\bar{J} = [n] \backslash J$, 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 $0 < |J| \leq q = 1/(8\theta)$ and $|b|_{(2)}/|b|_{(1)} \leq 1/2$, then the solution to the restricted problem

$$\|z^T X_{J,*}\|_1 \text{ subject to } (b')^T z = 1 \tag{5}$$

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.

In the full version, we show that it suffices that **(P1')** and **(P2')** hold instead of **(P1)** and **(P2)** to guarantee correctness of **ER-SpUD(DCv2)**. In particular, we show the following lemma.

▶ **Lemma 2.** *Suppose conditions* **(P0)**, **(P1')**, **(P2')**, *and* **(P3)** *all hold. Then* ***ER-SpUD*** *and* ***Greedy*** *from Figure 2 recover* (A', X') *such that* $X' = \Pi DX$ *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*).*

In the full version, we then 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 Lemma 2 implies that **ER-SpUD** has the desired correctness guarantee under this same regime for p, θ.

▶ **Theorem 3.** *For* $p \gtrsim n \log(n/\delta)$ *and* $1/n \lesssim \theta \lesssim 1/\sqrt{n}$,

$$\mathbb{P}(\neg(\mathbf{P0}) \vee \neg(\mathbf{P1'}) \vee \neg(\mathbf{P2'}) \vee \neg(\mathbf{P3})) < \delta \tag{6}$$

▶ Remark. Here we sketch just how **(P1')** is analyzed in Theorem 3, similarly to the discussion in [24, 17]. This reveals why our chaining result, Theorem 12, is relevant.

For **(P1')**, the analysis is almost identical to the proofs of [24, Lemma 11] and [17, 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:

$$\forall v \in \mathbb{R}^{|\bar{J}|},\ \|v^T X_{\bar{J},*}\|_1 - 2\|v^T X_{\bar{J},S}\|_1 > Cp\sqrt{\frac{\theta}{|\bar{J}|}}\|v\|_1 \tag{7}$$

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 $\Pi = X^T_{\bar{J},*}$ and applying Theorem 12 with our choice of p, with probability at least $1 - \delta/p^2$, for all $v \in B_1$,

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

where $(X_{\bar{J},*})_{*,1}$ clumsily denotes the first column of the matrix $X_{\bar{J},*}$. The last inequality follows from [24, 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 12 to $\Pi = (X')^T$, with probability at least $1 - \delta/p^2$,

$$\forall v \in B_1,\ \|v^T X'\|_1 \le \frac{3}{2}\,\mathbb{E}\,\|v^T X'\|_1 = \frac{3p}{8}\,\mathbb{E}\,|v^T X'_{*,1}| = \frac{3p}{8}\alpha(v). \tag{9}$$

Then by combining (8), (9) and scaling by $\|v\|_1$, we see that the left hand side of (7) is at least $\frac{p}{8}\alpha(v) \gtrsim p\sqrt{\frac{\theta}{|J|}}\|v\|_1$, with the inequality following from [24, Lemma 16].

We are now going to sketch how to show that with probability $1-\delta$ condition **(P3)** holds. The full proof is in the full version.

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. We prove the following in the full version.

▶ **Lemma 4.** *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.*

3 Chaining background

We now provide some preliminary definitions and results we will need to prove Theorem 12. As per Lemma 2 and the proof of Theorem 3, Theorem 12 fits in to show that **ER-SpUD** achieves correct recovery with probability $1-\delta$ for $p \gtrsim n\log(n/\delta)$ and $1/n \lesssim \theta \lesssim 1/\sqrt{n}$.

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 the 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.

▶ **Definition 5** (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 $|T_k| \le 2^{2^k}$ for $k \ge 1$ and $|T_0| = 1$.

▶ **Definition 6** (Gamma functionals)**.** For a metric space (T,d) we define

$$\gamma_\alpha(T,d) := \inf_{(T_k)} \sup_{x\in T} \sum_{k=0}^{\infty} 2^{k/\alpha} d(x,T_k) \tag{10}$$

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 7.** *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 8** (Generic chaining [26], 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: $\mathbb{P}\left(|X_u - X_v| > \lambda\right) \lesssim \exp\left(-\frac{\lambda^2}{d_1(u,v)^2}\right) + \exp\left(-\frac{\lambda}{d_2(u,v)}\right)$. Then $\mathbb{E} \sup_{u\in T} |X_u| \lesssim \gamma_2(T, d_1) + \gamma_1(T, d_2)$.*

▶ **Theorem 9** (Dirksen, [11])**.** *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 $\Delta(T, d) := \sup_{u,v\in T} d(u, v)$.

▶ **Theorem 10** (Majorizing measures [26], 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 $\mathbb{E} \sup_{t\in T} \langle g, t\rangle \simeq \gamma_2(T, d_2)$, where d_p is the metric induced by the ℓ_p norm.*

▶ **Theorem 11** ([26], 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 $\mathbb{E} \sup_{t\in T} \langle t, x\rangle \simeq \gamma_2(T, d_2) + \gamma_1(T, d_\infty)$*

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 12.** *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 $m = \Omega(\varepsilon^{-2} n \log \frac{n}{\delta})$,*

$$\mathop{\mathbb{P}}_{\Pi}\left(\sup_{v\in B_1} |\|\Pi v\|_1 - \mathbb{E}\, \|\Pi v\|_1| > \varepsilon \cdot \mathbb{E}\, \|\Pi v\|_1\right) < \delta \tag{11}$$

We now prove the theorem. Define $B_1 := \{t \in \mathbb{R}^n : \|t\|_1 \leq 1\}$. For each $v \in B_1$, consider $\tilde{X}_v := \|\Pi v\|_1 - \mathbb{E}\, \|\Pi v\|_1$. We wish to prove that with high probability over Π we have $\sup_{v\in B_1} |\tilde{X}_v| \leq \varepsilon\mu_{min}$, where $\mu_{min} := m\sqrt{\frac{\theta}{n}}$ is such that for every $v \in B_1$ we have $\mathbb{E}\, \|\Pi v\|_1 \geq \mu_{min}$ (see [24, Lemma 16] for a proof). Let $\pi_1, \dots \pi_m$ be the rows of matrix Π. With each $v \in B_1$ we associate another random variable $X_v := \sum_{i=1}^m \sigma_i |\langle \pi_i, v\rangle|$, with the σ_i being independent Rademachers.

▶ **Lemma 13.** *For every integer p we have*

$$\| \sup_{v\in B_1} |\tilde{X}_v| \|_p \lesssim \| \sup_{v\in B_1} |X_u| \|_p . \tag{12}$$

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

have

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

Now each summand $|\langle\pi_i,v\rangle| - |\langle\pi_i,v\rangle|$ is symmetric random variable, and they are independent. We can thus introduce independent random signs σ_i without altering the distribution:

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

◀

We will first analyze tail behavior of the random variable $\sup_{v\in B_1}|X_v|$, and then use Lemma 13 together with [15, Lemma 4.10] to obtain tail bounds for the random variable of original interest $\sup_{v\in B_1}|\tilde{X}_v|$.

In order to use Theorem 9 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 14.** *For every pair of points $u, v \in B_1$, we have*

$$\mathbb{P}(|X_u - X_v| > \lambda) \lesssim \exp\left(-\frac{\lambda^2}{2m\theta\|u-v\|_2^2}\right) + \exp\left(-\frac{\lambda}{\|u-v\|_\infty}\right) \tag{13}$$

Proof. We can write

$$X_u - X_v = \sum_{i=1}^m \sigma_i(|\langle\pi_i,u\rangle| - |\langle\pi_i,v\rangle|) \tag{14}$$

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 $|Q_i| = ||\langle\pi_i,u\rangle| - |\langle\pi_i,v\rangle|| \le |\langle\pi_i,u-v\rangle|$. Observe that each π_{ij} is $(\sqrt{2\theta},1)$-subgamma. Here we say a random variable Z is (σ,B)-subgamma if $\mathbb{E}\,Z = 0$ and $\psi_Z(\lambda) \le \lambda^2\sigma^2/(2(1-B\lambda))$ for all $|\lambda| < 1/|B|$, where $\psi_Z(\lambda) = \ln \mathbb{E}\,e^{\lambda Z}$. By basic properties of subgamma random variables (see [9, Section 2.4]), 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 $|Q_i| \le |\langle\pi_i,u-v\rangle|$ always, 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 [9, Section 2.4] implies the tail bound

$$\mathbb{P}\left(\left|\sum_{i=1}^m Q_i\right| > \lambda\right) \lesssim \exp\left(\frac{\lambda^2}{2m\theta\|u-v\|_2^2}\right) + \exp\left(\frac{\lambda}{\|u-v\|_\infty}\right) \tag{15}$$

◀

With this lemma in hand, we can use Theorem 9, to deduce the tail bound for supremum of $|X_v|$.

$$\mathbb{P}\left(\sup_{v\in B_1}|X_u| > M + \sqrt{u}D_1 + uD_2\right) < e^{-u} \tag{16}$$

Where $M := C_1(\gamma_2(B_1, \sqrt{2m\theta}d_2) + \gamma_1(B_1, d_\infty))$, $D_1 := C_2\Delta(B_1, \sqrt{2m\theta}d_2)$, and $D_2 := C_3\Delta(B_1, d_\infty)$, with d_2, d_∞ being 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

$$\mathbb{P}\left(\sup_{v\in B_1}|\tilde{X}_u| > L(M + \sqrt{u}D_1 + uD_2)\right) < e^{-u} \tag{17}$$

for some universal constant L.

Indeed it is known (see [15, Lemma 4.10]) that tail bounds of the form Eq. (16) imply moment bounds of the form $\|\sup_{v\in B_1}|X_v|\|_p \lesssim M + \sqrt{p}D_1 + pD_2$. By Lemma 13, the same (up to a constant) p-norm bounds are true for $\sup_{v\in B_1}|\tilde{X}_v|$. Finally, [15, Lemma 4.10] also implies similar tail behavior of the random variable $\sup_{v\in B_1}\tilde{X}_v$, as in Eq. (17).

If we set $u := \log\frac{1}{\delta}$ in Eq. (17), 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 $\gamma_1(B_1, d_\infty)$. We prove the following in the full version, using Theorem 10 and Theorem 11.

▶ **Fact 15.** *$\gamma_2(B_1, d_2) \lesssim \sqrt{\log n}$ and $\gamma_1(B_1, d_\infty) \lesssim \log n$.*

Fact 15 together with previous discussion yield an upper bound $M \lesssim \sqrt{m\theta\log n} + \log n$.

Moreover, as $d_2(u, v) \le d_1(u, v)$ for any $u, v \in \mathbb{R}^n$, where d_1 is the metric induced by the 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

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

and similarly $\Delta(B_1, d_\infty) = 2$. Altogether, we have following inequalities: $M \lesssim \sqrt{m\theta\log n} + \log n$, $D_1 \lesssim \sqrt{m\theta}$, and $D_2 \lesssim 1$. Plugging this back to Eq. (17), we have

$$\mathbb{P}\left(sup_{v\in B_1}|\tilde{X}_v| < L_2\left(\sqrt{m\theta(\log n + \log\frac{1}{\delta})} + \log n + \log\frac{1}{\delta}\right)\right) < \delta \tag{18}$$

where again L_2 is some constant.

The following inequalities are equivalent: $L_2\sqrt{m\theta\log\frac{n}{\delta}} \le \frac{1}{2}\varepsilon\mu_{min}$, $L_2\sqrt{m\theta\log\frac{n}{\delta}} \le \frac{1}{2}\varepsilon\sqrt{\frac{\theta}{n}}m$, and $\frac{4L_2^2}{\varepsilon^2}n\log\frac{n}{\delta} \le m$. Similarly, the assumption $\theta \ge \frac{1}{n}$ implies that if $m > \frac{2L_2}{\varepsilon}n\log\frac{n}{\delta}$, then also $L_2\log\frac{n}{\delta} \le \frac{1}{2}\varepsilon\mu_{min}$, so once m is larger than both those values, Eq. (18) implies $\mathbb{P}\left(\sup_{v\in B_1}|\tilde{X}_v| > \varepsilon\mu_{min}\right) < \delta$, as desired.

Acknowledgments. We thank John Wright for pointing out to us the recent independent work [1].

References

1. 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.
2. 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.
3. 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.
4. Sanjeev Arora, Aditya Bhaskara, Rong Ge, and Tengyu Ma. More algorithms for provable dictionary learning. *CoRR*, abs/1401.0579, 2014.
5. 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.
6. 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.
7. 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.
8. 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.
9. Stephane Boucheron, Gabor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.
10. 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.
11. Sjoerd Dirksen. Tail bounds via generic chaining. *Electron. J. Probab.*, 20(53):1–29, 2015.
12. 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.
13. 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.
14. 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.
15. Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, 1991.
16. 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. doi:10.1109/MSP.2013.2296790.
17. 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.
18. 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.
19. 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.

20 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.

21 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.

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

23 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.

24 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`.

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

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

27 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.