Average Case Analysis of Multichannel Sparse Recovery Using Convex Relaxation

Yonina C. Eldar, *Senior Member, IEEE* and Holger Rauhut

***Abstract*—In this paper, we consider recovery of jointly sparse multichannel signals from incomplete measurements. Several approaches have been developed to recover the unknown sparse vectors from the given observations, including thresholding, simultaneous orthogonal matching pursuit (SOMP), and convex relaxation based on a mixed matrix norm. Typically, worst-case analysis is carried out in order to analyze conditions under which the algorithms are able to recover any jointly sparse set of vectors. However, such an approach is not able to provide insights into why joint sparse recovery is superior to applying standard sparse reconstruction methods to each channel individually. Previous work considered an average case analysis of thresholding and SOMP by imposing a probability model on the measured signals. In this paper, our main focus is on analysis of convex relaxation techniques. In particular, we focus on the mixed $\ell_{2,1}$ approach to multichannel recovery. We show that under a very mild condition on the sparsity and on the dictionary characteristics, measured for example by the coherence, the probability of recovery failure decays exponentially in the number of channels. This demonstrates that most of the time, multichannel sparse recovery is indeed superior to single channel methods. Our probability bounds are valid and meaningful even for a small number of signals. Using the tools we develop to analyze the convex relaxation method, we also tighten the previous bounds for thresholding and SOMP.**

Key Words: Multichannel sparse recovery, mixed-norm optimization, average performance, thresholding, simultaneous orthogonal matching pursuit

I. Introduction

Recovery of sparse signals from a small number of measurements is a fundamental problem in many different signal processing tasks such as image denoising [8], analog-to-digital conversion [31], [19], [32], radar, compression, inpainting, and many more. The recent framework of compressed sensing (CS), founded in the works of Donoho [15], Candès, Romberg and Tao [8], studies acquisition methods as well as efficient computational algorithms that allow reconstruction of a sparse vector x from linear measurements $y = Ax$, where $A \in \mathbb{R}^{n\times N}$ is referred to as the measurement matrix. The key observation is that y can be relatively short, so that $n < N$, and still contain enough information to recover x.

Yonina C. Eldar is with the Technion—Israel Institute of Technology, Haifa Israel. Email: yonina@ee.technion.ac.il. Holger Rauhut is with the Hausdorff Center for Mathematics and Institute for Numerical Simulation, University of Bonn, Germany. Email: rauhut@hcm.uni-bonn.de.

The work of Y. Eldar was supported by the Israel Science Foundation under Grant no. 1081/07 and by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (contract no. 216715). H. Rauhut acknowledges support by the WWTF project SPORTS (MA 07-004) and by the Hausdorff Center for Mathematics.

Determining the sparsest vector x consistent with the data $y = Ax$ is generally an NP-hard problem [14]. To determine x in practice, a multitude of efficient algorithms have been proposed, [14], [18], [43], [7], [9], which achieve high recovery rates. The basis pursuit (BP), or ℓ_1-minimization approach, is the most extensively studied recovery method [12], [8], [15], [35]. The use of general purpose or specialized convex optimization techniques [26], [18] allows for efficient reconstruction using this strategy. Although greedy methods, such as simple thresholding or orthogonal matching pursuit (OMP), are faster in practice, BP provides significantly better recovery guarantees. In particular, there exist measurement matrices $A \in \mathbb{R}^{n\times N}$ that allow for stable recovery of all k-sparse vectors as long as $n \geq Ck\log(N/k)$ where C is a constant. Such uniform recovery is not possible for simple thresholding or OMP [16], [36]. (We note, however, that the recent greedy algorithms CoSaMP [33] and ROMP [34] are able to provide such uniform guarantees.) In practice, the recovery rate of BP when averaged over all random sparse vectors is typically better than that predicted by the theory. This is due to the fact that existing analysis considers the ability of BP to recover all vectors x. On the other hand, in random simulations, the worst-case instance of x typically does not occur. Therefore, considering the behavior of various recovery methods over random x often leads to more characteristic behavior.

The BP principle as well as greedy approaches have been extended to the multichannel setup where the signal consists of several channels with joint sparsity support [47], [45], [22], [13], [11], [30], [20], [21]. In [2] the buzzword distributed compressed sensing was coined for this setup. An alternative approach is to first reduce the problem to a single channel problem that preserves the sparsity pattern, and recover the signal support set; given the support, the measurements can be inverted to recover the input [30]. A variety of different recovery results have been established that provide conditions ensuring that the output of the proposed efficient algorithms coincides with the true signals. In [11] a recovery result was derived for a mixed $\ell_{p,1}$ program in which the objective is to minimize the sum of the ℓ_p-norms of the rows of the estimated matrix whose columns are the unknown vectors. Recovery results for the more general problem of block-sparsity were developed in [21] based on the block restricted isometry property (RIP), and in [20] based on mutual coherence. In practice, multichannel reconstruction techniques perform much better than recovering each channel individually. However, the theoretical equivalence results predict no performance gain. The reason is that these results apply to all possible input signals, and are therefore worst-case measures. Clearly, if we input the

same signal to each channel, then no additional information on the joint support is provided from multiple measurements. Therefore, in this worst-case scenario there is no advantage for multiple channels.

In order to capture more closely the true underlying behavior of existing algorithms and observe a performance gain when using several channels, we consider an average-case analysis. In this setting, the inputs are considered to be random variables. The idea is to develop conditions on the measurement matrix A such that the inputs can be recovered with high probability given a certain input distribution.

Recently, there have been several papers that consider sparse recovery with random ensembles. In [46] random subdictionaries of A are considered and analyzed. This allows to obtain average results for BP with a single input channel. In [40], average-case performance of single channel thresholding was studied. In [25], [24] extensions to two multichannel recovery algorithms were developed: thresholding and simultaneous OMP (SOMP) [25], [24]. Under a mild condition on the sparsity and on the matrix A, the probability of reconstruction failure decays exponentially with the number of channels. In the present paper we contribute to this line of research by analyzing the average-case performance of multichannel BP, *i.e.,* mixed $\ell_{2,1}$-minimization [45], [22], [21], [20]. The tools we derive in this context are then also used to slightly improve previous bounds on average performance of multichannel thresholding and SOMP.

The theoretical average-case results we develop for multichannel BP are superior to the average bounds developed on thresholding and SOMP. For an equally mild or even milder condition on the sparsity and on the matrix A, we obtain faster exponential decay of the failure probability with respect to the number of channels. Thus, in this sense, the extension of BP to the multichannel case is superior to existing greedy algorithms, just as in the single channel setting. Moreover, our recovery results are applicable also in the single channel case whereas previous results [25] require a large number of channels to yield meaningful (*i.e.,* positive) probability bounds (although our new bound for thresholding generalizing the one in [40] does not suffer from this drawback). Note, however, that in simulations SOMP often exhibits the best performance. This may be explained by the fact that the bounds are not tight (at least for SOMP).

To develop our probability bounds, we rely on a new sufficient condition that ensures recovery of the exact signal set via $\ell_{2,1}$-minimization. This condition generalizes a result of [44], [23] to the multichannel setting, and is weaker than existing multichannel recovery conditions. Our average-case analysis is then carried out assuming that the elements of the input signal are drawn at random. We prove that under a certain restriction on A and the sparsity set S, the sufficient condition we develop is satisfied with high probability. The restriction we impose is that the ℓ_2-norm of $A_S^\dagger a_\ell$ over all ℓ not in the set S is bounded, where a_ℓ is the ℓth column of A, and $A_S^\dagger$ is the pseudo inverse of the restriction of A to the columns in S. This is an improvement over known worst-case recovery conditions which require a bound on the ℓ_1-norm [11], [20], and are therefore stronger. Loosely speaking, we will show that while worst-case results based on the coherence limit the sparsity level to order $\sqrt{n}$, average-case analysis shows that sparsity up to order n may enable recovery with high probability. In terms of RIP, instead of bounding the restricted isometry constant for sparsity sets of size $2k$, we will only need to consider sets of size $k+1$.

The remaining of the paper is organized as follows. In Section II we introduce our problem and briefly summarize known equivalence results between the $\ell_{2,1}$ approach for multichannel recovery and NP-hard combinatorial optimization that recovers the true signals. A new recovery condition is derived in Section III, which is weaker than previous results, and will be instrumental in developing our average-case analysis in Section IV. Since the probability bounds we develop depend on the 2-norm of $A_S^\dagger a_\ell$, in Section V we derive several upper bounds on this norm. In Section VI we use the tools developed in the previous section to derive new bounds on the average performance of thresholding and SOMP, that are tighter than existing results and also applicable to a broader set of problems. We then compare our bounds on multichannel BP to these results. Finally, in Section VII we present several simulations demonstrating the behavior of the different methods.

Throughout the paper, we denote by A_S the submatrix of A consisting of the columns indexed by $S \subset \{1,\ldots,N\}$, while X^S is the submatrix of X consisting of the rows of X indexed by S. The ℓth column of A is denoted by a_ℓ or A_ℓ. For a matrix A, $\|A\|_2$ is the spectral norm of A, *i.e.,* the largest singular value, and A^* is its conjugate transpose. The unit sphere in $\mathbb{R}^L$ is defined by $S^{L-1} = \{x \in \mathbb{R}^L, \|x\|_2 = 1\}$; the complex counterpart is denoted $S_{\mathbb{C}}^{L-1} = \{x \in \mathbb{C}^L, \|x\|_2 = 1\}$.

II. Multichannel ℓ_1-Minimization

A. Problem Formulation

We consider multichannel signal recovery where our goal is to recover a jointly-sparse matrix $X \in \mathbb{C}^{N\times L}$ from n linear measurements per channel. Here N denotes the signal length and L the number of channels, *i.e.,* the number of signals. We assume that X is jointly k-sparse, meaning that there are at most k rows in the matrix X that are not identically zero. More formally, we define the support of the matrix X as

$$\operatorname{supp} X = \bigcup_{\ell=1}^{L} \operatorname{supp} X_\ell, \tag{1}$$

where the support of the ℓth column is

$$\operatorname{supp} X_\ell = \{j, X_{j\ell} \neq 0\}. \tag{2}$$

Our assumption is that $\|X\|_0 := |\operatorname{supp} X| \leq k$. The measurements are given by

$$Y = AX, \quad Y \in \mathbb{C}^{n\times L}, \tag{3}$$

where $A \in \mathbb{C}^{n\times N}$ is a given measurement matrix. Each measurement vector $Y_\ell = AX_\ell$ corresponds to a measurement of the corresponding signal X_ℓ.

The natural approach to determine X given Y is to solve

the ℓ_0-minimization problem

$$\min_X \|X\|_0 \quad \text{s.t.} \quad AX = Y. \tag{4}$$

However, (4) is NP hard in general [14]. Several alternative methods have been proposed, that have polynomial complexity [47], [45], [22], [13], [11], [30], [20], [21], [30]. A variety of different equivalence results between the solution of the ℓ_0-problem and the output of the proposed efficient algorithm. In [11] an equivalence result was derived for a mixed $\ell_{p,1}$ program in which the objective is to minimize the sum of the ℓ_p-norms of the rows of the estimated matrix whose columns are the unknown vectors. The condition is based on mutual coherence, and turns out to be the same as that obtained from a single measurement problem, so that the joint sparsity pattern does not lead to improved recovery capabilities as judged by this condition. Recovery results for the more general problem of block-sparsity were developed in [21] based on the RIP, and in [20] based on mutual coherence. Reducing these results to the multiple measurement vectors (MMV) setting leads again to conditions that are the same as in the single measurement case. An exception is the work in [25], [24] which considers average-case performance of thresholding and SOMP. Under a mild condition on the sparsity and on the matrix A, the probability of reconstruction failure decays exponentially with the number of channels L. In Section VI we slightly improve on these bounds using the tools developed in this paper.

In Section IV we follow a similar approach and treat the average behavior of the mixed $\ell_{2,1}$-minimization program [45], [22], [21] defined by

$$\min \|X\|_{2,1} = \sum_{j=1}^{N} \|X^j\|_2, \quad \text{subject to } AX = Y, \tag{5}$$

which promotes joint sparsity, as argued for instance in [22]. In the single channel case $L = 1$ this is the usual BP principle. Therefore, our results can also be used to deduce the average-case behavior of the BP method. This is in contrast to [25], in which the recovery results derived are not applicable to the single channel case. As we discuss in Section VI, our theoretical results are superior to the previous average-case analysis of [25] in the sense that we use an equally mild or even milder condition on the sparsity and on the matrix A, but at the same time get a faster exponential decay of the failure probability with respect to the number of channels L.

B. Recovery Results

Recovery results for the program (5) were considered in [11], [21], [20]. In particular, the lemma below is derived in [11] and follows also from [20] where the more general case of block sparsity is considered.

Proposition 2.1: Let $S \subset \{1, \dots, N\}$ and suppose that

$$\|A_S^\dagger a_\ell\|_1 < 1 \quad \text{for all } \ell \notin S, \tag{6}$$

with $A_S^\dagger = (A_S^* A_S)^{-1} A_S^*$ denoting the pseudo-inverse of A_S. Then (5) recovers all $X \in \mathbb{C}^{N \times L}$ with $\operatorname{supp} X = S$ from $Y = AX$.

Note, that the condition above does not depend on the number of channels. In the next section we will derive a condition similar to (6) that involves the 2-norm instead of the 1-norm, and is therefore weaker (namely, easier to satisfy).

Assuming the columns of A are normalized, $\|a_\ell\|_2 = 1$, we can guarantee that (6) holds as long as the coherence μ of A is small enough, where [17]

$$\mu = \max_{j \neq \ell} |\langle a_j, a_\ell \rangle|. \tag{7}$$

The following result follows from [20] by noting that the block coherence in this setting is equal to μ/d.

Proposition 2.2: Assume that

$$(2k - 1)\mu < 1. \tag{8}$$

Then (5) recovers all X with $\|X\|_0 \leq k$ from $Y = AX$.

Under the same conditions as in Propositions 2.1 and 2.2, it is shown in [43] that BP will recover a single k-sparse vector. Therefore, if (6) holds, then instead of solving (5) we could also use BP on each of the columns of Y.

The coherence is lower bounded by [41]

$$\mu \geq \sqrt{\frac{N - n}{n(N - 1)}}. \tag{9}$$

The lower bound behaves like $1/\sqrt{n}$ for large N, which limits the Proposition 2.2 to maximal sparsities $k = \mathcal{O}(\sqrt{n})$. To improve on this we can generalize existing recovery results [8], [6] based on RIP to the multichannel setup. The restricted isometry constant δ_k of a matrix A is defined to be the smallest constant δ_k such that

$$(1 - \delta_k)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k)\|x\|_2^2, \tag{10}$$

for all k-sparse vectors x. The next proposition follows from [21].

Proposition 2.3: Assume $A \in \mathbb{C}^{n \times N}$ with $\delta_{2k} < \sqrt{2} - 1$ Let $X \in \mathbb{C}^{N \times L}$, $Y = AX$, and let $\overline{X}$ be the minimizer of (5). Then

$$\|X - \overline{X}\|_{\mathrm{F}} \leq C k^{-1/2} \|X - \hat{X}^{(k)}\|_{2,1}$$

where C is a constant, $\|X\|_{\mathrm{F}} = \sqrt{\operatorname{Tr}(X^* X)}$ is the Frobenius norm of X and $\hat{X}^{(k)}$ denotes the best k-term approximation of X, *i.e.*, $\operatorname{supp} \hat{X}^{(k)}$ consists of the indices corresponding to the k largest row norms $\|X^\ell\|_2$. In particular, recovery is exact if $|\operatorname{supp} X| \leq k$.

It is well known that Gaussian and Bernoulli random matrices $A \in \mathbb{R}^{n \times N}$ satisfy $\delta_{2k} \leq \sqrt{2} - 1$ with high probability as long as [1], [10]

$$n \geq Ck \log(N/k). \tag{11}$$

For random partial Fourier matrices the respective condition is $n \geq ck \log^4(N)$ [37], [39]. Therefore, Proposition 2.3 allows for a smaller number of measurements. However, there is still no dependency on the number of channels. Indeed, under the same RIP condition BP will recover a single k-sparse vector and therefore, as before, BP may as well be applied to each of the columns of Y individually.

We conclude this overview by stressing that known equivalence results do not improve on those for single channel sparse recovery. In [21], [20] equivalence results are derived for a mixed $\ell_{2,1}$ program when different measurement matrices

A_i are used on each channel. In this case, even worst-case analysis shows improvement over $L = 1$. However, when all measurement matrices are equal, the recovery conditions do not show any advantage with multiple signals.

III. A Recovery Condition

Before turning to analyze the average-case behavior of (5), we first develop a new condition on A that allows for perfect recovery. This formulation will be useful in deriving the average-case results.

In the following theorem we give a sufficient condition on the minimizers of (5). This theorem generalizes a result of [44], [23] for the $L = 1$ case. To this end we denote by $\operatorname{sgn}(X) \in \mathbb{C}^{N\times L}$ the matrix with entries

$$\operatorname{sgn}(X)_{\ell j} = \begin{cases} \frac{X_{\ell j}}{\|X^\ell\|_2}, & \|X^\ell\|_2 \neq 0; \\ 0, & \|X^\ell\|_2 = 0. \end{cases} \tag{12}$$

In this definition, each element of X is normalized by the norm of the corresponding row. When $L = 1$, $\operatorname{sgn}(X)$ reduces to the sign of the elements of the vector x.

Theorem 3.1: Let $X \in \mathbb{C}^{N\times L}$ with $\operatorname{supp} X = S$ and assume A_S to be non-singular. If there exists a matrix $H \in \mathbb{C}^{n\times L}$ such that

$$A_S^* H = \operatorname{sgn}(X^S), \tag{13}$$

and

$$\|H^* a_\ell\|_2 < 1 \quad \text{for all } \ell \notin S \tag{14}$$

then X is the unique solution of (5).

Before proving the theorem we note that the two conditions on H easily imply that

$$\|H^* a_\ell\|_2 \leq 1, \quad \text{for all } \ell. \tag{15}$$

Proof: The proof follows the ideas of [44], with appropriate modifications to account for the mixed $\ell_{2,1}$ norm that replaces the ℓ_1 norm.

Let $Y = AX$, and assume there exists a matrix H such that X, H satisfy (13) and (14). Let X' be an alternative matrix satisfying $Y = AX'$. Our goal is to show that $\|X\|_{2,1} < \|X'\|_{2,1}$. To this end, we note that

$$\|X\|_{2,1} = \|X^S\|_{2,1} = \operatorname{Tr}\left(\operatorname{sgn}(X^S)(X^S)^*\right), \tag{16}$$

where Tr denotes the trace. Substituting $A_S^* H = \operatorname{sgn}(X^S)$ into (16), and using the cyclicity of the trace we have

$$\begin{aligned} \|X\|_{2,1} &= \operatorname{Tr}\left(H^* A_S X^S\right) = \operatorname{Tr}\left(H^* A X'\right) \\ &= \operatorname{Tr}\left(X'^{S'} H^* A_{S'}\right), \end{aligned} \tag{17}$$

where we used the fact that $A_S X^S = Y = AX'$ and S' denotes the support of X'. We next rely on the following lemma.

Lemma 3.2: Let A, B be matrices such that AB is defined. Then $|\operatorname{Tr}(BA)| \leq \|B\|_{2,1} \max_\ell \|A_\ell\|_2$, with strict inequality if $\|A_\ell\|_2 < \max_\ell \|A_\ell\|_2$ for some value of ℓ for which $\|B^\ell\|_2 \neq 0$.

Proof: The proof follows from noting that

$$\begin{aligned} |\operatorname{Tr}(BA)| &\leq \sum_\ell |B^\ell A_\ell| \leq \sum_\ell \|B^\ell\|_2 \|A_\ell\|_2 \\ &\leq \max_\ell \|A_\ell\|_2 \sum_\ell \|B^\ell\|_2 = \max_\ell \|A_\ell\|_2 \|B^\ell\|_{2,1}, \end{aligned}$$

where the second inequality is a result of applying Cauchy-Schwartz. Under the condition of the lemma, we have strict inequality in the last inequality. ∎

Applying Lemma 3.2 to (17), leads to

$$\begin{aligned} \|X\|_{2,1} &\leq \|X'^{S'}\|_{2,1} \max_{\ell\in S'} \|H^* A_\ell\|_2 \leq \|X'^{S'}\|_{2,1} \\ &= \|X'\|_{2,1}, \end{aligned} \tag{18}$$

where the last inequality follows from (15). We have strict inequality in the first inequality of (18) as long as the values $\|H^* A_\ell\|_2$ for $\ell \in S'$ are not all equal since $\|X'^\ell\|_2 \neq 0$ for all $\ell \in S'$ be definition of the support.

Suppose to the contrary that $\|H^* A_\ell\|_2 = a$ for all $\ell \in S'$. Clearly, S' must contain at least one index ℓ that is not contained in S; otherwise $S' \subset S$, which would contradict the hypothesis that A_S is non-singular, $A_{S'} X' = A_S X$ and $X \neq X'$. By our assumption $\|H^* a_\ell\|_2 < 1$, which then implies that $a < 1$ or $\|H^* A_\ell\|_2 < 1, \ell \in S'$. The inequalities in (18) then become

$$\|X\|_{2,1} \leq \|X'^{S'}\|_{2,1} \max_{\ell\in S'} \|H^* A_\ell\|_2 < \|X'^{S'}\|_{2,1} = \|X'\|_{2,1}. \tag{19}$$

Thus, we have shown that $\|X'\|_{2,1} > \|X\|_{2,1}$ for any X' such that $Y = AX'$, and therefore (5) recovers the true sparse matrix X. ∎

Choosing $H = (A_S^\dagger)^* \operatorname{sgn}(X_S)$ in Theorem 3.1 results in the following corollary.

Corollary 3.3: Let $X \in \mathbb{C}^{N\times L}$ with $\operatorname{supp} X = S$ and assume A_S to be non-singular. If

$$\|\operatorname{sgn}(X^S)^* A_S^\dagger a_\ell\|_2 < 1 \quad \text{for all } \ell \notin S, \tag{20}$$

then X is the unique minimizer of (5).

This corollary will be instrumental in proving the average-case performance of (5). It can easily be seen that Corollary 3.3 implies Proposition 2.1. This follows from the triangle inequality,

$$\begin{aligned} \left\|\operatorname{sgn}(X^S)^* A_S^\dagger a_\ell\right\|_2 &= \left\|\sum_{j\in S} (A_S^\dagger a_\ell)_j \operatorname{sgn}(X^j)^*\right\|_2 \\ &\leq \sum_{j\in S} |(A_S^\dagger a_\ell)_j| \ \|\operatorname{sgn}(X^j)\|_2 = \|A_S^\dagger a_\ell\|_1, \end{aligned}$$

where we used the fact that $\|\operatorname{sgn}(X^j)\|_2 = 1$.

IV. Average Case Analysis

Intuitively, we would expect multichannel sparse recovery to perform better than single channel recovery. However, in the worst case setting this is not true as already suggested by the results of Section II. The reason is very simple. If each channel carries the same signal, $X_\ell = x$ for $\ell = 1, \ldots, L$, then also the components of $Y = AX$ are all the same and we do not have more information on the support of X than

provided by a single component Y_ℓ. The following proposition establishes formally that if BP fails for a given measurement matrix A, then multichannel optimization (5) will fail as well so that in the worst-case, adding channels will not improve performance.

Proposition 4.1: Suppose there exists a k-sparse vector $x \in \mathbb{C}^N$ that ℓ_1-minimization is not able to recover from $y = Ax$. Then $\ell_{2,1}$-minimization fails to recover $X = (x|x|\cdots|x) \in \mathbb{C}^{N\times L}$ from $Y = AX$.

Proof: If ℓ_1-recovery fails on some k-sparse x then necessarily $\|x'\|_1 \leq \|x\|_1$ for some x' satisfying $Ax' = Ax$. Clearly $X = (x|x|\cdots|x)$ is (jointly) k-sparse and $AX = AX'$ for $X' = (x'|x|\cdots|x')$. Furthermore,

$$\|X'\|_{2,1} = \sqrt{L}\|x'\|_1 \leq \sqrt{L}\|x\|_1 = \|X\|_{2,1}$$

and therefore X is not the unique minimizer of the $\ell_{2,1}$-minimization problem. ∎

Realizing that (5) is not more powerful than usual BP in the worst case, we seek an average-case analysis. This means that we impose a probability model on the k-sparse X. In particular, as in [25], we will assume that on the support S of size k the coefficients of X are chosen at random. We then show that under a suitable probability model on the non-zero elements of X, the condition given by Corollary 3.3 is satisfied with high probability, which depends on L.

We follow the probability model used in [25]: let S be the joint support of cardinality k. On S the coefficients are given by

$$X^S = \Sigma\Phi \tag{21}$$

where $\Sigma = \operatorname{diag}(\sigma_j, j \in S) \in \mathbb{R}^{k\times k}$ is an arbitrary diagonal matrix with positive diagonal elements σ_j. The matrix Φ will be chosen at random according to one of the following models.

- **Real Gaussian:** each entry of $\Phi \in \mathbb{R}^{k\times L}$ is chosen independently from a standard normal distribution.
- **Real spherical**: the rows of $\Phi \in \mathbb{R}^{k\times L}$ are chosen independently and uniformly at random from the real sphere S^{L-1}.
- **Complex Gaussian**: the real and imaginary parts of each entry of $\Phi \in \mathbb{C}^{k\times L}$ are chosen independently according to a standard normal distribution.
- **Complex spherical**: the rows of $\Phi \in \mathbb{C}^{k\times L}$ are chosen independently and uniformly at random from the complex sphere $S^{L-1}_{\mathbb{C}}$.

Note that taking Σ to be the identity matrix results in a standard Gaussian random matrix X^S, while taking arbitrary non-zero σ_j's on the diagonal of Σ allows for different variances. The matrix Σ may be deterministic or random. In particular, choosing Σ to be the matrix with diagonal elements given by the inverse ℓ_2-norm of the rows of Φ in the real (complex) Gaussian model, leads to a matrix X^S with a real (complex) spherical distribution.

In Theorems 4.4 and 4.5 below we develop conditions under which (5) recovers X from $Y = AX$ with probability that decays exponentially with L. The condition in both theorems is given in terms of an upper bound on $\|A_S^\dagger a_\ell\|_2$ for ℓ not in S. This is in contrast to the worst-case result of Proposition 2.1 that is given in terms of $\|A_S^\dagger a_\ell\|_1$ and therefore stronger. The essential idea in both proofs is to show that if the bound on $\|A_S^\dagger a_\ell\|_2$ is satisfied, then the sufficient condition of Corollary 3.3 holds with high probability.

Before stating the first theorem, we derive the following result on the norm of sums of independent random vectors, uniformly distributed on a sphere.

Theorem 4.2: Let $a \in \mathbb{C}^k$ and let Z_j, $j = 1,\ldots,k$, be a sequence of independent random vectors which are uniformly distributed on the real sphere S^{L-1}. Then for any $u > 1$

$$\mathbb{P}\left(\left\|\sum_{j=1}^k a_j Z_j\right\|_2 \geq u\|a\|_2\right) \leq \exp\left(-\frac{L}{2}(u^2 - \log(u^2) - 1)\right).$$

Proof: See Appendix A. ∎

Theorem 4.2 generalizes the Bernstein inequality for Steinhaus sequences in [46, Theorem 13] to higher dimensions. We may extend the estimate easily to random vectors uniformly distributed on complex unit spheres.

Corollary 4.3: Let $a \in \mathbb{C}^k$ and let Z_j, $j = 1,\ldots,k$, be a sequence of independent random vectors which are uniformly distributed on the complex sphere $S^{L-1}_{\mathbb{C}}$. Then for any $u > 1$

$$\mathbb{P}\left(\left\|\sum_{j=1}^k a_j Z_j\right\|_2 \geq u\|a\|_2\right) \leq \exp\left(-L(u^2 - \log(u^2) - 1)\right).$$

Proof: First observe that $a_j Z_j$ has the same distribution as $|a_j| Z_j$. We may therefore assume without loss of generality that $a_j \in \mathbb{R}$. Next, a random vector $Z \in S^{L-1}_{\mathbb{C}}$ is uniformly distributed on $S^{L-1}_{\mathbb{C}}$ if and only if $(\operatorname{Re}(Z)^T, \operatorname{Im}(Z)^T)^T$ is uniformly distributed on the real sphere S^{2L-1}. Applying Theorem 4.2 with L replaced by $2L$ yields the statement. ∎

With this tool at hand we can now easily prove the following average-case recovery theorem.

Theorem 4.4: Let $S \subset \{1,\ldots,N\}$ be a set of cardinality k and suppose

$$\|A_S^\dagger a_\ell\|_2 \leq \alpha < 1 \quad \text{for all } \ell \notin S. \tag{22}$$

Let $X \in \mathbb{R}^{N\times L}$ with $\operatorname{supp} X \subset \{1,\ldots,N\}$ such that the coefficients on S are given by (21) with some diagonal matrix $\Sigma \in \mathbb{R}^{k\times k}$ and $\Phi \in \mathbb{R}^{k\times L}$ chosen from the real Gaussian or spherical probability. Then with probability at least

$$1 - N\exp\left(-\frac{L}{2}(\alpha^{-2} - \log(\alpha^{-2}) - 1)\right) \tag{23}$$

(5) recovers X from $Y = AX$.

If the real probability model is replaced by one of the two complex models then $L/2$ can be replaced by L in (23).
For $\alpha < 1$ we are guaranteed that the exponent in (23) has a negative argument, and therefore the error decays exponentially in L.

Proof: First observe that by the rotational invariance of Gaussian random vectors the columns of $\operatorname{sgn}(X^S)^* = \operatorname{sgn}(\Phi^*)$ are independent and uniformly distributed on the real sphere, and the same is also true if we use the real

spherical random model. Denote $b^{(\ell)} = A_S^\dagger a_\ell$ for $\ell \notin S$ and by Z_j, $j = 1, \dots, k$ a sequence of independent random vectors that are uniformly distributed on the sphere S^{L-1}. Using the sufficient recovery condition of Corollary 3.3, the union bound and Theorem 4.2 we can estimate the probability that $\ell_{2,1}$ minimization fails to recover X by

$$\begin{aligned}
&\mathbb{P}(\max_{\ell \notin S} \| \operatorname{sgn}(X^S)^* b^{(\ell)} \|_2 > 1) \\
&\quad \le \sum_{\ell \notin S} \mathbb{P}(\| \operatorname{sgn}(X^S)^* b^{(\ell)} \|_2 > 1) \\
&\quad \le \sum_{\ell \notin S} \mathbb{P}\left(\left\| \sum_{j=1}^k b_j^{(\ell)} Z_j \right\|_2 > \alpha^{-1} \|b^{(\ell)}\|_2 \right) \\
&\quad \le (N-k) \exp\left(-\frac{L}{2}(\alpha^{-2} - \log(\alpha^{-2}) - 1) \right).
\end{aligned}$$

The complex case follows analogously using Corollary 4.3. ■

For $L = 1$, Theorem 4.4 is contained implicitly in [46, Theorem 13]. The appearance of the 2-norm in (24) instead of the 1-norm as in (6) makes the condition of the theorem weaker than worst-case estimates (recall that $\|x\|_2 \le \|x\|_1 \le \sqrt{k}\|x\|_2$ for any length-k vector x). In Section V this will be made more evident when we consider conditions on the coherence μ and the RIP constant to allow for recovery with high probability. The requirement we obtain on μ is weaker than that of Proposition 2.2 and allows for recovery with k on the order of n, while the worst-case results limit recovery to order $\sqrt{n}$. Furthermore, in contrast to the worst-case results which depend on δ_{2k}, we will show that high-probability recovery is possible as long as δ_{k+1} is small enough.

It is evident from (23) that the failure probability decays exponentially with growing number of channels L. Moreover, the bound is also useful for small L, and in particular for the monochannel case $L = 1$. Indeed, a simple algebraic manipulation shows that the failure probability is less than ϵ provided $\|A_S^\dagger a_\ell\|_2 \le \alpha$ for all $\ell \notin S$ with α satisfying

$$\alpha^{-2} - \log(\alpha^{-2}) \ge \frac{2\log(N/\epsilon)}{L} + 1.$$

This provides a useful average-case analysis even for $L = 1$.

For completeness, we also state an alternative recovery result below which provides a slightly better probability estimate than Theorem 4.4 for very large values of N. However, the required condition on $\|A_S^\dagger a_\ell\|_2$ is stronger.

Theorem 4.5: Let $S \subset \{1, \dots, N\}$ be a set of cardinality k, and let $X \in \mathbb{R}^{N \times L}$ be random sparse coefficients with $\operatorname{supp} X = S$ given by the real Gaussian probability model. If

$$\|A_S^\dagger a_\ell\|_2 < \frac{A_L}{3\sqrt{L} + 2\sqrt{k}} =: \gamma \sim \frac{1}{3 + 2\sqrt{k/L}} \tag{24}$$

for all $\ell \notin S$, where

$$A_L = \sqrt{2}\frac{\Gamma((L+1)/2)}{\Gamma(L/2)} \sim \sqrt{L}, \tag{25}$$

and Γ denotes the Gamma function, then with probability at least

$$P = 1 - \exp(-L/8) - k\exp(-A_L^2/8)$$

(5) recovers X from $Y = AX$.

It follows from Stirling's formula $\Gamma(z) \sim \sqrt{2\pi z} z^{z-1/2} e^{-z}$, that

$$\begin{aligned}
A_L &= \sqrt{2}\frac{\Gamma((L+1)/2)}{\Gamma(L/2)} \sim \sqrt{2}\frac{((L+1)/2)^{L/2} e^{-(L+1)/2}}{(L/2)^{(L-1)/2} e^{-L/2}} \\
&= e^{-1/2}\frac{(L+1)^{L/2}}{L^{(L-1)/2}} = e^{-1/2}\left(L(1+1/L)^L\right)^{1/2} \sim \sqrt{L}.
\end{aligned}$$

Moreover, for all $L \ge 1$ it holds that $\sqrt{L} \ge A_L \ge \sqrt{\frac{2}{\pi}}\sqrt{L} \approx 0.797\sqrt{L}$.

Note that $\gamma = \frac{A_L}{3\sqrt{L}+2\sqrt{k}}$ is monotonically increasing in L. In addition, the probability P is also increasing (towards 1) in L. Therefore, more channels increase the probability of success and in addition relax the requirements on the matrix A.

Proof: To prove the theorem we show that if (24) is satisfied, then condition (20) of Corollary 3.3 holds with probability P.

To this end, let $\Phi \in \mathbb{R}^{k \times L}$ denote a random matrix with independent standard normal distributed entries, and define D as the $k \times k$ diagonal matrix with diagonal elements $1/s_j, j \in S$, where $s_j = \|\Phi^j\|_2 = \sqrt{\sum_{\ell=1}^L |\Phi_{j\ell}|^2}$. We can then express $\operatorname{sgn}(X^S) = \operatorname{sgn}(\Sigma\Phi) = \operatorname{sgn}(\Phi) = D\Phi$. (This equation also means that the diagonal matrix Σ does not play any role.) Denoting $b_j = A_T^\dagger a_j$ for $j \notin S$,

$$\| \operatorname{sgn}(X_S)^* b_j \|_2 = \|\Phi^* D b_j\|_2 \le \|\Phi\|_2 \|D\|_2 \|b_j\|_2.$$

By the assumption of the theorem $\|b_j\|_2 < \gamma$ where γ is defined by (24). It therefore remains to bound $\|\Phi\|_2$ and $\|D\|_2$. From [10, equation (4.35)], see also [42], the operator norm of Φ satisfies

$$\|\Phi\|_2 \le \sqrt{L} + \sqrt{k} + r \tag{26}$$

with probability at least $1 - \exp(-r^2/2)$.

Next we consider $\|D\|_2$. Observe that the s_j^2 are $\chi^2(L)$ distributed. Therefore, denoting a $\chi^2(L)$-variable by Y,

$$\begin{aligned}
\mathbb{E}[s_j] = \mathbb{E}[\sqrt{Y}] &= \frac{1}{2^{L/2}\Gamma(L/2)} \int_0^\infty \sqrt{x} x^{L/2} e^{x/2} dx \\
&= \sqrt{2}\frac{\Gamma((L+1)/2)}{\Gamma(L/2)} = A_L \sim \sqrt{L}.
\end{aligned}$$

As a function of Φ^j the s_j are Lipschitz continuous, *i.e.*, $s_j(\Phi^j - \Psi^j) \le \|\Phi^j - \Psi^j\|_2$. Using these two observations we rely on the following standard concentration of measure result, see e.g. [28, eq. (2.35)] or [29, eq. (1.6)].

Theorem 4.6: Let f be a Lipschitz function on $\mathbb{R}^L$, *i.e.*, $|f(x) - f(y)| \le B\|x - y\|_2$ for all $x, y \in \mathbb{R}^L$. Further assume that $Z = (Z_1, Z_2, \dots, Z_L)$ is a vector of independent standard Gaussian random variables. Then

$$\begin{aligned}
\mathbb{P}(f(Z) \ge \mathbb{E}[f(Z)] + t) &\le \exp\left(-\frac{t^2}{2B^2}\right), \\
\mathbb{P}(f(Z) \le \mathbb{E}[f(Z)] - t) &\le \exp\left(-\frac{t^2}{2B^2}\right).
\end{aligned}$$

Our goal is to show that $\|D\|_2$ is bounded from above, which is equivalent to bounding the smallest value of s_j from below.

Applying Theorem 4.6 to s_j,

$$\mathbb{P}(s_j < A_L(1-t)) \le \exp(-t^2 A_L^2/2),$$

where we used the fact that $B = 1$ and $\mathbb{E}[s_j] = A_L$. Using a union bound over all j, we obtain

$$\begin{aligned}\mathbb{P}(s_j < A_L(1-t), \forall j) &= \mathbb{P}\left(\min_{j=1,\dots,k} s_j < A_L(1-t)\right) \\ &\le \sum_{j\in S} \mathbb{P}(s_j < A_L(1-t)) \le k\exp(-t^2 A_L^2/2).\end{aligned}$$

Assuming that $\min_{j\in S} s_j \ge A_L(1-t)$ holds, $\|D\|_2 \le 1/(A_L(1-t))$. Combining this bound with (26) for $r = \sqrt{L}s$ we have

$$\begin{aligned}\|\operatorname{sgn}(X_S)A_S^\dagger a_j\|_2 &\le \frac{\sqrt{k}+\sqrt{L}+s\sqrt{L}}{A_L(1-t)}\gamma \\ &= \frac{(s+1+\sqrt{k/L})\gamma\sqrt{L}}{(1-t)A_L}.\end{aligned}$$

Choosing $s = t = 1/2$,

$$\|\operatorname{sgn}(X_S)^* A_S^\dagger a_j\|_2 \le (3+2\sqrt{k/L})\gamma\sqrt{L}/A_L < 1. \quad (27)$$

From (27) and Corollary 3.3, X is recoverable using (5).

The probability that (27) does not hold can be computed by applying a union bound to the probabilities that the spectral norms of each of the matrices Φ and D are not bounded. This shows that (27) does not hold with probability at most $\exp(-L/8) + k\exp(-A_L^2/8)$ completing the proof of the theorem. ∎

V. Bounded Norm Condition

Both Theorems 4.4 and 4.5 state that X can be recovered with high probability from Y, as long as $\|A_S^\dagger a_\ell\|_2$ is bounded. In this section we develop several different conditions under which this holds.

Proposition 5.1: Let $A \in \mathbb{C}^{n\times N}$ have unit-norm columns and coherence μ, and let $S \subset \{1,\dots,N\}$ be a set of cardinality k. Assume that

$$(\sqrt{k} + (k-1)\delta)\mu < \delta \quad (28)$$

for some $\delta > 0$. Then $\|A_S^\dagger a_\ell\|_2 \le \delta$ for all $\ell \notin S$.

Proof: Gershgorin's disk theorem implies that the smallest eigenvalue $\lambda_{\min}$ of $A_S^* A_S$ is bounded from below by $1-(k-1)\mu$. In particular, $A_S^* A_S$ is invertible provided $(k-1)\mu < 1$. Further,

$$\|A_S^* a_\ell\|_2 = \sqrt{\sum_{j\in S} |\langle a_\ell, a_j\rangle|^2} \le \sqrt{k}\mu,$$

since by definition, $|\langle a_\ell, a_j\rangle| \le \mu$. Now, using the fact that $A_S^\dagger = (A_S^* A_S)^{-1} A_S^*$,

$$\begin{aligned}\|A_S^\dagger a_\ell\|_2 &\le \|(A_S^* A_S)^{-1}\|_2 \|A_S^* a_\ell\|_2 \\ &\le (1-(k-1)\mu)^{-1}\sqrt{k}\mu < \delta,\end{aligned}$$

where the last inequality follows from the fact that (28) implies $\delta > \sqrt{k}/(1-(k-1)\mu)^{-1}$. ∎

Condition (28) is slightly weaker than (8) as long as $\delta > 1/\sqrt{k}$. This follows from the 2-norm that replaced the 1-norm in the upper bound. However, (28) still suffers the square-root bottleneck $k = \mathcal{O}(\sqrt{n})$. To improve on this result, we next provide a condition based on the following refinement of the RIP of A. For a set $S \subset \{1,\dots,N\}$ we let

$$\delta(S) = \|A_S^* A_S - I\|_2.$$

The restricted isometry constant δ_k of (10) satisfies $\delta_k = \max_{|S|\le k} \|A_S^* A_S - I\|_2$ so that if S has cardinality k then $\delta(S) \le \delta_k$. We further define

$$\delta^*(S) = \max_{\ell\notin S} \delta(S\cup\{\ell\}). \quad (29)$$

Clearly, $\delta(S) \le \delta^*(S) \le \delta_{k+1}$. Finally, we make use of the following "local" 2-coherence function,

$$\mu_2(S) = \max\left\{\max_{\ell\notin S}\|A_S^* a_\ell\|_2, \max_{\ell\in S}\|A_{S\setminus\ell}^* a_\ell\|_2\right\} \quad (30)$$

for a subset $S \subset \{1,\dots,N\}$, where $S\setminus\ell$ denotes the elements in S excluding the ℓth one. From the definition of the coherence it follows immediately that

$$\mu_2(S) \le \sqrt{|S|}\mu, \quad (31)$$

since the magnitude of each element $|\langle a_\ell, a_j\rangle|$ of the vector $A_S^* a_\ell$ is bounded above by μ. In addition,

$$\mu_2(S) \le \delta^*(S). \quad (32)$$

This is a result of the fact that $A_S^* a_\ell$ is a submatrix of $A_{S\cup\{\ell\}}^* A_{S\cup\{\ell\}} - I$ for $\ell \notin S$, while $A_{S\setminus\{\ell\}}^* a_\ell$ is a submatrix of $A_S^* A_S - I$ for $\ell \in S$. (They both consist of a subcolumn of the respective matrix, that "leaves" out the diagonal element.) We now use these definitions to bound $\|A_S^\dagger a_\ell\|_2$:

Proposition 5.2: Let $S \subset \{1,\dots,N\}$. Then:

(a) If A satisfies $\delta^*(S) \le \delta < 1/2$ then

$$\|A_S^\dagger a_\ell\|_2 \le \frac{\delta}{1-\delta} < 1 \quad \text{for all } \ell \notin S.$$

(b) If A satisfies $\delta(S) \le \delta < 1$ and $\mu_2(S) \le \eta$ then

$$\|A_S^\dagger a_\ell\|_2 \le \frac{\eta}{1-\delta}.$$

Proof: Denoting by λ an eigenvalue of $A_S^* A_S$, the definition of $\delta(S) \le \delta^*(S) \le \delta$ implies that $|1-\lambda| \le \delta$. Consequently, the smallest eigenvalue of $A_S^* A_S$ is bounded from below by $1-\delta$ and therefore

$$\|(A_S^* A_S)^{-1}\|_2 \le \frac{1}{1-\delta}.$$

For (a), as already noted above, $A_S^* a_\ell$ for $\ell \notin S$ is a $k\times 1$ submatrix of $A_{T\cup\ell}^* A_{T\cup\ell} - I$. Therefore, $\|A_S^* a_\ell\|_2 \le \|A_{T\cup\ell}^* A_{T\cup\ell} - I\|_2 \le \delta$, and

$$\begin{aligned}\|A_S^\dagger a_\ell\|_2 &\le \|(A_S^* A_S)^{-1} A_S^* a_\ell\|_2 \\ &\le \|(A_S^* A_S)^{-1}\|_2 \|A_S^* a_\ell\|_2 \le \frac{\delta}{1-\delta}.\end{aligned}$$

The proof of (b) follows from the fact that $\|A_S^* a_\ell\|_2 \le \mu_2(S)$. A similar estimate as above yields $\|A_S^\dagger a_\ell\|_2 \le (1-\delta)^{-1}\eta$. ∎

Proposition 5.2 applies if δ_{k+1} is small while in contrast Theorem 2.3 works with δ_{2k}, which is generally larger than δ_{k+1}. By (11) the condition $\delta_{k+1} \le \delta$ can be satisfied if $n \ge$

$C_\delta k \log(N/k)$. Working with $\delta^*(S)$ instead of δ_{k+1} allows to improve on the bound (11) for Gaussian, Bernoulli and random spherical matrices.

Proposition 5.3: Let $S \subset \{1,\dots,N\}$ be a set of cardinality k and suppose that $A = \frac{1}{\sqrt{n}}\Phi \in \mathbb{R}^{n\times N}$, where Φ is drawn at random according to a standard Gaussian or Bernoulli distribution (with expectation 0 and variance $1/n$). Then $\delta^*(S) \le \delta$ with probability at least $1-\epsilon$ provided that

$$n \ge C_1 \delta^{-2} \max\{k \log(1/\delta), \log(N/\epsilon)\} \tag{33}$$

for a suitable constant.

The same statement holds as well (with possibly a different constant) for a random matrix whose columns are chosen independently at random according to the uniform distribution on a sphere.

Proof: See Appendix B. ∎

A straightforward extension of the proof, as in [1], also shows that a random matrix $A \in \mathbb{R}^{n\times N}$ with independent columns drawn from the uniform distribution on the sphere satisfies RIP, $\delta_k \le \delta$ with probability at least $1-\epsilon$ provided $n \ge C\delta^{-2}(k\log(N/k) + \log(\epsilon^{-1}))$. Although this fact seems to be known [48], we are not aware of reference where this is rigorously proven.

The next result relies on a theorem by Tropp [46, Theorem B] that uses random support sets S and allows to work with the coherence μ alone. Note that choosing S at random is perfectly in line with an average-case analysis.

Theorem 5.4: Let $A \in \mathbb{C}^{n\times N}$ have unit norm columns and coherence μ. Let $S \subset \{1,\dots,N\}$ be a set of cardinality $k \ge 4$ chosen uniformly at random. Let $\delta, \epsilon \in (0,1)$ and assume that

$$\mu^2 k \log(\epsilon^{-1}) \le c\delta^2, \tag{34}$$

$$\frac{k}{N}\|A\|_2^2 \le \frac{\delta}{4e^{1/4}}, \tag{35}$$

where $c = \frac{\log(2)e^{-1/2}}{4\cdot 144\log(3)} \approx 6.64\cdot 10^{-4}$. Then

$$\|A_S^\dagger a_\ell\|_2 \le \frac{\sqrt{c}\,\delta}{(1-\delta)\sqrt{\log(\epsilon^{-1})}} \quad \text{for all } \ell \notin S$$

with probability at least $1-\epsilon$.

Proof: The proof relies on [46, Theorem 12]. The formulation below follows from [46] by setting $s = \log(\epsilon^{-1})/\log(k/2)$ and estimating $\log(k/2+1)/\log(k/2) \le \log(3)/\log(2)$ for $k \ge 4$.

Theorem 5.5: Assume $A \in \mathbb{C}^{n\times N}$ has unit norm columns and coherence μ. Let $S \subset \{1,\dots,N\}$ be a set of cardinality $k \ge 4$ chosen uniformly at random. The condition

$$\sqrt{144\log(3)\log(2)^{-1}\mu^2 k\log(\epsilon^{-1})} + \frac{k}{N}\|A\|_2^2 \le e^{-1/4}\delta \tag{36}$$

implies

$$\mathbb{P}(\|A_S^* A_S - I\| \ge \delta) \le \epsilon.$$

Using (34) and the value of c, the square-root in (36) becomes $\delta/(2e^{1/4})$. Combining this with (35) shows that (36) is satisfied. Therefore, $\|A_S^* A_S - I\|_2 \le \delta$ with probability at least $1-\epsilon$, which implies that

$$\|(A_S^* A_S)^{-1}\|_2 \le \frac{1}{1-\delta}.$$

Finally,

$$\|A_S^\dagger a_\ell\|_2 \le \|(A_S^* A_S)\|_2 \|A_S^* a_\ell\|_2 \le \frac{1}{1-\delta}\sqrt{k}\mu \le \frac{\sqrt{c}\,\delta}{(1-\delta)\sqrt{\log(\epsilon^{-1})}}$$

by using condition (34) once more. ∎

A. Comparison With Worst-Case Results

Our average-case analysis depends on $\|A_S^\dagger a_\ell\|_2$, while the classical condition (6) of Proposition 2.1 depends on $\|A_S^\dagger a_\ell\|_1$ and is therefore significantly stronger. Proposition 5.2 establishes that the 2-norm condition can be satisfied as long as $\delta_{k+1} < 1/2$. This is clearly weaker than the worst case condition $\delta_{2k} < \sqrt{2}-1 \approx 0.41$ of Proposition 2.3.

Let us now compare worst-case and average-case results based on the coherence μ, by relying on Theorem 5.4. For simplicity, we consider the case in which A is a unit-norm tight frame, for which $\|A\|_2^2 = \frac{N}{n}$. In this case, (35) is equivalent to $k \le \frac{\delta}{4e^{1/4}}n$. If additionally $\mu = c/\sqrt{n}$, then conditions (34) and (35) are both satisfied for fixed ϵ, δ provided

$$k \le C' n.$$

This beats the square-root bottleneck and even removes the log-factor present in estimates for the restricted isometry constants, see (11). Moreover, we have the additional advantage that the coherence is much easier to estimate than the restricted isometry constants.

Combining Theorem 5.4 with the average-case analysis of Theorems 4.4 and 4.5 shows that for a unit norm tight frame A of coherence μ multichannel sparse recovery by (5) can be ensured in the average-case provided $k \le C\mu^{-2}$, which can be as small as $k \le Cn$. Moreover, the failure probability decays exponentially in the number of channels.

In then next section we provide further examples when we discuss particular choices of the matrix A.

VI. Comparison with Multichannel Greedy Algorithms

We now compare our results regarding $\ell_{2,1}$ optimization to those obtained for the greedy algorithms p-thresholding and p-SOMP [25]. These are multichannel versions of simple thresholding and orthogonal matching pursuit. For $1 \le p \le \infty$ they produce a k-sparse signal $\hat{X}$ from measurements $Y = AX$ using a greedy search. To this end, we improve slightly on previous average-case performance results in [25] for these algorithms in the noiseless setting.

A. Greedy Methods

In p-thresholding, we select a set S of k indices whose p-correlation with Y are among the k largest:

$$\|a_\ell^* Y\|_p \ge \|a_j^* Y\|_p, \quad \forall \ell \in S, \forall j \notin S. \tag{37}$$

After the support S is determined, the non-zero coefficients of $\hat{X}$ are computed via an orthogonal projection: $\hat{X}^S = A_S^\dagger Y$.

The p-SOMP algorithm is an iterative procedure. At each iteration, an atom index ℓ_m is selected, and a residual is updated. At the first iteration the residual is simply $Y_0 = Y$. After M iterations, the set of selected atoms being $S_M = \{\ell_m\}_{m=1}^M$, the new residual is computed as $Y_M = Y - A_{S_M} X_M = (I - P_{S_M})Y$ where $X_M = A_{S_M}^\dagger Y$ and $P_{S_M} = A_{S_M} A_{S_M}^\dagger$ is the orthogonal projection onto the linear span of the selected atoms. The next selected atom k_{M+1} is the one which maximizes the p-correlation with the residual Y_M,

$$\|a_{\ell_{M+1}}^* Y_M\|_p = \max_{1\le \ell \le N} \|a_\ell^* Y_M\|_p. \tag{38}$$

Using the real Gaussian probability model (21) average-case recovery theorems for p-thresholding and p-SOMP have been proven in [25], [24, Theorems 4,6,7,8]. We improve slightly on these in the following. (Note, however, that [25] also treats the noisy case.) Our first result generalizes the one in [40] to the multichannel setup.

Theorem 6.1: Let $A \in \mathbb{C}^{n\times N}$ have unit norm columns and local 2-coherence function $\mu_2(S)$ defined in (30). Let $X \in \mathbb{R}^{N\times L}$ with $\operatorname{supp} X \subset S$ where $S \subset \{1,\dots,N\}$, and such that the coefficients on S are given by (21), $X^S = \Sigma\Phi$, where we choose the real spherical model for Φ. Set $Y = AX$ and $R = \max_j \sigma_j / \min_j \sigma_j$, where $\Sigma = \operatorname{diag}(\sigma_j, j \in S)$. If

$$\theta = R\mu_2(S) < 1, \tag{39}$$

then the probability that 2-thresholding applied to Y fails to recover X is bounded by

$$N \exp\left(-L/2(\theta^{-2} - \log(\theta^{-2}) - 1)\right).$$

If we use the complex spherical model instead of the real spherical model then $L/2$ in the above probability estimate may be replaced by L.

The probability bound of Theorem 6.1 is similar to that of Theorem 4.4. However, in contrast to our results for $\ell_{2,1}$-minimization, success of thresholding suffers a dependency on the diagonal matrix Σ. The larger the ratio R, the stronger the condition (39) on the maximal allowed sparsity k, and the larger the probability of error.

Proof: We proceed similarly as in [40]. We denote by Θ the event that 2-thresholding fails. Clearly,

$$\begin{aligned}\mathbb{P}(\Theta) &= \mathbb{P}(\min_{i\in S} \|a_i^* Y\|_2 < \max_{\ell\notin S} \|a_\ell^* Y\|_2) \\ &\le \mathbb{P}(\min_{i\in S} \|a_i^* Y\|_2 < \rho) + \mathbb{P}(\max_{\ell\notin S} \|a_\ell^* Y\|_2 > \rho),\end{aligned}$$

where ρ will be specified later. Denote by Z_j, $j \in S$, a sequence of independent random vectors which are uniformly distributed on the unit sphere of $\mathbb{R}^L$. Then,

$$\mathbb{P}(\min_{i\in S} \|a_i^* Y\|_2 < \rho) = \mathbb{P}\left(\min_{i\in S} \left\|\sum_{j\in S} a_i^* a_j \sigma_j Z_j^*\right\|_2 < \rho\right). \tag{40}$$

Now,

$$\begin{aligned}\left\|\sum_{j\in S} a_i^* a_j \sigma_j Z_j^*\right\|_2 &= \left\|\sigma_i Z_i^* + \sum_{j\in S, j\neq i} a_i^* a_j \sigma_j Z_j^*\right\|_2 \\ &\ge |\sigma_{\min}| - \left\|\sum_{j\in S, j\neq i} \sigma_j \langle a_i, a_j\rangle Z_j^*\right\|_2.\end{aligned}$$

Substituting into (40),

$$\begin{aligned}&\mathbb{P}(\min_{i\in S} \|a_i^* Y\|_2 < \rho) \\ &\quad\le \sum_{i\in S} \mathbb{P}\left(\left\|\sum_{j\in S, j\neq i} \sigma_j \langle a_i, a_j\rangle Z_j^*\right\|_2 \ge \sigma_{\min} - \rho\right).\end{aligned}$$

Choosing $\rho = \sigma_{\min}/2$ and applying Theorem 4.2 we obtain

$$\begin{aligned}&\mathbb{P}(\min_{i\in S} \|a_i^* Y\|_2 < \rho) \\ &\quad\le k \exp(-L/2(\theta^{-2} - \log(\theta^{-2}) - 1))\end{aligned}$$

where we used the definition of θ and $\mu_2(S)$. Similarly we estimate

$$\begin{aligned}&\mathbb{P}(\max_{\ell\notin S} \|a_\ell^* Y\|_2 > \sigma_{\min}/2) \\ &\quad\le (N-k) \exp(-L/2(\theta^{-2} - \log(\theta^{-2}) - 1)).\end{aligned}$$

Combining the two estimates completes the proof for the real case. Choosing the vectors Z_j, $j \in S$, from the complex unit sphere $S_{\mathbb{C}}^L$ and using Corollary 4.3 yields the statement for the complex case. ∎

We now state the corresponding result for 2-SOMP, which slightly improves the one in [25] for the noiseless case. (Note that we restrict to $p = 2$ here, although the theorem is easily extended to general values of p.)

Theorem 6.2: Let A be a matrix with unit norm columns and constants $\delta(S), \mu_2(S) < 1$ where $S \subset \{1,\dots,N\}$. Assume that

$$\frac{\mu_2(S)^2 + (1+\epsilon)(1-\epsilon)^{-1}\mu_2(S)}{1-\delta(S)} \le 1 \tag{41}$$

for some $\epsilon \in (0,1)$. Let X be a random coefficient matrix with support S that is selected according to the real Gaussian probability model, see (21), and let $Y = AX$. Then 2-SOMP applied to Y recovers X in k steps with probability at least

$$1 - N2^k \exp(-\epsilon^2 A_L^2), \tag{42}$$

where $A_L \sim \sqrt{L}$ is given by (25).

If we use the complex Gaussian model instead of the real Gaussian model then the same conclusion holds with A_L replaced by A_{2L} in (42).

Proof: See Appendix C. ∎

Remark 6.3: (a) Due to the factor 2^k the probability bound (42) becomes effective only when the number of channels becomes comparable to the sparsity k. This drawback is very likely due to the analysis and is not observed in practice. However, it seems to be very difficult to remove this factor by a more sophisticated proof technique.

(b) We require $\epsilon < 1$, so that the probability decay of (42) is potentially slower than that given by Theorem 4.4.
(c) With $\delta = \epsilon = 1/2$ condition (41) is satisfied if $\mu_2(\Lambda) \leq 1/7$ while the probability estimate (42) behaves like $1 - N2^k \exp(-L/4)$.
(d) With the estimates $\delta(S) \leq \delta^*(S)$ and $\mu_2(S) \leq \delta^*(S)$, (41) with $\epsilon = 3/11$ is implied by

$$\delta^*(S) < 1/3.$$

(e) By Proposition 5.2 the condition $\delta^*(S) < 1/3$ implies $\|A_S^\dagger a_\ell\|_2 \leq 1/2$ for all $\ell \notin S$, i.e., the bounded norm condition (22) of the average case recovery result for mixed $\ell_{2,1}$. In other words, the condition in (d) for SOMP is slightly stronger than the one for $\ell_{2,1}$.

B. Comparison

We now compare the average-case recovery conditions for mixed $\ell_{2,1}$, thresholding and SOMP for the following choices of the matrix A which we will also use in the numerical experiments:

1) Random spherical ensemble;
2) Union of Dirac and Fourier;
3) Time-Frequency shifts of the Alltop window.

1) Random spherical ensemble: Assume that the random columns of $A \in \mathbb{R}^{n\times N}$ are independent and uniformly distributed on the sphere S^{n-1}. Let S be a support set of size k. Then according to Proposition 5.2 the condition $\|A_S^\dagger a_\ell\|_2 \leq \alpha < 1$ of Theorem 4.4 is implied by $\delta^*(S) \leq \frac{\alpha}{1+\alpha} < 1/2$, while by Proposition 5.3 the latter holds with probability at least $1 - \epsilon$ provided

$$n \geq \max\{C_1(\alpha)k, C_2(\alpha)\log(N/\epsilon)\}. \tag{43}$$

Assuming, for example, $\alpha = 1/4$, under the probability model (21), the probability that reconstruction by $\ell_{2,1}$ fails is bounded from above by $N\exp(-L/2(15-\log(16))) + \epsilon = N\exp(-cL) + \epsilon$ with $c \approx 6.1137$.

We now compare this result with the condition of Theorem 6.1 concerning thresholding. As noted in (32), $\mu_2(S) \leq \delta^*(S)$. Therefore, by Proposition 5.3 we have

$$\theta = 2R\mu_2(S) \leq 2R\delta^*(S) < 1$$

with probability at least $1 - \epsilon$ provided

$$n \geq C\frac{R^2}{\theta^2}\max\{k\log(R/\theta), \log(N/\epsilon)\} \tag{44}$$

and the failure probability of thresholding is bounded by $N\exp(-L/2(\theta^{-2} - \log(\theta^{-2}) - 1)) + \epsilon$.

Let us finally consider Theorem 6.2 for SOMP. By Proposition 5.3 the condition $\delta^*(S) < 1/3$ in Remark 6.3 is satisfied with probability at least $1 - \epsilon$ provided

$$n \geq \max\{C_1 k, C_2\log(N/\epsilon)\} \tag{45}$$

and the failure probability of SOMP is bounded by

$$N2^k\exp(-9/121\, A_L^2) + \epsilon \tag{46}$$

with $A_L^2 \sim L$ if the real Gaussian probability model is used.

Conditions (43), (44), (45) for $\ell_{2,1}$, thresholding and SOMP are rather similar. However, condition (44) for thresholding involves the ratio R. If R is large then thresholding behaves much worse compared to $\ell_{2,1}$ and SOMP. The probability estimate (46) is the worst compared to the other two algorithms due to the factor 2^k. Therefore, $\ell_{2,1}$ gives the best known theoretical average case result.

2) Union of Dirac and Fourier: Consider the $n \times 2n$ matrix $A = (I|F)$, where I is the $n \times n$ identity matrix and F is the normalized $n \times n$ Fourier matrix. The coherence of A is easily seen to be $\mu = 1/\sqrt{n}$. By Proposition 5.1 condition (22), $\|A_S^\dagger a_\ell\|_2 \leq \alpha$ with $\alpha = 1/2$ is satisfied for all support sets S of cardinality at most k provided

$$\sqrt{\frac{k}{n}} + \frac{k-1}{2\sqrt{n}} < \frac{1}{2}.$$

If S is chosen at random then a much better bound (up to constants) is obtained using Theorem 5.4. In our special case, however, further improvement is possible. A reformulation of a result of [5], see also [46, Proposition 3] shows the following. If the support S consists of k_1 arbitrary elements of $\{1,\dots,n\}$ and k_2 random elements of $\{n+1,\dots,2n\}$ then with probability at least $1-\epsilon$ we have $\delta(S) \leq 1/2$ provided

$$k = k_1 + k_2 \leq \frac{cn}{\sqrt{\log\left(\frac{\epsilon}{Cn}\right) + \log(n)}}, \tag{47}$$

with $c = 0.25$. In particular $k \leq n/4$ and the same reasoning as in the proof of Theorem 5.4 yields

$$\|A_S^\dagger a_\ell\|_2 \leq \alpha = 1/2.$$

Using one of the complex probability models in Theorem 4.4, the failure probability of $\ell_{2,1}$-minimization is bounded by $N\exp(-L(4-\log(4)-1)) = N\exp(-cL)$ with $c \approx 1.61$.

To compute the performance of thresholding, note that condition (39), $2R\mu_2(S) \leq 2R\mu\sqrt{k} \leq \theta < 1$, is satisfied provided

$$n \geq \frac{4R^2}{\theta^2}k. \tag{48}$$

Assuming that the non-zero rows of the matrix Φ in the probability model (21) on the coefficients are independent and uniformly distributed on the complex unit sphere $S_{\mathbb{C}}^{L-1}$, the failure probability of thresholding is bounded by $N\exp(-L(\theta^{-2} - \log(\theta^{-2}) - 1))$.

Assuming $\delta(S) \leq \delta = 1/2$ and $\mu\sqrt{k} \leq 1/7$, i.e.,

$$n \geq 49\, k, \tag{49}$$

the condition of Remark 6.3(c) is satisfied since by (32), $\mu_2(S) \leq \mu\sqrt{k} \leq 1/7$. Then by Theorem 6.2 SOMP fails with probability at most $N2^k\exp(-A_{2L}^2/4)$ assuming the complex Gaussian probability model. Assuming as in the discussion of $\ell_{2,1}$ that the support set is such that k_1 arbitrary elements of $\{1,\dots,n\}$ and k_2 random elements of $\{n+1,\dots,2n\}$ are chosen with $k = k_1 + k_2$ then the assumed condition $\delta(S) \leq 1/2$ is true with probability at least $1-\epsilon$ provided (47) holds.

Similar conclusions on the comparison of the three algorithms as in the previous example apply. We note, however,

that in contrast to $\ell_{2,1}$ and SOMP, the performance bound for thresholding does not require a probability model on the support set S.

3) Time-Frequency shifts of Alltop window: Let $n \geq 5$ be a prime. Denote by $(T_r g)_\ell = g_{\ell - r \mod n}$ and $(M_s g)_\ell = e^{2\pi i s \ell / n} g_\ell$ the cyclic shift and modulation operator, respectively. Then $T_r M_s$, $r, s = 0, \ldots, n-1$ forms the set of time-frequency shifts. Let $g_\ell = \frac{1}{\sqrt{n}} e^{2\pi i \ell^3 / n}$ be the so-called Alltop window. Then define A to be the $n \times n^2$ matrix with columns being the time-frequency shifts $T_r M_s g$, $r, s = 0, \ldots, n-1$. The coherence of A is $\mu = 1/\sqrt{n}$ [41].

As in the Fourier-Dirac case, under condition (48) and the complex spherical probability model, thresholding fails with probability at most $N \exp(-L(\theta^{-2} - \log(\theta^{-2}) - 1))$ by Theorem 6.1.

For the analysis of $\ell_{2,1}$ and SOMP we assume that the support S is chosen uniformly at random. As A is the union of n orthonormal bases we have $\|A\|_2^2 = n$. Then choosing $\delta = 3/4$ in Theorem 5.4 yields that under the condition

$$n \geq Ck \log(\epsilon^{-1})$$

with a constant C (which also implies (35)) we have

$$\|A_S^\dagger a_\ell\|_2 \leq 3\sqrt{c} \log^{-1/2}(\epsilon^{-1}) \leq \alpha \quad \text{for all } \ell \notin S$$

with probability at least $1 - \epsilon$ where $\alpha = 3\sqrt{c} \approx 0.0773$. By Theorem 4.4, using one of the complex probability models, the failure probability of $\ell_{2,1}$ is then bounded by $N \exp(-c_2 L) + \epsilon$ with $c_2 = \alpha^{-2} - \log(\alpha^{-2}) - 1$.

For the analysis of SOMP we choose $\delta = 1/2$ in Theorem 5.5. Assuming that the square-root in (36) is less than $\frac{9}{10} e^{-1/4} \frac{1}{2}$ is equivalent to

$$n \geq Ck \log(\epsilon^{-1}) \tag{50}$$

with an appropriate C, and condition (36) is satisfied. Then with probability at least $1 - \epsilon$ we have $\delta^*(S) \leq 1/2$. Furthermore, as suggested by Remark 6.3(b) the condition $\mu_2(S) \leq 1/12$ is also implied by (50) since $\mu_2(S) \leq \sqrt{k}\mu = \sqrt{\frac{k}{n}}$. Assuming the complex Gaussian probability model on the non-zero coefficients of X the failure probability of SOMP is bounded by $N 2^k \exp(-A_{2L}^2/2) + \epsilon$ due to Theorem 6.2.

VII. Numerical Simulations

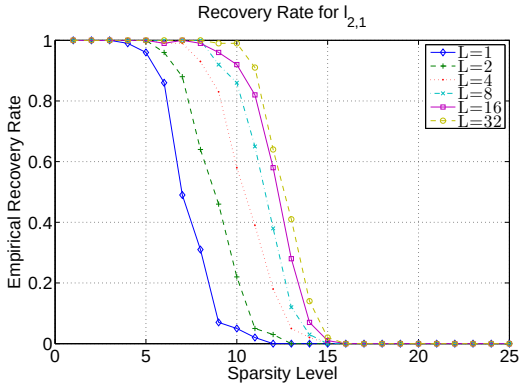
We tested the three algorithms $\ell_{2,1}$ minimization, thresholding and SOMP using the three different types of matrices indicated in the previous section. The support set S of the sparse coefficient matrices X was always selected uniformly at random while the non-zero coefficients were selected at random using one of the following choices of the probability model (21), $X^S = \Sigma\Phi$:

1) Φ is chosen at random according to the real Gaussian model; Σ has independent diagonal entries with standard normal distribution.
2) Φ is chosen at random according to the complex Gaussian model; Σ equals the identity.
3) Φ is chosen at random according to the complex spherical model; Σ equals the identity.

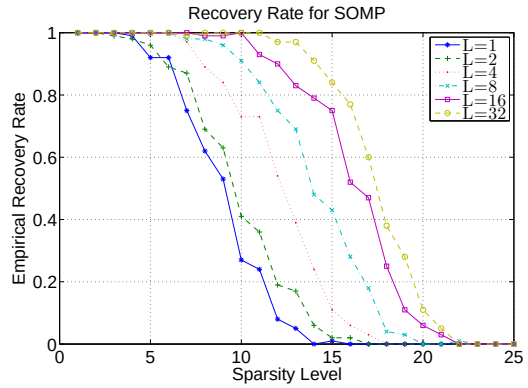
Note that $\Sigma = I$ is favorable for thresholding, while the choice of Σ should have no influence on the performance of $\ell_{2,1}$ and only a mild influence on SOMP.

In the following figures the results of various simulation runs are plotted (we always used 100 simulations for each choice of parameters).

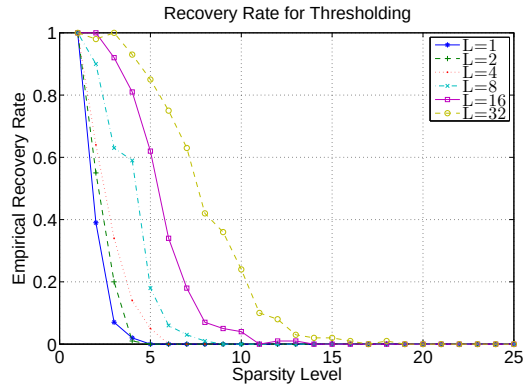
In Fig. 1 we plot the results when choosing A from a random spherical ensemble of size $n = 32$ columns and $N = 256$ rows for L between 1 and 16. The matrix X was generated according to model (1). The improvement with increasing L is clearly evident.



(a)



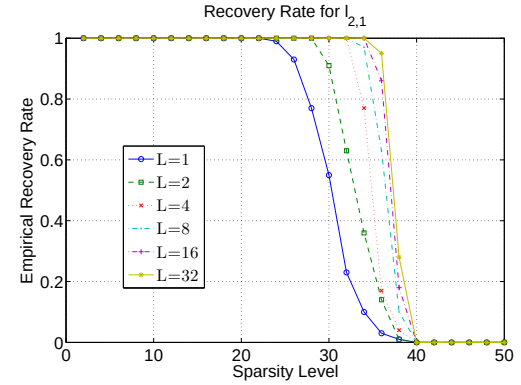
(b)



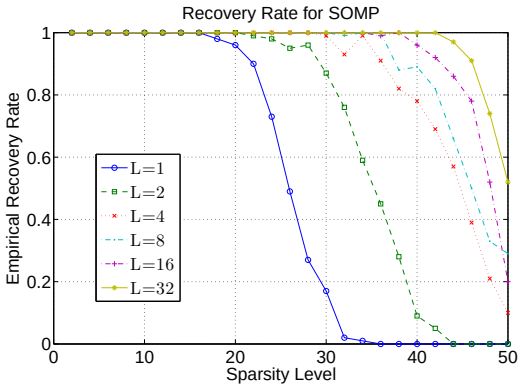
(c)

Fig. 1. Multichannel recovery with X generated according to model (1) and A chosen from a random spherical ensemble, (a) $\ell_{2,1}$, (b) SOMP, (c) Thresholding.

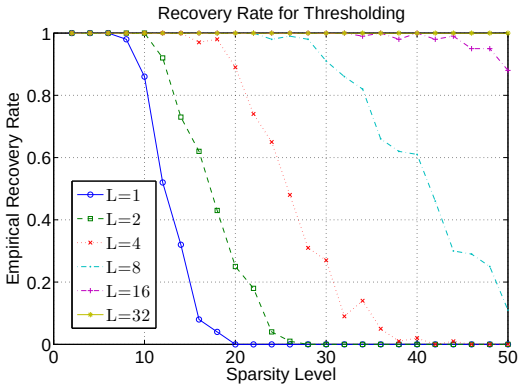
In Fig. 2 we consider all three methods when A is a union of Dirac and Fourier bases, each with 32 elements. Therefore, $n = 64$ and $N = 128$. The matrix X was generated according to model (3). Our simulations show that depending on the number of channels L the three algorithms behave differently. For small values of $L \leq 4$ the mixed norm program $\ell_{2,1}$ shows the best performance. For intermediate number of channels SOMP shows the best recovery results, while, quite surprisingly, for large values of $L \geq 16$ actually thresholding exhibits the best recovery performance.



(a)



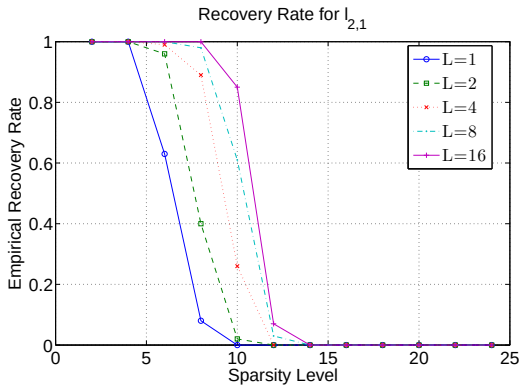
(b)



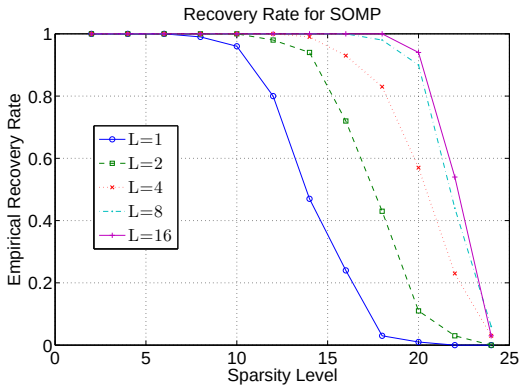
(c)

Fig. 2. Multichannel recovery with X generated according to model (3) and A a union of the Dirac and Fourier bases, (a) $\ell_{2,1}$, (b) SOMP, (c) Thresholding.

Finally, in Fig. 3 we plot the results when using time-frequency shifts of the Alltop window with $n = 29$ and $N = 29^2 = 841$. Here the results of thresholding are extremely poor and therefore not plotted.



(a)



(b)

Fig. 3. Multichannel recovery with X generated according to model (2) and A chosen as time-frequency shifts of the Alltop window, (a) $\ell_{2,1}$, (b) SOMP.

In all examples the three recovery methods show clear performance advantage with increasing L.

VIII. Conclusion

In this paper we analyzed the average-case performance of $\ell_{2,1}$ recovery of multichannel signals. Our main result is that under mild conditions on the sparsity and measurement matrix, the probability of failure decays exponentially with the number of channels. To develop this result we assumed a probability model on the non-zero coefficients of a jointly sparse signal. The results we obtained appear to be the best-known theoretical results on multichannel recovery. Using the tools we developed for analyzing the $\ell_{2,1}$ approach, we also improved slightly on previous performance bounds for thresholding and SOMP.

Appendix A
Proof of Theorem 4.2

The proof uses the following extension of Khintchine's inequality to higher dimensions stated in [27],

$$\mathbb{E}\left\|\sum_{j=1}^{k} a_j Z_j\right\|_2^p \leq \left(\frac{2}{L}\right)^{p/2} \frac{\Gamma\left(\frac{L+p}{2}\right)}{\Gamma\left(\frac{L}{2}\right)} \|a\|_2^p$$

for all $p \geq 2$ and all vectors $a \in \mathbb{R}^k$. By splitting in real and imaginary parts it easily follows that this inequality also holds for all $a \in \mathbb{C}^k$. We may assume without loss of generality that $\|a\|_2 = 1$. Then an application of Markov's inequality yields

$$\begin{aligned}
&\mathbb{P}\left(\left\|\sum_{j=1}^k a_j Z_j\right\|_2 \geq u\right) \\
&= \mathbb{P}\left(\exp\left(\lambda L/2 \left\|\sum_j a_j Z_j\right\|_2^2\right) \geq \exp(\lambda L u^2/2)\right) \\
&\leq \exp(-\lambda L u^2/2)\mathbb{E}\left[\exp\left(\lambda L/2 \left\|\sum_j a_j Z_j\right\|_2^2\right)\right] \\
&= \exp(-\lambda L u^2/2)\sum_{i=0}^\infty (\lambda L/2)^i \mathbb{E}\left\|\sum_{j=1}^k a_j Z_j\right\|_2^{2i} \\
&\leq \exp(-\lambda L u^2/2)\sum_{i=0}^\infty \lambda^i \frac{\Gamma(L/2+i)}{i!\Gamma(L/2)} \\
&= \exp(-\lambda L u^2/2)\sum_{i=0}^\infty \frac{(L/2)_i}{i!}\lambda^i \\
&= \exp(-\lambda L u^2/2)\frac{1}{(1-\lambda)^{L/2}}, \qquad (51)
\end{aligned}$$

where $(a)_i = a(a+1)(a+2)\cdots(a+i-1)$ denotes the Pochhammer symbol. The last equation is due to the fact that $\sum_{i=0}^\infty \frac{(a)_i}{i!}\lambda^i$ is the Taylor series of $(1-\lambda)^{-a}$, which converges for $\lambda < 1$. Minimizing (51) with respect to λ gives $\lambda = 1 - u^{-2}$. Inserting this value yields the statement of the theorem.

Appendix B
Proof of Proposition 5.3

Consider first the case of Gaussian or Bernoulli matrices. According to Theorem 2.1 in [38] (see also Lemma 5.1 in [1]), we have $\|A_S^* A_S - I\|_2 \geq \delta$ with probability at most $2(1 + 12/\delta)^k \exp(-c_0/9n\delta^2)$ with $c_0 = 7/18$. A similar estimate holds for $\|A_{S\cup\ell}^* A_{S\cup\ell} - I\|_2$ with $\ell \notin S$. A union bound over all $\ell \notin S$ yields $\delta^*(S) \geq \delta$ with probability at most $2N(1 + 12/\delta)^k \exp(-c_0/9n\delta^2)$. This term is less than ϵ if (33) holds.

Now consider a random matrix $\Psi \in \mathbb{R}^{n\times N}$ with independent columns that are uniformly distributed on the sphere S^{n-1}. Then Ψ has the same distribution as DA, where A is Gaussian matrix as above, $D = \operatorname{diag}(s_1^{-1}, \ldots, s_N^{-1})$ and $s_j = \sqrt{n}\|\Phi_j\|_2$ where $\Phi_j \in \mathbb{R}^n$ is a vector of independent standard normally-distributed random variables. We now use the following measure concentration inequality [3, Corollary (2.3)] or [4, eq. (2.6)] for a standard Gaussian vector $Z \in \mathbb{R}^n$,

$$\begin{aligned}
\mathbb{P}(\|Z\|_2^2 \geq \frac{n}{1-\gamma}) &\leq \exp(-\gamma^2 n/4), \\
\mathbb{P}(\|Z\|_2^2 \leq (1-\gamma)n) &\leq \exp(-\gamma^2 n/4).
\end{aligned}$$

By a union bound this implies that

$$\begin{aligned}
&\mathbb{P}\left(1-\gamma \leq \min_{j=1,\ldots,N} s_j^2 \leq \max_{j=1,\ldots,N} s_j^2 \leq \frac{1}{1-\gamma}\right) \\
&\quad \geq 1 - 2N\exp(-\gamma^2 n/4). \qquad (52)
\end{aligned}$$

By the above reasoning, we have $(1-\delta/3)\|x\|_2^2 \leq \|Ax\|_2 \leq (1+\delta/3)\|x\|_2^2$ for all x with $\operatorname{supp} x \subset S \cup \{\ell\}$ for some $\ell \notin S$ with probability at least $1-\epsilon$ provided (33) holds with a suitable constant. If additionally $1-\gamma \leq \min_{j=1,\ldots,N} s_j^2 \leq \max_{j=1,\ldots,N} s_j^2 \leq \frac{1}{1-\gamma}$ for $\gamma = \delta/4$ then $(1-\delta)\|x\|_2^2 \leq \|DAx\|_2^2 = \|\Psi x\|_2^2 \leq (1+\delta)\|x\|_2^2$ for all x with $\operatorname{supp} x \subset S \cup \{\ell\}$ for some $\ell \notin S$. By a union bound and (52) this holds with probability at least $1-2\epsilon$ provided (33) holds and $2N\exp(-\delta^2 n/64) \leq \epsilon$, the latter being equivalent to $n \geq 64\delta^2 \log(2N/\epsilon)$. Adjusting the constant in (33) completes the proof.

Appendix C
Proof of Theorem 6.2

We assume that until a certain step SOMP has selected only correct indices, collected in $J \subset S$. Let us first estimate the probability that it selects a correct element of $S \setminus J$ also in the next step.

We denote by $P_J = A_J A_J^\dagger$ the orthogonal projection onto the span of the columns of A in J, and $Q_J = I - P_J$. The residual at the current iteration is given by $Y_M = Q_J Y = Q_J A_S X = Q_J A_S \Sigma\Phi$. SOMP selects a correct index in $S \setminus J$ in the next step if

$$\max_{\ell\in S\setminus J} \|a_\ell^* Q_J A_S \Sigma\Phi\|_2 > \max_{\ell\notin S} \|a_\ell^* Q_J A_S \Sigma\Phi\|_2. \qquad (53)$$

By Theorem 11 in [25] (which is proven using Theorem 4.6; note that there is a slight error in [25] in the computation of the constant A_L) we have the following concentration of measure inequalities

$$\begin{aligned}
&\mathbb{P}\Big(\max_{\ell\in S\setminus J} \|a_\ell^* Q_J A_S \Sigma\Phi\|_2 < (1+\epsilon)C_2(L) \\
&\qquad \times \max_{\ell\in S\setminus J} \|a_\ell^* Q_J A_S \Sigma\|_2\Big) \leq \exp(-\epsilon^2 A_L^2), \\
&\mathbb{P}\Big(\max_{\ell\notin S} \|a_\ell^* Q_J A_S \Sigma\Phi\|_2 > (1-\epsilon)C_2(L) \\
&\qquad \times \max_{\ell\notin S} \|a_\ell^* Q_J A_S \Sigma\|_2\Big) \leq |S^c|\exp(-\epsilon^2 A_L^2),
\end{aligned}$$

where A_L is the constant in (25) and $C_2(L) = \mathbb{E}\|Z\|_2$ with $Z = (Z_1, \ldots, Z_L)$ being a vector of independent standard normal variables. Now we assume that

$$\begin{aligned}
&(1+\epsilon)C_2(L)\max_{\ell\in S\setminus J} \|a_\ell^* Q_J A_S \Sigma\|_2 \\
&\quad \geq (1-\epsilon)C_2(L)\max_{\ell\notin S} \|a_\ell^* Q_J A_S \Sigma\|_2. \qquad (54)
\end{aligned}$$

Then by the above and a union bound the probability that SOMP fails can be bounded by

$$\begin{aligned}
&\mathbb{P}(\max_{\ell\in S\setminus J} \|a_\ell^* Q_J A_S \Sigma\Phi\|_2 \leq \max_{\ell\notin S} \|a_\ell^* Q_J A_S \Sigma\Phi\|_2) \\
&\quad \leq (|S^c|+1)\exp(-\epsilon^2 A_L^2). \qquad (55)
\end{aligned}$$

Let us consider now the maximum on the right hand side of (54). First note that $P_J a_\ell = a_\ell$ for all $\ell \in J$, in other words $Q_J a_\ell = 0$. Hence, we can estimate

$$\begin{aligned}\max_{\ell\notin S} \|a_\ell^* Q_J A_S \Sigma\|_2^2 &= \max_{\ell\notin S} \|\Sigma_{S\setminus J} A_{S\setminus J}^* Q_J a_\ell\|_2^2 \\ &\le \max_{\ell\notin S} \sum_{j\in S\setminus J} \sigma_j^2 |\langle Q_J a_j, a_\ell\rangle|^2 \\ &\le \max_{i\in S\setminus J} \sigma_i^2 \max_{\ell\notin S} \sum_{j\in S\setminus J} |\langle Q_J a_j, a_\ell\rangle|^2.\end{aligned}$$

Furthermore, for $\ell \notin S$ we have

$$\begin{aligned}&\left(\sum_{j\in S\setminus J} |\langle Q_J a_j, a_\ell\rangle|^2\right)^{1/2} = \|A_{S\setminus J}^* Q_J a_\ell\|_2 \\ &= \|A_{S\setminus J}^*(I-P_J)a_\ell\|_2 \\ &\le \|A_{S\setminus J}^* a_\ell\|_2 + \|A_{S\setminus J}^* A_J (A_J^* A_J)^{-1} A_J^* a_\ell\|_2 \\ &\le \mu_2(S\setminus J) + \|A_{S\setminus J}^* A_J\|_2 \|(A_J^* A_J)^{-1}\|_2 \|A_J^* a_\ell\|_2 \\ &\le \mu_2(S) + \frac{\delta(S)}{1-\delta(S)}\mu_2(S) = \frac{\mu_2(S)}{1-\delta(S)},\end{aligned}$$

where we used the fact that $A_{S\setminus J}^* A_J$ is a submatrix of $A_S^* A_S - I$.

Next we consider the maximum on the left hand side of (54). We can estimate

$$\begin{aligned}\max_{\ell\in S\setminus J} \|a_\ell^* Q_J A_S \Sigma\|_2^2 &= \max_{\ell\in S\setminus J} \sum_{j\in S\setminus J} \sigma_j^2 |\langle Q_J a_\ell, a_j\rangle|^2 \\ &\ge \max_{\ell\in S\setminus J} \sigma_\ell^2 \inf_{j\in S\setminus J} |\langle Q_J a_j, a_j\rangle|^2.\end{aligned}$$

Furthermore, for $j \in S\setminus J$

$$\begin{aligned}|\langle Q_J a_j, a_j\rangle| &= |\langle (I-P_J)a_j, a_j\rangle| \\ &= |1 - a_j^* A_J (A_J^* A_J)^{-1} A_J^* a_j| \\ &\ge 1 - \|A_J^* a_j\|^2 \|(A_J^* A_J)^{-1}\|_2 \\ &\ge 1 - \mu_2(S)^2 (1-\delta(S))^{-1}.\end{aligned}$$

Combining the above estimates, condition (54) is satisfied if

$$(1+\epsilon)\frac{\mu_2(S)}{1-\delta(S)} \ge (1-\epsilon)\left(1 - \frac{\mu_2(S)^2}{1-\delta(S)}\right),$$

which is equivalent to (41).

In order to complete the proof, we note that OMP successfully recovers the correct signal if (54) holds for all $J \subset S$. By a union bound of (55) over all those 2^k subsets this is true with probability at least $1 - N2^k \exp(-\epsilon^2 A_L^2)$ provided condition (41) holds.

The extension to the complex valued case is straightforward.

REFERENCES

[1] R. G. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, 2008.

[2] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham, and R. G. Baraniuk. Distributed compressed sensing. *preprint*, 2005.

[3] A. Barvinok. Measure concentration, 2005. lecture notes.

[4] E. Candés and B. Recht. Exact matrix completion via convex optimization. *preprint*, 2008.

[5] E. Candès and J. Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Found. Comput. Math.*, 6(2):227–254, 2006.

[6] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci. Paris S'er. I Math.*, 346:589–592, 2008.

[7] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, Feb. 2006.

[8] E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[9] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Trans. Inform. Theory*, 51(12):4203–4215, Dec. 2005.

[10] E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inform. Theory*, 52(12):5406–5425, 2006.

[11] J. Chen and X. Huo. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Trans. Signal Processing*, 54(12):4634–4643, Dec. 2006.

[12] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by Basis Pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1999.

[13] S. F. Cotter, B. D. Rao, K. Engan, and K. Kreutz-Delgado. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Processing*, 53(7):2477–2488, July 2005.

[14] G. Davis, S. Mallat, and M. Avellaneda. Adaptive greedy approximations. *Constr. Approx.*, 13(1):57–98, 1997.

[15] D. L. Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.

[16] D. L. Donoho. For most large underdetermined systems of linear equations the minimal l^1 solution is also the sparsest solution. *Commun. Pure Appl. Anal.*, 59(6):797–829, 2006.

[17] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions Info. Theory*, 47(7):2845–2862, 2001.

[18] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004.

[19] Y. C. Eldar. Compressed sensing of analog signals in shift-invariant spaces. to appear in *IEEE Trans. Signal Processing*.

[20] Y. C. Eldar and H. Bölcskei. Block-sparsity: Coherence and efficient recovery. to appear in *ICASSP09*.

[21] Y. C. Eldar and M. Mishali. Robust recovery of signals from a union of subspaces. submitted to *IEEE Trans. Inf. Theory*.

[22] M. Fornasier and H. Rauhut. Recovery algorithms for vector valued data with joint sparsity constraints. *SIAM J. Numer. Anal.*, 46(2):577–613, 2008.

[23] J. J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inform. Theory*, 50(6):1341–1344, 2004.

[24] R. Gribonval, B. Mailhe, H. Rauhut, K. Schnass, and P. Vandergheynst. Average case analysis of multichannel thresholding. In *Proc. IEEE Intl. Conf. Acoust. Speech Signal Process.*, 2007.

[25] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *J. Fourier Anal. Appl.*, 14(5):655–687, 2008.

[26] S. Kim, K. Ksh, M. Lustig, S. Boyd, and D. Gorinevsky. A method for large-scale l1-regularized least squares problems with applications in signal processing and statistics. *IEEE J. Sel. Top. Signal Proces.*, 4(1):606–617, 2007.

[27] H. König and S. Kwapień. Best Khintchine type inequalities for sums of independent, rotationally invariant random vectors. *Positivity*, 5(2):115–152, 2001.

[28] M. Ledoux. *The Concentration of Measure Phenomenon*. AMS, 2001.

[29] M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes.*, volume 23. Springer-Verlag, Berlin, Heidelberg, NewYork, 1991.

[30] M. Mishali and Y. C. Eldar. Reduce and boost: Recovering arbitrary sets of jointly sparse vectors. *IEEE Trans. Signal Process.*, 56(10):4692–4702, Oct. 2008.

[31] M. Mishali and Y. C. Eldar. Blind multiband signal reconstruction: Compressed sensing for analog signals. *IEEE Trans. Signal Process.*, 57:993–1009, Mar. 2009.

[32] M. Mishali and Y. C. Eldar. From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals. *arXiv 0902.4291;* submitted to *IEEE Selected Topics on Signal Process.*, 2009.

[33] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. submitted, 2008.

[34] D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. submitted, 2008.
[35] H. Rauhut. Random sampling of sparse trigonometric polynomials. *Appl. Comput. Harmon. Anal.*, 22(1):16–42, 2007.
[36] H. Rauhut. On the impossibility of uniform sparse reconstruction using greedy methods. *Sampl. Theory Signal Image Process.*, 7(2):197–215, 2008.
[37] H. Rauhut. Stability results for random sampling of sparse trigonometric polynomials. *IEEE Trans. Information Theory*, 54(12):5661–5670, 2008.
[38] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Trans. Inform. Theory*, 54(5):2210 – 2219, 2008.
[39] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045, 2008.
[40] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Letters*, 14(11):828–831, Nov. 2007.
[41] T. Strohmer and R. W. Heath. Grassmannian frames with applications to coding and communication. *Appl. Comput. Harmon. Anal.*, 14(3):257–275, 2003.
[42] S. J. Szarek. Condition numbers of random matrices. *J. Complexity*, 7:131–149, 1991.
[43] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, 50(10):2231–2242, 2004.
[44] J. A. Tropp. Recovery of short, complex linear combinations via l_1 minimization. *IEEE Trans. Inform. Theory*, 51(4):1568–1570, 2005.
[45] J. A. Tropp. Algorithms for simultaneous sparse approximation. Part II: Convex relaxation. *Signal Processing*, 86(3):589 – 602, 2006.
[46] J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, to appear.
[47] J. A. Tropp, A. C. Gilbert, and M. J. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86(3):572 – 588, 2006.
[48] P. Wojtaszczyk. Stability and instance optimality for Gaussian measurements in compressed sensing. *preprint*, 2008.