Sparse Recovery with Fusion Frames

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Ulaş Ayaz

aus

İskenderun, Türkei

Bonn, 2014

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der
Rheinischen Friedrich-Wilhelms Universität Bonn

1. Gutachter: Prof. Dr. Holger Rauhut
2. Gutachter: Prof. Dr. Patrik Ferrari

Tag der Promotion: 26.06.2014

Erscheinungsjahr: 2014

Aileme.

Abstract

Sparse signal structures have become increasingly important in signal processing applications as the technology progresses and a plethora of data needs to be handled. It has been shown in practice that various signals have fewer degrees of freedom compared to their actual sizes, i.e., can be expressed in terms of a small number of elements from some dictionary. Alongside an increase in applications, a recent theory of sparse and compressible signal recovery has been recently developed under the name of Compressed Sensing (CS). This approach states that a sparse signal can be efficiently recovered from a small number of random linear measurements.

Another powerful tool in signal processing is frames which provide redundant representations for signals. Such redundancy is desirable in many applications where resilience to errors and losses in data is important. The increase in data has also significantly increased the demand to model applications requiring distributed processing which goes beyond the classical frames. The recent theory of Fusion Frames, which can be regarded as a generalization of classical frames, satisfies those needs by analyzing signals by projecting them onto multidimensional subspaces. Fusion frames provide a suitable mathematical framework to model large data systems such as sensor networks, transmission of data of communication networks, etc.

In this thesis, we combine these two recent theories and consider the recovery of signals that have a sparse representation in a fusion frame. As in the classical CS, a sparse signal from a fusion frame can be sampled using very few random projections and efficiently recovered using a convex optimization that minimizes the mixed ℓ_1/ℓ_2-norm. This problem has close connections with other similar recovery problems studied in CS literature such as block sparsity and joint sparsity problems. A key contribution in this thesis is to exploit the incoherence of the fusion frame subspaces in order to enhance the existing recovery results by incorporating this structure. In particular, we derive upper and lower bounds for the number of measurements required for the sparse recovery and the error derived by convex optimization. Aside from our results in the fusion frame setup, we also present results in the classical CS where we focus on improving constants appearing in the number of measurements required and prove optimal constants in the nonuniform setting with rather concise and simple proofs.

Acknowledgements

I am very grateful to my advisor Holger Rauhut for his constant support during the whole period of my Ph.D. He has been very generous with his time and ideas during this process and helped me immensely develop and improve as a researcher and writer. His devotion to producing a very high quality research in terms of the content as well as the presentation has inspired me as a young scientist at the beginning of his research journey. I also appreciate the way he helped me to reach the resources and to learn the material on my own and the right mixture of pressure and freedom he provided.

In addition, I would like to thank several fellow students and postdocs for their support and help during my Ph.D. In particular, many thanks to Tino Ullrich, Felix Krahmer, Sjoerd Dirksen, Maryia Kabanava, Željka Stojanac and Max Hügel. During my stay in Bonn and travels within Europe, I am grateful to many other friends who made my time with them fun and memorable.

Throughout my stay at the University of Bonn and at RWTH Aachen University in my last year of Ph.D., I appreciate the all help from the staff there, especially from Karen Bingel and Elke Zimmermann which made things go much smoother for me. I am grateful to RWTH Aachen University for welcoming me during my last year. I also acknowledge funding through the WWTF project SPORTS (MA07-004) and the ERC Starting Grant StG 258926 during my studies.

I would like to thank Prof. Patrik Ferrari, Prof. Herbert Koch and Prof. Marek Karpinski for agreeing to be on my defense committee.

Finally, very special thanks to my family who has supported me unconditionally and has given me the energy to go all the way to the end of my studies. They never failed to make me feel their love despite the distances and the time between us. I am grateful to my father for passing me the love for mathematics and science, encouraging me for pursuing my talents and always showing his confidence in me when I hesitated. My mother and brother have always been there for me and made life full of joy and and laughter around them. I dedicate this thesis to all my family.

Contents

List of Figures

CHAPTER 1

Introduction

Many technological advances deal with the problem of measuring a signal and reconstructing it from measured data. This is one of the main goals of signal and image processing. The conventional way of data acquisition relies on Shannon's celebrated sampling theorem, which dictates that a signal can be perfectly reconstructed from its samples if the sampling rate is at least twice the maximum frequency present in the signal (the so-called Nyquist rate). Following this principle, in most devices used in consumer audio and visual communication, medical imaging, radar etc., the data acquisition typically works as follows: massive amounts of samples are collected through the measurement device, and then a compression stage is invoked in which most of this data is discarded and only the meaningful part is stored and transmitted. For instance, our modern digital cameras which use JPEG standard [**112**] acquire data with millions of pixels and compress it into a picture of only few hundred kilobytes. This is possible because many man-made and natural images are 'sparse' in the sense that most of its components in an appropriate basis are zero or close to zero. However this measurement process, where a lot of data is collected to be only thrown away at the end, seems like a waste of energy and resources. Compressed sensing (CS) initiated by the seminal papers by E. Candès, J. Romberg, T. Tao [**18**, **20**] and by D. Donoho [**38**] merges the two steps of acquiring and compressing data into one step by exploiting the sparseness inherent in many signals. Since the appearance of these papers, a large research activity has been carried out at the interface of mathematics, engineering and computer science with a lot of potential applications such as MRI (magnetic resonance imaging) [**66**, **81**], radar imaging [**67**, **69**], single-pixel cameras [**46**]. The motivation in each of these applications for using the ideas of CS may vary. For instance, as anybody who has sat through an imaging session in an MRI machine knows, the sensing process might be quite slow. To speed up this process, applying undersampling techniques from CS becomes very promising.

In mathematical terms, given a signal $x \in \mathbb{C}^N$, many measurement schemes in signal acquisition can be modelled by the linear system of equations

$$y = Ax$$

where $y \in \mathbb{C}^m$ and $A \in \mathbb{C}^{m \times N}$. Recovering x from its measurements y is not possible if the system is underdetermined, i.e., $m < N$. CS shows that if x is sparse, i.e., has few nonzero entries, then it can be recovered efficiently when m is much smaller than N.

This thesis comprises two parts. In the first part we provide our results in CS for the nonuniform setting, where we deduce the optimal number of measurements m in terms of the constants appearing in the lower bounds. This has importance for the practical use of CS results. In the second part, we focus on fusion frames which are generalizations of classical frames. We extend

the concepts of CS to fusion frames and improve existing results by taking a certain structure in the fusion frames into account.

1.1. Compressed sensing

When the acquisition system is linear, the information of a signal $f(t)$ is obtained by linear functionals recording the values $y_k = \langle f, \theta_k \rangle,\ k = 1, \dots, m$. This is a standard setup. For instance linear functionals θ_k can be Dirac delta functions taking samples of f, can be indicator functions of pixels as in a digital camera, or can be sinusoids which yield Fourier coefficients of f. After an appropriate discretization step, we can consider our signal to be a vector $f \in \mathbb{C}^N$. Then the measurement scheme can be written as

$$y = \Theta f$$

where $\Theta \in \mathbb{C}^{m \times N}$ consists of the rows $\theta_1, \dots, \theta_m \in \mathbb{C}^N$. Most of the results in CS are in this discrete setting which is conceptually simple and can be potentially extended to continuous time signals as well. As mentioned earlier, many signals have a concise representation with respect to a suitable basis. Mathematically speaking, the vector $f \in \mathbb{C}^N$ can be expanded into an orthonormal basis $\Phi = [\phi_1, \dots, \phi_N]$ as

$$f = \sum_{i=1}^{N} x_i \phi_i,$$

where $x \in \mathbb{C}^N$ is a coefficient sequence. For instance, if sparsity is with respect to the discrete Fourier transform (DFT) which is used in MRI, the ϕ_i are sinusoid functions. Then the measurement can be written as

$$y = \Theta \Phi x = Ax \tag{1.1}$$

with $A \in \mathbb{C}^{m \times N}$. Before continuing, for the sake of simplicity, we will assume Φ to be the identity matrix. The basic goal is to recover x from its measurements y. When $m < N$, (1.1) becomes underdetermined, therefore there are infinitely many solutions to this system. In other words, it is impossible to find the original x without additional information. CS shows that if x is sparse, i.e., has few nonzero entries, then it can be recovered efficiently even when m is much smaller than N. The sparsity assumption is related to the empirical observation that many real world signals are approximately sparse. The CS problem consists of two main questions: *a)* how to design the measurement matrix A, *b)* how to reconstruct x from y in an efficient and tractable way. In Section 1.2, we will give an overview of the answers to these questions. We would like to mention an interesting point of CS: often randomness plays an important role in the design of the measurement schemes. All known provably optimal results of CS are obtained by such random matrices, which are also a central theme of our theoretical results.

1.2. Recovery results for compressed sensing

The field of CS emerged after the appearance of the seminal papers by Candès, Romberg and Tao [**18**] and Donoho [**38**]. There have been earlier works on the effectiveness of ℓ_1-minimization, however the seminal works [**18**,**38**] have combined the random masurements with ℓ_1-minimization in order to achieve near-optimal recovery results and have also pointed towards important applications in signal processing. We have seen that the central mathematical question in CS is solving

the underdetermined system (1.1) ($m < N$) under a sparsity assumption for x. A vector $x \in \mathbb{C}^N$ is called s-sparse if $\|x\|_0 := \text{card}\{\ell \in [N], x_\ell \neq 0\} \leq s$, where $[N] = \{1, \dots, N\}$. We note that $\|\cdot\|_0$ is called ℓ_0-norm although it is not a norm. As a first approach one is led to solve the combinatorial problem

$$\min_{z \in \mathbb{C}^N} \|z\|_0 \text{ subject to } Az = y,$$

where $y = Ax$. Unfortunately, this problem is NP-hard in general. One of the initial ideas of CS was to replace ℓ_0-norm by ℓ_1-norm which leads to the convex optimization problem

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \text{ subject to } Az = y, \tag{1.2}$$

where $\|z\|_1 = \sum_i |z_i|$. This problem is also called the *basis pursuit* and can be solved in polynomial time. There are other ℓ_1-type minimizations, e.g. LASSO, Dantzig selector, and greedy algorithms that have been investigated for solving the original CS problem, but in this thesis we will only focus on ℓ_1-minimization. A classical result of CS is that $m \geq Cs \ln(N/s)$ many measurements are sufficient for recovery via ℓ_1-minimization. This result is achieved by random measurement matrices such as Gaussian or Bernoulli (± 1 entries). Up to today, it is an open and very hard problem to come up with deterministic constructions for A which gives the same result. It is also known that $m \geq Cs \ln(N/s)$ measurements are necessary for recovery of an s-sparse vector via any reconstruction method. This implies that CS allows us to reduce the number of measurements almost to the sparsity level s which is potentially far less than the dimension N of the signal. Since we do not know a priori where the s nonzero entries of x are, we pay the factor of $\ln(N/s)$ as a price to find those entries.

It is of interest to find what type of measurement matrices are in general suitable for CS. One observation is that the measurement matrix needs to be incoherent with the sparsifying basis matrix, which we chose to be the identity earlier. The *coherence* of a matrix $A \in \mathbb{C}^{m \times N}$ is defined as

$$\mu := \max_{j \neq k} |\langle a_j, a_k \rangle|,$$

where the a_j are the ℓ_2-normalized columns of A. Analysis of various recovery algorithms including ℓ_1-minimization yields the sufficient condition

$$(2s - 1)\mu < 1 \tag{1.3}$$

for sparse recovery. This suggests to choose A with low coherence to obtain good recovery guarantees. However the coherence obeys the lower bound

$$\mu \geq \sqrt{\frac{N - m}{m(N - 1)}}.$$

For large N, the right hand side scales like $1/\sqrt{m}$. There are also deterministic constructions of matrices achieving this lower bound such as equiangular tight frames. However, this bound together with (1.3) only gives us the sufficient condition

$$m \geq Cs^2$$

for recovery. This quadratic scaling of the number m of samples in terms of sparsity s is only suboptimal. In order to overcome this quadratic bottleneck, a stronger concept to measure the

quality of a measurement matrix was given by Candès and Tao in [**20**,**24**] under the name *restricted isometry property (RIP)*. The RIP constant $\tilde{\delta}_s$ of a matrix $A \in \mathbb{C}^{m\times N}$ is defined as the smallest $\delta \geq 0$ such that

$$(1-\delta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1+\delta)\|x\|_2^2 \tag{1.4}$$

for all s-sparse x. The matrix A is said to satisfy RIP if $\tilde{\delta}_s$ is small for sufficiently large s. In other words, A needs to be almost an isometry on its column submatrices of size s. The matrices which satisfy RIP turn out to be suitable for CS. For instance it has been shown that if $\tilde{\delta}_s < 1/3$ [**15**], then (1.2) recovers all s-sparse vectors from their measurements. The constant $1/3$ is optimal. The next question is: which matrices satisfy the RIP? There has been a great amount of research in finding such matrices with good RIP constants. As mentioned before, most of the theoretical results involve some degree of randomness in the design of the matrix and we will give an overview of these results in Section 1.2.1. Proving RIP results is generally not easy and involves advanced probabilistic techniques that have been successfully implemented and improved over the course of CS. In this thesis, we will also use some of those techniques in proving our results.

In real world applications, one often measures a signal which is not exactly sparse but only approximately sparse. These are vectors with small best s-term approximation error in ℓ_1-norm,

$$\sigma_s(x)_1 := \inf_{\|z\|_0 \leq s} \|x - z\|_1,$$

also called 'compressible' vectors. In addition, noise is often present in the measurement process due to the circuit imperfections which is modeled by

$$y = Ax + e,$$

with $\|e\|_2 \leq \eta$ (in some settings the error is assumed to be stochastic as well). Reconstruction by ℓ_1-minimization and other CS techniques is also robust with respect to such nonidealities. In particular, if the matrix A satisfies the RIP with constant $\tilde{\delta}_{2s} \leq 1/\sqrt{2}$ [**16**], the solution $\hat{x}$ to $\min_z \|z\|_1$ s.t. $y = Az + e$ and $\|e\|_2 \leq \eta$ satisfies

$$\|x - \hat{x}\|_2 \leq C\frac{\sigma_s(x)_1}{\sqrt{s}} + D\eta, \quad \text{and} \quad \|x - \hat{x}\|_1 \leq C\sigma_s(x)_1 + D\sqrt{s}\eta, \tag{1.5}$$

with absolute constants $C, D > 0$. For random measurement matrices, one can speak of two types of recovery results: *uniform* and *nonuniform*. A *uniform* recovery guarantee means that, with high probability on the choice of a random matrix, one can recover all s-sparse vectors. The RIP gives sufficient conditions for this type of recovery conditions. In *nonuniform* recovery, we deal with specific sparse vectors rather than with all vectors with a given sparsity. A nonuniform result states that a fixed sparse vector can be recovered with high probability with a random draw of the matrix. We give a precise formulation of these two type of results in Section 2.1.1. In CS, a significant theory dealing with the nonuniform recovery of sparse signals has been developed. In particular, one requires weaker conditions than RIP in order to guarantee nonuniform recovery. In Sections 2.2.1 and 2.3.1, we present sufficient conditions for nonuniform sparse recovery.

Our contribution: Practical experience suggests that constants appearing in the lower bounds of the necessary number m of measurements are quite pessimistic, thus there is a huge gap between theoretical guarantees and practical performance of CS techniques. Our main contribution is to

bridge this gap by showing that subgaussian matrices provide explicit and good constants for nonuniform recovery. In particular, our nonuniform analysis in Chapter 2 yields roughly that $m \geq 2s\ln(N)$ Gaussian or Bernoulli measurements are enough to guarantee exact and stable recovery of s-sparse vectors with a rather concise and simpler proof than the other similar results in the literature. This is interesting theoretically, since the constant 2 is shown to be optimal (although the optimal term $\ln(N/s)$ is replaced by $\ln(N)$) and our results hold for finite values of m, s, N in contrast to some asymptotic results in the literature, see Section 2.4 for a review.

Next, we will state various important results of CS obtained in the last 10 years starting with the pioneering CS papers and continuing with the subsequent line of theory.

1.2.1. Overview of RIP results

We begin by reviewing the relevant works of Candès and co-workers. The first nonasymptotic results in CS were derived in [**18**, **38**], showing that $m \gtrsim s\ln(N)$ random Fourier measurements are sufficient for exact recovery of a fixed s-sparse vector. This is a nonuniform result. Candès and Tao introduced the concept of the restricted isometry property (RIP) in [**20**, **24**] for the first time. They showed that the RIP implies uniform sparse recovery and proved exact and stable recovery results for Gaussian, partial Fourier and Bernoulli random matrices. In particular, they gave the sufficient condition $\tilde{\delta}_{2s} + \tilde{\delta}_{3s} < 1$ for exact recovery via ℓ_1-minimization (basis pursuit). Cai and Zhang [**16**] recently improved this condition to $\tilde{\delta}_{2s} < 1/\sqrt{2}$ and this was shown to be optimal by Davies and Gribonval [**33**]. Another optimal result concerning RIP for exact recovery is $\tilde{\delta}_s < 1/3$ shown by Cai and Zhang [**15**]. A uniform recovery result for Gaussian matrices was first given by Candès and Tao [**20**] which requires $m \gtrsim s\ln(N/s)$. This was also shown to be optimal up to constants by using the works of Kashin [**71**] and Garnaev and Gluskin [**56**, **59**]. Before CS became a field, they gave sharp upper and lower bounds for the *Gelfand widths* of the ℓ_1-ball B_1^N in $\mathbb{R}^N$ and provided substantial material in order to derive the RIP results for Gaussian and Bernoulli matrices. A similar optimal result $m \gtrsim s\ln(N/s)$ for subgaussian matrices was given with a rather short proof by Baraniuk et. al. [**3**] and independently by Mendelson et. al. [**84**]. The stability and robustness of sparse reconstruction via basis pursuit as stated in (1.5) were also established in earlier CS papers by Candès, Romberg and Tao [**19**] where they derived the condition $\tilde{\delta}_{3s} + 3\tilde{\delta}_{4s} < 2$. This was later improved several times. In particular, Candès gave the condition $\tilde{\delta}_{2s} < \sqrt{2} - 1$ in [**17**] which we follow similarly for proving stable recovery in the fusion frame setup in Section 3.2.7. We also note that the condition $\tilde{\delta}_{2s} < 1/\sqrt{2}$ derived in [**16**] is optimal for stable recovery.

While subgaussian random matrices provide optimal measurement matrices for CS, using a fully random matrix may not be very desirable in some applications for a variety of reasons. First of all, when the matrix is too large, it may be an issue to store the matrix. Secondly, in some applications there are a priori contraints on the structure of the measurement matrix which do not allow a liberal design of the matrix. In addition, some structure in the matrix allows for fast matrix multiplications such as fast Fourier transform (FFT). Since, up to date, it has not been possible to design fully deterministic matrices with optimal guarantees for CS, there has been a significant body of research focused on other types of measurements ensembles, such as structured random matrices, showing that they satisfy the RIP with high probability. It turns out that the additional structure makes it harder to prove the RIP when compared to entirely random matrices such as

Gaussian and Bernoulli. The RIP for partial random Fourier matrices was first analyzed in [**20**], where the authors obtained the bound $m \gtrsim s \ln^5(N)$ on the number of measurements. This was later improved by Rudelson and Verhsynin [**95**] by one log-factor by using involved probabilistic analysis including Dudley's inequality. A similar version of their method is also used in proving our results given in Chapter 3. The proof techniques of [**20**, **95**] apply also to general discrete orthogonal systems. Krahmer, Mendelson and Rauhut [**75**] considered partial random circulant matrices and obtained the sufficient condition $m \gtrsim s \ln^2(s) \ln^2(N)$ for uniform recovery via RIP. It seems that extra log-factors appearing in the conditions are the price of having extra structure in the measurement process when compared to fully random matrices.

While there has been a lot of work in ℓ_1-minimization based algorithms, e.g., basis pursuit, LASSO [**102**] and the Dantzig selector [**21**], other types of reconstruction methods such as *greedy algorithms* have also been considered for recovering sparse vectors from undersampled measurements. A simple greedy algorithm is the orthogonal matching pursuit (OMP) which has initially appeared in [**34**, **82**]. This algorithm builds up the support set of the sparse vector by adding one index to the target support set at each step. Gilbert and Tropp [**57**] demonstrated that a fixed s-sparse vector can be recovered from $\mathcal{O}(s \ln(N))$ Gaussian measurements via OMP in s iterations. However, it is not possible to achieve uniform recovery of s-sparse signals via s OMP iterations under RIP conditions [**41**, **92**]. This issue was solved by Needell and Vershynin [**85**, **86**] with the regularized OMP algorithm. This method is stable to noise under RIP conditions, although an additional log-factor appeared in the condition on the restricted isometry constant. This log-factor was later removed by the compressive sampling matching pursuit (CoSaMP) introduced by Needell and Tropp [**104**]. In particular they gave the condition $\tilde{\delta}_{4s} < 0.17$ for stable and robust recovery, see also [**54**] where a slightly better condition appears.

1.2.2. Null space conditions

The null space property (NSP) is another important ingredient in CS theory. It gives necessary and sufficient conditions for exact recovery of s-sparse vetors via ℓ_1-minimization.

Definition. *A matrix $A \in \mathbb{C}^{m \times N}$ is said to satisfy the null space property of order s if for all subsets $S \subset [N]$ with $|S| = s$ it holds that*

$$\|v_S\|_1 < \|v_{\overline{S}}\|_1 \quad \text{for all} \quad v \in \ker A \setminus \{0\}.$$

NSP basically requires that every vector in the null space of A is far from being sparse. It first appeared in the CS literature implicitly in works of Donoho and Elad [**42**], of Donoho and Huo [**43**] and of Elad and Bruckstein [**48**], and explicitly in the work of Gribonval and Nielsen [**64**]. Later on, stable and robust versions of NSP were also given in the context of recovery via basis pursuit (see [**54**, Chapter 4] for a review). The next result indicates the link between the null space property and exact recovery of sparse vectors via basis pursuit.

Theorem 1.1. *Given a matrix $A \in \mathbb{C}^{m \times N}$, every s-sparse vector $x \in \mathbb{C}^N$ is the unique solution of* (1.2) *with $y = Ax$ if and only if A satisfies the null space property of order s.*

Even though NSP provides a necessary and sufficient condition for sparse recovery via ℓ_1-minimization, it is hard to verify it by a direct computation. This is why many researchers have instead worked with sufficient RIP conditions in order to prove their results. However there has

been some work towards showing the null space property directly for Gaussian matrices. Foucart and Rauhut [**54**, Theorem 9.29] used the escape through the mesh theorem due to Gordon [**62**] in order to show NSP, see also [**70**]. This yields that $m \gtrsim s\ln(N/s)$ Gaussian measurements are sufficient for stable recovery of s-sparse vectors. Similar ideas were also used in [**95**, **98**]. In Section 3.2.8 we follow these ideas in order to prove NSP for fusion frames.

1.3. Compressed sensing meets fusion frames

Before introducing fusion frames, we would like to start with explaining 'classical frames'. A frame for a vector space equipped with an inner product can be seen as a generalization of a basis to a linearly dependent set of vectors. One can obtain a frame by adding some more elements to a basis. Formally speaking, a set of vectors $\{e_k\}_{k=1}^M \subset \mathbb{R}^N$ is called a *frame* if

$$A\|x\|_2^2 \leq \sum_{j=1}^{M} |\langle e_k, x\rangle|^2 \leq B\|x\|_2^2$$

for all $x \in \mathbb{R}^N$. Here $A, B > 0$ are called frame constants and $\|\cdot\|_2$ denotes the Euclidean norm. Frames were first introduced by Dustin and Schaeffer [**47**] in the context of nonharmonic Fourier series. They typically provide non-unique representations of each vector in the space [**29**]. Due to this redundancy, classical frames are nowadays a standard notion for modeling applications, in which redundancy plays a vital role such as filter bank theory, sigma-delta quantization, signal and image processing and wireless communication. During the transfer of data (i.e., frame expansion coefficients of a signal), the overcompleteness of frames offers resilience to errors such as data loss, stability to noise, and freedom to capture important signal characteristics.

Fusion frames are a relatively new concept which is potentially useful when a single frame system is not sufficient to represent the whole sensing mechanism efficiently. Fusion frames were first introduced in [**26**] under the name of 'frames of subspaces' (see also the survey [**25**]). They analyze signals by projecting them onto multidimensional subspaces, in contrast to frames which consider only one-dimensional projections. We define a *fusion frame* for $\mathbb{R}^d$ as a collection of N subspaces $W_j \subset \mathbb{R}^d$ and associated weights v_j that satisfy

$$A\|x\|_2^2 \leq \sum_{j=1}^{N} v_j^2 \|P_j x\|_2^2 \leq B\|x\|_2^2 \tag{1.6}$$

for all $x \in \mathbb{R}^d$ and for some universal fusion frame bounds $0 < A \leq B < \infty$, where $P_j \in \mathbb{R}^{d\times d}$ denotes the orthogonal projection onto the subspace W_j. If $A = B$, then $(W_j)_j$ is called a *tight* fusion frame. The initial idea to construct a fusion frame [**26**] was to obtain a frame for the whole space $\mathbb{R}^N$ by first building 'smaller' frames, i.e., for subspaces W_j, and then fuse them together. This idea allows managing the data by first projecting them to subspaces and then process them within each subspace and finally collect the locally computed objects. There have been a number of applications where fusion frames have proven to be useful practically, such as distributed processing [**27**, **76**], parallel processing [**5**, **88**], packet encoding [**6**]. In Section 3.1.2, we give more details about these applications.

Sparse recovery in fusion frames. Often signals have additional a priori structure other than just pure sparsity. For instance the support set to be recovered can exhibit a certain structure,

i.e., only specific support sets are allowed. Suppose that we have block vectors consisting of N blocks $x_1, \dots, x_N$, where each $x_j \in \mathbb{R}^d$. The sparsity concept for such vectors is 'block' sparsity which is the 'ℓ_0- block norm' defined as

$$\|\mathbf{x}\|_0 = \text{card}\{j \in [N] : x_j \neq 0\}.$$

Here we count the number of non-zero blocks instead of the number of non-zero entries. The problem of recovering a sparse block vector from linear measurements can be considered as an extension of the classical CS problem, where the vectors of interest now consist of blocks. The block sparsity model [**51**] along with other types of similar problems (such as joint [**52**] and group sparsity [**91**], see Section 3.1.4) has been already investigated in the literature. A usual proxy for the ℓ_0-minimization algorithm for recovering the block signal is the ℓ_1/ℓ_2-minimization [**51**].

Boufounos, Kutyniok and Rauhut in their paper [**9**] further extend ideas from CS to the fusion frame setup. In this paper, they consider N fusion frame subspaces $W_1, \dots, W_N$ in $\mathbb{R}^d$ and collect block vectors $\mathbf{x} \equiv (x_j)_{j=1}^N$ with $x_j \in W_j$ in the space

$$\mathcal{H} = \{(x_j)_{j=1}^N : x_j \in W_j,\ \forall j \in [N]\}.$$

Then the authors consider the recovery of s-sparse vectors from the set $\mathcal{H}$ from $m < N$ linear measurements. Details of the problem setup is given in Section 3.1.3. The novelty of [**9**] is the additional assumption that each block x_j belongs to one of the fusion frame subspaces W_j. Typically those subspaces are lower dimensional than the ambient space $\mathbb{R}^d$, say $\dim(W_j) = k < d$, and they satisfy an incoherence property which is quantified by the parameter

$$\lambda = \max_{i \neq j} \|P_i P_j\|_{2\to 2}, \quad i, j \in [N]$$

which gives the cosine of the least principle angle between the subspaces. In [**9**] it was shown that $m \gtrsim s \ln(N/s)$ measurements are sufficient for many random measurement ensembles as in the classical CS. This result is also parallel to other results in block sparsity problems in the literature, see Section 3.6.

Our contribution: In Chapter 3 we show that one can improve the condition $m \gtrsim s \ln(N/s)$ by using the incoherence property of the subspaces. This is mainly done by invoking the ℓ_1/ℓ_2-minimization program where the a priori information of the subspaces, in which the original vector lies, is encoded. In particular we show that as λ decreases, i.e., the incoherence of the subspaces increases, fewer measurements are needed for sparse recovery. Our results are three-fold. First we present a nonuniform result stating that

$$m \gtrsim (1 + \lambda s) \ln^\alpha(N)$$

measurements suffice for nonuniform recovery with Gaussian or Bernoulli matrices. Secondly, we show a similar result for uniform recovery in two separate ways giving the conditions

$$m \gtrsim (\lambda s + \sqrt{s}) \ln(Nd) \ln^2(sk) \left(\ln(N) + k \ln\left(s\sqrt{k}\right)\right), \qquad \text{and} \tag{1.7}$$

$$m \gtrsim d(1 + \lambda s) \ln(N), \tag{1.8}$$

respectively. Condition (1.7) is valid for subgaussian matrices and exhibits a slightly worse order in s and more log-factors. Condition (1.8) is valid only for Gaussian matrices and has the ambient dimension d as a linear factor, which we think is suboptimal. All nonuniform and uniform results

are also shown to be robust with respect to noise and stable under passing to approximately sparse signals. Finally we provide a necessary condition for sparse recovery via ℓ_1/ℓ_2-minimization. We show that

$$m \gtrsim \frac{s}{d} \ln\left(\frac{N}{s}\right) + \frac{ks}{d}$$

measurements are required for any type of linear measurement scheme – not necessarily random. We also provide a comparison between our results and show that our sufficient recovery conditions do not fall very short from the necessary ones in the special case when the subspaces are assumed to be equi-angular, see Section 3.2.5.

In Sections 3.3.3 and 3.6, we briefly explain certain standard approaches in CS that we have tried for proving recovery results for (sub-)Gaussian matrices but do not give the desired estimate that becomes better with decreasing λ. Therefore, we employ more recently developed tools [**75**] based on suprema of chaos processes and apply them to our problem in order to reach our results presented in Chapter 3. We would also like to note that our results do not necessarily assume that the subspaces satisfy the fusion frame property (1.6).

1.4. Organization

Each of the chapters below is self contained, including notation.

- **Chapter 2**: Here, we present nonuniform recovery results in the classical CS setting with subgaussian matrices which include Gaussian and Bernoulli matrices as examples. We use an RIP-less theory in order to achive our results. In many contributions the constant in the optimal scaling $m \geq Cs\ln(N/s)$ is either not provided explicitly or is rather large. In our work we give an explicit constant $C = 2$ for Gaussian and Bernoulli matrices (with a slightly suboptimal log-factor). The constant 2 is also shown to be optimal, and has been provided by different authors almost simultaneously. We also show the stability and robustness of our results. In the rest of the chapter, we give an overview and comparison of existing CS results in the nonuniform setting.
- **Chapter 3**: Here, we extend the concepts of CS to fusion frames and present results on the solution of the sparse recovery problem in fusion frames. Our model considers a block sparsity setup which is induced by the fusion frames subspaces with an additional incoherence property of these subspaces (quantified by the parameter λ). A sparse vector from the fusion frame is measured by a random matrix and then recovered by using ℓ_1/ℓ_2-minimization. It was shown earlier that $m \gtrsim s\ln(N/s)$ many measurements are sufficient for recovery. We improve this result to $m \gtrsim (1+\lambda s)\ln(N)$ by incorporating the structure of the subspaces. In Section 3.2.6, we present a number of numerical experiments in order to support our theoretical results. Our proofs use various techniques from CS such as an extension of RIP for fusion frames (see Section 3.2.7), the null space property (see Section 3.2.8) and several tools from Banach space geometry, coding theory and probability theory, in particular, random matrices and concentration inequalities.
- **Chapter 4**: Here, we give a brief summary of our results presented in the earlier chapters and discuss some open problems and future research directions.
- **Appendix**: In order not to break the flow of the thesis, we present some auxiliary results used in the main proofs in the appendix. These are mostly borrowed from probability theory.

CHAPTER 2

Nonuniform Recovery with Subgaussian Matrices

2.1. Introduction

Compressed sensing (CS) allows to reconstruct signals from far fewer measurements than what was considered necessary before. The seminal papers by E. Candès, J. Romberg, T. Tao [**18**, **20**] and by D. Donoho [**38**] on this subject have triggered a large research activity in mathematics, engineering and computer science with a lot of potential applications.

In mathematical terms we aim at solving the linear system of equations $y = Ax$ for $x \in \mathbb{C}^N$ when $y \in \mathbb{C}^m$ and $A \in \mathbb{C}^{m \times N}$ are given, and when $m \ll N$. Clearly, in general this task is impossible since even if A has full rank as there are infinitely many solutions to this equation. The situation dramatically changes if x is sparse, that is, $\|x\|_0 := \mathrm{card}\{\ell, x_\ell \neq 0\}$ is small.

As a first approach one is led to solve the optimization problem

$$\min_{z \in \mathbb{C}^N} \|z\|_0 \text{ subject to } Az = y,$$

where $y = Ax$. Unfortunately, this problem is NP-hard in general. It has become common to replace the ℓ_0-minimization problem by the ℓ_1-minimization problem

$$\min_{z \in \mathbb{C}^N} \|z\|_1 \text{ subject to } Az = y. \tag{2.1}$$

This problem can be solved by efficient convex optimization techniques [**11**]. As a key result of CS, under appropriate conditions on A and on the sparsity of x, ℓ_1-minimization indeed reconstructs the original x. Certain random matrices A are known to provide optimal recovery guarantees with high probability.

2.1.1. Nonuniform vs. uniform recovery

Throughout this thesis, we mainly talk about two types of recovery results:

- **Uniform recovery**: Such results state that with high probability on the draw of the random matrix A, *every* sparse vector can be reconstructed under appropriate conditions.
- **Nonuniform recovery**: Such results state that a given sparse vector x can be reconstructed with high probability on the the draw of the matrix A under appropriate conditions. The difference to uniform recovery is that nonuniform recovery does not imply that there is a matrix that recovers all x simultaneously. Or in other words, the small exceptional set of matrices for which recovery fails may depend on x.

Clearly, uniform recovery implies nonuniform recovery, but the converse is not true. One pursues different strategies in order to prove these different types of recovery results. In this chapter, we deal with nonuniform recovery with subgaussian matrices (see Section 2.1.3 for definition). In order to obtain such a result, a sufficient recovery condition for an individual vector is given by

constructing a 'dual certificate' which is mainly based on the Lagrange dual problem of (2.1), see Section 2.2.1.

2.1.2. Obtaining good constants for compressed sensing

Uniform recovery via ℓ_1-minimization is for instance satisfied if the by-now classical restricted isometry property (RIP) (1.4) holds for A with high probability [**17**, **19**]. A common choice is to take $A \in \mathbb{R}^{m\times N}$ as a *Gaussian* random matrix, that is, the entries of A are independent normal distributed random variables. If, for $\varepsilon \in (0,1)$,

$$m \geq C(s\ln(N/s) + \ln(2/\varepsilon)), \tag{2.2}$$

then with probability at least $1-\varepsilon$, we have uniform recovery of all s-sparse vectors $x \in \mathbb{R}^N$ using ℓ_1-minimization and A as measurement matrix, see e.g. [**20**, **40**, **84**]. The constant $C > 0$ is universal and estimates via the restricted isometry property give a value of around $C \approx 200$, which is significantly worse than what can be observed in practice. A direct analysis in [**95**] for the Gaussian case, which avoids the restricted isometry property, gives $C \approx 12$. Foucart and Rauhut in [**54**] follows a similar approach to obtain $C \approx 8$, see also [**70**]. This is still somewhat larger than the constants we obtain in the nonuniform setting. Our main results consider nonuniform sparse recovery using Gaussian and more general subgaussian random matrices in connection with ℓ_1-minimization and provide nonuniform recovery guarantees with an explicit and good constant $C = 2$. In particular, we also obtain the same constant for *Bernoulli* matrices whose entries are independent random variables taking the values $+1$ and -1 with equal probability. Moreover, our results can treat also the recovery of complex vectors and they extend to the stability of reconstruction when the vectors are only approximately sparse and measurements are perturbed.

Gaussian and subgaussian random matrices are very important for the theory of CS because they provide a model of measurement matrices which can be analyzed very accurately (as shown in this chapter). They are used in real-world sensing scenarios, for instance in the one-pixel camera [**46**]. Moreover, even if certain applications require more structure of the measurement matrix (leading to structured random matrices [**94**]), the empirically observed recovery performance of many types of matrices is very close to the one of (sub-)Gaussian random matrices [**44**], which underlines the importance of understanding subgaussian random matrices in CS.

2.1.3. Main results

- *The Gaussian case*

We say that an $m \times N$ random matrix A is Gaussian if its entries are independent and standard normal distributed random variables, that is, having mean zero and variance 1. Our nonuniform sparse recovery result for Gaussian matrices and ℓ_1-minimization reads as follows.

Theorem 2.1. *Let $x \in \mathbb{C}^N$ with $\|x\|_0 = s$. Let $A \in \mathbb{R}^{m\times N}$ be a randomly drawn Gaussian matrix, and let $\varepsilon \in (0,1)$. If*

$$m \geq s\left[\sqrt{2\ln(4N/\varepsilon)} + \sqrt{2\ln(2/\varepsilon)/s} + 1\right]^2, \tag{2.3}$$

then with probability at least $1-\varepsilon$ the vector x is the unique solution to the ℓ_1-minimization problem (2.1).

Remark. For large N and s, Condition (2.3) roughly becomes

$$m > 2s \ln(4N/\varepsilon). \tag{2.4}$$

Comparing with (2.2) we realize that the log-term falls slightly short of the optimal one $\log(N/s)$. However, we emphasize that our proof is short, and the constant is explicit and good. Indeed, when in addition s/N becomes very small (this is in fact the interesting regime) then we nevertheless reveal the conditions found by Donoho and Tanner [**39**,**40**], and in particular, the optimal constant 2. We note that Donoho and Tanner used methods from random polytopes, which are quite different from our proof technique. For a detailed overview see Section 2.4.

• *The subgaussian case*

We generalize our recovery result for matrices with entries that are independent subgaussian random variables. A random variable X is called *subgaussian* if there are constants $\beta, \theta > 0$ such that

$$\mathbb{P}(|X| \geq t) \leq \beta e^{-\theta t^2} \quad \text{for all } t > 0.$$

It can be shown [**111**] that X is subgaussian with $\mathbb{E}X = 0$ if and only if there exists a constant c (depending only on β and θ) such that

$$\mathbb{E}[\exp(tX)] \leq e^{ct^2} \text{ for all } t \in \mathbb{R}. \tag{2.5}$$

Important special cases of subgaussian mean-zero random variables are standard Gaussians, and Bernoulli (also called Rademacher) variables, that is, random variables that take the values ± 1 with equal probability. For both of these random variables the constant $c = 1/2$, see the Bernoulli case.

A random matrix with entries that are independent mean-zero subgaussian random variables with the same constant c in (2.5) is called a subgaussian random matrix. Note that the entries are not required to be identically distributed.

Theorem 2.2. *Let $x \in \mathbb{C}^N$ with $\|x\|_0 = s$. Let $A \in \mathbb{R}^{m \times N}$ be a random draw of a subgaussian matrix with constant c in* (2.5)*, and let $\varepsilon \in (0, 1)$. If*

$$m \geq s \left[\sqrt{4c \ln(4N/\varepsilon)} + \sqrt{C(3 + \ln(4/\varepsilon)/s)} \right]^2, \tag{2.6}$$

then with probability at least $1 - \varepsilon$ the vector x is the unique solution to the ℓ_1-minimization problem (2.1)*. The constant C in* (2.6) *only depends on c.*

More precisely, the constant $C = 1.646\tilde{c}^{-1}$, where $\tilde{c} = \tilde{c}(c)$ is the constant in (2.17).

• *The Bernoulli case*

We specialize the previous result for subgaussian matrices to Bernoulli matrices. We are then able to give explicit values for the constants appearing in the result of Theorem 2.2. If Y is a Bernoulli random variable, then by a Taylor series expansion

$$\mathbb{E}(\exp(tY)) = \frac{1}{2}\left(e^t + e^{-t}\right) \leq e^{\frac{1}{2}t^2}.$$

This shows that the subgaussian constant $c = 1/2$ in the Bernoulli case. Further, we have the following concentration inequality for a matrix $B \in \mathbb{R}^{m \times N}$ with independent entries taking values

of $\pm 1/\sqrt{m}$ with equal probability,

$$\mathbb{P}\left(\left|\|Bx\|_2^2 - \|x\|_2^2\right| > t\|x\|_2^2\right) \leq 2e^{-\frac{m}{2}(t^2/2 - t^3/3)}, \tag{2.7}$$

for all $x \in \mathbb{R}^N$, $t \in (0,1)$, see e.g. [**1**, **3**]. We can simply estimate $t^3 < t^2$ in (2.7) and obtain $\tilde{c} = 1/12$ in (2.17) and consequently $C = 1.646\tilde{c}^{-1} = 19.76$.

Corollary 2.3. *Let $x \in \mathbb{C}^N$ with $\|x\|_0 = s$. Let $A \in \mathbb{R}^{m\times N}$ be a matrix with entries that are independent Bernoulli random variables, and let $\varepsilon \in (0,1)$. If*

$$m \geq 2s\left[\sqrt{\ln(4N/\varepsilon)} + \sqrt{29.64 + 9.88\ln(4/\varepsilon)/s}\right]^2, \tag{2.8}$$

then with probability at least $1-\varepsilon$ the vector x is the unique solution to the ℓ_1-minimization problem (2.1).

Roughly speaking for large N and mildly large s the second square-root in (2.8) can be ignored and we arrive at $m \geq 2s\log(4N/\varepsilon)$.

- *Stable and robust recovery*

In this section we state extensions of our results for nonuniform recovery with Gaussian matrices that show stability when passing from sparse signals to only approximately sparse ones and that are robust under perturbations of the measurements. In this context we assume the noise model

$$y = Ax + e \in \mathbb{C}^m \text{ with } \|e\|_2 \leq \eta\sqrt{m}. \tag{2.9}$$

It it natural then to work with the noise constrained ℓ_1-minimization problem

$$\min_{z\in\mathbb{C}^N} \|z\|_1 \text{ subject to } \|Az - y\|_2 \leq \eta\sqrt{m}. \tag{2.10}$$

For the formulation of the next result we also recall the best s-term approximation error of x in ℓ_1-norm

$$\sigma_s(x)_1 = \inf_{\|z\|_0\leq s} \|x - z\|_1.$$

Theorem 2.4. *Let $x \in \mathbb{C}^N$ be an arbitrary fixed vector and let $S \subset \{1,2,\ldots,N\}$ denote the index set corresponding to its s largest absolute entries. Let $A \in \mathbb{R}^{m\times N}$ be a draw of a Gaussian random matrix. Suppose we take noisy measurements as in* (2.9). *If, for $\theta \in (0,1)$,*

$$m \geq s\left[\frac{\sqrt{2\ln(12N/\varepsilon)}}{1-\theta} + \sqrt{2\ln(6/\varepsilon)/s} + \sqrt{2}\right]^2, \tag{2.11}$$

then with probability at least $1-\varepsilon$, the solution $\hat{x}$ to minimization problem (2.10) *satisfies*

$$\|x - \hat{x}\|_2 \leq \frac{C_1}{\theta}\eta + \frac{C_2}{\theta}\frac{\sigma_s(x)_1}{\sqrt{s}}. \tag{2.12}$$

The constants $C_1, C_2 > 0$ are universal.

Remark. Condition (2.11) on the number of required measurements is very similar to (2.3) in the exactly sparse and noiseless case. When θ tends to 0 we almost obtain the same condition, but then the right hand side of the stability estimate (2.12) blows up. In other words, we need to take slightly more measurements than required for exact recovery in order to ensure stability and robustness of the reconstruction. Also, a version for subgaussian random matrices can be shown.

We note that the ℓ_2-estimate for the reconstruction error in (2.12) is as strong as the ℓ_2-error guarantee implied by the RIP in (1.5), as the coefficient of $\sigma_s(x)_1$ is $1/\sqrt{s}$ and the coefficient of η is a constant in both. In our proof of Theorem 2.4, we use a weaker concept of RIP introduced in Section 2.3.1 which allows us to obtain this strong error estimate and also have good constants in (2.11).

2.1.4. Notation

We now set up some notation needed in this chapter. Let $[N]$ denote the set $\{1, 2, \dots, N\}$. For a set $S \subset [N]$, $\overline{S}$ denotes the complement set $[N] \setminus S$. The column submatrix of a matrix A consisting of the columns indexed by S is written $A_S = (a_j)_{j \in S}$ and $a_j \in \mathbb{R}^m$, $j = 1, \dots, m$ denote the columns of A. Similarly $x_S \in \mathbb{C}^S$ denotes the vector $x \in \mathbb{C}^N$ restricted to the entries in S. We further need to introduce the sign vector $\operatorname{sgn}(x) \in \mathbb{C}^N$ having entries

$$\operatorname{sgn}(x)_j := \begin{cases} \frac{x}{|x_j|} & \text{if } x_j \neq 0, \\ 0 & \text{if } x_j = 0, \end{cases} \quad j \in [N].$$

The Moore-Penrose pseudo-inverse of a matrix B such that (B^*B) is invertible is given by $B^\dagger = (B^*B)^{-1}B^*$, so that $B^\dagger B = \mathrm{Id}$, where Id is the identity matrix.

2.2. Exact sparse recovery

2.2.1. Dual certificate

In this section we state some auxiliary results that are used in the proofs of the main theorems, directly or indirectly. The proofs of Theorems 2.1 and 2.2 require a condition for sparse recovery, which not only depends on the matrix A but also on the sparse vector $x \in \mathbb{C}^N$ to be recovered. The following theorem is due to J.J. Fuchs [**55**] in the real-valued case and was extended to the complex case by J. Tropp [**103**], see also [**94**, Theorem 2.8] for a slightly simplified proof.

Theorem 2.5. *Let $A \in \mathbb{C}^{m \times N}$ and $x \in \mathbb{C}^N$ with $S := \operatorname{supp}(x)$. Assume that A_S is injective and that there exists a vector $h \in \mathbb{C}^m$ such that*

$$A_S^* h = \operatorname{sgn}(x_S),$$
$$|(A^*h)_\ell| < 1, \ \ell \in [N] \setminus S.$$

Then x is the unique solution to the ℓ_1-minimization problem (2.1) with $y = Ax$.

Choosing the vector $h = \left(A_S^\dagger\right)^* \operatorname{sgn}(x_S)$ leads to the following corollary.

Corollary 2.6. *Let $A \in \mathbb{C}^{m \times N}$ and $x \in \mathbb{C}^N$ with $S := \operatorname{supp}(x)$. If the matrix A_S is injective and if*

$$|\langle (A_S)^\dagger a_\ell, \operatorname{sgn}(x_S)\rangle| < 1 \ \text{ for all } \ \ell \in [N] \setminus S,$$

then the vector x is the unique solution to the ℓ_1-minimization problem (2.1) with $y = Ax$.

2.2.2. Proof of Theorem 2.1

We set $S := \operatorname{supp}(x)$, which has a cardinality s. By Corollary 2.6, for recovery via ℓ_1-minimization, it is sufficient to show that

$$|\langle (A_S)^\dagger a_\ell, \operatorname{sgn}(x_S)\rangle| = |\langle a_\ell, (A_S^\dagger)^* \operatorname{sgn}(x_S)\rangle| < 1 \ \text{ for all } \ \ell \in [N] \setminus S.$$

Therefore, the probability of recovery failure is bounded by

$$\mathcal{P} := \mathbb{P}(\exists \ell \notin S \ |\langle (A_S)^\dagger a_\ell, \mathrm{sgn}(x_S)\rangle| \geq 1).$$

If we condition $X := \langle a_\ell, (A_S^\dagger)^* \mathrm{sgn}(x_S)\rangle$ on A_S, it is a Gaussian random variable. Further, $X = \sum_{j=1}^m (a_\ell)_j [(A_S^\dagger)^* \mathrm{sgn}(x_S)]_j$ is centered so its variance ν^2 can be estimated by

$$\begin{aligned}\nu^2 = \mathbb{E}(X^2) &= \sum_{j=1}^m \mathbb{E}[(a_\ell)_j^2][(A_S^\dagger)^* \mathrm{sgn}(x_S)]_j^2 \\ &= \|(A_S^\dagger)^* \mathrm{sgn}(x_S)\|_2^2 \leq \sigma_{\min}^{-2}(A_S)\|\mathrm{sgn}(x_S)\|_2^2 = \sigma_{\min}^{-2}(A_S)\, s,\end{aligned}$$

where $\sigma_{\min}$ denotes the smallest singular value. The last inequality uses the fact that $\|(A_S^\dagger)^*\|_{2\to 2} = \|A_S^\dagger\|_{2\to 2} = \sigma_{\min}^{-1}(A_S)$. The tail of a mean-zero Gaussian random variable X with variance σ^2 obeys

$$\mathbb{P}(|X| > t) \leq e^{-t^2/2\sigma^2}. \tag{2.13}$$

(See [**94**, Lemma 10.2].) Then it follows that

$$\begin{aligned}\mathcal{P} &\leq \mathbb{P}\left(\exists \ell \notin S \ |\langle (A_S)^\dagger a_\ell, \mathrm{sgn}(x_S)\rangle| \geq 1 \,\middle|\, \|(A_S^\dagger)^* \mathrm{sgn}(x_S)\|_2 < \alpha\right) \\ &\quad + \mathbb{P}(\|(A_S^\dagger)^* \mathrm{sgn}(x_S)\|_2 \geq \alpha) \\ &\leq 2N \exp(-1/2\alpha^2) + \mathbb{P}(\sigma_{\min}^{-1}(A_S)\sqrt{s} \geq \alpha).\end{aligned} \tag{2.14}$$

The inequality in (2.14) uses the tail estimate (2.13), the union bound, and the independence of a_ℓ and A_S. The first term in (2.14) is bounded by $\varepsilon/2$ if

$$\alpha \leq \frac{1}{\sqrt{2\ln(4N/\varepsilon)}}. \tag{2.15}$$

In order to estimate the second term in (2.14) we use (A.9) as follows

$$\begin{aligned}\mathbb{P}(\sigma_{\min}^{-1}(A_S)\sqrt{s} \geq \alpha) = \mathbb{P}(\sigma_{\min}(A_S) \leq \sqrt{s}/\alpha) &= \mathbb{P}\left(\sigma_{\min}(A_S/\sqrt{m}) \leq \frac{1}{\sqrt{m}}\frac{\sqrt{s}}{\alpha}\right) \\ &\leq \exp\left(\frac{-m(1-(\alpha^{-1}+1)\sqrt{s/m})^2}{2}\right).\end{aligned} \tag{2.16}$$

If we choose α that makes (2.15) an equality, plug it into condition (2.16), and require that (2.16) is bounded by $\varepsilon/2$ we arrive at the condition

$$m \geq s\left[\sqrt{2\ln(4N/\varepsilon)} + \sqrt{2\ln(2/\varepsilon)/s} + 1\right]^2,$$

which ensures recovery with probability at least $1-\varepsilon$. This concludes the proof of Theorem 2.1. □

2.2.3. Proof of Theorem 2.2

- *Conditioning of subgaussian matrices*

We start with a definition.

Definition. *Let Y be a random vector in $\mathbb{R}^N$. If $\mathbb{E}|\langle Y, x\rangle|^2 = \|x\|_2^2$ for all $x \in \mathbb{R}^N$, then Y is called isotropic. Furthermore, if for all $x \in \mathbb{R}^N$ with $\|x\|_2 = 1$, the random variable $\langle Y, x\rangle$ is subgaussian with subgaussian parameter c being independent of x, that is,*

$$\mathbb{E}[\exp(\theta\langle Y, x\rangle)] \leq \exp(c\theta^2), \quad \text{for all } \theta \in \mathbb{R}, \quad \|x\|_2 = 1,$$

then Y is called a subgaussian random vector.

While the following lemma is well-known in principle, the right scaling in δ has only recently appeared in [**54**], compare with [**3**, **84**]. We include its proof for the sake of completeness.

Lemma 2.7. *Let $S \subset [N]$ with $\mathrm{card}(S) = s$. Let A be an $m \times N$ random matrix with independent, isotropic, and subgaussian rows with the same parameter c as in (2.5). Then, for $\delta \in (0,1)$, normalized matrix $\tilde{A} = \frac{1}{\sqrt{m}} A$ satisfies*

$$\|\tilde{A}_S^* \tilde{A}_S - \mathrm{Id}\|_{2\to 2} \leq \delta$$

with probability at least $1 - \varepsilon$ provided

$$m \geq C\delta^{-2}(3s + \ln(2\varepsilon^{-1})),$$

where C depends only on c.

Proof. The following concentration inequality for subgaussian random variables appears, for instance, in [**1**, **84**].

$$\mathbb{P}(|\|\tilde{A}x\|_2^2 - \|x\|_2^2| > t\|x\|_2^2) \leq 2\exp(-\tilde{c}mt^2), \tag{2.17}$$

where $\tilde{c}$ depends only on c. We will combine the above concentration inequality with the net technique. Let $\rho \in (0, \sqrt{2} - 1)$ be a number to be determined later. According to a volumetric argument given in Proposition 3.47, see also [**94**, Proposition 10.1], there exists a finite subset U of the unit sphere $\mathcal{S} = \{x \in \mathbb{R}^{\mathbb{N}}, \mathrm{supp}(x) \subset S, \|x\|_2 = 1\}$, which satisfies

$$|U| \leq \left(1 + \frac{2}{\rho}\right)^s \qquad \text{and} \qquad \min_{u \in U} \|z - u\|_2 \leq \rho \quad \text{for all } z \in \mathcal{S}.$$

The concentration inequality (2.17) yields

$$\begin{aligned}
&\mathbb{P}\left(\left|\|\tilde{A}u\|_2^2 - \|u\|_2^2\right| > t\,\|u\|_2^2 \quad \text{for some } u \in \mathcal{U}\right) \\
&\leq \sum_{u \in U} \mathbb{P}\left(\left|\|\tilde{A}u\|_2^2 - \|u\|_2^2\right| > t\,\|u\|_2^2\right) \leq 2|U| \exp\left(-\tilde{c}t^2\, m\right) \\
&\leq 2\left(1 + \frac{2}{\rho}\right)^s \exp\left(-\tilde{c}t^2\, m\right).
\end{aligned}$$

The positive number t will be set later depending on δ and on ρ. Let us assume for now that the realization of the random matrix $\tilde{A}$ yields

$$\left|\|\tilde{A}u\|_2^2 - \|u\|_2^2\right| \leq t \qquad \text{for all } u \in U. \tag{2.18}$$

By the above, this occurs with probability exceeding

$$1 \;-\; 2\left(1 + \frac{2}{\rho}\right)^s \exp\left(-\tilde{c}t^2\, m\right).$$

Next we show that (2.18) implies $\left|\|\tilde{A}x\|_2^2 - \|x\|_2^2\right| \leq \delta$ for all $x \in \mathcal{S}$, that is $\|\tilde{A}_S^* \tilde{A}_S - \mathrm{Id}\|_{2\to 2} \leq \delta$ (when t is determined appropriately). Let $B = \tilde{A}_S^* \tilde{A}_S - \mathrm{Id}$, so that we have to show $\|B\|_{2\to 2} \leq \delta$. Note that (2.18) means that $|\langle Bu, u\rangle| \leq t$ for all $u \in U$. Now consider a vector $x \in \mathcal{S}$, for which we choose a vector $u \in U$ satisfying $\|x - u\|_2 \leq \rho < \sqrt{2} - 1$. We obtain

$$|\langle Bx, x\rangle| = |\langle B(u + x - u), u + x - u\rangle|$$

$$
\begin{aligned}
&= |\langle Bu, u\rangle + \langle B(x-u), x-u\rangle + 2\,\langle Bu, x-u\rangle| \\
&\le |\langle Bu, u\rangle| + |\langle B(x-u), x-u\rangle| + 2\,\|Bu\|_2\,\|x-u\|_2 \\
&\le t + \|B\|_{2\to 2}\,\rho^2 + 2\,\|B\|_{2\to 2}\,\rho.
\end{aligned}
$$

Taking the supremum over all $x \in \mathcal{S}$, we deduce that

$$
\|B\|_{2\to 2} \le t + \|B\|_{2\to 2}\left(\rho^2 + 2\rho\right), \qquad \text{i.e.,} \qquad \|B\|_{2\to 2} \le \frac{t}{2-(\rho+1)^2}.
$$

Note that the division by $2-(\rho+1)^2$ is justified by the assumption that $\rho < \sqrt{2}-1$. Then we choose

$$
t = t_{\delta,\rho} := \left(2-(\rho+1)^2\right)\delta,
$$

so that $\|B\|_{2\to 2} \le \delta$, and with our definition of t,

$$
\mathbb{P}\left(\|\tilde{A}_S^*\tilde{A}_S - \mathrm{Id}\|_{2\to 2} > \delta\right) \le 2\left(1+\frac{2}{\rho}\right)^s \exp\left(-\tilde{c}\delta^2(2-(\rho+1)^2)^2 m\right).
$$

Hence, $\|\tilde{A}_S^*\tilde{A}_S - \mathrm{Id}\|_{2\to 2} \le \delta$ with probability at least $1-\varepsilon$ provided

$$
m \ge \frac{1}{\tilde{c}(2-(\rho+1)^2)^2}\delta^{-2}\left(\ln(1+2/\rho)s + \ln(2\varepsilon^{-1})\right). \tag{2.19}
$$

Now we choose ρ such that $\ln(1+2/\rho) = 3$, that is, $\rho = 2/(e^3-1)$. Then (2.19) gives the condition

$$
m \ge C\delta^{-2}\left(3s + \ln(2\varepsilon^{-1})\right)
$$

with $C = 1.646\,\tilde{c}^{-1}$. This concludes the proof. □

• *Proof of Theorem 2.2 continued*

We follow a similar path as in the proof of the Gaussian case. First denote $S := \mathrm{supp}(x)$. The failure probability $\mathcal{P}$ can be bounded by

$$
\mathcal{P} \le \mathbb{P}\left(\exists \ell \notin S\ |\langle (A_S)^\dagger a_\ell, \mathrm{sgn}(x_S)\rangle| \ge 1 \,\middle|\, \|(A_S^\dagger)^*\mathrm{sgn}(x_S)\|_2 < \alpha\right) + \mathbb{P}(\|(A_S^\dagger)^*\mathrm{sgn}(x_S)\|_2 \ge \alpha). \tag{2.20}
$$

The first term in (2.20) can be bounded by using Lemma A.1. Conditioning on A_S and $\|(A_S^\dagger)^*\mathrm{sgn}(x_S)\|_2 < \alpha$, we obtain

$$
\mathbb{P}(|\langle (A_S)^\dagger a_\ell, \mathrm{sgn}(x_S)\rangle| \ge 1) = \mathbb{P}(|\sum_{j=1}^m (a_\ell)_j [(A_S^\dagger)^*\mathrm{sgn}(x_S)]_j| \ge 1) \le 2\mathrm{exp}(-1/(4c\alpha^2)).
$$

So by the union bound the first term in (2.20) can be estimated by $2N\mathrm{exp}(-1/(4c\alpha^2))$, which in turn is no larger than $\varepsilon/2$ provided

$$
\alpha \le \sqrt{1/(4c\ln(4N/\varepsilon))}. \tag{2.21}
$$

For the second term in (2.20), we have

$$
\begin{aligned}
&\mathbb{P}(\|(A_S^\dagger)^*\mathrm{sgn}(x_S)\|_2 \ge \alpha) \le \mathbb{P}(\sigma_{\min}^{-1}(A_S)\sqrt{s} \ge \alpha) \\
&= \mathbb{P}(\sigma_{\min}(A_S) \le \sqrt{s}/\alpha) = \mathbb{P}\left(\sigma_{\min}(A_S/\sqrt{m}) \le \frac{1}{\sqrt{m}}\frac{\sqrt{s}}{\alpha}\right),
\end{aligned}
$$

where $\sigma_{\min}$ denotes the smallest singular value. By Lemma 2.7 the normalized subgaussian matrix $\tilde{A}_S := A_S/\sqrt{m}$ satisfies

$$\mathbb{P}(\sigma_{\min}(\tilde{A}_S) < 1-\delta) < \mathbb{P}(\sigma_{\min}(\tilde{A}_S) < \sqrt{1-\delta}) < \mathbb{P}(\|\tilde{A}_S^*\tilde{A}_S - \mathrm{Id}\|_{2\to 2} \geq \delta) < \varepsilon/2$$

provided $m \geq C\delta^{-2}(3s + \ln(4\varepsilon^{-1}))$ and $\delta \in (0,1)$, where C depends on the subgaussian constant c. The choice $\frac{1}{\sqrt{m}}\frac{\sqrt{s}}{\alpha} = 1-\delta$ yields $\delta = 1 - \frac{\sqrt{s}}{\alpha\sqrt{m}}$. Combining these arguments and choosing α that makes (2.21) an equality, we can bound the failure probability by ε provided

$$m \geq C\left(1 - \frac{\sqrt{4cs\ln(4N/\varepsilon)}}{\sqrt{m}}\right)^{-2}(3s + \ln(4/\varepsilon)). \tag{2.22}$$

Solving (2.22) for m yields the condition

$$m \geq s\left[\sqrt{4c\ln(4N/\varepsilon)} + \sqrt{C(3+\ln(4/\varepsilon)/s)}\right]^2.$$

This condition also implies $\delta \in (0,1)$ which concludes the proof of Theorem 2.2. □

2.3. Stable and robust recovery

2.3.1. Weak RIP

The concept of weak restricted isometry property (weak RIP) was introduced in [**22**].

Definition. *(Weak RIP) Let $S \subset [N]$ be fixed with cardinality s and fix $\delta_1, \delta_2 > 0$. Then a matrix $A \in \mathbb{R}^{m\times N}$ is said to satisfy the weak RIP with parameters $(S, r, \delta_1, \delta_2)$ if*

$$(1-\delta_1)\|v\|_2^2 \leq \|Av\|_2^2 \leq (1+\delta_2)\|v\|_2^2$$

for all v supported on $S \cup R$ and all subsets R in $[N] \setminus S$ with cardinality $|R| \leq r$.

The weak RIP is a combination of the RIP and the local conditioning of A on the column set S. The next theorem derives conditions for a Gaussian matrix to satisfy the weak RIP for a given set S and r value.

Theorem 2.8. *Let $A \in \mathbb{R}^{m\times N}$ be a randomly drawn Gaussian matrix. Then the normalized matrix $\tilde{A} = A/\sqrt{m}$ satisfies the weak RIP with parameters $(S, r, \delta_1, \delta_2)$ for $r \leq N$ and $\delta_1, \delta_2 \in (0,1)$ with probability at least $1-\varepsilon$ provided that*

$$m \geq \min\left\{1-\sqrt{1-\delta_1}, \sqrt{1+\delta_2}-1\right\}^{-2}\left[\sqrt{s+r} + \sqrt{2r\ln(eN/r) + 2\ln(2/\varepsilon)}\right]^2. \tag{2.23}$$

Proof. Let us denote the event $E := \{\tilde{A}$ satisfies weak RIP with parameters $(S, r, \delta_1, \delta_2)\}$ and the set

$$D = \{x : \|x\|_2 = 1, \mathrm{supp}(x) \subset S \cup R, |R| \leq r, S \cap R = \emptyset\}.$$

Then

$$\sup_{x\in D}\|\tilde{A}x\|_2^2 \leq (1+\delta_2)\|x\|_2^2 \quad \text{and} \quad \inf_{x\in D}\|\tilde{A}x\|_2^2 \geq (1-\delta_1)\|x\|_2^2 \tag{2.24}$$

implies E. It is easy to see that (2.24) is equivalent to

$$\sup_{|R|\leq r}\sigma_{\max}(\tilde{A}_{S\cup R}) \leq \sqrt{1+\delta_2} \quad \text{and} \quad \inf_{|R|\leq r}\sigma_{\min}(\tilde{A}_{S\cup R}) \geq \sqrt{1-\delta_1},$$

where $\sigma_{\max}$ and $\sigma_{\min}$ denote the largest and smallest singular values, respectively. Therefore we obtain

$$\mathbb{P}(E^{\mathsf{c}}) \leq \mathbb{P}(\exists R,\ \sigma_{\max}(\tilde{A}_{S\cup R}) > \sqrt{1+\delta_2}) + \mathbb{P}(\exists R,\ \sigma_{\min}(\tilde{A}_{S\cup R}) < \sqrt{1-\delta_1}).$$

The terms on the right hand side of the inequality can be bounded by (A.10) which gives

$$\mathbb{P}(\sigma_{\max}(\tilde{A}_{S\cup R}) > 1 + \sqrt{(s+r)/m} + t_1) \leq \exp^{-mt_1^2/2}.$$

Equating $1 + \sqrt{(s+r)/m} + t_1 = \sqrt{1+\delta_2}$ yields $t_1 = \sqrt{1+\delta_2} - 1 - \sqrt{(s+r)/m}$. Similarly by (A.9)

$$\mathbb{P}(\sigma_{\min}(\tilde{A}_{S\cup R}) > 1 - \sqrt{(s+r)/m} - t_2) \leq \exp^{-mt_2^2/2}.$$

Now we choose $t_2 = 1 - \sqrt{(s+r)/m} - \sqrt{1-\delta_1}$. We have

$$\mathbb{P}(\sigma_{\max}(\tilde{A}_{S\cup R}) > \sqrt{1+\delta_2}) + \mathbb{P}(\sigma_{\min}(\tilde{A}_{S\cup R}) < \sqrt{1-\delta_1}) \leq \exp^{-mt_1^2/2} + \exp^{-mt_2^2/2}.$$

Taking the union bound over all subsets R, we have

$$\begin{aligned}\mathbb{P}(E^{\mathsf{c}}) &\leq \sum_{|R|\leq r} \Big(\mathbb{P}(\sigma_{\max}(\tilde{A}_{S\cup R}) > \sqrt{1+\delta_2}) + \mathbb{P}(\sigma_{\min}(\tilde{A}_{S\cup R}) < \sqrt{1-\delta_1})\Big) \\ &\leq \binom{N}{r}(\exp^{-mt_1^2/2} + \exp^{-mt_2^2/2}) \leq \left(\frac{eN}{r}\right)^r \exp^{-mt_1^2/2} + \left(\frac{eN}{r}\right)^r \exp^{-mt_2^2/2}. \qquad (2.25)\end{aligned}$$

The last estimate is due to [**54**, Lemma C.5]. Now we require that both terms on the right hand side of (2.25) are bounded by $\varepsilon/2$. For the first term this holds provided

$$m(\sqrt{1+\delta_2} - 1 - \sqrt{(s+r)/m})^2 \geq 2r\ln(eN/r) + 2\ln(2/\varepsilon).$$

Taking square roots on both sides and rearranging yields

$$\sqrt{m} \geq (\sqrt{1+\delta_2} - 1)^{-1}[\sqrt{s+r} + \sqrt{2r\ln(eN/r) + 2\ln(2/\varepsilon)}]. \qquad (2.26)$$

Similarly the second term in (2.25) is bound by $\varepsilon/2$ provided

$$\sqrt{m} \geq (1 - \sqrt{1-\delta_1})^{-1}[\sqrt{s+r} + \sqrt{2r\ln(eN/r) + 2\ln(2/\varepsilon)}]. \qquad (2.27)$$

Squaring (2.26) and (2.27), we arrive at condition (2.23). This completes the proof. □

Before proceeding, we state an auxiliary result from [**94**, Prop. 2.5(c)] that will be useful shortly.

Lemma 2.9. *Let $A \in \mathbb{R}^{m\times N}$ satisfy the RIP (1.4) with constants $\tilde{\delta}_s$ and let $u, v \in \mathbb{R}^N$ with disjoint supports,* $\operatorname{supp}(u) \cap \operatorname{supp}(v) = \emptyset$. *Let $s = |\operatorname{supp}(u)| + |\operatorname{supp}(v)|$. Then*

$$|\langle Au, Av\rangle| \leq \tilde{\delta}_s \|u\|_2 \|v\|_2.$$

The key to the proof of Theorem 2.4 is the following stable and robust version of the dual certificate based recovery condition of Theorem 2.5. Its proof follows a similar strategy as in [**22**] and [**69**, Theorem 3.1].

Lemma 2.10. *Let $x \in \mathbb{C}^N$ and $A \in \mathbb{R}^{m\times N}$. Let S be the set of indices of the s largest absolute entries of x. Assume that A satisfies weak RIP with parameters $(S, r, \delta_1, \delta_2)$, $r \leq N$ even and $\delta_1, \delta_2 \in (0,1)$, and there exists a vector $v \in \mathbb{C}^m$ such that, for some $\theta \in (0,1)$,*

$$A_S^* v = \operatorname{sgn}(x_S), \qquad (2.28)$$

$$|(A^*v)_\ell| < 1-\theta,\ \ell \in \overline{S}, \tag{2.29}$$

$$\|v\|_2 \le \beta\sqrt{s}. \tag{2.30}$$

Suppose we take noisy measurements $y = Ax + e \in \mathbb{C}^m$ with $\|e\|_2 \le \eta$. Then the solution $\hat{x}$ to

$$\min_{z\in\mathbb{C}^N} \|z\|_1 \text{ subject to } \|Az - y\| \le \eta$$

satisfies

$$\|x-\hat{x}\|_2 \le \frac{\sqrt{1+\delta_2}}{1-\delta_1}2\eta + \left(\frac{2\sqrt{2}\max\{\delta_1,\delta_2\}}{1-\delta_1} + \sqrt{2}\right)\left(\frac{2\beta}{\theta}\sqrt{\frac{s}{r}}\eta + \frac{2}{\theta}\frac{\sigma_s(x)_1}{\sqrt{r}}\right). \tag{2.31}$$

PROOF. Set $\hat{x} = x + h$. Our goal is to bound the norm of h. Due to (2.10) and the assumption on the noise level $\|e\|_2\eta$, we have

$$\|Ah\|_2 = \|Ax - y - (A\hat{x} - y)\|_2 \le \|Ax - y\|_2 + \|A\hat{x} - y\|_2 \le 2\eta. \tag{2.32}$$

Since $\hat{x}$ is a minimizer of (2.10),

$$\begin{aligned}\|x\|_1 \ge \|\hat{x}\|_1 &= \|(x+h)_S\|_1 + \|(x+h)_{\overline{S}}\|_1 \\ &\ge \mathrm{Re}\langle (x+h)_S, \mathrm{sgn}(x)_S\rangle + \|h_{\overline{S}}\|_1 - \|x_{\overline{S}}\|_1 \\ &= \|x\|_1 + \mathrm{Re}\langle h_S, \mathrm{sgn}(x)_S\rangle + \|h_{\overline{S}}\|_1 - 2\|x_{\overline{S}}\|_1.\end{aligned}$$

Rearranging gives

$$\|h_{\overline{S}}\|_1 \le |\mathrm{Re}\langle h_S, \mathrm{sgn}(x)_S\rangle| + 2\|x_{\overline{S}}\|_1. \tag{2.33}$$

Using assumptions (2.28), (2.29), (2.30) together with the Cauchy-Schwarz and Hölder's inequalities

$$\begin{aligned}|\mathrm{Re}\langle h_S, \mathrm{sgn}(x)_S\rangle| = |\mathrm{Re}\langle h_S, (A^*v)_S\rangle| &\le |\langle h, A^*v\rangle| + |\langle h_{\overline{S}}, (A^*v)_{\overline{S}}\rangle| \\ &\le \|Ah\|_2\|v\|_2 + \|h_{\overline{S}}\|_1\|(A^*v)_{\overline{S}}\|_\infty \le 2\eta\beta\sqrt{s} + (1-\theta)\|h_{\overline{S}}\|_1.\end{aligned}$$

Plugging this into (2.33) yields

$$\|h_{\overline{S}}\|_1 \le \frac{2\eta\beta}{\theta}\sqrt{s} + \frac{2}{\theta}\sigma_s(x)_1. \tag{2.34}$$

We have bounded the off-support part of h above. Now we proceed to bound the norm of h_S. We will use the weak RIP of the matrix A. As assumed r is even in the definition of the weak RIP. We begin by partitioning $\overline{S}$ into subsets $S_1, S_2, \ldots$ of length $r/2$. Let S_1 be the indices of the $r/2$ largest entries of $h_{\overline{S}}$, S_2 be those of the next $r/2$ largest and so on. Denote $S_{01} = S \cup S_1$. We first bound $\|h_{S_{01}}\|_2$. The weak RIP assumption gives

$$(1-\delta_1)\|h_{S_{01}}\|_2^2 \le \|A_{S_{01}}h_{S_{01}}\|_2^2 = \langle A_{S_{01}}h_{S_{01}}, Ah\rangle - \langle A_{S_{01}}h_{S_{01}}, A_{\overline{S_{01}}}h_{\overline{S_{01}}}\rangle. \tag{2.35}$$

The first part on the right hand side of the equality above can be bounded by the Cauchy-Schwarz and the weak RIP,

$$|\langle A_{S_{01}}h_{S_{01}}, Ah\rangle| \le \|A_{S_{01}}h_{S_{01}}\|_2\|Ah\|_2 \le 2\eta\sqrt{1+\delta_2}\|h_{S_{01}}\|_2. \tag{2.36}$$

We also used (2.32) here. For the second term, we again split into two further terms,

$$|\langle A_{S_{01}}h_{S_{01}}, A_{\overline{S_{01}}}h_{\overline{S_{01}}}\rangle| \le |\langle A_S h_S, A_{\overline{S_{01}}}h_{\overline{S_{01}}}\rangle| + |\langle A_{S_1}h_{S_1}, A_{\overline{S_{01}}}h_{\overline{S_{01}}}\rangle|.$$

We have

$$|\langle A_S h_S, A_{\overline{S_{01}}} h_{\overline{S_{01}}}\rangle| \leq \sum_{j\geq 2} |\langle A_S h_S, A_{S_j} h_{S_j}\rangle|.$$

According to Lemma 2.9, the weak RIP implies

$$|\langle A_S h_S, A_{S_j} h_{S_j}\rangle| \leq \max\{\delta_1, \delta_2\}\|h_S\|_2 \|h_{S_j}\|_2.$$

Therefore,

$$|\langle A_S h_S, A_{\overline{S_{01}}} h_{\overline{S_{01}}}\rangle| \leq \max\{\delta_1, \delta_2\}\|h_S\|_2 \sum_{j\geq 2} \|h_{S_j}\|_2.$$

We bound the sum following a well-known argument from [**21**, (3.10)]. First we note that for each $j \geq 2$,

$$\|h_{S_j}\|_2 \leq \sqrt{r/2}\|h_{S_j}\|_\infty \leq \frac{1}{\sqrt{r/2}}\|h_{S_{j-1}}\|_1,$$

where $\|\cdot\|_\infty$ gives the largest absolute entry, and thus

$$\sum_{j\geq 2} \|h_{S_j}\|_2 \leq \frac{1}{\sqrt{r/2}}(\|h_{S_1}\|_1 + \|h_{S_2}\|_1 + \ldots) \leq \frac{1}{\sqrt{r/2}}\|h_{\overline{S}}\|_1. \tag{2.37}$$

This yields

$$|\langle A_S h_S, A_{\overline{S_{01}}} h_{\overline{S_{01}}}\rangle| \leq \frac{\max\{\delta_1, \delta_2\}}{\sqrt{r/2}}\|h_S\|_2 \|h_{\overline{S}}\|_1.$$

Similarly we obtain

$$|\langle A_{S_1} h_{S_1}, A_{\overline{S_{01}}} h_{\overline{S_{01}}}\rangle| \leq \frac{\max\{\delta_1, \delta_2\}}{\sqrt{r/2}}\|h_{S_1}\|_2 \|h_{\overline{S}}\|_1.$$

Then,

$$|\langle A_{S_{01}} h_{S_{01}}, A_{\overline{S_{01}}} h_{\overline{S_{01}}}\rangle| \leq 2\sqrt{2}\frac{\max\{\delta_1, \delta_2\}}{\sqrt{r}}\|h_{S_{01}}\|_2 \|h_{\overline{S}}\|_1.$$

Plugging this and (2.36) into (2.35) yields

$$\|h_{S_{01}}\|_2 \leq \frac{\sqrt{1+\delta_2}}{1-\delta_1} 2\eta + \frac{2\sqrt{2}\max\{\delta_1, \delta_2\}}{(1-\delta_1)\sqrt{r}}\|h_{\overline{S}}\|_1.$$

This together with (2.34) leads us to the desired bound,

$$\begin{aligned}
\|h\|_2 &\leq \|h_{S_{01}}\|_2 + \sum_{j\geq 2}\|h_{S_j}\|_2 \leq \|h_{S_{01}}\|_2 + \frac{1}{\sqrt{r/2}}\|h_{\overline{S}}\|_1 \\
&\leq \frac{\sqrt{1+\delta_2}}{1-\delta_1} 2\eta + \left(\frac{2\sqrt{2}\max\{\delta_1, \delta_2\}}{1-\delta_1} + \sqrt{2}\right)\frac{1}{\sqrt{r}}\|h_{\overline{S}}\|_1 \\
&\leq \frac{\sqrt{1+\delta_2}}{1-\delta_1} 2\eta + \left(\frac{2\sqrt{2}\max\{\delta_1, \delta_2\}}{1-\delta_1} + \sqrt{2}\right)\left(\frac{2\eta\beta}{\theta}\sqrt{\frac{s}{r}} + \frac{2}{\theta}\frac{\sigma_s(x)_1}{\sqrt{r}}\right).
\end{aligned}$$

This completes the proof. □

2.3.2. Proof of Theorem 2.4

The proof relies on Lemma 2.10. We will set $\beta, \delta_1, \delta_2$ and r later on. First observe that (2.10) is equivalent to

$$\min_{z\in\mathbb{C}^N} \|z\|_1 \text{ subject to } \left\|\frac{1}{\sqrt{m}}Az - \frac{1}{\sqrt{m}}y\right\|_2 \leq \eta.$$

Our candidate for the dual certificate is $v = (\tilde{A}_S^\dagger)^* \mathrm{sgn}(x_S)$ as in the noiseless case, where $\tilde{A} = A/\sqrt{m}$ is the normalized matrix. We will check the assumptions of Lemma 2.10 for this choice. First, Theorem 2.8 says that $\tilde{A}$ satisfies weak RIP with parameters $(S, r, \delta_1, \delta_2)$ with at least probability $1 - \varepsilon$ provided

$$m \geq \min\left\{1 - \sqrt{1-\delta_1}, \sqrt{1+\delta_2} - 1\right\}^{-2} \left[\sqrt{s+r} + \sqrt{2r\ln(eN/r) + 2\ln(2/\varepsilon)}\right]^2. \tag{2.38}$$

Assumption (2.28) is satisfied by definition. In the proof of Theorem 2.1, it was shown that Condition (2.29) is satisfied for $\theta = 0$ with probability at least $1-\varepsilon$ provided (2.3) holds. Replacing 1 by $1 - \theta$, basically changes the condition (2.15) to

$$\alpha \leq \frac{1-\theta}{\sqrt{2\ln(4N/\varepsilon)}},$$

where α appears as a variable in the proof of Theorem 2.1. Then we can similarly show that (2.29) is satisfied provided

$$m \geq s\left[\frac{\sqrt{2\ln(4N/\varepsilon)}}{1-\theta} + \sqrt{2\ln(2/\varepsilon)/s} + 1\right]^2, \tag{2.39}$$

with probability at least $1 - \varepsilon$. Finally we seek a condition for (2.30) to hold. Observe that

$$\|v\|_2 = \|(\tilde{A}_S^\dagger)^* \mathrm{sgn}(x_S)\|_2 \leq \sigma_{\min}^{-1}(\tilde{A}_S)\|\mathrm{sgn}(x_S)\|_2 = \sigma_{\min}^{-1}(\tilde{A}_S)\sqrt{s}.$$

It remains to show that $\sigma_{\min}^{-1}(\tilde{A}_S) \leq \beta$ with high probability. In fact the assumed weak RIP implies such a bound but the following analysis gives tighter bound. We start this analysis by observing that

$$\mathbb{P}(\sigma_{\min}^{-1}(\tilde{A}_S) > \beta) = \mathbb{P}(\sigma_{\min}(\tilde{A}_S) < 1/\beta).$$

Using (A.9), we set $\frac{1}{\beta} = 1 - \sqrt{\frac{s}{m}} - r$. Then we have

$$\mathbb{P}(\sigma_{\min}^{-1}(\tilde{A}_S) > \beta) \leq \exp\left(-\frac{m}{2}\left(1 - \sqrt{\frac{s}{m}} - \frac{1}{\beta}\right)^2\right).$$

The previous expression is bounded by ε provided

$$\sqrt{m} \geq \frac{\beta}{\beta - 1}\left(\sqrt{s} + \sqrt{2\ln(1/\varepsilon)}\right). \tag{2.40}$$

Hence we conclude that if (2.38), (2.39) and (2.40) hold, the recovery error satisfies (2.31) with probability at least 3ε. Now we assign values to the variables. Set $r = s/8$, $\delta_1 = \delta_2 = 0.75$, and $\beta = 3.5$. Then all three conditions are implied by

$$m \geq s\left[\frac{\sqrt{2\ln(4N/\varepsilon)}}{1-\theta} + \sqrt{2\ln(2/\varepsilon)/s} + \sqrt{2}\right]^2.$$

Plugging these numbers into (2.31) and replacing ε by $\varepsilon/3$ completes the proof. □

2.4. Discussion

2.4.1. Comparison to related theoretical results

Asymptotic results and phase transitions: Donoho and Tanner [**39**, **40**] obtain sparse recovery results via ℓ_1-minimization with Gaussian matrices via methods from random polytopes.

They operate essentially in an asymptotic regime (although some of their results apply also for finite values of N, m, s). Donoho and Tanner consider the case that

$$m/N \to \delta,\ s/m \to \rho,\ \log(N)/m \to 0,\ N \to \infty,$$

where $\rho, \delta \in [0, 1]$ are some fixed values. Recovery conditions are then expressed in terms of ρ and δ in this asymptotic regime. In particular, they obtain transition curves that separate the regions in the $[0, 1]^2$ plane of ρ, δ in which the recovery succeeds and the recovery fails with probability 1 when $N \to \infty$. They distinguish nonuniform and uniform recovery with the terminology "weak phase transition" and "strong phase transition" respectively. They obtain a weak transition curve $\delta_W(\delta)$ and a strong transition curve $\delta_S(\delta)$ such that $\delta < \delta_W(\delta)$ implies recovery with high probability and $\delta > \delta_W(\delta)$ mean failure with high probability (as $N \to \infty$), and similarly for $\delta_S(\delta)$. Explicit formulas for these curves are not available, but implicit formulas are available that can be evaluated numerically in the limiting case $\delta \to 0$. As a consequence, translated back into the quantities N, m, s this gives roughly the *uniform recovery* condition

$$m > 2es \ln(N/(\sqrt{\pi}m)),$$

and the *nonuniform recovery* condition

$$m > 2s \ln(N/m)$$

in an asymptotic regime. We note that the nonuniform condition is very close to our condition (2.4). Donoho and Tanner provide precise statements about when ℓ_1-minimization fails in both cases which allows to deduce that the constants $2e$ and 2 are indeed optimal. Even though the results in the papers of Donoho and Tanner are derived for real vectors, combined with the fact that the real null space property are equivalent to the complex one [**54**], they apply also to the complex vectors. Their methods, however, do not cover stability issues and it is not clear whether their results can be extended to Bernoulli or subgaussian random matrices. In addition to Donoho and Tanner's work, Dossal et. al. [**45**] derive a recovery condition for Gaussian matrices of the form $m \geq cs \ln(N)$, where c approaches 2 in an asymptotic regime where they also obtain stability results for noisy measurements.

Nonasymptotic results: There have been several papers dealing with nonuniform recovery. Most of these papers consider only the Gaussian case. In [**22**], Candès and Plan give a rather general framework for nonuniform recovery which applies to measurement matrices with independent rows having bounded entries. In fact, they prove a recovery condition for such random matrices of the form $m \geq Cs \ln(N)$ for some constant C and also cover the stability of the reconstruction. However, they do not obtain explicit and good constants. The work [**23**] of Candès and Recht derives closely related results to ours. For Gaussian measurement matrices, they show that, for any $\beta > 1$, an s-sparse vector can be recovered with probability at least $1 - 2N^{-f(\beta,s)}$ if

$$m \geq 2\beta s \ln N + s,$$

where

$$f(\beta, s) = \left[\sqrt{\frac{\beta}{2s} + \beta - 1} - \sqrt{\frac{\beta}{2s}}\right]^2.$$

Their method uses the duality based recovery Theorem 2.5 due to Fuchs [**55**] like in our approach, but then proceeds differently. They are also able to derive a similar recovery condition for subgaussian matrices but only state it for the special case of Bernoulli matrices. Furthermore, they also work out recovery results in the context of block sparsity and low-rank recovery. However, they do not cover stability of the reconstruction.

Finally, Chandrasekaran et al. [**28**] use convex geometry in order to obtain nonuniform recovery results. They develop a rather general framework that applies also to low-rank recovery and further setups. However, they can only treat Gaussian measurements. They approach the recovery problem via Gaussian widths of certain convex sets. In particular, they estimate the number m of Gaussian measurements needed in order to recover an s sparse vector by $m \geq 2s(\ln(N/s-1)+1)$ which is essentially the optimal result, see also [**54**, Chapter 9]. Their method heavily relies on properties of Gaussian random vectors and therefore, it does not seem possible to extend it to more general subgaussian random matrices such as Bernoulli matrices.

CHAPTER 3

Sparse Recovery with Fusion Frames

3.1. Introduction

In this chapter, we consider the recovery of signals that have a sparse fusion frame representation. Signals of interest are collections of vectors from fusion frame subspaces which can be considered as 'block' signals. In other words, say we have N subspaces, then we have a collection of N vectors from those subspaces which is the (block) signal we wish to recover. We are allowed to observe linear measurements of the signal and we aim to reconstruct the original signal from those measurements. In order to do so, we use ideas from CS. We assume a block sparsity model on the signals to be recovered where a few of the vectors in the collection are assumed to be nonzero. For instance, we are not interested whether each vector is sparse or not within the subspace it belongs to. For the reconstruction, a mixed ℓ_1/ℓ_2-minimization is invoked in order to capture the structure that comes with the sparsity model.

This problem was studied before in [**9**] where the authors proved that it is sufficient for sparse recovery to take $m \gtrsim s\ln(N)$ random measurements. Here s is the sparsity level and N is the number of subspaces. Our setup reduces to the problem of recovering a block sparse vector if the subspaces are not assumed to satisfy the fusion frame property. It is worth emphasizing that our model assumes in addition that the signals of interest lie in particular subspaces which is not assumed in block sparsity problems. In this chapter, our goal is to improve the results in [**9**] when the subspaces have a certain incoherence property, i.e., they have nontrivial mutual angles, and we assume the knowledge of the subspaces.

3.1.1. Notation

We denote the Euclidean norm of vectors by $\|\cdot\|_2$, the spectral norm of matrices by $\|\cdot\|$ and the Frobenius norm by $\|\cdot\|_F$. Boldface notation refers to block vectors and matrices throughout this chapter. Let $[N] = \{1, 2, \dots, N\}$. For a block matrix $\mathbf{A} = (a_{ij}B_{ij})_{i\in[m],j\in[N]} \in \mathbb{R}^{m\cdot d\times N\cdot d}$ with blocks $B_{ij} \in \mathbb{R}^{d\times d}$, we denote the ℓ-th block column by $\mathbf{A}_\ell = (a_{i\ell}B_{i\ell})_{i\in[m]} \in \mathbb{R}^{m\cdot d\times d}$ and the column submatrix restricted to $S \subset [N]$ by $\mathbf{A}_S = (a_{ij}B_{ij})_{i\in[m],j\in S}$. Similarly for a block vector $\mathbf{x} = (x_i)_{i=1}^N \in \mathbb{R}^{N\cdot d}$ with $x_i \in \mathbb{R}^d$, we denote the vector $\mathbf{x}_S = (x_i)_{i\in S}$ restricted to S. The complement $[N] \setminus S$ is denoted $\overline{S}$. The $\ell_\infty \to \ell_\infty$ and $\ell_2 \to \ell_\infty$ norms of a matrix $A \in \mathbb{R}^{m\times n}$ are given as

$$\|A\|_\infty = \max_{i\in[m]} \sum_{j=1}^{n} A_{ij}, \quad \text{and} \quad \|A\|_{2,\infty} = \max_{i\in[m]} \left(\sum_{j=1}^{n} A_{ij}^2\right)^{1/2},$$

respectively. Furthermore, we define the $\ell_{2,\infty}$-norm of a block vector $\mathbf{x} = (x_i)_{i=1}^N$ with $x_i \in \mathbb{R}^d$ as

$$\|\mathbf{x}\|_{2,\infty} = \max_{i\in[N]} \|x_i\|_2.$$

Given a block vector $\mathbf{z}$ with N blocks from $\mathbb{R}^d$, we define $\operatorname{sgn}(\mathbf{z}) \in \mathbb{R}^{dN}$ analogously to the scalar case as

$$\operatorname{sgn}(\mathbf{z})_i = \begin{cases} \frac{z_i}{\|z_i\|_2} & : \quad \|z_i\|_2 \neq 0, \\ 0 & : \quad \|z_i\|_2 = 0. \end{cases}$$

3.1.2. Fusion frames and applications

A *fusion frame* for $\mathbb{R}^d$ is a collection of N subspaces $W_j \subset \mathbb{R}^d$ and associated weights v_j that satisfy

$$A\|x\|_2^2 \leq \sum_{j=1}^{N} v_j^2 \|P_j x\|_2^2 \leq B\|x\|_2^2 \tag{3.1}$$

for all $x \in \mathbb{R}^d$ and for some universal fusion frame bounds $0 < A \leq B < \infty$, where $P_j \in \mathbb{R}^{d\times d}$ denotes the orthogonal projection onto the subspace W_j. For simplicity we assume that the dimensions of the W_j are equal, that is $\dim(W_j) = k$. For a fusion frame $(W_j)_{j=1}^N$, let us define the space

$$\mathcal{H} = \{(x_j)_{j=1}^N : x_j \in W_j, \ \forall j \in [N]\} \subset \mathbb{R}^{d\times N},$$

where we denote $[N] = \{1, \ldots, N\}$. We define the *mixed* $\ell_{2,1}$*-norm* of a vector $\mathbf{x} \equiv (x_j)_{j=1}^N \in \mathcal{H}$ as

$$\|\mathbf{x}\|_{2,1} = \sum_{j=1}^{N} \|x_j\|_2.$$

Furthermore, the 'ℓ_0- block norm' of $\mathbf{x} \in \mathcal{H}$ is defined as

$$\|\mathbf{x}\|_0 = \operatorname{card}\{j \in [N] : x_j \neq 0\}.$$

We say that a vector $\mathbf{x}$ is s-sparse, if $\|\mathbf{x}\|_0 \leq s$. In this chapter we consider the recovery of sparse vectors from the set $\mathcal{H}$ which collects all vectors from fusion frame subspaces. However our results do not assume necessarily that the subspaces satisfy the fusion frame property (3.1) but rather assume they satisfy an incoherence property as explained in Section 3.1.5. We note that the fusion frame setup can be considered as a refinement of the 'block' sparsity setup [**51**], in which we assume that $W_j = \mathbb{R}^d$, in other words the vectors x_i are not assumed to lie in certain subspaces W_i of $\mathbb{R}^d$. This setup is explained in Section 3.1.4 in more detail.

- *Applications*

In this section we mention some applications of fusion frames.

- *Distributed Processing.* In distributed sensing, typically a large number of sensors are deployed in an area to measure a physical quantity such a s temperature, sound etc. Due to the physical or economical restraints it is easier to cluster them locally and process information within each of their regions. Such a large sensor network can be viewed as a redundant collection of subnetworks forming a set of subspaces. Each information produced by those local systems are then gathered and submitted to a central joint processor and by the help of fusion frame theory, the original vector would be reconstructed [**27**, **76**].
- *Parallel Processing.* Consider a classical frame system which models the application in hand. If this frame system is too large, it can be computationally too hard to handle. Then it is advantageous to split this system into many smaller frame systems where we also have redundancy within those systems. The information can be processed within the subsystems

parallelly in a way which is robust against errors due to the failure of a subsystem. Fusion frames provide a natural setting for such a splitting and reuniting subsystems [**5**, **88**].

- *Packet Encoding.* When some data is transferred over a communication network like the internet, it is often first encoded into a number of packets. After the transfer, these packets are decoded in the receiver in order to acquire the original data. By introducing redundancy in the encoding process, the communication scheme can be made resilient against corruption or even loss of some packets. We can also think of each packet as a vector from fusion frame subspaces, and encoding as a linear measurement of this vector. Therefore fusion frames provide a means to achieve such redundant subspace representations [**6**].
- *Music segmentation.* Consider a music piece performed by an orchestra where the sound we hear is a combination of many instruments playing at the same time. A note contributed by each instrument is not characterized by a single frequency, but by the subspace spanned by the fundamental frequency of the instrument and its harmonics [**31**]. Then the collection of these instruments can be thought as a fusion frame. Furthermore, depending on the instrument, certain harmonics might or might not be present in the subspace and also the subspaces occupied by distinct instruments may overlap with each other. This gives the redundancy in our fusion frame.

3.1.3. Problem formulation

Our measurement model is as follows. Let $\mathbf{x}^0 = (x_j^0)_{j=1}^N \in \mathcal{H}$ be s-sparse and we observe m linear combinations of those vectors

$$\mathbf{y} = (y_i)_{i=1}^m = \left(\sum_{j=1}^N a_{ij} x_j^0 \right)_{i=1}^m, \quad y_i \in \mathbb{R}^d,$$

for some scalars (a_{ij}). Note that, for all $i \in [m]$, each vector y_i is a measurement of the vectors $x_1^0, \dots, x_N^0$, thus the measurement vector $\mathbf{y}$ is itself a block vector like $\mathbf{x}^0$. Let us introduce the block matrices $\mathbf{A_I} = (a_{ij}\mathrm{Id})_{i\in[m],j\in[N]}$ and $\mathbf{A_P} = (a_{ij}P_j)_{i\in[m],j\in[N]}$ that consist of the blocks $a_{ij}\mathrm{Id} \in \mathbb{R}^{d\times d}$ and $a_{ij}P_j \in \mathbb{R}^{d\times d}$ respectively. Here Id is the identity matrix of size $d \times d$. Then we can formulate this measurement scheme as

$$\mathbf{y} = \mathbf{A_I}\mathbf{x}^0 = \mathbf{A_P}\mathbf{x}^0, \quad \text{for } \mathbf{x}^0 \in \mathcal{H}.$$

We replaced $\mathbf{A_I}$ by $\mathbf{A_P}$ since the relation $P_j x_j = x_j$ holds for all $j \in [N]$. We wish to recover $\mathbf{x}^0$ from those measurements. This problem can be formulated as the following optimization program

$$(L0) \quad \hat{\mathbf{x}} = \mathrm{argmin}_{\mathbf{x}\in\mathcal{H}} \|\mathbf{x}\|_0 \quad s.t. \quad \mathbf{A_P}\mathbf{x} = \mathbf{y}.$$

This optimization program is NP-hard. Instead we propose the following convex program

$$(L1) \quad \hat{\mathbf{x}} = \mathrm{argmin}_{\mathbf{x}\in\mathcal{H}} \|\mathbf{x}\|_{2,1} \quad s.t. \quad \mathbf{A_P}\mathbf{x} = \mathbf{y}.$$

We shall show that block sparse signals can be accurately recovered by solving $(L1)$. In the rest of the chapter, $\mathbf{P}$ denotes the $N \times N$ block diagonal matrix where the diagonals are projection matrices P_j, $j \in [N]$. $\mathbf{P}_S$ denotes the $s \times s$ block diagonal matrix with entries P_i, $i \in S$ with $|S| = s$.

3.1.4. Relation to other sparsity models

A special case of the sparse recovery problem above appears when all subspaces coincide with the ambient space $W_j = \mathbb{R}^d$ for all j. Then the problem reduces to the well studied *joint sparsity setup* [**52**] in which all the vectors have the same sparsity structure. In order to see this, say we have N vectors $\{x_1, \ldots, x_N\}$ in $\mathbb{R}^d$ and write them as rows of a matrix $X \in \mathbb{R}^{N\times d}$. Assuming only s of the rows are non-zero, the d vectors consisting of the columns of the matrix X have the same common support set, i.e., are *jointly* sparse. To be more precise, if $(X_i)_{i=1}^d$ denote the columns of X, $\operatorname{supp}(X_i) = \operatorname{supp}(X_j)$ for all $i \neq j$.

Furthermore, our problem is itself a special case of *block sparsity setup* [**51**], with significant additional structure that allows us to enhance existing results. Without the subspace structure, we would have N vectors $\{x_1, \ldots, x_N\}$ in $\mathbb{R}^d$ where only s of them are non-zero, i.e., $\operatorname{card}\{i \in [N] : x_i \neq 0\} = s$. The reason this is called block sparsity is because we count the vectors which are non-zero as a *block* rather than checking if its entries are sparse.

Another related concept is *group sparsity* [**91**] where each entry of the vector is assigned to a predefined group and sparsity counts the groups that are active in the support set of the vector. In mathematical notation, consider a vector x of length p and assume that its coefficients are grouped into sets $\{G_i\}_{i=1}^N$, such that for all $i \in [N]$, $G_i \subset [p]$. The vector x is group s-sparse if the non-zero coefficients lie in s of the groups $\{G_i\}_{i=1}^N$. Formally,

$$\operatorname{card}\{i \in [N] : \operatorname{supp}(x) \cap G_i \neq 0\} = s.$$

This is similar to the block sparsity with $p = Nd$ except that the groups may be allowed to overlap, i.e., $G_i \cap G_j \neq \emptyset$. We also note that, for sparse recovery, a natural proxy for group sparsity becomes the norm $\sum_i \|x_{G_i}\|_2$.

Finally in the case $d = 1$, the projections equal 1, and hence the problem reduces to the *standard CS recovery problem* $Ax = y$ with $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^m$.

3.1.5. Incoherent subspaces

In this section we adapt the notion of mutual incoherence of vectors to more general situation of fusion frames involving the angles between the fusion frame subspaces. Such a notion of incoherence will be useful for incorporating the a priori knowledge about the signal structure which in turn will provide additional information that can be exploited in the measurement process. We begin by defining an incoherence matrix Λ associated with a fusion frame

$$\Lambda = (\alpha_{ij})_{i,j\in[N]}, \tag{3.2}$$

where $\alpha_{ij} = \|P_iP_j\|$ for $i \neq j \in [N]$ and $\alpha_{ii} = 0$. Note that $\|P_iP_j\|$ equals the largest absolute value of the cosines of the principle angles between W_i and W_j. Then, for a given support set $S \subset [N]$, we denote Λ_S as the submatrix of Λ with columns restricted to S, and Λ^S as the submatrix with columns and rows restricted to S. We will use the following norms

$$\|\Lambda_S\|_{2,\infty} = \max_{i\in[N]} \left(\sum_{j\in S, j\neq i} \|P_iP_j\|^2 \right)^{1/2} \quad \text{and} \quad \|\Lambda^S\|_{2,\infty} = \max_{i\in S} \left(\sum_{j\in S, j\neq i} \|P_iP_j\|^2 \right)^{1/2},$$

and

$$\|\Lambda_S\|_\infty = \max_{i\in[N]} \sum_{j\in S, j\neq i} \|P_i P_j\| \qquad \text{and} \qquad \|\Lambda^S\|_\infty = \max_{i\in S} \sum_{j\in S, j\neq i} \|P_i P_j\|.$$

Moreover, we define the parameter

$$\lambda = \max_{i\neq j} \|P_i P_j\|, \quad i,j \in [N].$$

Clearly λ equals the largest entry of Λ. In addition it holds that

$$\|\Lambda_S\|_{2,\infty} \leq \|\Lambda_S\|_\infty \leq \lambda s \tag{3.3}$$

for any S with $|S| = s$. If the subspaces are all orthogonal to each other, i.e., $\lambda = 0$ and $\Lambda = 0$, then only one measurement $y = \sum_i a_i x_i$ suffices to recover $\mathbf{x}^0$. In fact, since $x_i \perp x_j$, we have

$$x_i = \frac{1}{a_i} P_i y.$$

This observation suggests that fewer measurements are necessary when λ becomes smaller. In this chapter our goal is to provide a solid theoretical understanding of this intuitive behavior.

3.2. Main results

3.2.1. Nonuniform recovery results

This section studies nonuniform recovery from fusion frame measurements. Our main results state that a fixed sparse signal can be recovered with high probability using a random draw of a Bernoulli or Gaussian random matrix $A \in \mathbb{R}^{m\times N}$. Also we assume for simplicity that the subspaces $(W_j)_{j=1}^N$ of the fusion frame in $\mathbb{R}^d$ has $\dim(W_j) = k$ for all j and use this assumption in our main results below. We remark that in the general case where the $\dim(W_j)$ are different than each other, our results follow similarly with slight modifications.

- *The Bernoulli case*

Our first nonuniform recovery result concerns Bernoulli matrices and involves the incoherence matrix Λ.

Theorem 3.1. *Let $\mathbf{x} \in \mathcal{H}$ be s-sparse and $S = \operatorname{supp}(\mathbf{x})$. Let $A \in \mathbb{R}^{m\times N}$ be a Bernoulli matrix and a fusion frame $(W_j)_{j=1}^N$ be given with the incoherence matrix Λ. Assume that*

$$m \geq C(1 + \|\Lambda_S\|_\infty) \ln(N) \ln(sk) \ln(\varepsilon^{-1}), \tag{3.4}$$

where $C > 0$ is a universal constant. Then with probability at least $1-\varepsilon$, $(L1)$ recovers $\mathbf{x}$ from $\mathbf{y} = \mathbf{A_P x}$.

If the subspaces are not equi-dimensional, one can replace the quantity sk in Condition (3.4) by $\sum_{i\in S} k_j$, where $\dim(W_j) = k_j$. We also give an alternative result for nonuniform recovery with Bernoulli matrices. This result involves the parameter λ instead of the matrix Λ.

Theorem 3.2. *Let $\mathbf{x} \in \mathcal{H}$ be s-sparse. Let $A \in \mathbb{R}^{m\times N}$ be a Bernoulli matrix and $(W_j)_{j=1}^N$ be given with parameter $\lambda \in [0,1]$. Assume that*

$$m \geq C(1 + \lambda s) \ln(Nsk) \ln(\varepsilon^{-1}), \tag{3.5}$$

where $C > 0$ is a universal constant. Then with probability at least $1 - \varepsilon$, $(L1)$ recovers $\mathbf{x}$ from $\mathbf{y} = \mathbf{A_P x}$.

The proof of this theorem is similar to the one of Theorem 3.1 with slight modifications.

Remark. Theorem 3.2 improves Theorem 3.1 in terms of the log-factors as $\log(sk)$ does not appear in Condition (3.5). Since $\|\Lambda_S\|_\infty \leq \lambda s$, Condition (3.4) is slightly better than (3.5) in terms of the incoherence parameter, at least if there is a true gap between $\|\Lambda\|_\infty$ and λs, which happens if the quantities $\|P_i P_j\|$ are not all close to their maximal value. The equality $\|\Lambda_S\|_\infty = \lambda s$ is achieved when the subspaces are equi-angular. In the case that they are not equi-angular, even if only two subspaces align, then $\lambda = 1$. In this case, (3.5) suggests that we should not expect any improvement in the recovery performance with respect to the standard block sparse case. However, intuitively the orientation of the other subspaces might still be effective in the recovery process. A more average measure of the incoherence of the subspaces is captured by $\|\Lambda\|_\infty$ in (3.4), so that Theorem 3.1 improves for general orientations of the subspaces up to a slight drawback in the log-factors. The numerical experiments which we present in Section 3.2.6 also support this result.

- *The Gaussian case*

We state a similar result for Gaussian random matrices. These matrices have independent entries from a standard normal distribution, i.e., $A_{ij} = g_{ij} \sim \mathcal{N}(0, 1)$.

Theorem 3.3. *Let $\mathbf{x} \in \mathcal{H}$ be s-sparse. Let $A \in \mathbb{R}^{m \times N}$ be a Gaussian matrix and $(W_j)_{j=1}^N$ be given with parameter $\lambda \in [0, 1]$ and $\dim(W_j) = k_j$ for all j. Assume that*

$$m \geq \tilde{C}(1 + \lambda s) \ln^2 \left(6N \sum_{j=1}^{N} k_j \right) \ln^2(\varepsilon^{-1}), \tag{3.6}$$

where $\tilde{C} > 0$ is a universal constant. Then with probability at least $1 - \varepsilon$, $(L1)$ recovers $\mathbf{x}$ from $\mathbf{y} = \mathbf{A_P x}$.

Presently our nonuniform result for Gaussian matrices involves the parameter λ. However it does not seem possible to derive a similar result to Theorem 3.1 which involves Λ with our proof methods used for Theorem 3.3.

- *Stable and robust recovery*

In this section we show that the nonuniform sparse recovery for fusion frames is stable and robust under presence of noise. In other words we allow our signal $\mathbf{x}$ to be approximately sparse (compressible) and the measurements $\mathbf{y}$ to be noisy. Our measurement model then becomes

$$\mathbf{y} = \mathbf{A_P x} + \mathbf{e} \quad \text{with} \quad \|\mathbf{e}\|_2 \leq \eta\sqrt{m} \tag{3.7}$$

for some $\eta \geq 0$. For the reconstruction we employ

$$(L1)^\eta \quad \hat{\mathbf{x}} = \operatorname{argmin}_{x \in \mathcal{H}} \|\mathbf{x}\|_{2,1} \quad s.t. \quad \|\mathbf{A_P x} - \mathbf{y}\|_2 \leq \eta\sqrt{m}.$$

The condition $\|\mathbf{e}\|_2 \leq \eta\sqrt{m}$ in (3.7) is natural for a vector $\mathbf{e} = (e_j)_{j=1}^m$. For instance, it is implied by the bound $\|e_j\|_2 \leq \eta$ for all $j \in [m]$. We define the best s-term approximation of a vector $\mathbf{x}$ as

$$\sigma_s(\mathbf{x})_1 := \inf_{\|\mathbf{z}\|_0 \leq s} \|\mathbf{x} - \mathbf{z}\|_{2,1}.$$

Compressible vectors are the ones with small $\sigma_s(\mathbf{x})_1$. For a vector $\mathbf{x} = (x_i)_{i=1}^N$, let S be the index set of s largest ℓ_2-normed entries $\|x_i\|_2$. Then it is evident that $\sigma_s(\mathbf{x})_1 = \|\mathbf{x}_{\overline{S}}\|_{2,1}$. Next we state a stable version of Theorem 3.1.

Theorem 3.4. *Let $\mathbf{x} \in \mathcal{H}$ and $S \subset [N]$ with $|S| = s$ be chosen as explained above. Let $A \in \mathbb{R}^{m\times N}$ be a Bernoulli matrix and $(W_j)_{j=1}^N$ be given with parameter $\lambda \in [0,1]$ and $\dim(W_j) = k$ for all j. Assume the measurement model in (3.7) and let $\hat{\mathbf{x}}$ be a solution to $(L1)^\eta$. Provided*

$$m \geq C(1 + \|\Lambda_S\|_\infty) \ln(N) \ln(sk) \ln(\varepsilon^{-1}), \tag{3.8}$$

then with probability at least $1 - \varepsilon$,

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_1 \sigma_s(\mathbf{x})_1 + C_2 \sqrt{s}\eta. \tag{3.9}$$

The constants $C, C_1, C_2 > 0$ are universal.

We prove this theorem in Section 3.3.7. We note that the stability estimate (3.9) falls short by a factor of $\sqrt{s}$ of the one achievable with the RIP, which is shown later in Section 3.2.7. As in the noiseless case, we can improve Theorem 3.4 with respect to one of the log-factors while using λ instead of the incoherence matrix Λ.

Theorem 3.5. *Let $\mathbf{x} \in \mathcal{H}$. Let $A \in \mathbb{R}^{m\times N}$ be a Bernoulli matrix and $(W_j)_{j=1}^N$ be given with parameter $\lambda \in [0,1]$ and $\dim(W_j) = k$ for all j. Assume the measurement model in (3.7) and let $\hat{\mathbf{x}}$ be a solution to $(L1)^\eta$. If*

$$m \geq C(1 + \lambda s) \ln(Nsk) \ln(\varepsilon^{-1}),$$

then with probability at least $1 - \varepsilon$,

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_1 \sigma_s(\mathbf{x})_1 + C_2 \sqrt{s}\eta.$$

The constants $C, C_1, C_2 > 0$ are universal.

The proof of this result is the same as the proof of Theorem 3.4, therefore we omit it. In particular one follows similar arguments as in Section 3.3.7 combined with the proof of Theorem 3.2. This is also the case for the following result which gives stability and robustness for the Gaussian case.

Theorem 3.6. *Let $\mathbf{x} \in \mathcal{H}$. Let $A \in \mathbb{R}^{m\times N}$ be a Gaussian matrix and $(W_j)_{j=1}^N$ be given with parameter $\lambda \in [0,1]$ and $\dim(W_j) = k$ for all j. Assume the measurement model in (3.7) and let $\hat{\mathbf{x}}$ be a solution to $(L1)^\eta$. If*

$$m \geq \tilde{C}(1 + \lambda s) \ln^2(6Nk) \ln^2(\varepsilon^{-1}),$$

then with probability at least $1 - \varepsilon$,

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_1 \sigma_s(\mathbf{x})_1 + C_2 \sqrt{s}\eta.$$

The constants $\tilde{C}, C_1, C_2 > 0$ are universal.

3.2.2. Uniform recovery results

So far, we have studied the nonuniform recovery for fusion frames, in which we consider a fixed sparse or compressible vector to be recovered from its random measurements with high probability.

We have shown that as the incoherence of the subspaces, where the signals lie in, increases, less measurements are required for the recovery. In this section, we search for uniform recovery conditions for fusion frames with random matrices. Those type of results assert that once the random matrix is chosen, then with high probability all sparse signals can be recovered using the same matrix.

We will present two uniform results. First of these results uses an extension of RIP to the fusion frame case and the second uses a version of the null space property (NSP) for fusion frames. In Section 1.2, we noted that these are two common ways to obtain uniform results.

- *Recovery with subgaussian matrices*

We defined a subgaussian random variable in Section 2.1.3 satisfying Condition (2.5). For our purposes in this section, we call a random variable ξ as c-subgaussian if $\mathbb{E}[\exp(t\xi)] \leq e^{ct^2}$. An $m \times N$ subgaussian matrix A takes the form

$$A = \xi_{jk}, \ j \in [m], k \in [N],$$

where the ξ_{jk} are independent, mean-zero, variance one, c-subgaussian random variables. Note that standard Gaussian and Bernoulli random variables are subgaussian with $c = 1/2$. As before we let the fusion frame $(W_j)_{j=1}^N$ be given with the incoherence parameter λ and, for simplicity, assume that the dimensions of the frame subspaces W_j are the same, say $k_j = k$ for all j.

Theorem 3.7. *Let $A \in \mathbb{R}^{m\times N}$ be a subgaussian matrix and $(W_j)_{j=1}^N$ be given with parameter $\lambda \in [0,1]$. Assume that*

$$m \geq C(\lambda s + \sqrt{s}) \ln(Nd) \ln^2(sk) \left(\ln(N) + k \ln\left(s\sqrt{k}\right)\right), \tag{3.10}$$

$$m \geq C' \ln(2\varepsilon^{-1}) \tag{3.11}$$

where $C, C' > 0$ depend only on the subgaussian parameter c. Then with probability at least $1-\varepsilon$, $(L1)$ recovers all s-sparse $\mathbf{x}$ from $\mathbf{y} = \mathbf{A_P x}$. Moreover, with probability at least $1-\varepsilon$, every vector $\mathbf{x} \in \mathcal{H}$ is approximated by a minimizer $\hat{\mathbf{x}}$ of $(L1)^\eta$ with $\mathbf{y} = \mathbf{A_P x} + \mathbf{e}$ and $\|\mathbf{e}\|_2 \leq \eta$ in the sense that

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_1 \frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}} + C_2\eta, \tag{3.12}$$

and

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_{2,1} \leq C_1 \sigma_s(\mathbf{x})_1 + C_2\sqrt{s}\eta,$$

where the constants $C_1, C_2 > 0$ are universal.

Remark. Presently, the estimate (3.10) on the number of required samples behaves slightly worse than the nonuniform one (3.5) for small λ and suffers from additional log-terms. We believe the linear factor k in (3.10) can be removed. As it is now, for the range $k \lesssim \ln(N)/\ln(s\sqrt{k})$, this factor can be ignored. The logarithmic factor $\ln(s\sqrt{k})$ in k is not worse than the other logarithmic terms, unless k is exponential in N. We can discard this case since it is not practically plausible. We also note that in the special case of classical frames $k = 1$, so that the factor of k in the number of measurements vanishes.

Even though the uniform result is slightly weaker in terms of the order of sparsity and log-factors it provides stronger stability results than the nonuniform case. In the latter, the reconstruction error (3.9) obeys

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_1 \sigma_s(\mathbf{x})_1 + C_2 \sqrt{s}\eta$$

which is weaker than (3.12) by a factor of $\sqrt{s}$. As we will see in Section 3.2.7 that this is a consequence of the RIP.

- *Recovery with Gaussian matrices*

Our second uniform result is achieved by establishing null space properties for the fusion frames. We introduce extensions of null space properties for fusion frames and show that they guarantee uniform recovery in Section 3.2.8. As in the previous section, two type of results concerning exact and stable recovery are presented. However, the results in this section are restricted to only Gaussian matrices, since the main ingredient of the proof, namely Gordon's lemma [**62**] heavily relies on the rotational invariance of the Gaussian vectors.

Theorem 3.8. *(Stable and robust recovery) Let $A \in \mathbb{R}^{m\times N}$ be a Gaussian matrix and $(W_j)_{j=1}^N$ in $\mathbb{R}^d$ be given with parameter $\lambda \in [0,1]$. Let $0 < \rho < 1$ and $\tau > 0$. Assume that*

$$\frac{m}{\sqrt{m+1}} \geq (1+\rho^{-1})\sqrt{c(1+\lambda s)\ d\ \ln(N)} + \sqrt{d}\ \left[\tilde{c} + 1/\tau + \sqrt{2\ln(\varepsilon^{-1})}\right], \tag{3.13}$$

where $c, \tilde{c} > 0$ are universal constants. Then with probability at least $1-\varepsilon$, every vector $\mathbf{x} \in \mathcal{H}$ is approximated by a minimizer $\hat{\mathbf{x}}$ of $(L1)^\eta$ with $\mathbf{y} = \mathbf{A_P x} + \mathbf{e}$ and $\|\mathbf{e}\|_2 \leq \eta$ in the sense that

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq \frac{C}{\sqrt{s}}\sigma_s(\mathbf{x})_1 + D\eta, \tag{3.14}$$

for some constants $C, D > 0$ depending on only ρ and τ.

Remark. Roughly speaking, that is, ignoring terms of lower order, all nearly s-sparse vectors are stably recovered with high probability from

$$m \geq c(1+\lambda s) d \ln(N) \tag{3.15}$$

noisy Gaussian measurements. We note that the dependence in d in (3.15) is suboptimal when compared to other recovery results presented earlier, which is probably an artifact of the proof. Theorem 3.8 also implies the recovery of exact sparse signals in the presence of no noise. The explicit constants in (3.14) are $C = \frac{(2\rho+1)(1+\rho)}{1-\rho}$ and $D = \frac{4+2\rho}{1-\rho}\tau$. In the exact sparse and noiseless case, $\sigma_s(\mathbf{x})_1 = \eta = 0$, this yields perfect reconstruction since $\rho < 1$. In order to make Condition (3.13) as tight as possible, one can choose ρ very close to 1 and τ a very large number. We have the following corollary.

Corollary 3.9. *(Exact recovery) Let $A \in \mathbb{R}^{m\times N}$ be a Gaussian matrix and $(W_j)_{j=1}^N$ in $\mathbb{R}^d$ be given with parameter $\lambda \in [0,1]$. Assume that*

$$m \geq c(1+\lambda s) d \left(\ln(N) + \ln(\varepsilon^{-1})\right),$$

for an absolute constant $c > 0$. Then with probability at least $1-\varepsilon$, any s-sparse $\mathbf{x} \in \mathcal{H}$ is a solution of $(L1)$ with $\mathbf{y} = \mathbf{A_P x}$.

3.2.3. Necessary conditions for sparse recovery

In this section, we investigate the minimal number of measurements required for sparse recovery with fusion frames via $\ell_{2,1}$-minimization. In the classical CS setup, there is a close relation between Gelfand widths and the worst case reconstruction error of CS methods [**54**, Section 10]. Particularly, establishing lower bounds for the Gelfand widths of ℓ_1-balls leads to the fact that the minimal number of measurements required for stable recovery via *any* method is $m \geq Cs\ln(eN/s)$. If we consider only $(L1)$ as the reconstruction method and restrict ourselves to the exactly sparse case, one can estimate the necessary number of measurements with a key result from [**53**, **54**]. We now provide an extension of such a result for the case of $\ell_{2,1}$-minimization in the fusion frame setup.

Theorem 3.10. *Given a matrix $A \in \mathbb{R}^{m\times N}$ and a fusion frame $(W_j)_{j=1}^N$ in $\mathbb{R}^d$ with $\dim(W_j) = k$, if every $4s$-sparse vector $\mathbf{x} \in \mathcal{H}$ is a minimizer of $\|\mathbf{z}\|_{2,1}$ subject to $\mathbf{A_P z} = \mathbf{A_P x}$, then*

$$m \geq c_1 \frac{s}{d} \ln\left(\frac{N}{c_2 s}\right) + c_3 \frac{ks}{d} \tag{3.16}$$

where $c_1 \approx 0.46$, $c_2 = 32$ and $c_3 \approx 0.18$.

At first sight, it is not clear how this result can be compared to the sufficient conditions provided in Theorem 3.2 and Theorem 3.7 earlier since (3.16) does not involve the incoherence parameter λ. However, in Section 3.2.4 we give theoretical lower bounds on λ which, in turn, allows us in Section 3.2.5 to compare the two type of results and show that they do not fall too far from each other.

3.2.4. Packing of subspaces

This section proposes lower bounds on the incoherence parameter λ appearing in the uniform recovery conditions, thus it will allow us to compare our sufficient conditions given in Sections 3.2.1 and 3.2.2 with the necessary condition derived in the previous section. We are interested in the question that given a number N of subspaces each of dimension k in the space $\mathbb{R}^d$, what is the smallest value that

$$\lambda := \max_{i\neq j} \|P_i P_j\| \text{ for } i, j \in [N]$$

can attain? For instance, if $k = 1$ and $N \leq d$, then we can choose the subspaces (i.e., lines through the origin) all orthogonal to each other, which yields $\lambda = 0$. When $N > d/k$, it is not anymore possible to choose the subspaces orthogonally, so that $\lambda > 0$. Next, we introduce two different metrics on Grassmannian subspaces and present packing bounds with respect to these metrics.

Principal angles: If W_i and W_j are two different subspaces each of dimension k, then we have k principal angles between them. We label those angles $\theta_{ij}^{(1)} \leq \theta_{ij}^{(2)} \leq \ldots \leq \theta_{ij}^{(k)}$ in nondecreasing order. The cosines of the principal angles coincide with the singular values of $P_i P_j$, so that

$$\lambda = \max_{i\neq j} \cos\theta_{ij}^{(1)}. \tag{3.17}$$

Metrics on Grassmannian manifolds: The collection of all k dimensional subspaces of $\mathbb{R}^d$ is called the real 'Grassmannian manifold' $\mathbb{G}(k, \mathbb{R}^d)$. On this space we define two metrics.

(1) The *chordal* distance between two subspaces W_i and W_j is given by

$$d_c(W_i, W_j) := \left(\sum_{\ell=1}^{k} \sin^2 \theta_{ij}^{(\ell)} \right)^{1/2}.$$

(2) The *spectral* distance between W_i and W_j is given by

$$d_s(W_i, W_j) := \min_{\ell} \sin \theta_{ij}^{(\ell)} = \sin \theta_{ij}^{(1)}.$$

Optimal packing bounds have been derived with respect to the chordal distance d_c in [**30**]. It also proves to be the right metric for some practical problems like robustness with respect to subspace erasures in the fusion frames [**76**]. The spectral distance d_s is directly related to the parameter λ in our work, see (3.17), as

$$\lambda^2 = \max_{i \neq j} \max_{\ell} \left(\cos \theta_{ij}^{(\ell)} \right)^2 = \max_{i \neq j} \left(1 - \min_{\ell} \left(\sin \theta_{ij}^{(\ell)} \right)^2 \right) = 1 - \min_{i \neq j} d_s^2(W_i, W_j). \tag{3.18}$$

Packing problem: Let the space $\mathbb{G}(k, \mathbb{R}^d)$ be endowed with a distance function d which is either d_c or d_s. Then the *packing diameter* of a finite subset $\mathcal{X} \subset \mathbb{G}(k, \mathbb{R}^d)$ is the minimum distance between a pair of distinct subspaces drawn from $\mathcal{X}$. That is,

$$\operatorname{pack}_d(\mathcal{X}) := \min_{i \neq j} d(W_i, W_j).$$

An optimal packing of N subspaces in $\mathcal{X}$ is an ensemble that solves the mathematical program

$$\max_{|\mathcal{X}|=N} \operatorname{pack}_d(\mathcal{X}) =: \operatorname{pack}_d.$$

This program is guaranteed to have a solution but it is very hard to find the optimal solution. It is even hard to find a set $\mathcal{X}$ of N subspaces which satisfies $\operatorname{pack}_d(\mathcal{X}) \geq \rho$ for a given ρ. There have been many studies that focus on designing algorithms to come up with subspaces which yield good packings. Our focus here is on the theoretical upper bounds for optimal packing diameters which will consequently allow us to acquire a lower bound on λ.

Bounds on the packing diameter: A standard reference about optimal packing bounds is the paper [**30**] by Conway and Sloane, which provides an upper bound for the optimal packing diameter with respect to the chordal distance.

Theorem 3.11. *The packing diameter of N subspaces in the Grassmannian manifold $\mathbb{G}(k, \mathbb{R}^d)$ equipped with the chordal distance is bounded above as*

$$\operatorname{pack}_{d_c}^2(\mathcal{X}) \leq \frac{k(d-k)}{d} \frac{N}{N-1}.$$

If the bound is met, all pairs of subspaces are equidistant. The bound is attainable only if $N \leq \frac{1}{2} d(d+1)$.

The upper bound is referred as the *simplex bound*. It is worth to note that, in general, it is not known whether optimal packings exist for given parameters N, d, k. As mentioned earlier, our results in this thesis do not exploit the fusion frame property (3.1). However, we would like to note an interesting connection between tight fusion frames with equi-dimensional subspaces and optimal packings. The following result is from [**76**].

Theorem 3.12. *Let $(W_i)_{i=1}^N$ be a fusion frame of equi-dimensional subspaces with equal pairwise chordal distance d_c. Then, the fusion frame is tight if and only if d_c^2 equals the simplex bound.*

In essence, this theorem says that equi-distant tight fusion frames are optimal Grassmannian packings. Next we move on to packings with respect to the spectral distance d_s. A subspace packing is said to be *equi-isoclinic* if all the principal angles between all pairs of subspaces are identical [**79**]. The following theorem appears in [**36**].

Theorem 3.13. *The packing diameter of N subspaces in the Grassmannian manifold $\mathbb{G}(k, \mathbb{R}^d)$ equipped with spectral distance is bounded above as*

$$\operatorname{pack}_{d_s}^2(\mathcal{X}) \leq \frac{(d-k)}{d} \frac{N}{N-1}. \tag{3.19}$$

If the bound is met, the packing is equi-isoclinic.

The proof of this result relies on the simple observation that for any two subspaces W_i, W_j it holds

$$\min_{\ell} \sin \theta_{ij}^{(\ell)} \leq \left[k^{-1} \sum_{\ell=1}^{k} \sin^2 \theta_\ell \right]^{1/2}.$$

In other words, $k d_s^2(W_i, W_j) \leq d_c^2(W_i, W_j)$. It is shown in [**79**] that the maximum number of equi-isoclinic k-dimensional subspaces in $\mathbb{R}^d$ cannot be greater than

$$\frac{1}{2}d(d+1) - \frac{1}{2}k(k+1) + 1.$$

We finally derive a bound on the incoherence parameter. By (3.18) we have $\lambda^2 = 1 - \operatorname{pack}_{d_s}^2$ so that (3.19) implies

$$\lambda^2 \geq 1 - \frac{(d-k)}{d} \frac{N}{N-1} = \frac{kN-d}{dN-d}. \tag{3.20}$$

When $k = 1$, this bound gives a lower bound on the coherence of ℓ_2-normalized N vectors in $\mathbb{R}^d$, which is known as the *Welch bound*.

3.2.5. Comparison of results

In this section we provide a comparison of the sufficient recovery conditions given in our nonuniform and uniform results and the necessary condition in Theorem 3.10 in the light of the results in Section 3.2.4.

Our first observation is the absence of the incoherence parameter λ in the necessary condition (3.16) in contrast to the recovery conditions (3.5) and (3.10). A comparison of those conditions becomes possible under a condition on λ, i.e., on the orientation of the subspaces. Since we have a lower bound (3.20) on λ, we might consider the case when equality holds in (3.20). Let us assume that λ achieves its theoretical lower bound, i.e.,

$$\lambda^2 = \frac{kN-d}{dN-d} \leq \frac{k}{d}.$$

Then for nonuniform recovery

$$m \gtrsim \sqrt{\frac{k}{d}} s \ln(N) \tag{3.21}$$

becomes a sufficient condition. There is a slight gap between (3.21) and the necessary condition (3.16), which is

$$m \gtrsim \frac{s}{d}\ln(N/s) + \frac{ks}{d}. \tag{3.22}$$

It is not clear to us from which direction this gap may be closed.

Finally we would like to compare our two uniform results. Recall that Condition (3.10) of Theorem 3.7 implies that

$$m \geq C(\lambda s + \sqrt{s})\ln(Nd)\ln^2(sk)\left(\ln(N) + k\ln\left(s\sqrt{k}\right)\right) \tag{3.23}$$

many subgaussian measurements are sufficient for uniform recovery of sparse fusion frame vectors. Here k is the dimension of the subspaces. We briefly note that if λ attains its theoretical lower bound, then (3.23) scales roughly like

$$m \gtrsim \left(\sqrt{\frac{k}{d}}s + \sqrt{s}\right)\ln^{\alpha}(Nd),$$

which is further from the necessary condition (3.22) than (3.21) is. Condition (3.23) gives an order of $\sqrt{s}$ in the limiting case $\lambda = 0$. On the other hand, Theorem 3.8 gives the condition

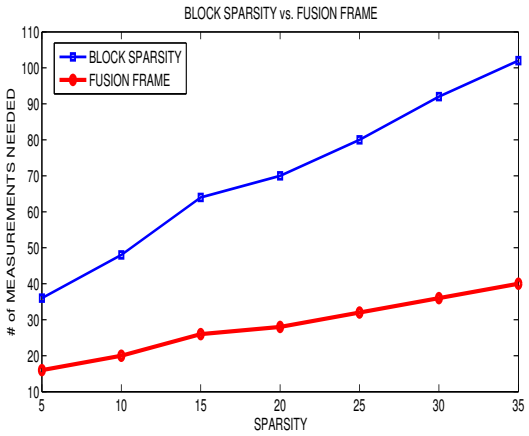
$$m \geq c(1 + \lambda s)d\ln(N) \tag{3.24}$$

which improves (3.23) in terms of log-factors and to the order of $(1 + \lambda s)$. However we would like to note that the ambient dimension d appears as a linear factor in (3.24), which is suboptimal. The latter result applies only to Gaussian matrices unlike the former one which also applies to subgaussian matrices.

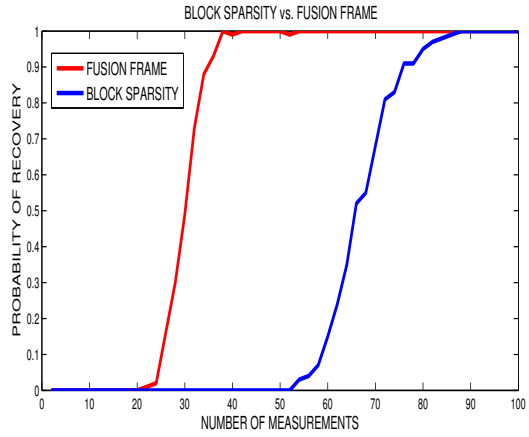
3.2.6. Numerical experiments

In this section, we present numerical experiments in order to highlight important aspects of the sparse reconstruction in the fusion frame (FF) setup. The experiments illustrate our theoretical results and show that when the subspaces are known, one can significantly improve the recovery of sparse vectors with respect to the standard block sparse recovery. In all of our experiments, we use SPGL1 [**107**, **108**] to solve the $\ell_{2,1}$-minimization problems.

General setup: We generate subspaces randomly, which allows us to generate fusion frames with different values of λ and Λ. Particularly, for N subspaces in $\mathbb{R}^d$ each with dimension k, we generate $N \cdot k$ random vectors from $\mathcal{N}(0, \mathrm{Id})$ and group them to form the bases for the subspaces. Every realization of random subspaces in this way yields a certain value for the parameter λ. In order to obtain different values of λ, it is enough to vary d or k. When N is fixed, λ increases with increasing k and decreasing d. This relation can be explained with an analogy: imagine we are given the task of placing N items of equal size k (i.e., subspaces) in a room of size d (i.e., the ambient space $\mathbb{R}^d$) as further away as possible from each other. If N and d are fixed, increasing the item size k leads to shorter distances between the items (i.e., larger λ) and vice versa. Similarly, if N and k are fixed, increasing the room size d allows us to place the items further away from each other (i.e., with smaller λ) and vice versa. A quantitative analysis of the packing of random subspaces was recently done in [**7**]. One drawback of our method of generating subspaces is that one cannot vary all parameters independently in this way. While our main goal is to control the parameter



(a)



(b)

FIGURE 1. Comparison between 'block' vs. 'fusion frame' sparsity. (a) Sparsity level vs. the number of necessary measurements for successful recovery ($N = 200$, success rate= 96%). (b) Sparsity is fixed and the number of measurements is plotted against the success rate ($N = 200, s = 20$).

λ so that we can observe its effect on the relation between the number of measurements m and the sparsity s, we have to also change d or k during this process. According to our nonuniform results, the role of the dimensional factors (d or k) is at most logarithmic, therefore it can be neglected for our purposes in this section. Another way to build fusion frames is deterministic constructions, e.g., Gabor and harmonic frames, however one typically needs to vary N largely in order to yield different values of λ. This is not desirable since we wish to keep N fixed and also it is computationally hard to work with very large values of N.

For the measurement matrices, we generate the normalized matrix $\tilde{A} = \frac{1}{\sqrt{m}}A$ where $A \in \mathbb{R}^{m\times N}$ is a Gaussian matrix. For a sparsity level s, sparse vectors are generated in the following way: we choose the support set S uniformly at random, then we sample a Gaussian vector in each subspace in this support set. The number N is kept fixed throughout the experiment at hand. In our experiments we work with the parameter $\|\Lambda_S\|_\infty$ introduced in Section 3.1.5. Since the random subspaces are not equiangular, this parameter reflects the linear relation between m and

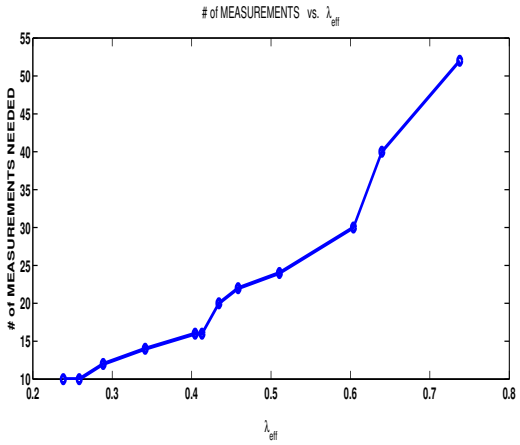


FIGURE 2. Exact signal recovery with fusion frames. For a fixed sparsity, the relation between the number of measurements m vs. the incoherence parameter λ_{eff} (3.25) is depicted ($N = 180, s = 25$, success rate= 96%).

s better than λ, see also Theorem 3.1. We work with the normalized parameter

$$\lambda_{\text{eff}} = \frac{\|\Lambda_S\|_\infty}{s}. \tag{3.25}$$

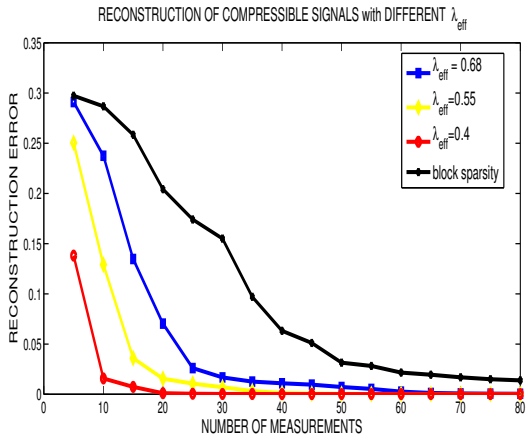
Exact sparse case: In Fig. 1, we show that the knowledge of the subspaces improves the recovery. To that end, we fix a fusion frame with $N = 200$ subspaces in $\mathbb{R}^d$ with $\lambda_{\text{eff}} \approx 0.6$. Then we vary the sparsity level s from 5 to 35, and generate an s-sparse vector $\mathbf{x}$ in the fusion frame. For each s, we vary the number of measurements m and compute empirical recovery rates via the programs

$$\text{(FF) } \hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}\in\mathcal{H}}\|\mathbf{x}\|_{2,1} \quad s.t. \quad \mathbf{A_P x} = \mathbf{y}, \tag{3.26}$$

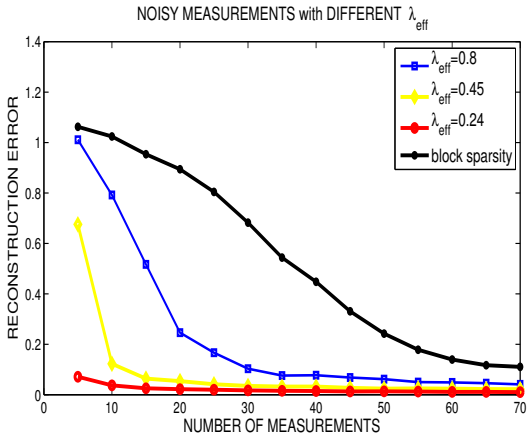
$$\text{(block) } \hat{\mathbf{x}} = \text{argmin}_{\mathbf{x}}\|\mathbf{x}\|_{2,1} \quad s.t. \quad \mathbf{A_I x} = \mathbf{y}. \tag{3.27}$$

For each sparsity level s, we leave the vector to be recovered fixed. Repeating this test 100 times with different random A for each choice of parameters (s, m, N) provides an empirical estimate of the success probability. In Fig. 1(a), we plot m which yields at least 96% success rate for each s. The difference in two plots is due to the incoherence of the subspaces. Fig. 1(b) ($d = 3$, $k = 1$) shows the transition from the unsuccessful regime to the successful regime for the sparsity level $s = 20$ for both cases (FF and block sparsity). The transition for the FF case occurs at a smaller value m which reflects Fig. 1(a) in a different way. As a consequence, the assumption $\mathbf{x} \in \mathcal{H}$ in (3.26) allows us to to recover $\mathbf{x}$ with far less measurements compared to (3.27) where such a constraint is not used.

Theorem 3.1 suggests that there is a linear relation between the number of measurements m and the parameter λ_{eff}. The experiment depicted in Fig. 2 is designed to reflect this relation. We generate fusion frames with $N = 180$ subspaces with various λ_{eff} which is managed by changing d and keeping $k = 3$ fixed. Then in each fusion frame, a vector $\mathbf{x}$ with sparsity $s = 25$ is generated and the number of measurements m that suffices for recovery is determined. The plot yields an almost linear relation in parallel to Theorem 3.1.



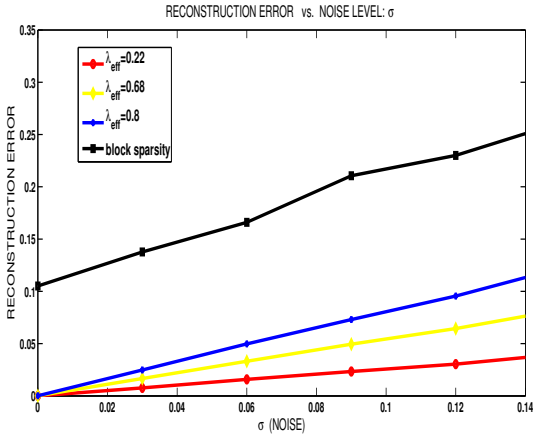
(a)



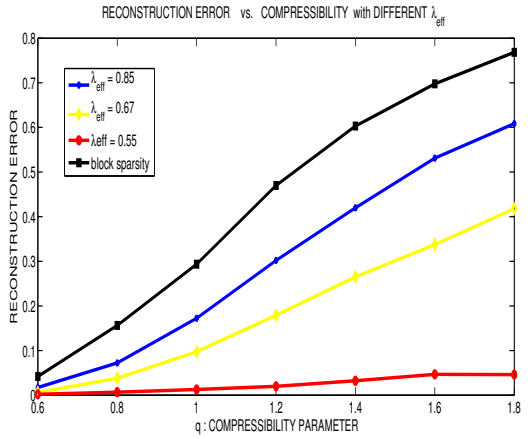
(b)

FIGURE 3. Stable and robust signal recovery with fixed noise and compressibility levels. (a) Recovery of compressible signals from fusion frames with different values of λ_{eff}. Reconstruction error vs. the number of measurements is plotted ($N = 180, s = 20, \theta = 0.12$). (b) Noisy measurements. Reconstruction error vs. the number of measurements is plotted for different values of λ_{eff} ($N = 200, s = 20, \sigma = 0.06$).

Stable case: In this part, we generate scenarios that allude to the conclusions of Theorems 3.4 and 3.6. In a fusion frame of $N = 200$ subspaces, we generate a signal $\mathbf{x}$ composed of $\mathbf{x}_S$, supported on an index set S, and a signal $\mathbf{z}_{\overline{S}}$ supported on $\overline{S}$. We then normalize $\mathbf{x}_S$ and $\mathbf{z}_{\overline{S}}$ so that $\|\mathbf{x}_S\|_{2,1} = \|\mathbf{z}_{\overline{S}}\|_{2,1} = 1$ and produce $\mathbf{x} = \mathbf{x}_S + \theta\mathbf{z}_{\overline{S}}$ where $\theta \in [0, 1]$. Then $\mathbf{x}$ is our compressible vector where compressibility is controlled with θ. For measurement, we choose the normalized Gaussian matrix $A \in \mathbb{R}^{m\times N}$. We measure $\mathbf{y} = \mathbf{A_P x}$ and then run the program $(L1)$ and measure the reconstruction error $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$. We repeat this test 20 times for a fixed $\mathbf{x}$ with $\theta = 0.12$ in order to obtain an average recovery error for different values of m. Fig. 3(a) reports the results of this experiment performed for different fusion frames with various values of λ_{eff} and also for the block sparsity case. The decrease in the reconstruction error with increasing m is natural even though it is not suggested directly by the theoretical results. Indeed, one would



(a)



(b)

FIGURE 4. Stable and robust signal recovery with varying noise and compressibility levels ($N = 200, m = 50, d = 35$). (a) Noise level σ vs. reconstruction error for different values of λ_{eff} ($s = 30$). (b) Compressibility q vs. reconstruction error.

expect that increasing the number of measurements would enhance the recovery conditions and yield an improved reconstruction.

For the noisy case, we similarly generate noisy observations $\mathbf{A_P x}_S + \sigma\mathbf{e}$, of a sparse signal $\mathbf{x}_S$ where $\|\mathbf{x}_S\|_2 = \|\mathbf{e}\|_2 = 1$ and $\sigma = 0.06$. Here, all entries of the noise vector $\mathbf{e}$ are chosen i.i.d from a $\mathcal{N}(0, \sigma^2)$ distribution. We then run the robust $(L1)^\eta$ program and measure the reconstruction error $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$. We plot the average of this error vs. the number of measurements in Fig 3(b) for different values of λ_{eff}.

Fig. 4(a) depicts the relation between the reconstruction error and the noise level σ for different values of λ_{eff}. In this setup, $N = 200$, $s = 30$ and $m = 50$ are fixed, and a sparse vector $\mathbf{x}$ in the fusion frame with specific value of λ_{eff} is generated. For each value of σ we plot the average reconstruction error. Results manifest the linear relation between σ and $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$ given in (3.9). Again, we obtain a better reconstruction quality when λ_{eff} is smaller.

Finally, we examine the relation between compressibility and the reconstruction error using a different model than described earlier. In Fig.4(b), we plot the results of an experiment in which we generate signals $\mathbf{x}$ in a fusion frame with $N = 200$, with sorted values of $\|x_j\|_2$ that decay according to some power law. In particular, for various values of $0 < q < 1$, we set $\|x_j\|_2 = cj^{-1/q}$ such that $\|\mathbf{x}\|_2 = 1$. We then measure $\mathbf{x}$ with Gaussian matrices A and compute the average reconstruction errors via the program $(L1)$. Note that the higher the value of q, the less compressible the signal is. The results indicate that the reconstruction error decreases when the compressibility of the signal increases as declared in (3.9). We can also see the improvement in the reconstruction when the subspaces are more incoherent, i.e., their parameter λ_{eff} is smaller.

3.2.7. RIP for fusion frames

The following definition of the restricted isometry property for fusion frames was given in [**9**].

Definition. *Let $A \in \mathbb{R}^{m\times N}$ and $(W_j)_{j=1}^N$ be a fusion frame for $\mathbb{R}^d$. The fusion restricted isometry constant δ_s is the smallest constant such that*

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A_P x}\|_2^2 \leq (1+\delta_s)\|\mathbf{x}\|_2^2 \tag{3.28}$$

for all $\mathbf{x} \in \mathcal{H}$, of sparsity $\|\mathbf{x}\|_0 \leq s$.

Informally, we say $(A, (W_j)_{j=1}^N)$ satisfies the fusion restricted isometry property (FRIP) if δ_s is small for reasonably large s. The following result was also shown in [**9**].

Theorem 3.14. *Let $(A, (W_j)_{j=1}^N)$ with FRIP constant $\delta_{2s} < 1/3$. Then $(L1)$ recovers all s-sparse $\mathbf{x}$ from $\mathbf{y} = \mathbf{A_P x}$.*

Boufounos, Kutyniok and Rauhut [**9**] have also stated -without a proof- an extension of Theorem 3.14 to the case where measurements are noisy and signals are well approximated by sparse fusion frame representation. Let us define the best s-term approximation of a vector $\mathbf{x}$ as follows

$$\sigma_s(\mathbf{x})_1 := \inf_{\|\mathbf{z}\|_0 \leq s} \|\mathbf{x}-\mathbf{z}\|_{2,1}.$$

This quantity measures the compressibility of the vector, in other words, how close it is to being s-sparse. We call vectors with small $\sigma_s(\mathbf{x})_1$ 'approximately sparse' or 'compressible'. The following stable version of FRIP condition is an analogous result to [**17**, Theorem 1.2].

Theorem 3.15. *Assume that the FRIP constant δ_{2s} of $(A, (W_j)_{j=1}^N)$ satisfies*

$$\delta_{2s} < \sqrt{2} - 1 \approx 0.41.$$

For $\mathbf{x} \in \mathcal{H}$, let noisy measurements $\mathbf{y} = \mathbf{A_P x} + \mathbf{e}$ be given with $\|\mathbf{e}\|_2 \leq \eta$. Let $\hat{\mathbf{x}}$ be the solution of the convex optimization problem

$$\min_{\mathbf{z}\in\mathcal{H}} \|\mathbf{z}\|_{2,1} \ s.t. \ \|\mathbf{A_P z} - \mathbf{y}\|_2 \leq \eta.$$

Then

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_2 \leq C_1 \frac{\sigma_s(\mathbf{x})_1}{\sqrt{s}} + C_2\eta, \tag{3.29}$$

and

$$\|\mathbf{x}-\hat{\mathbf{x}}\|_{2,1} \leq C_1\sigma_s(\mathbf{x})_1 + C_2\sqrt{s}\eta, \tag{3.30}$$

where the constants $C_1, C_2 > 0$ only depend on δ_{2s}.

We remark that this result implies Theorem 3.14 and improves it in terms of the FRIP constant by choosing $\mathbf{x}$ to be exactly s-sparse and $\eta = 0$. For the sake of completeness, the proof of Theorem 3.15 is presented at the end of this section. In Section 1.2.1, it was noted that for the classical RIP constant the optimal condition is $\tilde{\delta}_{2s} < 1/\sqrt{2} \approx 0.7071$ for both exact and stable recovery. We think that the condition for FRIP constants in Theorems 3.14 and 3.15 can also be improved to $\delta_{2s} < 1/\sqrt{2}$ by a similar proof to that in [**16**]. The results above show that given a fusion frame $(W_j)_{j=1}^N$ and matrix A, for uniform recovery it is enough to check whether the block matrix $\mathbf{A_P}$ satisfies the FRIP. The following result is from [**9**].

Proposition 3.16. *Let $A \in \mathbb{R}^{m\times N}$ satisfy classical RIP* (1.4)*. Then for an arbitrary fusion frame $(W_j)_{j=1}^N$ for $\mathbb{R}^d$, the associated matrix $\mathbf{A_P}$ satisfies FRIP.*

Recall from Theorem 3.14 that satisfying FRIP is enough for uniform recovery. This immediately suggests that $m \gtrsim s\ln(N/s)$ is sufficient for sparse recovery with many random measurement ensembles (up to some log-factors) as shown in Chapter 2. However we would like to improve this result by making use of the incoherence parameter λ.

For the proof of Theorem 3.7 we seek conditions on m which guarantees δ_s to be small with high probability for a subgaussian matrix A. The definition (3.28) is equivalent to

$$\left|\|\mathbf{A_P x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \leq \delta_s \|\mathbf{x}\|_2^2 \quad \text{for all } S \subset [N], |S| \leq s, \text{ for all } \mathbf{x} \in \mathcal{H}, \text{supp}(\mathbf{x}) \subset S.$$

In the term on the left hand side, taking the supremum over all $\mathbf{x} \in \mathcal{H}$ with $\text{supp}(\mathbf{x}) \subset S$, $|S| \leq s$ and unit norm $\|\mathbf{x}\|_2 = 1$ yields

$$\delta_s = \sup_{\mathbf{x}\in D_{s,N}} \left|\|\mathbf{A_P x}\|_2^2 - \|\mathbf{x}\|_2^2\right|, \tag{3.31}$$

where $D_{s,N} := \{\mathbf{x} \in \mathcal{H} : x_i \in W_i, \|\mathbf{x}\|_2 \leq 1, \|\mathbf{x}\|_0 \leq s\}$. The characterization (3.31) of the FRIP constant δ_s will be used in the proof of Theorem 3.7 in Section 3.4.1.

- *Proof of Theorem 3.15*

We mainly follow [**17**] for the proof. Some of the steps below are also parallel to the proof of Lemma 2.10. Set $\mathbf{h} = \hat{\mathbf{x}} - \mathbf{x}$. Similarly to (2.32), we observe that

$$\|\mathbf{A_P h}\|_2 \leq \|\mathbf{A_P}\hat{\mathbf{x}} - \mathbf{y}\|_2 + \|\mathbf{y} - \mathbf{A_P x}\|_2 \leq 2\eta. \tag{3.32}$$

We partition $[N]$ into sets $S_0, S_1, \ldots$ each of size s (except possibly the last set) such that S_0 contains the s largest components $\|\mathbf{h}_i\|_2$, S_1 has the second s largest components and so on. We first focus on bounding the size of $\mathbf{h}$ on the set $S_0 \cup S_1$. In an analogous fashion to the scalar case (2.37) (see also [**17**, Eq.(10)]), we can deduce

$$\sum_{j\geq 2} \|\mathbf{h}_{S_j}\|_2 \leq s^{-1/2} \|\mathbf{h}_{\overline{S_0}}\|_{2,1}. \tag{3.33}$$

In particular, it holds

$$\|\mathbf{h}_{\overline{S_0\cup S_1}}\|_2 = \|\sum_{j\geq 2} \mathbf{h}_{S_j}\|_2 \leq \sum_{j\geq 2} \|\mathbf{h}_{S_j}\|_2 \leq s^{-1/2} \|\mathbf{h}_{\overline{S_0}}\|_{2,1}. \tag{3.34}$$

The next step is to show that $\|\mathbf{h}_{\overline{S_0}}\|_{2,1}$ cannot be very large since $\|\hat{\mathbf{x}}\|_{2,1}$ is a minimum point. Indeed,

$$\begin{aligned}\|\mathbf{x}\|_{2,1} &\geq \|\mathbf{x}+\mathbf{h}\|_{2,1} = \|(\mathbf{x}+\mathbf{h})_{S_0}\|_{2,1} + \|(\mathbf{x}+\mathbf{h})_{\overline{S_0}}\|_{2,1} \\ &\geq \|\mathbf{x}_{S_0}\|_{2,1} - \|\mathbf{h}_{S_0}\|_{2,1} + \|\mathbf{h}_{\overline{S_0}}\|_{2,1} - \|\mathbf{x}_{\overline{S_0}}\|_{2,1},\end{aligned}$$

where we used the triangle inequality. Rearranging terms yields

$$\|\mathbf{h}_{\overline{S_0}}\|_{2,1} \leq \|\mathbf{h}_{S_0}\|_{2,1} + 2\|\mathbf{x}_{\overline{S_0}}\|_{2,1} \leq \sqrt{s}\|\mathbf{h}_{S_0}\|_2 + 2\sigma_s(\mathbf{x})_1, \tag{3.35}$$

by the Cauchy-Schwarz inequality. Combining (3.35) and (3.34) gives

$$\|\mathbf{h}_{\overline{S_0 \cup S_1}}\|_2 \leq \|\mathbf{h}_{S_0}\|_2 + 2s^{-1/2}\sigma_s(\mathbf{x})_1. \tag{3.36}$$

The next goal is to bound $\|\mathbf{h}_{S_0\cup S_1}\|_2$. To do this, observe that $\mathbf{A_P h}_{S_0\cup S_1} = \mathbf{A_P h} - \sum_{j\geq 2} \mathbf{A_P h}_{S_j}$, and therefore

$$\|\mathbf{A_P h}_{S_0\cup S_1}\|_2^2 = \langle \mathbf{A_P h}_{S_0\cup S_1}, \mathbf{A_P h}\rangle - \langle \mathbf{A_P h}_{S_0\cup S_1}, \sum_{j\geq 2} \mathbf{A_P h}_{S_j}\rangle.$$

The first term on the right hand side is estimated by using the FRIP along with (3.32)

$$|\langle \mathbf{A_P h}_{S_0\cup S_1}, \mathbf{A_P h}\rangle| \leq \|\mathbf{A_P h}_{S_0\cup S_1}\|_2 \|\mathbf{A_P h}\|_2 \leq 2\eta\sqrt{1+\delta_{2s}}\|\mathbf{h}_{S_0\cup S_1}\|_2.$$

Moreover, following an analogous result for the scalar case, Lemma 2.9, we have for $j \geq 2$

$$|\langle \mathbf{A_P h}_{S_0}, \mathbf{A_P h}_{S_j}\rangle| \leq \delta_{2s}\|\mathbf{h}_{S_0}\|_2\|\mathbf{h}_{S_j}\|_2,$$

and likewise for S_1 in place of S_0. This implies that

$$\langle \mathbf{A_P h}_{S_0\cup S_1}, \sum_{j\geq 2} \mathbf{A_P h}_{S_j}\rangle \leq \delta_{2s}(\|\mathbf{h}_{S_0}\|_2 + \|\mathbf{h}_{S_1}\|_2)\sum_{j\geq 2}\|\mathbf{h}_{S_j}\|_2 \leq \sqrt{2}\delta_{2s}\|\mathbf{h}_{S_0\cup S_1}\|_2 \sum_{j\geq 2}\|\mathbf{h}_{S_j}\|_2.$$

Then by the FRIP again we obtain

$$(1-\delta_{2s})\|\mathbf{h}_{S_0\cup S_1}\|_2^2 \leq \|\mathbf{A_P h}_{S_0\cup S_1}\|_2^2 \leq \|\mathbf{h}_{S_0\cup S_1}\|_2 \left(2\eta\sqrt{1+\delta_{2s}} + \sqrt{2}\delta_{2s}\sum_{j\geq 2}\|\mathbf{h}_{S_j}\|_2\right).$$

It follows from (3.33) that

$$\|\mathbf{h}_{S_0\cup S_1}\|_2 \leq \alpha\eta + \beta s^{-1/2}\|\mathbf{h}_{\overline{S_0}}\|_{2,1},$$

where $\alpha = \frac{2\sqrt{1+\delta_{2s}}}{1-\delta_{2s}}$ and $\beta = \frac{\sqrt{2}\delta_{2s}}{1-\delta_{2s}}$. We now conclude by using (3.35) along with the previous inequality that

$$\|\mathbf{h}_{S_0\cup S_1}\|_2 \leq \alpha\eta + \beta\|\mathbf{h}_{S_0\cup S_1}\|_2 + 2\beta s^{-1/2}\sigma_s(\mathbf{x})_1.$$

Rearranging terms yields

$$\|\mathbf{h}_{S_0\cup S_1}\|_2 \leq (1-\beta)^{-1}(\alpha\eta + 2\beta s^{-1/2}\sigma_s(\mathbf{x})_1). \tag{3.37}$$

Observe that $\beta < 1$ due to the assumption $\delta_{2s} < \sqrt{2}-1$. Finally using (3.36), we obtain

$$\begin{aligned}\|\mathbf{h}\|_2 &\leq \|\mathbf{h}_{S_0\cup S_1}\|_2 + \|\mathbf{h}_{\overline{S_0\cup S_1}}\|_2 \leq 2\|\mathbf{h}_{S_0\cup S_1}\|_2 + 2s^{-1/2}\sigma_s(\mathbf{x})_1 \\ &\leq 2(1-\beta)^{-1}(\alpha\eta + (1+\beta)s^{-1/2}\sigma_s(\mathbf{x})_1),\end{aligned}$$

which proves (3.29). In order to show (3.30), we use the first inequality in (3.35) to obtain that

$$\|\mathbf{h}\|_{2,1} = \|\mathbf{h}_{S_0}\|_{2,1} + \|\mathbf{h}_{\overline{S_0}}\|_{2,1} \leq 2\|\mathbf{h}_{S_0}\|_{2,1} + 2\sigma_s(\mathbf{x})_1. \tag{3.38}$$

Moreover, it follows from (3.37) and the Cauchy-Schwarz inequality that

$$\|\mathbf{h}_{S_0}\|_{2,1} \leq \sqrt{s}\|\mathbf{h}_{S_0}\|_2 \leq \sqrt{s}\|\mathbf{h}_{S_0 \cup S_1}\|_2 \leq \sqrt{s}(1-\beta)^{-1}(\alpha\eta + 2\beta s^{-1/2}\sigma_s(\mathbf{x})_1.$$

Plugging this into (3.38) yields the desired estimate (3.30). □

3.2.8. Null space properties for fusion frames

In this section we give notions of exact, stable and robust null space properties (NSP) for fusion frames and show in turn that they provide sufficient conditions for sparse recovery. We would like to note that Theorems 3.17 and 3.18 can be improved to stronger 'if and only if' statements, but we skip these parts here.

Definition. *(NSP of order s) Let $A \in \mathbb{R}^{m\times N}$ and $(W_j)_{j=1}^N$ be a fusion frame. Then the associated matrix $\mathbf{A_P} \in \mathbb{R}^{md\times Nd}$ is said to satisfy the null space property of order s if for all $S \subset [N], |S| = s$,*

$$\|\mathbf{v}_S\|_{2,1} < \|\mathbf{v}_{\overline{S}}\|_{2,1} \quad \textit{for all } \mathbf{v} \in \ker \mathbf{A}_{\mathbf{P}|\mathcal{H}} \setminus \{\mathbf{0}\}.$$

Here, $\mathbf{A}_{\mathbf{P}|\mathcal{H}}$ is the restriction of $\mathbf{A_P}$ to $\mathcal{H}$. We now show that the null space property implies exact recovery of sparse vectors from a fusion frame via $(L1)$.

Theorem 3.17. *Suppose $\mathbf{A_P}$ satisfies the null space property of order s. Then any s-sparse $\mathbf{x} \in \mathcal{H}$ is a solution of $(L1)$ with $\mathbf{y} = \mathbf{A_P x}$.*

One can use this result in order to prove Corollary 3.9 by establishing the null space property of order s. However, as we have remarked earlier, this is not necessary since choosing ρ and η in Theorem 3.8 appropriately does the job. Nevertheless, we prove Theorem 3.17 here for the sake of completeness. The following proof is analogous to [**54**, Theorem 4.4].

Proof. Assume that the null space property of order s holds. Given an s-sparse vector $\mathbf{x} \in \mathcal{H}$ and a vector $\mathbf{z} \neq \mathbf{x}$ such that $\mathbf{A_P z} = \mathbf{A_P x}$, we consider $\mathbf{v} = \mathbf{x} - \mathbf{z} \in \ker \mathbf{A}_{\mathbf{P}|\mathcal{H}} \setminus \{\mathbf{0}\}$. Then by the null space property, we have

$$\begin{aligned}\|\mathbf{x}\|_{2,1} &\leq \|\mathbf{x} - \mathbf{z}_S\|_{2,1} + \|\mathbf{z}_S\|_{2,1} = \|\mathbf{v}_S\|_{2,1} + \|\mathbf{z}_S\|_{2,1} \\ &< \|\mathbf{v}_{\overline{S}}\|_{2,1} + \|\mathbf{z}_S\|_{2,1} = \| - \mathbf{z}_{\overline{S}}\|_{2,1} + \|\mathbf{z}_S\|_{2,1} = \|\mathbf{z}\|_{2,1}.\end{aligned}$$

This establishes the minimality of $\|\mathbf{x}\|_{2,1}$. □

Stability and robustness. Theorem 3.17 is about the idealized situation when the signal is exactly sparse. In practice, one often faces signals with sparsity defects. Above we defined the best s-term approximation error $\sigma_s(\mathbf{x})_1$ for a vector $\mathbf{x}$ which is not exacty s-sparse. Another non-ideal situation is that often the measurement vector $\mathbf{y}$ is corrupted with some noise vector $\mathbf{e}$ which is bounded in ℓ_2-norm. The following definition gives a variant of the null space property adapted to these two scenarios.

Definition. *(Stable and robust NSP of order s) Let $A \in \mathbb{R}^{m\times N}$ and $(W_j)_{j=1}^N$ be a fusion frame. Then the associated matrix $\mathbf{A_P} \in \mathbb{R}^{md\times Nd}$ is said to satisfy the stable and robust null space*

property of order s with constants $0 < \rho < 1$ and $\tau > 0$, if for all $S \subset [N], |S| = s$, it holds

$$\|\mathbf{v}_S\|_{2,1} \leq \rho \|\mathbf{v}_{\overline{S}}\|_{2,1} + \tau \|\mathbf{A}_\mathbf{P}\mathbf{v}\|_2 \quad \text{for all } \mathbf{v} \in \mathcal{H}.$$

Theorem 3.18. *Suppose $\mathbf{A}_\mathbf{P}$ satisfies the stable null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then for any $\mathbf{x} \in \mathcal{H}$, a solution $\hat{\mathbf{x}}$ of $(L1)^\eta$ with $\mathbf{y} = \mathbf{A}_\mathbf{P}\mathbf{x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \eta$ satisfies*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_{2,1} \leq \frac{2(1+\rho)}{1-\rho}\sigma_s(\mathbf{x})_1 + \frac{4\tau}{1-\rho}\eta.$$

PROOF. Denote by $S \subset [N]$ an index set of s largest norms $\|x_i\|_2$ in $\mathbf{x}$, so that $\|\mathbf{x}_{\overline{S}}\|_{2,1} = \sigma_s(\mathbf{x})_1$. The inequalities below follow from the triangle inequality

$$\begin{aligned}
\|\mathbf{x}\|_{2,1} &= \|\mathbf{x}_{\overline{S}}\|_{2,1} + \|\mathbf{x}_S\|_{2,1} \leq \|\mathbf{x}_{\overline{S}}\|_{2,1} + \|\hat{\mathbf{x}}_S\|_{2,1} + \|(\mathbf{x} - \hat{\mathbf{x}})_S\|_{2,1} \\
\|(\mathbf{x} - \hat{\mathbf{x}})_{\overline{S}}\|_{2,1} &\leq \|\mathbf{x}_{\overline{S}}\|_{2,1} + \|\hat{\mathbf{x}}_{\overline{S}}\|_{2,1}.
\end{aligned}$$

Adding these two inequalities and using that $\|\hat{\mathbf{x}}\|_{2,1} \leq \|\mathbf{x}\|_{2,1}$ (since $\hat{\mathbf{x}}$ is a minimizer of $(L1)^\eta$) yields

$$\begin{aligned}
\|\mathbf{x}\|_{2,1} + \|(\mathbf{x} - \hat{\mathbf{x}})_{\overline{S}}\|_{2,1} &\leq 2\|\mathbf{x}_{\overline{S}}\|_{2,1} + \|\hat{\mathbf{x}}_S\|_{2,1} + \|\hat{\mathbf{x}}_{\overline{S}}\|_{2,1} + \|(\mathbf{x} - \hat{\mathbf{x}})_S\|_{2,1} \\
&= 2\|\mathbf{x}_{\overline{S}}\|_{2,1} + \|\hat{\mathbf{x}}\|_{2,1} + \|(\mathbf{x} - \hat{\mathbf{x}})_S\|_{2,1} \\
&\leq 2\|\mathbf{x}_{\overline{S}}\|_{2,1} + \|\mathbf{x}\|_{2,1} + \|(\mathbf{x} - \hat{\mathbf{x}})_S\|_{2,1}.
\end{aligned}$$

Simplifying above and setting $\mathbf{v} := \hat{\mathbf{x}} - \mathbf{x}$ gives

$$\|\mathbf{v}_{\overline{S}}\|_{2,1} \leq \|\mathbf{v}_S\|_{2,1} + 2\|\mathbf{x}_{\overline{S}}\|_{2,1} \leq \rho\|\mathbf{v}_{\overline{S}}\|_{2,1} + \tau\|\mathbf{A}_\mathbf{P}\mathbf{v}\|_2 + 2\|\mathbf{x}_{\overline{S}}\|_{2,1},$$

where we used the stable and robust null space property in the last inequality. After rearranging the terms, we obtain

$$\|\mathbf{v}_{\overline{S}}\|_{2,1} \leq \frac{2}{1-\rho}\|\mathbf{x}_{\overline{S}}\|_{2,1} + \frac{\tau}{1-\rho}\|\mathbf{A}_\mathbf{P}\mathbf{v}\|_2. \tag{3.39}$$

Observe that it holds

$$\|\mathbf{A}_\mathbf{P}\mathbf{v}\|_2 \leq \|\mathbf{A}_\mathbf{P}\mathbf{x} - \mathbf{y}\|_2 + \|\mathbf{y} - \mathbf{A}_\mathbf{P}\hat{\mathbf{x}}\|_2 \leq 2\eta \tag{3.40}$$

by the triangle inequality. Lastly, combining (3.39) and (3.40) yields

$$\begin{aligned}
\|\mathbf{v}\|_{2,1} &\leq \|\mathbf{v}_S\|_{2,1} + \|\mathbf{v}_{\overline{S}}\|_{2,1} \leq (1+\rho)\|\mathbf{v}_{\overline{S}}\|_{2,1} + \tau\|\mathbf{A}_\mathbf{P}\mathbf{v}\|_2 \\
&\leq \frac{2(1+\rho)}{1-\rho}\sigma_s(\mathbf{x})_1 + \frac{4\tau}{1-\rho}\eta,
\end{aligned}$$

which is the desired result. □

Next we generalize the previous theorem by replacing the $\ell_{2,1}$-error estimate by an ℓ_2-estimate. For this we need to strengthen the null space property.

Definition. *(ℓ_2-stable and robust NSP of order s) Let $A \in \mathbb{R}^{m \times N}$ and $(W_j)_{j=1}^N$ be a fusion frame. Then the associated matrix $\mathbf{A}_\mathbf{P} \in \mathbb{R}^{md \times Nd}$ is said to satisfy the ℓ_2-stable and robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$, if for all $S \subset [N], |S| = s$, it holds*

$$\|\mathbf{v}_S\|_2 \leq \frac{\rho}{\sqrt{s}}\|\mathbf{v}_{\overline{S}}\|_{2,1} + \tau\|\mathbf{A}_\mathbf{P}\mathbf{v}\|_2 \quad \text{for all } \mathbf{v} \in \mathcal{H}.$$

Theorem 3.19. *Suppose $\mathbf{A_P}$ satisfies the ℓ_2-stable and robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then for any $\mathbf{x} \in \mathcal{H}$, a solution $\hat{\mathbf{x}}$ of $(L1)^\eta$ with $\mathbf{y} = \mathbf{A_P x} + \mathbf{e}$ with $\|\mathbf{e}\|_2 \leq \eta$ satisfies*

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq \frac{C}{\sqrt{s}} \sigma_s(\mathbf{x})_1 + D\eta,$$

for some constants $C, D > 0$ depending on only ρ and τ.

PROOF. The ℓ_2-stable and robust null space property immediately implies that

$$\|\mathbf{v}_S\|_{2,1} \leq \rho \|\mathbf{v}_{\overline{S}}\|_{2,1} + \tau\sqrt{s}\|\mathbf{A_P v}\|_2 \quad \textit{for all } \mathbf{v} \in \mathcal{H}.$$

This, along with Theorem 3.18, gives

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_{2,1} \leq \frac{2(1+\rho)}{1-\rho} \sigma_s(\mathbf{x})_1 + \frac{4\tau\sqrt{s}}{1-\rho}\eta. \tag{3.41}$$

We briefly note here that, for a block vector $\mathbf{z} = (z_1, \ldots, z_N) \in \mathbb{R}^{N \cdot d}$ with block size d, if S is an index set of s largest ℓ_2-normed blocks in $\mathbf{z}$, then it holds that $\|\mathbf{z}_{\overline{S}}\|_2 \leq \frac{1}{2\sqrt{s}}\|\mathbf{z}\|_{2,1}$. This follows from an analogous result for the scalar case [**54**, Theorem 2.5]. Then, choosing S as an index set of s largest ℓ_2-normed blocks in $\mathbf{x} - \hat{\mathbf{x}}$, this result and the ℓ_2-stable and robust null space property yields

$$\begin{aligned}
\|\mathbf{x} - \hat{\mathbf{x}}\|_2 &\leq \|(\mathbf{x} - \hat{\mathbf{x}})_S\|_2 + \|(\mathbf{x} - \hat{\mathbf{x}})_{\overline{S}}\|_2 \\
&\leq \frac{\rho}{\sqrt{s}}\|(\mathbf{x} - \hat{\mathbf{x}})_{\overline{S}}\|_2 + \tau\|\mathbf{A_P}(\mathbf{x} - \hat{\mathbf{x}})\|_2 + \frac{1}{2\sqrt{s}}\|\mathbf{x} - \hat{\mathbf{x}}\|_{2,1} \\
&\leq \frac{2\rho + 1}{2\sqrt{s}}\|\mathbf{x} - \hat{\mathbf{x}}\|_{2,1} + 2\tau\eta.
\end{aligned}$$

Finally plugging (3.41) in the right hand side gives the desired inequality with $C = \frac{(2\rho+1)(1+\rho)}{1-\rho}$ and $D = \frac{4+2\rho}{1-\rho}\tau$. □

3.3. Proofs of nonuniform results

3.3.1. Preliminaries

In this section, we introduce some useful norms for random variables that will appear in our proofs. The following definition is taken from [**109**].

Definition. *(Orlicz norm) Let ψ be a nondecreasing convex function with $\psi(0) = 0$ and X be a random variable. The Orlicz norm of X is defined as*

$$\|X\|_\psi = \inf\left\{C > 0 : \mathbb{E}\psi\left(\frac{|X|}{C}\right) \leq 1\right\}.$$

The random variable X is called a ψ-variable if $\|X\|_\psi$ is finite. The best-known examples of Orlicz norms are those corresponding to the functions $x \mapsto x^p$ for $p \geq 1$; then the Orlicz norm is simply the L_p-norm

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p}.$$

Other interesting Orlicz norms are induced by the functions $\psi_\alpha(x) = e^{x^\alpha} - 1$, $\alpha \geq 1$. Particularly ψ_2 is related to subgaussian tails of X and ψ_1 is related to subexponential tails.

Remark. In [**111**], for a random variable X, the following alternative definitions are given

$$\|X\|'_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}, \tag{3.42}$$

$$\|X\|'_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}. \tag{3.43}$$

Recall that subgaussian random variables were introduced in Section 2.1.3. Indeed, it is shown in [**111**] that the norm in (3.43) is finite if and only if X is *subgaussian*. Particularly, a subgaussian random variable satisfies the tail bound

$$\mathbb{P}(|X| \geq t) \leq \exp\left(1 - \frac{ct^2}{(\|X\|'_{\psi_2})^2}\right),$$

for all $t > 0$ and an absolute constant $c > 0$. Similarly if (3.42) is finite, then X is called *subexponential*. We refer to the $\|\cdot\|'_{\psi_1}$ and $\|\cdot\|'_{\psi_2}$-norms as *subexponential* and *subgaussian norms* respectively. Moreover it is shown in [**111**, Lemma 5.5 and Section 5.2.4] that these norms are equivalent to Orlicz norms $\|\cdot\|_{\psi_1}$ and $\|\cdot\|_{\psi_2}$ respectively up to absolute constants.

The following lemma from [**111**] gives us a useful comparison of subexponential and subgaussian norms.

Lemma 3.20. *A random variable X is subgaussian if and only if X^2 is subexponential. Moreover,*

$$(\|X\|'_{\psi_2})^2 \leq \|X^2\|'_{\psi_1} \leq 2(\|X\|'_{\psi_2})^2.$$

Another result about Orlicz norms considers the ψ-norm of a maximum of finitely many random variables. The following lemma is due to [**109**, Lemma 2.2.2].

Lemma 3.21. *Let ψ_α for $0 < \alpha \leq 2$ be given as before. Then, for any random variables $X_1, \ldots, X_m$,*

$$\big\|\max_{1\leq i\leq m} X_i\big\|_{\psi_\alpha} \leq K(\psi_\alpha)^{-1}(m) \max_{1\leq i\leq m} \big\|X_i\big\|_{\psi_\alpha},$$

for a constant K depending only on ψ_α. Here $(\psi_\alpha)^{-1}$ denotes the inverse function of ψ_α.

In particular for $\alpha = 1$, $(\psi_1)^{-1}(m) = \ln(1+m)$. A similar inequality is also valid for $\psi(x) = x^p$, which induces the L_p-norm of a random variable. Using the fact that $\max_i |X_i|^p \leq \sum_i |X_i|^p$, one obtains

$$\big\|\max_{1\leq i\leq m} X_i\big\|_p = \left(\mathbb{E}\max_{1\leq i\leq m} |X_i|^p\right)^{1/p} \leq m^{1/p} \max_{1\leq i\leq m} \|X_i\|_p, \tag{3.44}$$

see [**109**] for a reference.

3.3.2. Inexact dual certificate

This section gives a sufficient condition for recovery of fixed sparse vectors based on an "inexact dual vector". Sufficient conditions involving an exact dual vector were given in [**55**, **103**]. The modified inexact version is due to Gross [**65**]. Below, $\mathbf{A}_{|\mathcal{H}}$ is the restriction of $\mathbf{A}$ to $\mathcal{H}$.

Lemma 3.22. *Let $A \in \mathbb{R}^{m\times N}, (W_j)_{j=1}^N$ be a fusion frame for $\mathbb{R}^d$ and $x \in \mathcal{H}$ with support S. Assume that*

$$\|[(\mathbf{A_P})_S^*(\mathbf{A_P})_S]^{-1}_{|\mathcal{H}}\| \leq 2 \quad \textit{and} \quad \max_{\ell\in\overline{S}} \|(\mathbf{A_P})_S^*(\mathbf{A_P})_\ell\| \leq 1. \tag{3.45}$$

Suppose there exists a block vector $\mathbf{u} \in \mathbb{R}^{Nd}$ *of the form* $\mathbf{u} = \mathbf{A_P}^*\mathbf{h}$ *with block vector* $\mathbf{h} \in \mathbb{R}^{md}$ *such that*

$$\|\mathbf{u}_S - \operatorname{sgn}(\mathbf{x}_S)\|_2 \leq 1/4 \quad \textit{and} \quad \max_{i \in \overline{S}} \|u_i\|_2 \leq 1/4. \tag{3.46}$$

Then x *is the unique minimizer of* $\|\mathbf{z}\|_{2,1}$ *subject to* $\mathbf{A_P z} = \mathbf{A_P x}$.

PROOF. The proof follows [**54**, Theorem 4.32] and generalizes it to the block vector case. For convenience we give the details here. Let $\hat{\mathbf{x}}$ be a minimizer of $\|\mathbf{z}\|_{2,1}$ subject to $\mathbf{A_P z} = \mathbf{A_P x}$. Then $\mathbf{v} = \hat{\mathbf{x}} - \mathbf{x} \in \mathcal{H}$ satisfies $\mathbf{A_P v} = \mathbf{0}$. We need to show that $\mathbf{v} = \mathbf{0}$. First we observe that

$$\begin{aligned}\|\hat{\mathbf{x}}\|_{2,1} &= \|\mathbf{x}_S + \mathbf{v}_S\|_{2,1} + \|\mathbf{v}_{\overline{S}}\|_{2,1} = \langle \operatorname{sgn}(\mathbf{x}_S + \mathbf{v}_S), (\mathbf{x}_S + \mathbf{v}_S)\rangle + \|\mathbf{v}_{\overline{S}}\|_{2,1} \\ &\geq \langle \operatorname{sgn}(\mathbf{x}_S), (\mathbf{x}_S + \mathbf{v}_S)\rangle + \|\mathbf{v}_{\overline{S}}\|_{2,1} \\ &= \|\mathbf{x}_S\|_{2,1} + \langle \operatorname{sgn}(\mathbf{x}_S), \mathbf{v}_S\rangle + \|\mathbf{v}_{\overline{S}}\|_{2,1}. \end{aligned} \tag{3.47}$$

For $\mathbf{u} = \mathbf{A_P}^*\mathbf{h}$ it holds

$$\langle \mathbf{u}_S, \mathbf{v}_S\rangle = \langle \mathbf{u}, \mathbf{v}\rangle - \langle \mathbf{u}_{\overline{S}}, \mathbf{v}_{\overline{S}}\rangle = \langle \mathbf{h}, \mathbf{A_P v}\rangle - \langle \mathbf{u}_{\overline{S}}, \mathbf{v}_{\overline{S}}\rangle = -\langle \mathbf{u}_{\overline{S}}, \mathbf{v}_{\overline{S}}\rangle.$$

Hence,

$$\begin{aligned}\langle \operatorname{sgn}(\mathbf{x}_S), \mathbf{v}_S\rangle &= \langle \operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S, \mathbf{v}_S\rangle + \langle \mathbf{u}_S, \mathbf{v}_S\rangle \\ &= \langle \operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S, \mathbf{v}_S\rangle - \langle \mathbf{u}_{\overline{S}}, \mathbf{v}_{\overline{S}}\rangle.\end{aligned}$$

The Cauchy-Schwarz inequality together with (3.46) yields

$$|\langle \operatorname{sgn}(\mathbf{x}_S), \mathbf{v}_S\rangle| \leq \|\operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 \|\mathbf{v}_S\|_2 + \max_{i \in \overline{S}} \|u_i\|_2 \|\mathbf{v}_{\overline{S}}\|_{2,1} \leq \frac{1}{4}\|\mathbf{v}_S\|_2 + \frac{1}{4}\|\mathbf{v}_{\overline{S}}\|_{2,1}.$$

Together with (3.47) this yields

$$\|\hat{\mathbf{x}}\|_{2,1} \geq \|\mathbf{x}_S\|_{2,1} - \frac{1}{4}\|\mathbf{v}_S\|_2 + \frac{3}{4}\|\mathbf{v}_{\overline{S}}\|_{2,1}.$$

We now bound $\|\mathbf{v}_S\|_2$. Since $\mathbf{A_P v} = \mathbf{0}$, we have $(\mathbf{A_P})_S \mathbf{v}_S = -(\mathbf{A_P})_{\overline{S}} \mathbf{v}_{\overline{S}}$ and

$$\begin{aligned}\|\mathbf{v}_S\|_2 &= \|[(\mathbf{A_P})_S^*(\mathbf{A_P})_S]_{|\mathcal{H}}^{-1}(\mathbf{A_P})_S^*(\mathbf{A_P})_S \mathbf{v}_S\|_2 = \| - [(\mathbf{A_P})_S^*(\mathbf{A_P})_S]_{|\mathcal{H}}^{-1}(\mathbf{A_P})_S^*(\mathbf{A_P})_{\overline{S}} \mathbf{v}_{\overline{S}}\|_2 \\ &\leq \|[(\mathbf{A_P})_S^*(\mathbf{A_P})_S]_{|\mathcal{H}}^{-1}\| \|(\mathbf{A_P})_S^*(\mathbf{A_P})_{\overline{S}} \mathbf{v}_{\overline{S}}\|_2 \leq 2 \left\| (\mathbf{A_P})_S^* \sum_{i \in \overline{S}} (\mathbf{A_P})_i v_i \right\|_2 \\ &\leq 2 \sum_{i \in \overline{S}} \|(\mathbf{A_P})_S^*(\mathbf{A_P})_i\| \|v_i\|_2 \leq 2\|\mathbf{v}_{\overline{S}}\|_{2,1}. \end{aligned} \tag{3.48}$$

Hereby, we used the second condition in (3.45). Then we have

$$\|\hat{\mathbf{x}}\|_{2,1} \geq \|\mathbf{x}\|_{2,1} + \frac{1}{4}\|\mathbf{v}_{\overline{S}}\|_{2,1}.$$

Since $\hat{\mathbf{x}}$ is an $\ell_{2,1}$-minimizer it follows that $\mathbf{v}_{\overline{S}} = \mathbf{0}$. Therefore $(\mathbf{A_P})_S \mathbf{v}_S = -(\mathbf{A_P})_{\overline{S}} \mathbf{v}_{\overline{S}} = \mathbf{0}$. Since $(\mathbf{A_P})_S$ is injective, it follows that $\mathbf{v}_S = \mathbf{0}$, so that $\mathbf{v} = \mathbf{0}$. □

The next statement makes Lemma 3.22 robust under noise and stable under passing from sparse to compressible vectors. It is an extension of [**54**, Theorem 4.33] and its proof is entirely analogous to the one in there.

Lemma 3.23. *Let $A \in \mathbb{R}^{m\times N}, (W_j)_{j=1}^N$ be a fusion frame for $\mathbb{R}^d$ and $x \in \mathcal{H}$. Let $S \subset [N]$ be the index set of the s largest ℓ_2-normed vectors x_i of $\mathbf{x}$. Assume that, for positive constants $\delta, \beta, \gamma, \theta \in (0,1)$ with $b := \theta + \beta\gamma/(1-\delta) < 1$ and*

$$\|(\mathbf{A_P})_S^* \mathbf{A_P}_S - \mathbf{P}_S\| \leq \delta, \tag{3.49}$$

$$\max_{\ell \in \overline{S}} \|(\mathbf{A_P})_S^* (\mathbf{A_P})_\ell\| \leq \beta. \tag{3.50}$$

Suppose there exists a block vector $\mathbf{u} \in \mathbb{R}^{Nd}$ of the form $\mathbf{u} = \mathbf{A_P}^*\mathbf{h}$ with block vector $\mathbf{h} \in \mathbb{R}^{md}$ such that

$$\|\mathbf{u}_S - \operatorname{sgn}(\mathbf{x}_S)\|_2 \leq \gamma, \tag{3.51}$$

$$\max_{i \in \overline{S}} \|u_i\|_2 \leq \theta, \tag{3.52}$$

$$\|\mathbf{h}\|_2 \leq \tau\sqrt{s}. \tag{3.53}$$

Let noisy measurements $\mathbf{y} = \mathbf{A_P x} + \mathbf{e}$ be given with $\|\mathbf{e}\|_2 \leq \eta$. Then the minimizer $\hat{\mathbf{x}}$ of

$$\min_{\mathbf{z}\in\mathcal{H}} \|\mathbf{z}\|_{2,1} \quad s.t. \quad \|\mathbf{A_P z} - \mathbf{y}\|_2 \leq \eta \tag{3.54}$$

satisfies

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C_1 \sigma_s(\mathbf{x})_1 + (C_2 + C_3\sqrt{s})\eta,$$

where

$$C_1 = \left(1 + \frac{\beta}{1-\delta}\right)\frac{2}{1-b}, \quad C_2 = 2\frac{\sqrt{1+\delta}}{1-\delta} + \left(1 + \frac{\beta}{1-\delta}\right)\frac{2\gamma\sqrt{1+\delta}}{(1-\delta)(1-b)}$$
$$C_3 = \left(1 + \frac{\beta}{1-\delta}\right)\frac{2\tau}{1-b}.$$

PROOF. Let $\hat{\mathbf{x}}$ be minimizer of the optimization program (3.54) and say $\hat{\mathbf{x}} = \mathbf{x} + \mathbf{v}$. Our goal is to bound $\|\mathbf{v}\|_{2,1}$. We have $\|\mathbf{x}\|_{2,1} \geq \|\hat{\mathbf{x}}\|_{2,1}$ by minimality. Then

$$\begin{aligned}\|\mathbf{x}\|_{2,1} &\geq \|\mathbf{x}+\mathbf{v}\|_{2,1} = \|(\mathbf{x}+\mathbf{v})_S\|_{2,1} + \|(\mathbf{x}+\mathbf{v})_{\overline{S}}\|_{2,1} \\ &\geq \langle (\mathbf{x}+\mathbf{v})_S, \operatorname{sgn}(\mathbf{x}_S)\rangle + \|\mathbf{v}_{\overline{S}}\|_{2,1} - \|\mathbf{x}_{\overline{S}}\|_{2,1} \\ &= \|\mathbf{x}_S\|_{2,1} + \langle \mathbf{v}_S, \operatorname{sgn}(\mathbf{x}_S)\rangle + \|\mathbf{v}_{\overline{S}}\|_{2,1} - \|\mathbf{x}_{\overline{S}}\|_{2,1}.\end{aligned}$$

Rearranging and using that $\|\mathbf{x}\|_{2,1} = \|\mathbf{x}_S\|_{2,1} + \|\mathbf{x}_{\overline{S}}\|_{2,1}$ yields

$$\|\mathbf{v}_{\overline{S}}\|_{2,1} \leq |\langle \mathbf{v}_S, \operatorname{sgn}(\mathbf{x}_S)\rangle| + 2\|\mathbf{x}_{\overline{S}}\|_{2,1}. \tag{3.55}$$

We further estimate

$$\begin{aligned}|\langle \mathbf{v}_S, \operatorname{sgn}(\mathbf{x}_S)\rangle| &\leq |\langle \mathbf{v}_S, \operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S\rangle| + |\langle \mathbf{v}_S, \mathbf{u}_S\rangle| \\ &\leq \gamma\|\mathbf{v}_S\|_2 + |\langle \mathbf{v}, \mathbf{u}\rangle| + |\langle \mathbf{v}_{\overline{S}}, \mathbf{u}_{\overline{S}}\rangle|\end{aligned} \tag{3.56}$$

where we used the triangle inequality and the Cauchy-Schwarz inequality together with (3.51). Next we will estimate three terms on the right hand side of the inequality above. It follows from the assumption on the noise vector $\mathbf{e}$ and the constraint in the optimization problem (3.54) that

$$\|\mathbf{A_P v}\|_2 = \|\mathbf{A_P}(\hat{\mathbf{x}} - \mathbf{x})\|_2 \leq \|\mathbf{A_P}\hat{\mathbf{x}} - \mathbf{y}\|_2 + \|\mathbf{y} - \mathbf{A_P x}\|_2 \leq 2\eta.$$

The well-conditionedness assumption (3.49) implies that $\|(\mathbf{A_P})_S\| \leq \sqrt{1+\delta}$ and $\|[(\mathbf{A_P})_S^*(\mathbf{A_P})_S]_{|\mathcal{H}}^{-1}\| \leq (1-\delta)^{-1}$. Then the first term in (3.56) can be estimated as follows

$$\begin{aligned}\|\mathbf{v}_S\|_2 &= \|[(\mathbf{A_P})_S^*(\mathbf{A_P})_S]_{|\mathcal{H}}^{-1}(\mathbf{A_P})_S^*(\mathbf{A_P})_S\mathbf{v}_S\|_2 \leq \frac{1}{1-\delta}\|(\mathbf{A_P})_S^*(\mathbf{A_P})_S\mathbf{v}_S\|_2 \\ &\leq \frac{1}{1-\delta}\|(\mathbf{A_P})_S^*\mathbf{A_P}\mathbf{v}\|_2 + \frac{1}{1-\delta}\|(\mathbf{A_P})_S^*(\mathbf{A_P})_{\overline{S}}\mathbf{v}_{\overline{S}}\|_2 \\ &\leq 2\frac{\sqrt{1+\delta}}{1-\delta}\eta + \frac{\beta}{1-\delta}\|\mathbf{v}_{\overline{S}}\|_{2,1}, \end{aligned} \tag{3.57}$$

where the last step follows from (3.50) in the same way as in (3.48). For the second term we use Condition (3.53) to derive

$$|\langle \mathbf{v}, \mathbf{u}\rangle| = |\langle \mathbf{v}, \mathbf{A_P}^*\mathbf{h}\rangle| = |\langle \mathbf{A_P}\mathbf{v}, \mathbf{h}\rangle| \leq \|\mathbf{A_P}\mathbf{v}\|_2\|\mathbf{h}\|_2 \leq 2\tau\eta\sqrt{s}.$$

Finally Hölder's inequality with (3.52) yields

$$|\langle \mathbf{v}_{\overline{S}}, \mathbf{u}_{\overline{S}}\rangle| \leq \theta\|\mathbf{v}_{\overline{S}}\|_{2,1}.$$

Plugging these estimates into (3.55), we obtain

$$\|\mathbf{v}_{\overline{S}}\|_{2,1} \leq \left(2\gamma\frac{\sqrt{1+\delta}}{1-\delta} + 2\tau\sqrt{s}\right)\eta + \left(\theta + \frac{\beta\gamma}{1-\delta}\right)\|\mathbf{v}_{\overline{S}}\|_{2,1} + 2\|\mathbf{x}_{\overline{S}}\|_{2,1}.$$

Since $\theta + \beta\gamma/(1-\delta) = b < 1$ and $\|\mathbf{x}_{\overline{S}}\|_{2,1} = \sigma_s(\mathbf{x})_1$ a rearrangement yields

$$\|\mathbf{v}_{\overline{S}}\|_{2,1} \leq \frac{2\gamma\frac{\sqrt{1+\delta}}{1-\delta} + 2\tau\sqrt{s}}{1-b}\eta + \frac{2}{1-b}\sigma_s(\mathbf{x})_1.$$

Finally, this inequality together with (3.57) yields

$$\begin{aligned}\|\mathbf{v}\|_2 &\leq \|\mathbf{v}_S\|_2 + \|\mathbf{v}_{\overline{S}}\|_2 \leq \|\mathbf{v}_S\|_2 + \|\mathbf{v}_{\overline{S}}\|_{2,1} \\ &\leq 2\frac{\sqrt{1+\delta}}{1-\delta}\eta + \left(1 + \frac{\beta}{1-\delta}\right)\|\mathbf{v}_{\overline{S}}\|_{2,1} \\ &\leq 2\frac{\sqrt{1+\delta}}{1-\delta}\eta + \left(1 + \frac{\beta}{1-\delta}\right)\left(\frac{2\gamma\frac{\sqrt{1+\delta}}{1-\delta} + 2\tau\sqrt{s}}{1-b}\eta + \frac{2}{1-b}\sigma_s(\mathbf{x})_1\right) \\ &= C_1\sigma_s(\mathbf{x})_1 + (C_2 + C_3\sqrt{s})\eta \end{aligned}$$

with the claimed values of the constants. □

3.3.3. Proof of Theorem 3.1

We introduce the rescaled matrix $\tilde{\mathbf{A}}_\mathbf{P} = \frac{1}{\sqrt{m}}\mathbf{A_P}$. The term $\|[(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S]_{|\mathcal{H}}^{-1}\|$ in (3.45) will be treated with Theorem 3.24 by noticing that $\|(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S\| \leq \delta$ implies $\|[(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S]_{|\mathcal{H}}^{-1}\| \leq (1-\delta)^{-1}$. The other terms in Lemma 3.22 will be estimated by the lemmas in the next section. Throughout the proof, we use the notation $\mathbf{E}_{jj}(A)$ to denote the $s \times s$ block diagonal matrix with the matrix $A \in \mathbb{R}^{d\times d}$ in its j-th diagonal entry and 0 elsewhere.

- *Auxiliary results*

We use the matrix Bernstein inequality in Theorem A.3 due to [**106**] in order to bound $\|(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S\|$. Recall the definition (3.2) of the incoherence matrix Λ from Section 3.1.5.

Theorem 3.24. *Let $A \in \mathbb{R}^{m\times N}$ be a measurement matrix whose entries are i.i.d. Bernoulli random variables and $(W_j)_{j=1}^N$ be a fusion frame with the associated matrix Λ and $\dim(W_j) = k$. Then, for $\delta \in (0,1)$, the block matrix $\tilde{\mathbf{A}}_{\mathbf{P}}$ satisfies*

$$\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\| \le \delta$$

with probability at least $1-\varepsilon$ provided

$$m \ge \delta^{-2}\left(2\|\Lambda^S\|_{2,\infty}^2 + \frac{2}{3}\max\{\|\Lambda^S\|, 1\}\right)\ln(2sk/\varepsilon).$$

We note that the relation $\|\Lambda^S\|_{2,\infty} \le \|\Lambda^S\|$ holds by definition of the norms, see Section 3.1.5.

Proof. Denote $\mathbf{Y}_\ell = (\epsilon_{\ell j}P_j)_{j\in S}$ for $\ell \in [m]$ as the ℓ-th block column vector of $(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*$. Observing that $\mathbb{E}(\mathbf{Y}_\ell\mathbf{Y}_\ell^*)_{j,k} = \mathbb{E}(\epsilon_{\ell j}P_j\epsilon_{\ell k}P_k) = \delta_{jk}P_jP_k$, where $\delta_{jk} = 1$ for $j = k$ and 0 otherwise, we have $\mathbb{E}\mathbf{Y}_\ell\mathbf{Y}_\ell^* = \mathbf{P}_S$. Therefore, we can write

$$(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S = \frac{1}{m}\sum_{\ell=1}^m(\mathbf{Y}_\ell\mathbf{Y}_\ell^* - \mathbb{E}\mathbf{Y}_\ell\mathbf{Y}_\ell^*).$$

This is a sum of independent self-adjoint random matrices. We define the block matrices $\mathbf{X}_\ell := \frac{1}{m}(\mathbf{Y}_\ell\mathbf{Y}_\ell^* - \mathbb{E}\mathbf{Y}_\ell\mathbf{Y}_\ell^*)$ which have mean zero. Moreover,

$$\begin{aligned}\|\mathbf{X}_\ell\| &= \frac{1}{m}\max_{\|\mathbf{x}\|_2=1,\mathbf{x}\in\mathcal{H}}|\langle\mathbf{Y}_\ell\mathbf{Y}_\ell^*\mathbf{x},\mathbf{x}\rangle - \langle\mathbf{P}_S\mathbf{x},\mathbf{x}\rangle| = \frac{1}{m}\max_{\|\mathbf{x}\|_2=1,\mathbf{x}\in\mathcal{H}}\left|\|\mathbf{Y}_\ell^*\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \\ &\le \frac{1}{m}\max\left\{\max_{\|\mathbf{x}\|_2=1,\mathbf{x}\in\mathcal{H}}\|\mathbf{Y}_\ell^*\mathbf{x}\|_2^2 - 1, 1\right\} = \frac{1}{m}\max\left\{\|\mathbf{Y}_\ell\|^2 - 1, 1\right\}.\end{aligned}$$

We further bound the spectral norm of the block matrix $\mathbf{Y}_\ell^*$. We separate a vector $\mathbf{x} \in \mathbb{R}^{S\cdot d}$ into s blocks of length d and denote $\mathbf{x} = (x_i)_{i\in S}$. Defining the vector $\beta \in \mathbb{R}^S$ with $\beta_i = \|x_i\|_2$ we have

$$\begin{aligned}\|\mathbf{Y}_\ell^*\|^2 &= \max_{\|\mathbf{x}\|_2=1}\left\|\sum_{i\in S}\epsilon_iP_ix_i\right\|^2 = \max_{\|\mathbf{x}\|_2=1}\sum_{i,j\in S}\epsilon_i\epsilon_j\langle P_ix_i, P_jx_j\rangle \\ &\le \max_{\|\mathbf{x}\|_2=1}\sum_{i,j\in S}|\langle P_iP_jx_j, x_i\rangle| \le \max_{\|\mathbf{x}\|_2=1}\sum_{i,j\in S}\|P_iP_j\|\|x_i\|_2\|x_j\|_2 \\ &\le \max_{\|\mathbf{x}\|_2=1}\sum_{j\in S}\|x_j\|_2^2 + \max_{\|\beta\|_2=1}\sum_{i\ne j}\|P_iP_j\|\beta_i\beta_j \\ &\le 1 + \max_{\|\beta\|_2=1}\langle\beta,\Lambda\beta\rangle \le 1 + \|\Lambda^S\|. \end{aligned} \tag{3.58}$$

This implies the estimate

$$\|\mathbf{X}_\ell\| \le \frac{\max\{\|\Lambda^S\|, 1\}}{m}.$$

Furthermore,

$$\begin{aligned}\mathbb{E}\mathbf{X}_\ell^2 &= \frac{1}{m^2}\mathbb{E}\left(\mathbf{Y}_\ell\mathbf{Y}_\ell^*\mathbf{Y}_\ell\mathbf{Y}_\ell^* + \mathbf{P}_S - \mathbf{Y}_\ell\mathbf{Y}_\ell^*\mathbf{P}_S - \mathbf{P}_S\mathbf{Y}_\ell\mathbf{Y}_\ell^*\right) \\ &= \mathbb{E}\frac{1}{m^2}\mathbf{Y}_\ell\left(\sum_{j\in S}P_j\right)\mathbf{Y}_\ell^* + \frac{1}{m^2}\mathbf{P}_S - \frac{1}{m^2}\mathbb{E}(\mathbf{Y}_\ell\mathbf{Y}_\ell^*)\mathbf{P}_S - \frac{1}{m^2}\mathbf{P}_S\mathbb{E}(\mathbf{Y}_\ell\mathbf{Y}_\ell^*) \\ &= \frac{1}{m^2}\sum_{i\in S}\mathbf{E}_{ii}\left(P_i\left(\sum_{j\in S}P_j\right)P_i\right) - \frac{1}{m^2}\mathbf{P}_S.\end{aligned}$$

In the first equality above, we used the independence of $\epsilon_{\ell j}$ for $j \in S$ and the fact that $\epsilon_{\ell j}^2 = 1$. The last inequality follows from the relation $\mathbb{E}\epsilon_{\ell j}\epsilon_{\ell k} = 0$ for $j \neq k$, which yields the block diagonal matrix in the last line. Next, we estimate the variance parameter appearing in the noncommutative Bernstein inequality as

$$\begin{aligned}\sigma^2 := \left\|\sum_{\ell=1}^m \mathbb{E}(\mathbf{X}_\ell^2)\right\| &= \frac{1}{m}\left\|\sum_{i\in S}\mathbf{E}_{ii}\left(P_i\left(\sum_{j\in S}P_j\right)P_i\right) - \mathbf{P}_S\right\| \\ &= \frac{1}{m}\left\|\sum_{i\in S}\mathbf{E}_{ii}\left(P_i\left(\sum_{j\in S, j\neq i}P_j\right)P_i\right)\right\| = \frac{1}{m}\max_{i\in S}\left\|P_i\left(\sum_{j\in S, j\neq i}P_j\right)P_i\right\|.\end{aligned}$$

We further estimate,

$$\max_{i\in S}\left\|P_i\left(\sum_{j\in S, j\neq i}P_j\right)P_i\right\| = \max_{i\in S}\left\|\sum_{j\in S, j\neq i}P_iP_jP_i\right\| \leq \max_{i\in S}\sum_{\substack{j\in S\\ j\neq i}}\|P_iP_j\|\|P_jP_i\| = \max_{i\in S}\sum_{\substack{j\in S\\ j\neq i}}\|P_iP_j\|^2.$$

Finally we arrive at

$$\sigma^2 \leq \frac{\|\Lambda^S\|_{2,\infty}^2}{m}.$$

We are now in the position of applying Theorem A.3. This gives

$$\begin{aligned}&\mathbb{P}\left(\|(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S\| > \delta\right) = \mathbb{P}\left(\|\sum_{\ell=1}^m \mathbf{X}_\ell\| > \delta\right) \\ &\leq 2sk\exp\left(-\frac{\delta^2 m/2}{\|\Lambda^S\|_{2,\infty}^2 + \max\{\|\Lambda^S\|,1\}\delta/3}\right) \leq 2sk\exp\left(-\frac{\delta^2 m}{2\|\Lambda^S\|_{2,\infty}^2 + \frac{2}{3}\max\{\|\Lambda^S\|,1\}}\right),\end{aligned} \tag{3.59}$$

where we used that $\delta \in (0,1)$. The careful reader may have noticed that $2sk$ appears in front of the exponential instead of the dimension of $\mathbf{X}_\ell \in \mathbb{R}^{sd\times sd}$ as asked by Theorem A.3. In fact, Theorem A.3 gives a better estimate if the matrices $\mathbb{E}\mathbf{X}_\ell^2$ are not full rank, see (A.2) and the remark after Theorem A.3. Indeed, in our case since $\operatorname{rank}(P_j) = \dim(W_j) = k$, we have $\operatorname{rank}(\mathbb{E}\mathbf{X}_\ell^2) = sk$ which appears in (3.59). Bounding the right hand side of (3.59) by ε completes the proof. □

We now provide modified versions of the auxiliary lemmas in [**54**, Section 12.4].

Lemma 3.25. *Let S be a subset of $[N]$ with cardinality s and $\mathbf{v} \in \mathbb{R}^{S\times d}$ be a block vector of size s with $v_j \in W_j$ for $j \in S$. Assume that $m \geq \|\Lambda_S\|_{2,\infty}^2$ and $\max_{i\in S}\|v_i\|_2 \leq \kappa \leq 1$. Then, for $t > 0$,*

$$\begin{aligned}&\mathbb{P}\left(\max_{\ell\in\overline{S}}\|(\tilde{\mathbf{A}}_\mathbf{P})_\ell^*(\tilde{\mathbf{A}}_\mathbf{P})_S\mathbf{v}\|_2 \geq \frac{\kappa\|\Lambda_S\|_{2,\infty}}{\sqrt{m}} + t\right) \\ &\qquad\qquad \leq N\exp\left(-\frac{t^2 m}{2\kappa^2\|\Lambda_S\|_{2,\infty}^2 + 4\kappa^2\|\Lambda_S\|_\infty + t\kappa\|\Lambda_S\|_\infty}\right).\end{aligned}$$

Proof. Fix $\ell \in \overline{S}$. Observe that for $i \in [m]$, $\epsilon_{i\ell}$ are independent from ϵ_{ij} for $j \in S$. For simplicity we denote the corresponding matrices as $\mathbf{B} = (\tilde{\mathbf{A}}_\mathbf{P})_\ell^*$ and $\mathbf{C} = (\tilde{\mathbf{A}}_\mathbf{P})_S$. The i-th block column and i-th block row are denoted as $\mathbf{B}_i$ and $\mathbf{B}^i$ respectively. Note that

$$(\tilde{\mathbf{A}}_\mathbf{P})_\ell^*(\tilde{\mathbf{A}}_\mathbf{P})_S\mathbf{v} = \sum_{i=1}^m \mathbf{B}_i\mathbf{C}^i\mathbf{v} = \sum_{i=1}^m\sum_{j\in S}\frac{1}{m}\epsilon_{i\ell}\epsilon_{ij}P_\ell P_j v_j \tag{3.60}$$

for $\ell \in \overline{S}$. For convenience we introduce

$$\mathbf{B}_i\mathbf{C}^i = \frac{1}{m}\begin{pmatrix}\epsilon_{i\ell}\epsilon_{i1}P_\ell P_1 & \epsilon_{i\ell}\epsilon_{i2}P_\ell P_2 & \cdots & \epsilon_{i\ell}\epsilon_{is}P_\ell P_s\end{pmatrix},$$

where we assumed $S = [s]$ for simplifying the notation. The sum of independent vectors in (3.60) will be bounded in ℓ_2-norm using the vector valued Bernstein inequality Lemma A.5. Observe that the vectors $\mathbf{B}_i\mathbf{C}^i\mathbf{v}$ have mean zero. Furthermore,

$$\begin{aligned} m\mathbb{E}\|\mathbf{B}_i\mathbf{C}^i\mathbf{v}\|_2^2 &= \frac{1}{m}\mathbb{E}\sum_{j,k\in S}\epsilon_{i\ell}^2\epsilon_{ij}\epsilon_{ik}\langle P_\ell P_j v_j, P_\ell P_k v_k\rangle \\ &= \frac{1}{m}\sum_{j\in S}\|P_\ell P_j v_j\|_2^2 \le \frac{1}{m}\sum_{j\in S}\|P_\ell P_j\|^2\|v_j\|_2^2 \le \frac{\kappa^2}{m}\|\Lambda_S\|_{2,\infty}^2 \end{aligned}$$

where we used $\|v_j\|_2 \le \kappa$. We bound σ^2 appearing in Lemma A.5 simply due to (A.5) by

$$m\sigma^2 \le m\mathbb{E}\|\mathbf{B}_i\mathbf{C}^i\mathbf{v}\|_2^2 \le \frac{\kappa^2}{m}\|\Lambda_S\|_{2,\infty}^2.$$

For the uniform bound, observe that

$$\|\mathbf{B}_i\mathbf{C}^i\mathbf{v}\|_2 = \frac{1}{m}\left\|\sum_{j\in S}\epsilon_{i\ell}\epsilon_{ij}P_\ell P_j v_j\right\|_2 \le \frac{1}{m}\sum_{j\in S}\|P_\ell P_j\|\|v_j\|_2 \le \frac{\kappa}{m}\|\Lambda_S\|_\infty.$$

Then the vector valued Bernstein inequality (A.4) yields

$$\mathbb{P}\left(\|(\tilde{\mathbf{A}}_\mathbf{P})_\ell^*(\tilde{\mathbf{A}}_\mathbf{P})_S\mathbf{v}\|_2 \ge \frac{\kappa\|\Lambda_S\|_{2,\infty}}{\sqrt{m}} + t\right) \le \exp\left(-\frac{t^2/2}{\frac{\kappa^2\|\Lambda_S\|_{2,\infty}^2}{m} + \frac{2\kappa\|\Lambda_S\|_\infty}{m}\frac{\kappa\|\Lambda_S\|_{2,\infty}}{\sqrt{m}} + \frac{t}{3}\frac{\kappa\|\Lambda_S\|_\infty}{m}}\right).$$

Taking the union bound over $\ell \in \overline{S} \subset [N]$ and using that $\frac{\|\Lambda_S\|_{2,\infty}}{\sqrt{m}} \le 1$ yields

$$\mathbb{P}\left(\max_{\ell\in\overline{S}}\|(\tilde{\mathbf{A}}_\mathbf{P})_\ell^*(\tilde{\mathbf{A}}_\mathbf{P})_S\mathbf{v}\|_2 \ge \frac{\kappa\|\Lambda_S\|_{2,\infty}}{\sqrt{m}} + t\right) \le N\exp\left(-\frac{t^2 m}{2\kappa^2\|\Lambda_S\|_{2,\infty}^2 + 4\kappa^2\|\Lambda_S\|_\infty + t\kappa\|\Lambda_S\|_\infty}\right).$$

This completes the proof. □

Next, we prove a similar auxiliary result.

Lemma 3.26. *Let S be subset of $[N]$ with cardinality s and $\mathbf{v} \in \mathbb{R}^{S\times d}$ be a block vector of size s with $v_j \in W_j$ for $j \in S$. Assume that $m \ge \|\Lambda^S\|_{2,\infty}^2$. Then, for $t > 0$,*

$$\begin{aligned}\mathbb{P}\left(\|[(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S]\mathbf{v}\|_2 \ge \left(\frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}} + t\right)\|\mathbf{v}\|_2\right) \\ \le \exp\left(-\frac{mt^2}{8 + 4\|\Lambda^S\|_\infty + 2\|\Lambda^S\|_{2,\infty}^2 + t(\frac{4}{3} + \frac{2}{3}\|\Lambda^S\|_\infty)}\right).\end{aligned}$$

Proof. As in the proof of Theorem 3.24, we rewrite the term that we need to bound as

$$[(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S]\mathbf{v} = \frac{1}{m}\sum_{\ell=1}^m(\mathbf{Y}_\ell\mathbf{Y}_\ell^* - \mathbf{P}_S)\mathbf{v} \tag{3.61}$$

where $\mathbf{Y}_\ell = (\epsilon_{\ell i}P_i)_{i\in S}$ is the ℓ-th block row of $(\tilde{\mathbf{A}}_\mathbf{P})_S$. We use the vector valued Bernstein inequality, Lemma A.5 once again, in order to estimate the ℓ_2-norm of this sum. Observe that

$\mathbb{E}(\mathbf{Y}_\ell \mathbf{Y}_\ell^* - \mathbf{P}_S)\mathbf{v} = \mathbf{0}$ as in the proof of Theorem 3.24. Furthermore, denoting

$$Z = \left\| \frac{1}{m} \sum_{\ell=1}^{m} (\mathbf{Y}_\ell \mathbf{Y}_\ell^* - \mathbf{P}_S)\mathbf{v} \right\|_2,$$

we have

$$\begin{aligned}
\mathbb{E}Z^2 &= m\mathbb{E} \left\| \frac{1}{m} (\mathbf{Y}_1 \mathbf{Y}_1^* - \mathbf{P}_S)\mathbf{v} \right\|_2^2 \\
&= \frac{1}{m} \mathbb{E} \langle (\mathbf{Y}_1 \mathbf{Y}_1^* - \mathbf{P}_S)\mathbf{v}, (\mathbf{Y}_1 \mathbf{Y}_1^* - \mathbf{P}_S)\mathbf{v} \rangle \\
&= \frac{1}{m} \mathbb{E} (\langle \mathbf{Y}_1 \mathbf{Y}_1^* \mathbf{v}, \mathbf{Y}_1 \mathbf{Y}_1^* \mathbf{v} \rangle - 2\langle \mathbf{Y}_1 \mathbf{Y}_1^* \mathbf{v}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle) \\
&= \frac{1}{m} \mathbb{E} (\langle \mathbf{Y}_1 \mathbf{Y}_1^* \mathbf{Y}_1 \mathbf{Y}_1^* \mathbf{v}, \mathbf{v} \rangle - 2\|\mathbf{Y}_1^* \mathbf{v}\|_2^2 + 1).
\end{aligned}$$

We now estimate the first two terms in the last line above. First observe that due to $\mathbf{Y}_1^* \mathbf{Y}_1 = \sum_{i \in S} P_i$, it holds

$$\begin{aligned}
\mathbb{E}\langle \mathbf{Y}_1 \mathbf{Y}_1^* \mathbf{Y}_1 \mathbf{Y}_1^* \mathbf{v}, \mathbf{v} \rangle &= \langle \mathbb{E}(\mathbf{Y}_1 \sum_{i \in S} P_i \mathbf{Y}_1^*)\mathbf{v}, \mathbf{v} \rangle = \langle \sum_{j \in S} \mathbf{E}_{jj} (P_j \sum_{i \in S} P_i P_j)\mathbf{v}, \mathbf{v} \rangle \\
&= \sum_{j \in S} \langle P_j \sum_{i \in S} P_i P_j v_j, v_j \rangle = \sum_{i,j \in S} \langle P_j P_i P_j v_j, v_j \rangle \\
&\leq \sum_{i,j, i \neq j} \|P_i P_j\|^2 \|v_j\|_2^2 + \sum_{j \in S} \|v_j\|_2^2 \leq \left(\sum_{j \in S} \|v_j\|_2^2 \sum_{i \neq j} \|P_i P_j\|^2 \right) + 1 \\
&\leq 1 + \left(\max_{j \in S} \sum_{i \neq j} \|P_i P_j\|^2 \right) \sum_{j \in S} \|v_j\|_2^2 \leq 1 + \|\Lambda^S\|_{2,\infty}^2,
\end{aligned}$$

where we used that Λ^S is symmetric and $\|\mathbf{v}\|_2 = 1$. Secondly, since

$$\mathbb{E}\|\mathbf{Y}_1^* \mathbf{v}\|_2^2 = \mathbb{E}\| \sum_{i \in S} \epsilon_i P_i v_i \|_2^2 = \mathbb{E} \sum_{i,j \in S} \epsilon_i \epsilon_j \langle v_i, v_j \rangle = \sum_{i \in S} \|v_i\|_2^2 = 1,$$

we obtain

$$\mathbb{E}Z^2 \leq \frac{\|\Lambda^S\|_{2,\infty}^2}{m}.$$

For the uniform bound, we have

$$\begin{aligned}
\frac{1}{m} \|(\mathbf{Y}_\ell \mathbf{Y}_\ell^* - \mathbf{P}_S)\mathbf{v}\|_2 &\leq \frac{1}{m} \|\mathbf{Y}_\ell \mathbf{Y}_\ell^*\| \|\mathbf{v}\|_2 + \frac{1}{m} \|\mathbf{v}\|_2 \\
&= \frac{1}{m} \|\mathbf{Y}_\ell\|^2 + \frac{1}{m} \leq \frac{2 + \|\Lambda^S\|_\infty}{m}.
\end{aligned}$$

The last inequality follows from (3.58) and $\|\Lambda^S\| \leq \|\Lambda^S\|_\infty$, see [**54**, Lemma A.8] for a reference. Finally we estimate the weak variance simply by the strong variance

$$m\sigma^2 \leq \mathbb{E}Z^2 \leq \frac{\|\Lambda^S\|_{2,\infty}^2}{m}.$$

Then the ℓ_2-valued Bernstein inequality (A.4) yields

$$\mathbb{P}\left(\|[(\tilde{\mathbf{A}}_\mathbf{P})_S^* (\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S]\mathbf{v}\|_2 \geq \left(\frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}} + t \right) \|\mathbf{v}\|_2 \right)$$

$$\leq \exp\left(-\frac{t^2/2}{\frac{\|\Lambda^S\|_{2,\infty}^2}{m}+\frac{(4+2\|\Lambda^S\|_\infty)}{m}\frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}}+\frac{t(2+\|\Lambda^S\|_\infty)}{3m}}\right).$$

Using that $\frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}} \leq 1$, we obtain

$$\mathbb{P}\left(\|[(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S]\mathbf{v}\|_2 \geq \left(\frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}}+t\right)\|\mathbf{v}\|_2\right)$$
$$\leq \exp\left(-\frac{mt^2}{8+4\|\Lambda^S\|_\infty+2\|\Lambda^S\|_{2,\infty}^2+t(\frac{4}{3}+\frac{2}{3}\|\Lambda^S\|_\infty)}\right).$$

This completes the proof. □

Lemma 3.26 shows that the multiplication with $(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S$ decreases the ℓ_2-norm of the vectors with high probability. The next lemma shows that this is true for the $\ell_{2,\infty}$-norm as well.

Lemma 3.27. *Assume the conditions of Lemma 3.26. Then, for $t > 0$,*

$$\mathbb{P}\left(\|[(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S]\mathbf{v}\|_{2,\infty} \geq \left(\frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}}+t\right)\|\mathbf{v}\|_{2,\infty}\right)$$
$$\leq s\cdot\exp\left(-\frac{mt^2}{4\|\Lambda^S\|_\infty+2\|\Lambda^S\|_{2,\infty}^2+\frac{2}{3}t\|\Lambda^S\|_\infty}\right).$$

Proof. We assume that $\|\mathbf{v}\|_{2,\infty} = \max_{i\in S}\|v_i\|_2 = 1$ by normalizing $\mathbf{v}$ by $\|\mathbf{v}\|_{2,\infty}$. As in (3.61), we write

$$\mathbf{Z} := [(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S]\mathbf{v} = \frac{1}{m}\sum_{\ell=1}^m(\mathbf{Y}_\ell\mathbf{Y}_\ell^* - \mathbf{P}_S)\mathbf{v},$$

where $\mathbf{Y}_\ell = (\epsilon_{\ell i}P_i)_{i\in S}$ is the ℓ-th row of $(\tilde{\mathbf{A}}_{\mathbf{P}})_S$. We can further write for $i \in S$

$$Z_i = \sum_{\ell=1}^m\frac{1}{m}\sum_{\substack{j\in S\\ j\neq i}}\epsilon_{\ell i}\epsilon_{\ell j}P_iP_jv_j =: \sum_\ell X_\ell.$$

The vectors X_ℓ are independent, thus we use Lemma A.5 in order to bound $\|Z_i\|_2$. Since we have done similar estimations in the previous proofs, we skip some steps and obtain

$$\mathbb{E}\|Z_i\|_2^2 = m\mathbb{E}\|X_\ell\|_2^2 \leq \frac{1}{m}\sum_{\substack{j\in S\\ j\neq i}}\|P_iP_j\|^2\|v_j\|_2^2 \leq \frac{1}{m}\|\Lambda^S\|_{2,\infty}^2,$$

where we used that $\|v_j\|_2 \leq 1$. Furthermore, $m\sigma^2 \leq \frac{1}{m}\|\Lambda^S\|_{2,\infty}^2$. For any $\ell \in [m]$ we have the uniform bound

$$\|X_\ell\|_2 = \frac{1}{m}\left\|\sum_{j\in S,j\neq i}\epsilon_{\ell i}\epsilon_{\ell j}P_iP_jv_j\right\|_2 \leq \frac{1}{m}\sum_{j\in S,j\neq i}\|P_iP_j\|\|v_j\|_2 \leq \frac{1}{m}\|\Lambda^S\|_\infty.$$

Combining these with Lemma A.5 and taking the union bound yield

$$\mathbb{P}\left(\max_{i\in S}\|Z_i\|_2 \geq \frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}}+t\right) \leq s\cdot\exp\left(-\frac{mt^2}{4\|\Lambda^S\|_\infty+2\|\Lambda^S\|_{2,\infty}^2+\frac{2}{3}t\|\Lambda^S\|_\infty}\right).$$

□

Lastly we present the following lemma before the proof of our main result.

Lemma 3.28. *For $t \in (0, \frac{3}{2})$,*

$$\mathbb{P}(\max_{i\in\overline{S}} \|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\| \geq t) \leq 2(s+1)Nk \exp\left(-\frac{t^2 m}{3\|\Lambda_S\|_{2,\infty}^2}\right).$$

PROOF. Fix $i \in \overline{S}$. Similarly as before, we write $(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i$ as a sum of independent matrices,

$$(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i = \frac{1}{m}\sum_{\ell=1}^m \begin{pmatrix} \epsilon_{\ell 1}\epsilon_{\ell i} P_1 P_i \\ \epsilon_{\ell 2}\epsilon_{\ell i} P_2 P_i \\ \vdots \\ \epsilon_{\ell s}\epsilon_{\ell i} P_s P_i \end{pmatrix} =: \frac{1}{m}\sum_{\ell=1}^m \mathbf{Y}_\ell, \tag{3.62}$$

where we assumed $S = [s]$ for simplifying the notation. Above we introduced the block column vectors $\mathbf{Y}_\ell \in \mathbb{R}^{sd\times d}$ which are independent and identically distributed rectangular matrices. Observe also that $\mathbb{E}\mathbf{Y}_\ell = \mathbf{0}$. In order to estimate the norm of the sum in (3.62) we will employ Theorem A.4 [**106**] which is a version of the noncommutative Bernstein inequality for rectangular matrices. We first bound the variance parameter

$$\sigma^2 = \max\left\{ \|\sum_{\ell=1}^m \frac{1}{m^2}\mathbb{E}[\mathbf{Y}_\ell \mathbf{Y}_\ell^*]\|, \|\sum_{\ell=1}^m \frac{1}{m^2}\mathbb{E}[\mathbf{Y}_\ell^* \mathbf{Y}_\ell]\| \right\}. \tag{3.63}$$

We write $\mathbb{E}[\mathbf{Y}_\ell \mathbf{Y}_\ell^*] = \sum_{j\in S} \mathbf{E}_{jj}(P_j P_i P_j)$. The first term on the right hand side of (3.63) is estimated as

$$\|\frac{1}{m^2}\sum_{\ell=1}^m \mathbb{E}[\mathbf{Y}_\ell \mathbf{Y}_\ell^*]\| = \frac{1}{m}\max_{j\in S}\|P_j P_i P_j\| \leq \frac{1}{m}\max_{j\in S}\|P_j P_i\|\|P_i P_j\| \leq \frac{\lambda^2}{m}. \tag{3.64}$$

We used that $P_i^2 = P_i$ and $i \notin S$. Furthermore, $\mathbb{E}[\mathbf{Y}_\ell^* \mathbf{Y}_\ell] = \sum_{j\in S} P_i P_j P_i$ and

$$\frac{1}{m^2}\|\sum_{\ell=1}^m \mathbb{E}[\mathbf{Y}_\ell^* \mathbf{Y}_\ell]\| = \frac{1}{m}\|\sum_{j\in S} P_i P_j P_i\| \leq \frac{1}{m}\sum_{j\in S}\|P_i P_j\|\|P_j P_i\| \leq \frac{\|\Lambda_S\|_{2,\infty}^2}{m}. \tag{3.65}$$

Since (3.65) dominates (3.64), we have

$$\sigma^2 \leq \frac{\|\Lambda_S\|_{2,\infty}^2}{m}.$$

For the uniform bound we obtain

$$\begin{aligned}\|\mathbf{Y}_\ell\|^2 &= \sup_{\substack{\|x\|_2\leq 1\\ x\in\mathbb{R}^d}} \|\mathbf{Y}_\ell x\|_2^2 = \sup_{\substack{\|x\|_2\leq 1\\ x\in\mathbb{R}^d}} \sum_{j\in S}\|P_j P_i x\|_2^2 \\ &\leq \sup_{\substack{\|x\|_2\leq 1\\ x\in\mathbb{R}^d}} \sum_{j\in S}\|P_j P_i\|^2\|x\|_2^2 \leq \|\Lambda_S\|_{2,\infty}^2.\end{aligned}$$

We conclude that $\frac{1}{m}\|\mathbf{Y}_\ell\| \leq \frac{\|\Lambda_S\|_{2,\infty}}{m}$. Combining these estimates, Theorem A.4 yields

$$\mathbb{P}(\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\| \geq t) \leq 2(s+1)k \exp\left(-\frac{t^2/2}{\frac{\|\Lambda_S\|_{2,\infty}^2}{m} + \frac{t}{3}\frac{\|\Lambda_S\|_{2,\infty}}{m}}\right).$$

Taking the union bound over $i \in \overline{S} \subset [N]$ and using that $t \in (0, \frac{3}{2})$ yields

$$\mathbb{P}(\max_{i \in \overline{S}} \|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\| \geq t) \leq 2(s+1)Nk \exp\left(-\frac{t^2 m}{3\|\Lambda_S\|_{2,\infty}^2}\right).$$

Above, the dimension of the subspaces k appears instead the ambient dimension d for the same reasons explained in the proof of Theorem 3.24. This completes the proof. □

• *Proof of Theorem 3.1 continued*

Essentially we follow the arguments in [**54**, Section 12.4]. We will construct an inexact dual vector as in Lemma 3.22 satisfying the conditions there. To this end, we will use the so-called *golfing scheme* due to Gross [**65**]. We partition the m independent (block) rows of $\mathbf{A_P}$ into L disjoint blocks of sizes $m_1, \ldots, m_L$ and L to be specified later with $m = \sum_{j=1}^{L} m_j$. These blocks correspond to row submatrices of $\mathbf{A_P}$ which are denoted by $\mathbf{A_P}^{(1)} \in \mathbb{R}^{m_1 d \times Nd}, \ldots, \mathbf{A_P}^{(L)} \in \mathbb{R}^{m_L d \times Nd}$, i.e.,

$$\mathbf{A_P} = \begin{pmatrix} \mathbf{A_P}^{(1)} \\ \mathbf{A_P}^{(2)} \\ \vdots \\ \mathbf{A_P}^{(L)} \end{pmatrix} \begin{matrix} \}m_1 \\ \}m_2 \\ \vdots \\ \}m_L \end{matrix}$$

Set $S = \mathrm{supp}(x)$. The golfing scheme starts with $\mathbf{u}^{(0)} = \mathbf{0}$ and then inductively defines

$$\mathbf{u}^{(n)} = \frac{1}{m_n}(\mathbf{A_P}^{(n)})^*(\mathbf{A_P}^{(n)})_S(\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S^{(n-1)}) + \mathbf{u}^{(n-1)},$$

for $n = 1, \ldots, L$. The vector $\mathbf{u} = \mathbf{u}^{(L)}$ will serve as a candidate for the inxeact dual vector in Lemma 3.22. Thus, we need to check if it satisfies the two conditions in (3.46). By construction $\mathbf{u}$ is in the row space of $\mathbf{A_P}$, i.e., $\mathbf{u} = \mathbf{A_P}^*\mathbf{h}$ for some vector $\mathbf{h}$ as required in Lemma 3.22. To simplify the notation we introduce $\mathbf{w}^{(n)} = \mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S^{(n)}$. Observe that

$$\begin{aligned} \mathbf{u}_S^{(n)} - \mathbf{u}_S^{(n-1)} &= \frac{1}{m_n}(\mathbf{A_P}^{(n)})_S^*(\mathbf{A_P}^{(n)})_S(\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S^{(n-1)}) \\ \mathbf{w}^{(n-1)} - \mathbf{w}^{(n)} &= \frac{1}{m_n}(\mathbf{A_P}^{(n)})_S^*(\mathbf{A_P}^{(n)})_S\mathbf{w}^{(n-1)} \\ \mathbf{w}^{(n)} &= \left[\mathbf{P}_S - \frac{1}{m_n}(\mathbf{A_P}^{(n)})_S^*(\mathbf{A_P}^{(n)})_S\right]\mathbf{w}^{(n-1)}. \end{aligned} \tag{3.66}$$

Above we used that $\mathbf{P}_S\mathbf{w}^{(n)} = \mathbf{w}^{(n)}$. Furthermore we have

$$\begin{aligned} \mathbf{u}^{(n)} - \mathbf{u}^{(n-1)} &= \frac{1}{m_n}(\mathbf{A_P}^{(n)})^*(\mathbf{A_P}^{(n)})_S\mathbf{w}^{(n-1)} \\ \mathbf{u} = \mathbf{u}^{(L)} &= \sum_{n=1}^{L} \frac{1}{m_n}(\mathbf{A_P}^{(n)})^*(\mathbf{A_P}^{(n)})_S\mathbf{w}^{(n-1)}, \end{aligned} \tag{3.67}$$

where the last line follows by a telescopic sum. We will later show that the matrices

$$\mathbf{P}_S - \frac{1}{m_k}(\mathbf{A_P}^{(k)})_S^*(\mathbf{A_P}^{(k)})_S$$

are contractions and the norm of the residual vector $\mathbf{w}^{(n)}$ decreases geometrically fast, thus $\mathbf{u}^{(n)}$ becomes close to $\mathrm{sgn}(\mathbf{x}_S)$ on its support set S. Particularly, we will prove that $\|\mathbf{w}^{(L)}\|_2 \leq 1/4$ for a suitable choice of L. In addition we also need that the off-support part of $\mathbf{u}$ remains small as well, satisfying the condition $\max_{i \in \overline{S}} \|u_i\|_2 \leq 1/4$.

For these tasks, we will use the lemmas proven above. For the moment, we assume that the following holds for each n with high probability

$$\|\mathbf{w}^{(n)}\|_{2,\infty} \leq \left(\frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}} + q_n \right) \|\mathbf{w}^{(n-1)}\|_{2,\infty}, \quad n \in [L]. \tag{3.68}$$

Let $q_n' := \frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}} + q_n$. Since $\|\mathbf{w}^{(0)}\|_{2,\infty} = \|\mathrm{sgn}(\mathbf{x}_S)\|_{2,\infty} = 1$, we have

$$\|\mathbf{w}^{(n)}\|_{2,\infty} \leq \prod_{j=1}^{n} q_j' =: h_n.$$

Further assume that the following inequalities hold for each n with high probability,

$$\|\mathbf{w}^{(n)}\|_2 \leq \left(\frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}} + r_n \right) \|\mathbf{w}^{(n-1)}\|_2, \quad n \in [L], \tag{3.69}$$

$$\max_{i \in \overline{S}} \left\| \frac{1}{m_n} (\mathbf{A_P}^{(n)})_i^* (\mathbf{A_P}^{(n)})_S \mathbf{w}^{(n-1)} \right\|_2 \leq \frac{h_n \|\Lambda_S\|_{2,\infty}}{\sqrt{m}} + t_n, \quad n \in [L]. \tag{3.70}$$

The parameters q_n, r_n, t_n will be specified later. Now let $r_n' := \frac{\|\Lambda^S\|_{2,\infty}}{\sqrt{m}} + r_n$ and $t_n' := \frac{h_n \|\Lambda_S\|_{2,\infty}}{\sqrt{m}} + t_n$. Then the relations in (3.66) and (3.69) yield

$$\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 = \|\mathbf{w}^{(L)}\|_2 \leq \|\mathrm{sgn}(\mathbf{x}_S)\|_2 \prod_{n=1}^{L} r_n' \leq \sqrt{s} \prod_{n=1}^{L} r_n'.$$

Furthermore, (3.67) and (3.70) give

$$\begin{aligned} \max_{i \in \overline{S}} \|u_i\|_2 &= \max_{i \in \overline{S}} \left\| \sum_{n=1}^{L} \frac{1}{m_n} (\mathbf{A_P}^{(n)})_i^* (\mathbf{A_P}^{(n)})_S \mathbf{w}^{(n-1)} \right\|_2 \\ &\leq \sum_{n=1}^{L} \max_{i \in \overline{S}} \left\| \frac{1}{m_n} (\mathbf{A_P}^{(n)})_i^* (\mathbf{A_P}^{(n)})_S \mathbf{w}^{(n-1)} \right\|_2 \\ &\leq \sum_{n=1}^{L} t_n'. \end{aligned}$$

Next we define the probabilities $p_0(n)$, $p_1(n)$ and $p_2(n)$ that (3.68), (3.69) and (3.70) do not hold respectively. Then by Lemma 3.27 and independence of the blocks,

$$p_0(n) \leq \varepsilon,$$

provided

$$m_n \geq \left(\frac{4\|\Lambda^S\|_\infty + 2\|\Lambda^S\|_{2,\infty}^2}{q_n^2} + \frac{2\|\Lambda^S\|_\infty}{3q_n} \right) \ln(s/\varepsilon). \tag{3.71}$$

Also by Lemma 3.26 and independence of the blocks,

$$p_1(n) \leq \varepsilon$$

provided

$$m_n \geq \left(\frac{8 + 4\|\Lambda^S\|_\infty + 2\|\Lambda^S\|_{2,\infty}^2}{r_n^2} + \frac{4/3 + (2/3)\|\Lambda^S\|_\infty}{r_n} \right) \ln(\varepsilon^{-1}). \tag{3.72}$$

Similarly, due to Lemma 3.25 and independence of the blocks,

$$p_2(n) \leq \varepsilon,$$

provided

$$m_n \geq \left(\frac{2h_n^2 \|\Lambda_S\|_{2,\infty}^2 + 4h_n^2 \|\Lambda_S\|_\infty}{t_n^2} + \frac{h_n \|\Lambda_S\|_\infty}{t_n} \right) \ln(N/\varepsilon). \tag{3.73}$$

We now set the parameters L, m_n, t_n, r_n, q_n for $n \in [L]$ such that $\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 \leq 1/4$ and $\max_{i \in \overline{S}} \|u_i\|_2 \leq 1/4$ as required in the Lemma 3.22. We choose

$$\begin{aligned} L &= \lceil \ln(s)/\ln\ln(N) \rceil + 3, \\ m_n &\geq c(1 + \|\Lambda_S\|_\infty) \ln(N) \ln(2L\varepsilon^{-1}), \\ r_n &= \frac{1}{4\sqrt{\ln(N)}}, \\ t_n &= \frac{1}{2^{n+3}}, \\ q_n &= \frac{1}{8}. \end{aligned}$$

We can estimate each of $\|\Lambda^S\|_{2,\infty}^2, \|\Lambda_S\|_{2,\infty}^2, \|\Lambda^S\|_\infty$ by $\|\Lambda_S\|_\infty$ from above, due to the relation (3.3) in Section 3.1.5. Then by definitions of r_n', t_n', h_n, q_n', we have $r_n' \leq \frac{1}{2\sqrt{\ln N}}$, $q_n' \leq \frac{1}{2}$, $h_n \leq \frac{1}{2^n}$ and $t_n' \leq \frac{1}{2^{n+2}}$ for $n = 1, \ldots, L$. Furthermore,

$$\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 \leq \sqrt{s} \prod_{n=1}^{L} r_n' \leq \frac{1}{4}, \tag{3.74}$$

and

$$\max_{i \in \overline{S}} \|u_i\|_2 \leq \sum_{n=1}^{L} t_n' \leq \frac{1}{4}. \tag{3.75}$$

Next we bound the failure probabilities according to our choices of parameters above. Considering also the conditions (3.71), (3.72) and (3.73), we have $p_0(n), p_1(n), p_2(n) \leq \varepsilon/L$. These yield

$$\sum_{n=1}^{L} p_0(n) + p_1(n) + p_2(n) \leq 3\varepsilon.$$

The overall number of samples obey

$$m = \sum_{n=1}^{L} m_n \geq 2c(1 + \|\Lambda_S\|_\infty) L \ln(N) \ln(L/\varepsilon). \tag{3.76}$$

This is already very close to the proposed condition in the statement of our theorem. We will strengthen this condition later. Next we look into the first part of Condition (3.45) of Lemma 3.22. By Theorem 3.24 $\|(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S\| \leq 1/2$ with probability at least $1 - \varepsilon$ provided

$$m \geq \left(8\|\Lambda^S\|_{2,\infty}^2 + \frac{8}{3} \max\{\|\Lambda^S\|, 1\} \right) \ln(2sk/\varepsilon). \tag{3.77}$$

This implies that $\|[(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S]_{|\mathcal{H}}^{-1}\| \leq 2$. For the second part of Condition (3.45) we use Lemma 3.28. It says that

$$\mathbb{P}(\max_{i\in\overline{S}} \|(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_i\| \geq t) \leq 2(s+1)Nk \exp\left(-\frac{t^2 m}{3\|\Lambda_S\|_{2,\infty}^2}\right).$$

Taking $t = 1$ implies that

$$\max_{i\in\overline{S}} \|(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_i\| \leq 1$$

with probability at least $1-\varepsilon$ provided

$$m \geq 6\|\Lambda_S\|_{2,\infty}^2 \ln(N(s+1)k/\varepsilon). \tag{3.78}$$

Altogether we have shown that Conditions (3.45) and (3.46) of Lemma 3.22 hold simultaneously with probability at least $1-5\varepsilon$ provided Conditions (3.76), (3.77) and (3.78) hold. Replacing ε by $\varepsilon/5$, the main condition of Theorem 3.1

$$m \geq C(1+\|\Lambda_S\|_\infty)\ln(N)\ln(sk)\ln(\varepsilon^{-1})$$

implies all three conditions above with an appropriate constant C since $\|\Lambda^S\| \leq \|\Lambda^S\|_\infty \leq \|\Lambda_S\|_\infty$ since Λ is symmetric. This ends the proof of our theorem. □

Remark. The inexact dual method yields a relatively long and technical proof for sparse recovery and involves several auxiliary results. Other methods used in the CS literature prove to be hard to apply for our particular case where we work with the block matrix $\mathbf{A_P}$ which is more structured than a purely random Gaussian matrix. The main difficulty is to obtain the parameter λ appearing in the recovery condition. Here we name a few of other methods that fail to do so. The first one is the exact dual approach developed by J.-J. Fuchs [**55**] that is used in Chapter 2 for subgaussian matrices. In particular, the exact dual approach involves taking the pseudo-inverse of $\mathbf{A_P}$ in order to obtain the dual certificate (see Section 2.2.1) which loses the structure given by projection matrices P_i. This structure is crucial because it allows us to prove our results involving the incoherence parameter λ. As we will see in the next section, in the inexact dual approach, the candidate for the dual vector is defined via an iterative process involving the row submatrices of $\mathbf{A_P}$ rather than inversing it.

Another approach to prove nonuniform recovery is using convex geometry [**28**] which is briefly reviewed in Section 2.4. This method gives a rather general result with respect to an 'atomic norm' approach which covers many different problems other than sparse recovery. In particular, it uses Gordon's lemma [**62**] and is restricted to only Gaussian matrices. After adapting Gordon's lemma to our particular case with the block matrix $\mathbf{A_P}$, even though one could obtain the optimal scaling λs in the number of measurements m, the ambient dimension d appears as a linear factor in m, which is suboptimal. A similar approach is used for the uniform recovery for fusion frames to obtain the result in Theorem 3.8, see Section 3.2.2.

3.3.4. Proof of Theorem 3.2

• *Auxiliary results*

We recall the rescaled matrix $\tilde{\mathbf{A}}_\mathbf{P} = \frac{1}{\sqrt{m}}\mathbf{A_P}$. The following result follows as a corollary of Theorem 3.24 by observing that $\|\Lambda^S\|_{2,\infty}^2 \leq \lambda^2 s$ and $\|\Lambda^S\| \leq \lambda s$ with $\lambda \in [0,1]$.

Corollary 3.29. *Let $A \in \mathbb{R}^{m\times N}$ be a measurement matrix whose entries are i.i.d. Bernoulli random variables and $(W_j)_{j=1}^N$ be a fusion frame with $\dim(W_j) = k$. Then, for $\delta \in (0,1)$, the block matrix $\tilde{\mathbf{A}}_{\mathbf{P}}$ defined above satisfies*

$$\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\| \le \delta$$

with probability at least $1-\varepsilon$ provided

$$m \ge \frac{8}{3}\delta^{-2}(1+\lambda s)\ln(2sk/\varepsilon).$$

Lemma 3.30. *Let S be subset of $[N]$ with cardinality s and $\mathbf{v} \in \mathbb{R}^{S\times d}$ be a block vector of size s with $v_j \in W_j$ for $j \in S$. Assume that $m \ge \lambda s$. Then, for $t > 0$,*

$$\mathbb{P}\left(\max_{\ell\in\overline{S}} \|(\tilde{\mathbf{A}}_{\mathbf{P}})_\ell^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S\mathbf{v}\|_2 \ge \left(\frac{\lambda}{\sqrt{m}} + t\right)\|\mathbf{v}\|_2\right) \le N\exp\left(-\frac{t^2 m}{6\lambda + t\lambda\sqrt{s}}\right).$$

PROOF. Fix $\ell \in \overline{S}$. We may assume without loss of generality that $\|\mathbf{v}\|_2 = 1$. We see the block matrices multiplied in the following matrix diagram.

$$(\tilde{\mathbf{A}}_{\mathbf{P}})_\ell^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S = \frac{1}{m}\begin{pmatrix}\tilde{\epsilon}_1 P_\ell & \tilde{\epsilon}_2 P_\ell & \cdots & \tilde{\epsilon}_m P_\ell\end{pmatrix}\begin{pmatrix}\epsilon_{11}P_1 & \epsilon_{12}P_2 & \dots & \epsilon_{1s}P_s\\ \epsilon_{21}P_1 & \epsilon_{22}P_2 & \dots & \\ & \ddots & & \\ \epsilon_{m1}P_1 & \epsilon_{m2}P_2 & \dots & \epsilon_{ms}P_s\end{pmatrix}$$

where we assumed $S = [s]$ for simplifying the notation. Here $\tilde{\epsilon}$ denotes $(\epsilon_{i\ell})_{i=1}^m$. Observe that $\tilde{\epsilon}_i$ for $i \in [m]$ are independent from ϵ_{ij}. For simplicity we label those matrices as $\mathbf{B} = (\tilde{\mathbf{A}}_{\mathbf{P}})_\ell^*$ and $\mathbf{C} = (\tilde{\mathbf{A}}_{\mathbf{P}})_S$. We also denote the ℓ-th block column and ℓ-th block row as $\mathbf{B}_i$ and $\mathbf{B}^i$ respectively. Note that

$$(\tilde{\mathbf{A}}_{\mathbf{P}})_\ell^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S\mathbf{v} = \sum_{i=1}^m \mathbf{B}_i\mathbf{C}^i\mathbf{v} = \sum_{i=1}^m\sum_{j\in S}\frac{1}{m}\tilde{\epsilon}_i\epsilon_{ij}P_\ell P_j v_j \tag{3.79}$$

for $\ell \in \overline{S}$. For convenience we give the following

$$\mathbf{B}_i\mathbf{C}^i = \frac{1}{m}\begin{pmatrix}\tilde{\epsilon}_i\epsilon_{i1}P_\ell P_1 & \tilde{\epsilon}_i\epsilon_{i2}P_\ell P_2 & \cdots & \tilde{\epsilon}_i\epsilon_{is}P_\ell P_s\end{pmatrix}.$$

The sum of independent vectors in (3.79) will be bounded in ℓ_2-norm by the vector valued Bernstein inequality Lemma A.5. Observe that the vectors $\mathbf{B}_i\mathbf{C}^i\mathbf{v}$ have mean zero. Furthermore,

$$\begin{aligned} m\mathbb{E}\|\mathbf{B}_i\mathbf{C}^i\mathbf{v}\|_2^2 &= \frac{1}{m}\mathbb{E}\sum_{j,k=1}^s \tilde{\epsilon}_i^2\epsilon_{ij}\epsilon_{ik}\langle P_\ell P_j v_j, P_\ell P_k v_k\rangle \\ &= \frac{1}{m}\sum_{j\in S}\|P_\ell P_j v_j\|_2^2 \le \frac{1}{m}\sum_{j\in S}\|P_\ell P_j\|^2\|v_j\|_2^2 \le \frac{\lambda^2}{m}. \end{aligned}$$

Here we used the assumption that $\|P_\ell P_j\| \le \lambda$ for $j \in S$ and $\ell \in \overline{S}$. We bound σ^2 appearing in Lemma A.5 simply due to (A.5),

$$m\sigma^2 \le m\mathbb{E}\|\mathbf{B}_i\mathbf{C}^i\mathbf{v}\|_2^2 \le \frac{\lambda^2}{m}.$$

For the uniform bound, observe that

$$\|\mathbf{B}_i\mathbf{C}^i\mathbf{v}\|_2 = \frac{1}{m}\left\|\sum_{j\in S}\tilde{\epsilon}_i\epsilon_{ij}P_\ell P_j v_j\right\|_2 \le \frac{1}{m}\sum_{j\in S}\|P_\ell P_j\|\|v_j\|_2 \le \frac{\lambda\sqrt{s}}{m}.$$

The last inequality follows from the fact that $\|\mathbf{v}\|_{2,1} \le \sqrt{s}\|\mathbf{v}\|_2$. Then the vector valued Bernstein inequality (A.4) yields

$$\mathbb{P}\left(\|(\tilde{\mathbf{A}}_\mathbf{P})_\ell^*(\tilde{\mathbf{A}}_\mathbf{P})_S\mathbf{v}\|_2 \ge \left(\frac{\lambda}{\sqrt{m}}+t\right)\|\mathbf{v}\|_2\right) \le \exp\left(-\frac{t^2/2}{\frac{\lambda^2}{m}+\frac{2\lambda\sqrt{s}}{m}\frac{\lambda}{\sqrt{m}}+\frac{t}{3}\frac{\lambda\sqrt{s}}{m}}\right).$$

Taking the union bound over $\ell \in \overline{S} \subset [N]$ and using that $\frac{\sqrt{\lambda s}}{\sqrt{m}} \le 1$ and $\lambda \in [0,1]$ yields

$$\mathbb{P}\left(\max_{\ell\in\overline{S}}\|(\tilde{\mathbf{A}}_\mathbf{P})_\ell^*(\tilde{\mathbf{A}}_\mathbf{P})_S\mathbf{v}\|_2 \ge \left(\frac{\lambda}{\sqrt{m}}+t\right)\|\mathbf{v}\|_2\right) \le N\exp\left(-\frac{t^2m}{6\lambda+t\lambda\sqrt{s}}\right).$$

This completes the proof. □

The next lemma is a corollary of Lemma 3.26.

Lemma 3.31. *Assume the conditions of Lemma 3.30. Then for $t > 0$*

$$\mathbb{P}\left(\|[(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S]\mathbf{v}\|_2 \ge \left(\frac{\lambda\sqrt{s}}{\sqrt{m}}+t\right)\|\mathbf{v}\|_2\right) \le \exp\left(-\frac{mt^2}{8+6\lambda s+t(\frac{4}{3}+\lambda s)}\right).$$

We now state another corollary, which follows from Lemma 3.28.

Lemma 3.32. *For $t \in (0, \frac{3}{2})$,*

$$\mathbb{P}(\max_{i\in\overline{S}}\|(\tilde{\mathbf{A}}_\mathbf{P})_S^*(\tilde{\mathbf{A}}_\mathbf{P})_i\| \ge t) \le 2(s+1)Nk\exp\left(-\frac{t^2m}{3\lambda s}\right).$$

• *Proof of Theorem 3.2 continued*

We follow very similar steps and notation as in the proof of Theorem 3.1. We define the vectors $\mathbf{u}$ and $\mathbf{w}$ as before and recall that

$$\mathbf{u}^{(n)} = \frac{1}{m_n}(\mathbf{A_P}^{(n)})^*(\mathbf{A_P}^{(n)})_S(\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S^{(n-1)}) + \mathbf{u}^{(n-1)}, \tag{3.80}$$

$$\mathbf{w}^{(n)} = \left[\mathbf{P}_S - \frac{1}{m_n}(\mathbf{A_P}^{(n)})_S^*(\mathbf{A_P}^{(n)})_S\right]\mathbf{w}^{(n-1)}, \tag{3.81}$$

for $n = 1, \dots, L$, and

$$\mathbf{u} = \mathbf{u}^{(L)} = \sum_{n=1}^{L}\frac{1}{m_n}(\mathbf{A_P}^{(n)})^*(\mathbf{A_P}^{(n)})_S\mathbf{w}^{(n-1)}. \tag{3.82}$$

According to Lemma 3.22, we have the tasks of showing that $\|\mathbf{w}^{(L)}\|_2 \le 1/4$ and $\max_{i\in\overline{S}}\|u_i\|_2 \le 1/4$. For the moment, we assume that the following inequalities hold for each n with high probability,

$$\|\mathbf{w}^{(n)}\|_2 \le \left(\frac{\lambda\sqrt{s}}{\sqrt{m}}+r_n\right)\|\mathbf{w}^{(n-1)}\|_2, \quad n \in [L], \tag{3.83}$$

$$\max_{i\in\overline{S}}\left\|\frac{1}{m_n}(\mathbf{A_P}^{(n)})_i^*(\mathbf{A_P}^{(n)})_S\mathbf{w}^{(n-1)}\right\|_2 \le \left(\frac{\lambda}{\sqrt{m}}+t_n\right)\|\mathbf{w}^{(n-1)}\|_2, \quad n \in [L], \tag{3.84}$$

where the parameters r_n, t_n will be specified later. Now let $r'_n := \frac{\lambda\sqrt{s}}{\sqrt{m}} + r_n$ and $t'_n := \frac{\lambda}{\sqrt{m}} + t_n$. Then the relations in (3.81) and (3.83) yield

$$\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 = \|\mathbf{w}^{(L)}\|_2 \leq \|\mathrm{sgn}(\mathbf{x}_S)\|_2 \prod_{n=1}^{L} r'_n \leq \sqrt{s} \prod_{n=1}^{L} r'_n. \tag{3.85}$$

Furthermore, (3.82),(3.84) and (3.85) give

$$\begin{aligned}
\max_{i\in\overline{S}} \|u_i\|_2 &= \max_{i\in\overline{S}} \left\| \sum_{n=1}^{L} \frac{1}{m_n} (\mathbf{A_P}^{(n)})_i^* (\mathbf{A_P}^{(n)})_S \mathbf{w}^{(n-1)} \right\|_2 \\
&\leq \sum_{n=1}^{L} \max_{i\in\overline{S}} \left\| \frac{1}{m_n} (\mathbf{A_P}^{(n)})_i^* (\mathbf{A_P}^{(n)})_S \mathbf{w}^{(n-1)} \right\|_2 \\
&\leq \sum_{n=1}^{L} t'_n \|\mathbf{w}^{(n-1)}\|_2 \leq \sqrt{s} \sum_{n=1}^{L} t'_n \prod_{j=1}^{n-1} r'_j
\end{aligned}$$

with the understanding that $\prod_{j=1}^{n-1} r'_j = 1$ if $n = 1$. Next we define the probabilities that (3.83) and (3.84) do not hold as $p_1(n)$ and $p_2(n)$ respectively. Then by Lemma 3.31 and independence of the blocks,

$$p_1(n) \leq \varepsilon$$

provided

$$m_n \geq \left(\frac{8 + 6\lambda s}{r_n^2} + \frac{4/3 + \lambda s}{r_n} \right) \ln(\varepsilon^{-1}). \tag{3.86}$$

Similarly, due to Lemma 3.30 and independence of the blocks,

$$p_2(n) \leq \varepsilon$$

provided

$$m_n \geq \left(\frac{6\lambda}{t_n^2} + \frac{\lambda\sqrt{s}}{t_n} \right) \ln(N/\varepsilon). \tag{3.87}$$

We now set the parameters L, m_n, t_n, r_n for $n \in [L]$ such that $\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 \leq 1/4$ and $\max_{i\in\overline{S}} \|u_i\|_2 \leq 1/4$ as required in the Lemma 3.22. We choose

$$\begin{aligned}
&L = \lceil \ln(s)/2 \rceil + 2, \\
&m_1, m_2 \geq c(1 + \lambda s) \ln(N) \ln(2\varepsilon^{-1}), \quad \text{and} \quad m_n \geq c(1 + \lambda s) \ln(2L\varepsilon^{-1}), \\
&r_1 = r_2 = \frac{1}{4\sqrt{\ln(N)}} \quad \text{and} \quad r_n = \frac{1}{4}, \\
&t_1 = t_2 = \frac{1}{16\sqrt{s}}, \quad \text{and} \quad t_n = \frac{\ln(N)}{16\sqrt{s}}, \quad n = 3, \ldots, L,
\end{aligned}$$

where $c = 1536$. Then safely assuming that $\frac{\lambda s \ln(N)}{m_n} \leq \frac{1}{256}$ and by definitions of r'_n, t'_n, we have

$$r'_1, r'_2 \leq \frac{1}{2\sqrt{\ln(N)}} \text{ and } r'_n \leq 1/2,$$

and

$$t'_1, t'_2 \leq \frac{1}{8\sqrt{s}} \text{ and } t'_n \leq \frac{\ln(N)}{8\sqrt{s}}$$

for $n = 3, \ldots, L$. Furthermore,

$$\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 \leq \sqrt{s} \prod_{n=1}^{L} r'_n \leq \frac{\sqrt{s}}{4} 2^{-\ln(s)/2} \leq \frac{1}{4},$$

and

$$\max_{i \in \overline{S}} \|u_i\|_2 \leq \sqrt{s} \sum_{n=1}^{L} t'_n \prod_{j=1}^{n-1} r'_j \leq \frac{1}{8} \sum_{k=0}^{L-1} \frac{1}{2^k} \leq \frac{1}{4}.$$

Next we bound the failure probabilities according to our choices of parameters above. Considering also the conditions (3.86) and (3.87), $p_1(1), p_1(2), p_2(1), p_2(2) \leq \varepsilon/2$ and $p_1(n), p_2(n) \leq \varepsilon/(2L)$. These yield

$$\sum_{n=1}^{L} p_1(n) \leq 2\varepsilon \quad \text{and} \quad \sum_{n=1}^{L} p_2(n) \leq 2\varepsilon.$$

The overall number of samples obey

$$\begin{aligned} m &= \sum_{n=1}^{L} m_n = m_1 + m_2 + \sum_{n=3}^{L} m_n \\ &\geq 2c(1+\lambda s)\ln(N)\ln(2\varepsilon^{-1}) + c(1+\lambda s)\lceil \ln(s)/2 \rceil \ln(2\lceil \ln(s)/2 \rceil \varepsilon^{-1}). \end{aligned} \tag{3.88}$$

This is already very close to the proposed condition in the statement of our theorem. We will strengthen this condition later. Next we look into first part of Condition (3.45) of Lemma 3.22. By Corollary 3.29 $\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\| \leq 1/2$ with probability at least $1 - \varepsilon$ provided

$$m \geq \frac{32}{3}(1+\lambda s)\ln(2sk/\varepsilon). \tag{3.89}$$

This implies that $\|[(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S]_{|\mathcal{H}}^{-1}\| \leq 2$.
For the second part of Condition (3.45) we use Lemma 3.32. It says that

$$\mathbb{P}(\max_{i \in \overline{S}} \|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\| \geq t) \leq 2(s+1)Nk \exp\left(-\frac{t^2 m}{3\lambda s}\right).$$

Taking $t = 1$ implies that

$$\max_{i \in \overline{S}} \|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\| \leq 1$$

with probability at least $1 - \varepsilon$ provided

$$m \geq 6\lambda s \ln(N(s+1)k/\varepsilon). \tag{3.90}$$

Altogether we have shown that Conditions (3.45) and (3.46) of Lemma 3.22 hold simultaneously with probability at least $1 - 6\varepsilon$ provided Conditions (3.88), (3.89) and (3.90) hold. Replacing ε by $\varepsilon/6$, we notice that if

$$\ln(s)\ln(\ln(s)) \leq c' \ln(N), \tag{3.91}$$

the main condition of Theorem 3.2

$$m \geq C(1+\lambda s)\ln(Nsk)\ln(\varepsilon^{-1})$$

implies all three conditions above with an appropriate constant C. We are almost done with the proof of our theorem.

Finally, we use a nice trick due to Gross [**65**] in order to remove the mild condition (3.91). We assume the basic structure of our proof. The idea is to sample slightly more row blocks ${\mathbf{A}_\mathbf{P}}^{(n)}$ of the matrix $\mathbf{A}_\mathbf{P}$ than in the previous argument. Then we use only the part which obey the conditions (3.83) and (3.84). The probability of that is higher than the probability that these conditions hold simultaneously for all sampled blocks. This process of sampling more rows does not increase the total number of measurements m, on the contrary, it decreases since each block size m_n can be chosen smaller.

More precisely, we choose a number of row submatrices $L' > L$ to be determined later. As before, we set $\mathbf{u}^{(0)} = 0$ and define recursively $\mathbf{u}^{(1)}$ and $\mathbf{u}^{(2)}$ (for $n = 1, 2$ we do not allow replacements) via (3.80). Next we define $\mathbf{u}^{(n)}$ recursively but at each step we check if $\mathbf{w}^{(n)} = \operatorname{sgn}(\mathbf{x}_S) - \mathbf{u}_S$ satisfies (3.83) and (3.84). If not, we discard this particular n and replace ${\mathbf{A}_\mathbf{P}}^{(n)}$ with ${\mathbf{A}_\mathbf{P}}^{(n+1)}$ (and also all subsequent ${\mathbf{A}_\mathbf{P}}^{(\ell)}$ by ${\mathbf{A}_\mathbf{P}}^{(\ell+1)}$, $\ell > n$). Then we redefine $\mathbf{u}^{(n)}$ and $\mathbf{w}^{(n)}$ by using the modified ${\mathbf{A}_\mathbf{P}}^{(n)}$. We continue this way and discard all n when conditions (3.83) and (3.84) are not satisfied and until $n = L$. We will estimate the probability that this actually happens below. Since the ${\mathbf{A}_\mathbf{P}}^{(n)}$ are independent, all such events for different n are also independent.
With respect to the previous part, we use a slightly different definition of m_n, $n \geq 3$,

$$m_n \geq c(1 + \lambda s) \ln(2\rho^{-1}),$$

for some $\rho \in (0, 1)$ to be defined later. The remaining quantities L, m_1, m_2, r_n, t_n are defined in the same way as before. We still have $p_1(1), p_1(2), p_2(1), p_2(2) \leq \varepsilon/2$. We need to determine the probability that (3.83) and (3.84) hold for at least $L - 2$ choices of $n \in \{3, 4, \ldots, L'\}$. Due to the relations (3.86) and (3.87) together with the modified definition of m_n, we have $p_1(n), p_2(n) \leq \rho/2$, $n \geq 3$. We define the probability that both events (3.83) and (3.84) hold as E_n. Then $\mathbb{P}(E_n) \leq 1-\rho$ for $n \geq 3$. The event that E_n occurs for at least $L - 2$ choices of n has probability larger than the event that

$$\sum_{n=3}^{L'} X_n \geq L - 2,$$

where the X_n are independent random variables that take the value 1 with probability $1 - \rho$ and the value 0 with probability ρ. This sum is also called a binomial random variable. We estimate the probability of this last event via Hoeffding's inequality. Clearly $\mathbb{E}X_n = 1-\rho$ and $X - \mathbb{E}X_n \leq 1$ for all n. Set $J := L' - 2$. Then

$$\mathbb{P}\left(\sum_{n=3}^{L'}(X_n - \mathbb{E}X_n) < -\sqrt{J}t\right) = \mathbb{P}\left(\sum_{n=3}^{L'} X_n < (1-\rho)J - \sqrt{J}t\right) \leq e^{-t^2/2}.$$

Setting $L = (1 - \rho)J - \sqrt{J}t$ and solving for t yields

$$\mathbb{P}\left(\sum_{n=3}^{L'} X_n < L\right) \leq \exp\left(-\frac{((1-\rho)J - L)^2}{2J}\right).$$

This probability is bounded by $\tilde{\varepsilon}$ if

$$J = \left\lceil \frac{2}{1-\rho} L + \frac{2}{(1-\rho)^2} \ln(\tilde{\varepsilon}^{-1}) \right\rceil.$$

This implies that the event E_n occurs at least L times (we actually only need $L-2$ times) with probability at least $1-\tilde{\varepsilon}$. The overall number of samples satisfies

$$\begin{aligned} m &= m_1 + m_2 + \sum_{n=3}^{L'} m_n \\ &\geq 2c(1+\lambda s)\ln(N)\ln(2\varepsilon^{-1}) + Jc(1+\lambda s)\ln(2\rho^{-1}) \\ &= 2c(1+\lambda s)\ln(N)\ln(2\varepsilon^{-1}) + \left\lceil \frac{2}{1-\rho}L + \frac{2}{(1-\rho)^2}\ln(\tilde{\varepsilon}^{-1}) \right\rceil c(1+\lambda s)\ln(2\rho^{-1}) \\ &= 2c(1+\lambda s)\ln(N)\ln(2\varepsilon^{-1}) + \left\lceil \frac{2}{1-\rho}(\lceil \ln(s)/2 \rceil + 2) + \frac{2}{(1-\rho)^2}\ln(\tilde{\varepsilon}^{-1}) \right\rceil c(1+\lambda s)\ln(2\rho^{-1}). \end{aligned}$$

Choosing $\rho = 1/2$ and $\tilde{\varepsilon} = \varepsilon$, this condition is implied by

$$m \geq 2c(1+\lambda s)\ln(N)\ln(2\varepsilon^{-1}) + \ln(4)c(1+\lambda s)[2\ln(s) + 8\ln(\varepsilon^{-1}) + 16]. \tag{3.92}$$

We have finally proved that $(L1)$ recovers $\mathbf{x}$ with probability at least $1-5\varepsilon$. Replacing ε by $\varepsilon/5$, we observe that our assumption (3.5) on m implies (3.89) and (3.90) as discussed before as well as (3.92) by an appropriate choice of C. This concludes the proof. □

3.3.5. Proof of Theorem 3.3

We recall the rescaled matrix $\tilde{\mathbf{A}}_{\mathbf{P}} = \frac{1}{\sqrt{m}}\mathbf{A}_{\mathbf{P}}$. Throughout the proof, we use the notation $\mathbf{E}_{ij}(A)$ to denote a block matrix with blocks each of size $d \times d$ where $A \in \mathbb{R}^{d\times d}$ is at the intersection of i-th block row and j-th block column, and is 0 elsewhere. The number of blocks depends on the context and is not specified unless there is a chance of confusion. In addition the Kronecker delta

$$\delta_{ij} = \begin{cases} 1 & : \quad i = j, \\ 0 & : \quad i \neq j, \end{cases}$$

is used at various instances.

- *Auxiliary results*

We will mainly use Lemma A.8, in order to bound $\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\|$.

Theorem 3.33. *Let $A \in \mathbb{R}^{m\times N}$ be a measurement matrix whose entries are i.i.d. Gaussian random variables and $(W_j)_{j=1}^N$ be a fusion frame given with parameter $\lambda \in [0,1]$ and $\dim(W_j) = k_j$. Then, for $\delta > 0$, the block matrix $\tilde{\mathbf{A}}_{\mathbf{P}}$ satisfies*

$$\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\| \leq \delta$$

with probability at least $1-\varepsilon$, provided that

$$m \geq 100e^2\delta^{-2}(1+\sqrt{\lambda s}/5)^2 \ln^2\left(6s^2\varepsilon^{-1}\sum_{j\in S} k_j\right).$$

Proof. We first observe that

$$(\tilde{\mathbf{A}}_{\mathbf{P}})_S = \frac{1}{\sqrt{m}} \sum_{i\in[m], j\in S} g_{ij}\mathbf{E}_{ij}(P_j)$$

$$(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S = \frac{1}{m} \sum_{i\in[m],j\in S} \sum_{k\in[m],\ell\in S} g_{ij}g_{k\ell}\mathbf{E}_{ij}^*(P_j)\mathbf{E}_{k\ell}(P_\ell)$$

$$= \frac{1}{m} \sum_{(i,j);(k,\ell)} g_{ij} g_{k\ell} \delta_{ik} \mathbf{E}_{j\ell}(P_j P_\ell).$$

In the third step we used $\mathbf{E}_{ij}(P)\mathbf{E}_{k\ell}(Q) = \delta_{jk}\mathbf{E}_{i\ell}(PQ)$ and $\mathbf{E}^*_{ij}(X) = \mathbf{E}_{ji}(X^*)$. We will use these relations many times later on. Note that $i, k \in [m]$ and $j, \ell \in S$ in the rest of the proof. Denote $\tau := (i,j), \xi := (k,\ell)$ and $D_{\tau,\xi} := \delta_{ik}\mathbf{E}_{j\ell}(P_j P_\ell)$. We separate $(\tilde{\mathbf{A}}_\mathbf{P})^*_S(\tilde{\mathbf{A}}_\mathbf{P})_S$ into two parts (i.e., diagonal and off-diagonal):

$$\begin{aligned}(\tilde{\mathbf{A}}_\mathbf{P})^*_S(\tilde{\mathbf{A}}_\mathbf{P})_S &= \frac{1}{m}\sum_{\tau\neq\xi} g_\tau g_\xi D_{\tau,\xi} + \frac{1}{m}\sum_{\tau=\xi} g_\tau g_\xi D_{\tau,\xi} \\ &= \frac{1}{m}\sum_{\tau\neq\xi} g_\tau g_\xi D_{\tau,\xi} + \frac{1}{m}\sum_{i,j} g_{ij}^2 \mathbf{E}_{jj}(P_j).\end{aligned}$$

Furthermore we write $\mathbf{P}_S = \frac{1}{m}\sum_{i,j}\mathbf{E}_{jj}(P_j)$ and obtain

$$\begin{aligned}(\tilde{\mathbf{A}}_\mathbf{P})^*_S(\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S &= \frac{1}{m}\sum_{\tau\neq\xi} g_\tau g_\xi D_{\tau,\xi} + \frac{1}{m}\sum_{i,j}(g_{ij}^2 - 1)\mathbf{E}_{jj}(P_j) \\ &=: \mathcal{B} + \mathcal{C}.\end{aligned} \tag{3.93}$$

It follows from Minkowski's inequality that

$$\left(\mathbb{E}\|(\tilde{\mathbf{A}}_\mathbf{P})^*_S(\tilde{\mathbf{A}}_\mathbf{P})_S - \mathbf{P}_S\|^p\right)^{1/p} \leq (\mathbb{E}\|\mathcal{B}\|^p)^{1/p} + (\mathbb{E}\|\mathcal{C}\|^p)^{1/p}, \tag{3.94}$$

for an integer p. We first bound the first term on the right hand side of (3.94) by applying Lemma A.7 and Lemma A.8 and obtain for an integer n

$$\begin{aligned}\mathbb{E}\|\mathcal{B}\|^{2n} \leq \mathbb{E}\|\mathcal{B}\|^{2n}_{S_{2n}} &= \frac{1}{m^{2n}}\mathbb{E}\left\|\sum_{\tau\neq\xi} g_\tau g_\xi D_{\tau,\xi}\right\|^{2n}_{S_{2n}} \leq \frac{4^{2n}}{m^{2n}}\mathbb{E}\left\|\sum_{\tau\neq\xi} g_\tau g'_\xi D_{\tau,\xi}\right\|^{2n}_{S_{2n}} \\ \leq \frac{4^{2n}}{m^{2n}} \times 2C_n^2 \times \max&\left\{\left\|\left(\sum_{\tau\neq\xi} D_{\tau,\xi}D^*_{\tau,\xi}\right)^{1/2}\right\|^{2n}_{S_{2n}}, \left\|\left(\sum_{\tau\neq\xi} D^*_{\tau,\xi}D_{\tau,\xi}\right)^{1/2}\right\|^{2n}_{S_{2n}}, \|F\|^{2n}_{S_{2n}}, \|\tilde{F}\|^{2n}_{S_{2n}}\right\},\end{aligned} \tag{3.95}$$

where $F, \tilde{F}$ are the block matrices $F = (D_{\tau,\xi})_{\tau,\xi}$, $\tilde{F} = (D^*_{\tau,\xi})_{\tau,\xi}$ with the assumption that $D_{\tau,\tau} = 0$ and $C_n = \frac{(2n)!}{2^n n!}$. The Schatten norm $\|\cdot\|_{S_{2n}}$ is defined in Appendix A.2. We label the terms in the maximum above, from *(a)* to *(d)*. For a matrix A it holds $\|A\|^{2n}_{S_{2n}} = \mathrm{Tr}((AA^*)^n)$, see (A.6).

a) Let us denote $H := \left(\sum_{\tau\neq\xi} D_{\tau,\xi}D^*_{\tau,\xi}\right)$. Then $\|H^{1/2}\|^{2n}_{S_{2n}} = \mathrm{Tr}((H^{1/2}H^{1/2})^n) = \mathrm{Tr}(H^n)$. Observe that

$$\sum_{\tau\neq\xi} D_{\tau,\xi}D^*_{\tau,\xi} = \sum_{(i,j)\neq(k,\ell)} \delta_{ik}\mathbf{E}_{j\ell}(P_jP_\ell)\mathbf{E}_{\ell j}(P_\ell P_j) = m\sum_{j\neq\ell}\mathbf{E}_{jj}(P_jP_\ell P_j).$$

Next we calculate the trace

$$\mathrm{Tr}(H^n) = m^n \mathrm{Tr}\sum_{\substack{j_1,\ldots,j_n \\ \ell_1,\ldots,\ell_n \\ j_i\neq\ell_i\ \forall i}} \mathbf{E}_{j_1j_1}(P_{j_1}P_{\ell_1}P_{j_1})\mathbf{E}_{j_2j_2}(P_{j_2}P_{\ell_2}P_{j_2})\ldots\mathbf{E}_{j_nj_n}(P_{j_n}P_{\ell_n}P_{j_n})$$

$$= m^n \mathrm{Tr} \sum_{\substack{j,\ell_1,\ldots,\ell_n \\ j\neq\ell_i\ \forall i}} \mathbf{E}_{jj}(P_j P_{\ell_1} P_j P_{\ell_2} P_j \ldots P_j P_{\ell_n} P_j) = m^n \mathrm{Tr} \sum_{\substack{j,\ell_1,\ldots,\ell_n \\ j\neq\ell_i\ \forall i}} P_j P_{\ell_1} P_j P_{\ell_2} P_j \ldots P_j P_{\ell_n} P_j$$

$$\leq m^n \sum_{\substack{j,\ell_1,\ldots,\ell_n \\ j\neq\ell_i\ \forall i}} \mathrm{Tr}(P_j)\|P_{\ell_1} P_j P_{\ell_2} P_j \ldots P_j P_{\ell_n} P_j\| \leq m^n \sum_{\substack{j,\ell_1,\ldots,\ell_n \\ j\neq\ell_i\ \forall i}} \mathrm{Tr}(P_j)\|P_{\ell_1} P_j\|\|P_{\ell_2} P_j\| \ldots \|P_{\ell_n} P_j\|$$

$$\leq m^n \sum_{\substack{j,\ell_1,\ldots,\ell_n \\ j\neq\ell_i\ \forall i}} k_j \lambda^n \leq \lambda^n s^n m^n \sum_j k_j.$$

In the first inequality above, we used $|\mathrm{Tr}(A^*B)| \leq \|A\|_{S_1}\|B\|_{S_\infty}$, an analogy of Hölder's inequality, see [**68**, page 186].

b) It yields the same result since indices j and ℓ are symmetric, i.e.,

$$\left\| \left(\sum_{\tau\neq\xi} D^*_{\tau,\xi} D_{\tau,\xi} \right)^{1/2} \right\|^{2n}_{S_{2n}} \leq \lambda^n s^n m^n \sum_j k_j.$$

c) Recall that we need to estimate the Schatten-$2n$ norm of $F = (D_{\tau,\xi})_{\tau,\xi}$. We have

$$\|F\|^{2n}_{S_{2n}} = \mathrm{Tr}((FF^*)^n) = \mathrm{Tr} \sum_{\substack{\tau_1,\ldots,\tau_n \\ \xi_1,\ldots,\xi_n \\ \tau_i\neq\xi_i\ \forall i}} D^*_{\tau_1\xi_1} D_{\tau_1\xi_2} D^*_{\tau_2\xi_2} D_{\tau_2\xi_3} D^*_{\tau_3\xi_3} \ldots D^*_{\tau_n\xi_n} D_{\tau_n\xi_1}$$

$$= \mathrm{Tr} \sum_{\substack{i_1,\ldots,i_n \\ j_1,\ldots,j_n \\ k_1,\ldots,k_n \\ \ell_1,\ldots,\ell_n \\ \tau_i\neq\xi_i\ \forall i}} \delta_{i_1k_1} \mathbf{E}^*_{j_1\ell_1}(P_{j_1}P_{\ell_1}) \delta_{i_1k_2} \mathbf{E}_{j_1\ell_2}(P_{j_1}P_{\ell_2}) \delta_{i_2k_2} \mathbf{E}^*_{j_2\ell_2}(P_{j_2}P_{\ell_2}) \delta_{i_2k_3} \mathbf{E}_{j_2\ell_3}(P_{j_2}P_{\ell_3}) \ldots$$

$$\ldots \delta_{i_nk_n} \mathbf{E}^*_{j_n\ell_n}(P_{j_n}P_{\ell_n}) \delta_{i_nk_1} \mathbf{E}_{j_n\ell_1}(P_{j_n}P_{\ell_1})$$

$$= m\mathrm{Tr} \sum_{\substack{j_1,\ldots,j_n \\ \ell_1,\ldots,\ell_n \\ j_i\neq\ell_i\ \forall i}} \mathbf{E}_{\ell_1 j_1}(P_{\ell_1}P_{j_1}) \mathbf{E}_{j_1\ell_2}(P_{j_1}P_{\ell_2}) \mathbf{E}_{\ell_2 j_2}(P_{\ell_2}P_{j_2}) \mathbf{E}_{j_2\ell_3}(P_{j_2}P_{\ell_3}) \ldots$$

$$= m\mathrm{Tr} \sum_{\substack{j_1,\ldots,j_n \\ \ell_1,\ldots,\ell_n \\ j_i\neq\ell_i\ \forall i}} \mathbf{E}_{\ell_1\ell_2}(P_{\ell_1}P_{j_1}P_{\ell_2}) \mathbf{E}_{\ell_2\ell_3}(P_{\ell_2}P_{j_2}P_{\ell_3}) \ldots \mathbf{E}_{\ell_n\ell_1}(P_{\ell_n}P_{j_n}P_{\ell_1})$$

$$= m\mathrm{Tr} \sum_{\substack{j_1,\ldots,j_n \\ \ell_1,\ldots,\ell_n \\ j_i\neq\ell_i\ \forall i}} \mathbf{E}_{\ell_1\ell_1}(P_{\ell_1}P_{j_1}P_{\ell_2}P_{j_2} \ldots P_{\ell_n}P_{j_n}P_{\ell_1}) = m\mathrm{Tr} \sum_{\substack{j_1,\ldots,j_n \\ \ell_1,\ldots,\ell_n \\ j_i\neq\ell_i\ \forall i}} P_{\ell_1}P_{j_1}P_{\ell_2}P_{j_2} \ldots P_{\ell_n}P_{j_n}P_{\ell_1}$$

$$= m\mathrm{Tr} \left(\sum_{j\neq\ell} P_j P_\ell \right)^n = m \left\| \left(\sum_{j\neq\ell} P_j P_\ell \right)^n \right\|_{S_1}.$$

In the third line above, the implication of $i_1 = \ldots = i_n = k_1 = \ldots = k_n$ adds a factor of m. In the next series of equations we use the fact that, for a matrix A with rank r, it holds $\|A\|_{S_p} \leq r^{1/p}\|A\|$, see (A.7). Moreover, in the second line below, the relation $\mathrm{rank}(A^n) \leq \mathrm{rank}(A)$, the subadditivity of the *rank* function and $\mathrm{rank}(P_jP_\ell) \leq \min\{\mathrm{rank}(P_j), \mathrm{rank}(P_\ell)\}$ are used,

$$\|F\|^{2n}_{S_{2n}} = m \left\| \left(\sum_{j\neq\ell} P_j P_\ell \right)^n \right\|_{S_1} \leq m\ \mathrm{rank} \left[\left(\sum_{j\neq\ell} P_j P_\ell \right)^n \right] \left\| \left(\sum_{j\neq\ell} P_j P_\ell \right)^n \right\|$$

$$\leq m\left(\sum_{j\neq\ell}\operatorname{rank}(P_j)\right)\left\|\sum_{j\neq\ell}P_jP_\ell\right\|^n \leq ms\left(\sum_j k_j\right)\left(\sum_{j\neq\ell}\|P_jP_\ell\|\right)^n \leq ms\left(\sum_j k_j\right)\lambda^n s^{2n}.$$

d) We follow similarly for $\tilde{F}=(D^*_{\tau,\xi})_{\tau,\xi}$,

$$\begin{aligned}
\|\tilde{F}\|_{S_{2n}}^{2n} &= \operatorname{Tr}\sum_{\substack{\tau_1,\dots,\tau_n\\ \xi_1,\dots,\xi_n\\ \tau_i\neq\xi_i\ \forall i}} D_{\tau_1\xi_1}D^*_{\tau_1\xi_2}D_{\tau_2\xi_2}D^*_{\tau_2\xi_3}D_{\tau_3\xi_3}\dots D_{\tau_n\xi_n}D^*_{\tau_n\xi_1}\\
&= \operatorname{Tr}\sum_{\substack{i_1,\dots,i_n\\ j_1,\dots,j_n\\ k_1,\dots,k_n\\ \ell_1,\dots,\ell_n\\ \tau_i\neq\xi_i\ \forall i}} \delta_{i_1k_1}\mathbf{E}_{j_1\ell_1}(P_{j_1}P_{\ell_1})\delta_{i_1k_2}\mathbf{E}_{\ell_2j_1}(P_{\ell_2}P_{j_1})\delta_{i_2k_2}\mathbf{E}_{j_2\ell_2}(P_{j_2}P_{\ell_2})\delta_{i_2k_3}\mathbf{E}_{\ell_3j_2}(P_{\ell_3}P_{j_2})\dots\\
&\qquad\qquad \dots\delta_{i_nk_n}\mathbf{E}_{j_n\ell_n}(P_{j_n}P_{\ell_n})\delta_{i_nk_1}\mathbf{E}_{\ell_1j_n}(P_{\ell_1}P_{j_n})\\
&= m\operatorname{Tr}\sum_{\substack{j_1,\dots,j_n\\ \ell_1,\dots,\ell_n\\ j_i\neq\ell_i\ \forall i}} \delta_{\ell_1\ell_2}\mathbf{E}_{j_1j_1}(P_{j_1}P_{\ell_1}P_{\ell_2}P_{j_1})\delta_{\ell_2\ell_3}\mathbf{E}_{j_2j_2}(P_{j_2}P_{\ell_2}P_{\ell_3}P_{j_2})\dots\delta_{\ell_n\ell_1}\mathbf{E}_{j_nj_n}(P_{j_n}P_{\ell_n}P_{\ell_1}P_{j_n})\\
&= m\operatorname{Tr}\sum_{j\neq\ell}\mathbf{E}_{jj}(\underbrace{P_jP_\ell}_{n-times}P_jP_\ell\dots P_jP_\ell P_j) = m\sum_{j\neq\ell}\operatorname{Tr}[(P_jP_\ell P_j)^n] = m\sum_{j\neq\ell}\operatorname{Tr}[(P_jP_\ell P_\ell P_j)^n]\\
&= m\sum_{j\neq\ell}\operatorname{Tr}[(P_jP_\ell)(P_jP_\ell)^*]^n = m\sum_{j\neq\ell}\|P_jP_\ell\|_{S_{2n}}^{2n}.
\end{aligned}$$

In the third line above, the matrices are multiplied in consecutive pairs. Below, using $\|A\|_{S_{2n}}^{2n}\leq \operatorname{rank}(A)\|A\|^{2n}$ and $\operatorname{rank}(P_jP_\ell)\leq\min\{\operatorname{rank}(P_j),\operatorname{rank}(P_\ell)\}$, we further obtain

$$\|\tilde{F}\|_{S_{2n}}^{2n}\leq m\sum_{j\neq\ell}\operatorname{rank}(P_jP_\ell)\|P_jP_\ell\|^{2n}\leq m\sum_{j\neq\ell}\operatorname{rank}(P_jP_\ell)\lambda^{2n}\leq ms\lambda^{2n}\sum_j\operatorname{rank}(P_j)=ms\lambda^{2n}\sum_j k_j.$$

Finally we determine the maximum of the terms in (3.95),

$$a=b\leq m^ns^n\lambda^n\sum_{j\in S}k_j,\ \ c\leq ms\lambda^ns^{2n}\sum_{j\in S}k_j,\ \ d\leq ms\lambda^{2n}\sum_{j\in S}k_j.$$

Since $s\leq m$ it follows that

$$\max\{a,b,c,d\}\leq s^2m^ns^n\lambda^n\sum_{j\in S}k_j.$$

Putting this final estimate into (3.95), we have

$$\mathbb{E}\|\mathcal{B}\|^{2n}\leq\frac{4^{2n}}{m^{2n}}s^2(2C_n^2)\ m^ns^n\lambda^n\sum_{j\in S}k_j.$$

We generalize this to any power p by interpolation similarly to [**94**, page 83]. Stirling's formula (A.16) for the factorial yields the estimation

$$C_n=\frac{(2n)!}{2^nn!}\leq\sqrt{2}(2/e)^nn^n. \tag{3.96}$$

Let $p=2n+2\theta=(1-\theta)2n+\theta(2n+2)$ with $\theta\in(0,1)$. Applying Hölder's inequality and (3.96), we obtain

$$\mathbb{E}\|\mathcal{B}\|^{2n+2\theta}\leq(\mathbb{E}\|\mathcal{B}\|^{2n})^{1-\theta}(\mathbb{E}\|\mathcal{B}\|^{2n+2})^{\theta}$$

$$\leq 4s^2 \left(\sum_{j\in S} k_j\right) 4^{2n+2\theta}(2/e)^{2n+2\theta}\frac{(\lambda s)^{n+\theta}}{m^{n+\theta}}[n^{n(1-\theta)}(n+1)^{(n+1)\theta}]^2$$

$$= 4s^2 \left(\sum_{j\in S} k_j\right) 4^{2n+2\theta}(2/e)^{2n+2\theta}\frac{(\lambda s)^{n+\theta}}{m^{n+\theta}}(n^{1-\theta}(n+1)^{\theta})^{2n+2\theta}\left(\frac{n+1}{n}\right)^{2\theta(1-\theta)}$$

$$\leq 4s^2 \left(\sum_{j\in S} k_j\right) 4^{2n+2\theta}(2/e)^{2n+2\theta}\frac{(\lambda s)^{n+\theta}}{m^{n+\theta}}(n+\theta)^{2n+2\theta}2^{1/2}. \tag{3.97}$$

In the last line above, we used the inequality of geometric and arithmetic mean and also the relations $(n+1)/n \leq 2$ and $\theta(1-\theta) \leq 1/4$. Replacing $2n+2\theta$ by p and taking the p-th root of both sides yields

$$(\mathbb{E}\|\mathcal{B}\|^p)^{1/p} \leq (4p/e)\left(2^{5/2}s^2\left(\sum_{j\in S} k_j\right)\right)^{1/p}(\lambda s/m)^{1/2}. \tag{3.98}$$

Next we bound the second term on the right hand side of (3.94), i.e., $\mathbb{E}\|\mathcal{C}\|^p$. Recall that $\mathcal{C} = \frac{1}{m}\sum_{i,j}(g_{ij}^2-1)\mathbf{E}_{jj}(P_j)$. Now define the random variable $\Phi^{(m)} := \sum_{i=1}^m (g_i^2-1)$ where the g_i's are i.i.d. Gaussian random variables. Then we can estimate the spectral norm

$$\left\|\frac{1}{m}\sum_{i,j}(g_{ij}^2-1)\mathbf{E}_{jj}(P_j)\right\| = \left\|\frac{1}{m}\sum_j \Phi_j^{(m)}\mathbf{E}_{jj}(P_j)\right\| \leq \frac{1}{m}\max_{j\in S}\|\Phi_j^{(m)}P_j\| = \frac{1}{m}\max_{j\in S}|\Phi_j^{(m)}|,$$

where the $\Phi_j^{(m)}$ are independent. Taking the p-th moment of both sides yields

$$(\mathbb{E}\mathcal{C}^p)^{1/p} = \frac{1}{m}(\mathbb{E}\max_{j\in S}|\Phi_j^{(m)}|^p)^{1/p} \leq \frac{1}{m}s^{1/p}\max_{j\in S}(\mathbb{E}|\Phi_j^{(m)}|^p)^{1/p} = \frac{1}{m}s^{1/p}(\mathbb{E}|\Phi^{(m)}|^p)^{1/p}. \tag{3.99}$$

The inequality above follows from (3.44). Observe that (g_i^2-1) is a centered subexponential random variable, and

$$\|g_i^2-1\|'_{\psi_1} \leq 2\|g_i^2\|'_{\psi_1} \leq 4(\|g_i\|'_{\psi_2})^2 \leq 4,$$

where we used the triangle inequality and Lemma 3.20. Then by Proposition A.2 we can bound the tail probability of $\Phi^{(m)}$, i.e., the sum of m independent subexponential random variables, as

$$\mathbb{P}(|\Phi^{(m)}| \geq t) \leq 2\exp\left(-\frac{t^2}{4m+8t}\right).$$

Integrating this tail estimate, we obtain the following bound for the p-th moment

$$(\mathbb{E}|\Phi^{(m)}|^p)^{1/p} \leq 2^{1/p}(2\sqrt{m}\sqrt{p}+8p). \tag{3.100}$$

Combining this with (3.99) and (3.98) yields

$$\left(\mathbb{E}\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\|^p\right)^{1/p} \leq (4p/e)\left(2^{5/2}s^2\left(\sum_{j\in S}k_j\right)\right)^{1/p}(\lambda s/m)^{1/2} + (2s)^{1/p}\left(\frac{2\sqrt{p}}{\sqrt{m}}+\frac{8p}{m}\right)$$

$$\leq \left(2^{5/2}s^2\left(\sum_{j\in S}k_j\right)\right)^{1/p}\frac{10(1+\sqrt{\lambda s}/5)}{\sqrt{m}}p.$$

Now we use Proposition A.10 in order to have an estimate for the tail probability. Setting $\gamma = 1$, $\beta = 2^{5/2}s^2\left(\sum_{j\in S} k_j\right), \alpha = \frac{10(1+\sqrt{\lambda s}/5)}{\sqrt{m}}$, we have

$$\mathbb{P}\left(\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\| \geq \frac{10e(1+\sqrt{\lambda s}/5)}{\sqrt{m}}u\right) \leq 2^{5/2}s^2\left(\sum_{j\in S} k_j\right)e^{-u}.$$

For $\delta := \frac{10e(1+\sqrt{\lambda s}/5)}{\sqrt{m}}u$, we have

$$\mathbb{P}(\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\| \geq \delta) \leq 2^{5/2}s^2\left(\sum_{j\in S} k_j\right)\exp\left(-\frac{\delta\sqrt{m}}{10e(1+\sqrt{\lambda s}/5)}\right).$$

This proves our theorem. □

The following result bounds the coherence of $(\tilde{\mathbf{A}}_{\mathbf{P}})_S$ with the off-support columns.

Corollary 3.34. *Assume the setting of Theorem 3.33. Then, for $\delta > 0$, the block matrix $\tilde{\mathbf{A}}_{\mathbf{P}}$ satisfies*

$$\max_{i\in\overline{S}} \|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\| \leq \delta$$

with probability at least $1-\varepsilon$, provided that

$$m \geq 100e^2\delta^{-2}(1+\sqrt{\lambda(s+1)}/5)^2\ln^2\left(6N(s+1)^2\varepsilon^{-1}\sum_{j\in[N]} k_j\right).$$

Proof. Fix $i \in \overline{S}$. Then we consider the block matrix $(\tilde{\mathbf{A}}_{\mathbf{P}})_{S\cup\{i\}}$. Furthermore we observe that $(\tilde{\mathbf{A}}_{\mathbf{P}})_{S\cup\{i\}}^*(\tilde{\mathbf{A}}_{\mathbf{P}})_{S\cup\{i\}} - \mathbf{P}_{S\cup\{i\}}$ includes $(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i$ as a submatrix. Therefore it holds that

$$\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\| \leq \|(\tilde{\mathbf{A}}_{\mathbf{P}})_{S\cup\{i\}}^*(\tilde{\mathbf{A}}_{\mathbf{P}})_{S\cup\{i\}} - \mathbf{P}_{S\cup\{i\}}\|.$$

Applying Theorem 3.33 for $S\cup\{i\}$, we bound the failure probability by

$$\begin{aligned}\mathbb{P}(\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\| \geq \delta) &\leq \mathbb{P}(\|(\tilde{\mathbf{A}}_{\mathbf{P}})_{S\cup\{i\}}^*(\tilde{\mathbf{A}}_{\mathbf{P}})_{S\cup\{i\}} - \mathbf{P}_{S\cup\{i\}}\| \geq \delta)\\ &\leq 2^{5/2}(s+1)^2\left(\sum_{j\in S\cup\{i\}} k_j\right)\exp\left(-\frac{\delta\sqrt{m}}{10e(1+\sqrt{\lambda(s+1)}/5)}\right).\end{aligned}$$

Taking the union bound over $i \in \overline{S}$ gives

$$\begin{aligned}\mathbb{P}(\max_{i\in\overline{S}}\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\|) \geq \delta) &\leq 2^{5/2}(s+1)^2\left(\sum_{i\in\overline{S}}\sum_{j\in S\cup\{i\}} k_j\right)\exp\left(-\frac{\delta\sqrt{m}}{10e(1+\sqrt{\lambda(s+1)}/5)}\right)\\ &\leq 2^{5/2}N(s+1)^2\left(\sum_{j\in[N]} k_j\right)\exp\left(-\frac{\delta\sqrt{m}}{10e(1+\sqrt{\lambda(s+1)}/5)}\right).\end{aligned}$$

Bounding this probability by ε yields the result of this corollary. □

We continue by deriving the analogue of Lemma 3.25 for Gaussian matrices.

Lemma 3.35. *Let S be subset of $[N]$ with cardinality s and $\mathbf{v} \in \mathbb{R}^{S\times d}$ be a block vector of size s with $v_j \in W_j$ for $j \in S$. Then, for $t > 0$,*

$$\mathbb{P}(\max_{\ell\in\overline{S}}\|(\tilde{\mathbf{A}}_{\mathbf{P}})_\ell^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S\mathbf{v}\|_2 \geq t\|\mathbf{v}\|_2) \leq 2e^{1/12}sN\exp\left(-\frac{t^2m}{e\lambda^2+2\sqrt{e}\lambda t}\right).$$

PROOF. We fix $i \in \overline{S}$ and assume without loss of generality that $\|\mathbf{v}\|_2 = 1$. We can write as in (3.60) that

$$(\tilde{\mathbf{A}}_{\mathbf{P}})_\ell^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S\mathbf{v} = \sum_{i=1}^m \sum_{j\in S} \frac{1}{m}\tilde{g}_i g_{ij} P_\ell P_j v_j. \tag{3.101}$$

Here $\tilde{g}$ denotes $(g_{i\ell})_{i=1}^m$. We observe that $\tilde{g}_i$ is independent from g_{ij} for $i \in [m], j \in S$. Conditioning on $\tilde{g}$, we have a sum of independent vectors in (3.101). We will treat these vectors as matrices and bound the ℓ_2-norm first by the Schatten norm and then apply Lemma A.6. Denote $E := \|(\tilde{\mathbf{A}}_{\mathbf{P}})_\ell^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S\mathbf{v}\|_2$, $H_{\ell j} := P_\ell P_j v_j$. and fix $n \in \mathbb{N}$. Then we have

$$\begin{aligned}
E^{2n} &\le \mathbb{E}_{\tilde{g}}\mathbb{E}_g \|\sum_{i=1}^m \sum_{j\in S} \frac{1}{m}\tilde{g}_i g_{ij} H_{\ell j}\|_{S_{2n}}^{2n} \\
&\le \frac{C_n}{m^{2n}}\mathbb{E}_{\tilde{g}} \max\left\{ \left\| \left(\sum_{i,j} (\tilde{g}_i)^2 H_{\ell j} H_{\ell j}^* \right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{i,j} (\tilde{g}_i)^2 H_{\ell j}^* H_{\ell j} \right)^{1/2} \right\|_{S_{2n}}^{2n} \right\} \\
&= \frac{C_n}{m^{2n}}\mathbb{E}\left(\sum_{i=1}^m \tilde{g}_i^2 \right)^n \max\left\{ \left\| \left(\sum_{j\in S} H_{\ell j} H_{\ell j}^* \right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{j\in S} H_{\ell j}^* H_{\ell j} \right)^{1/2} \right\|_{S_{2n}}^{2n} \right\},
\end{aligned} \tag{3.102}$$

where $C_n = \frac{(2n)!}{2^n n!}$. Now we estimate the first term in the maximum argument above,

$$\begin{aligned}
\left\| \left(\sum_{j\in S} H_{\ell j} H_{\ell j}^* \right)^{1/2} \right\|_{S_{2n}}^{2n} &= \operatorname{Tr}\left(\sum_{j\in S} H_{\ell j} H_{\ell j}^* \right)^n = \left\| \left(\sum_{j\in S} H_{\ell j} H_{\ell j}^* \right)^n \right\|_{S_1} \\
&\le \operatorname{rank}\left(\sum_{j\in S} H_{\ell j} H_{\ell j}^* \right)^n \left\| \sum_{j\in S} H_{\ell j} H_{\ell j}^* \right\|^n \\
&\le \operatorname{rank}\left(\sum_{j\in S} (P_\ell P_j v_j)(P_\ell P_j v_j)^* \right) \left(\sum_{j\in S} \|(P_\ell P_j v_j)(P_\ell P_j v_j)^*\| \right)^n \\
&\le s\left(\sum_{j\in S} \|P_\ell P_j\|^2 \|v_j\|^2 \right)^n \le s\lambda^{2n}.
\end{aligned}$$

In the last step we used $\|\mathbf{v}\|_2 = 1$ and $j \neq \ell$. Similarly,

$$\begin{aligned}
\left\| \left(\sum_{j\in S} H_{\ell j}^* H_{\ell j} \right)^{1/2} \right\|_{S_{2n}}^{2n} &= \left\| \left(\sum_{j\in S} \|P_\ell P_j v_j\|_2^2 \right)^{1/2} \right\|_{S_{2n}}^{2n} \le \left\| \left(\sum_{j\in S} \|P_\ell P_j\|^2 \|v_j\|_2^2 \right)^{1/2} \right\|_{S_{2n}}^{2n} \\
&\le \left\| \left(\sum_{j\in S} \lambda^2 \|v_j\|_2^2 \right)^{1/2} \right\|_{S_{2n}}^{2n} \le \lambda^{2n}.
\end{aligned}$$

Next we estimate the expectation in (3.102). The distribution of a sum of the squares of m independent standard Gaussian random variables, i.e., $Q := \sum_{i=1}^m \tilde{g}_i^2$, is called the χ^2-distribution with m degrees of freedom. We need the n-th moment of this random variable which is well studied

and can be found in [**96**] for instance. We have

$$\mathbb{E}Q^n = 2^n \frac{\Gamma(n+m/2)}{\Gamma(m/2)}.$$

By using the estimations in (A.15) for the Gamma function, it follows that

$$\mathbb{E}Q^n \leq 2^n e^{1/12}(n+m/2)^n.$$

Plugging our estimates into (3.102), we obtain

$$E^{2n} \leq e^{1/12} s \frac{C_n}{m^{2n}} 2^n \lambda^{2n} (n+m/2)^n.$$

Then by interpolating for a general moment p and following similar steps in (3.97) we obtain

$$\begin{aligned}(E^p)^{1/p} &\leq (2e^{1/12}s)^{1/p}(\lambda/\sqrt{e})\frac{\sqrt{p}}{m}\sqrt{p+m} \\ &\leq (2e^{1/12}s)^{1/p}(\lambda/\sqrt{e})\frac{\sqrt{p}}{m}(\sqrt{p}+\sqrt{m}) \\ &= (2e^{1/12}s)^{1/p}\left(\frac{\lambda}{\sqrt{e}m}p + \frac{\lambda}{\sqrt{e}\sqrt{m}}\sqrt{p}\right).\end{aligned}$$

In the second inequality we used $\sqrt{a+b} \leq \sqrt{a}+\sqrt{b}$ for $a, b > 0$. Setting $\beta = 2e^{1/12}s$, $\alpha_1 = \frac{\lambda}{\sqrt{e}m}$ and $\alpha_2 = \frac{\lambda}{\sqrt{e}\sqrt{m}}$, we can estimate the tail bound of E by using Proposition A.11 as follows

$$\mathbb{P}(\|(\tilde{\mathbf{A}}_{\mathbf{P}})_\ell^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S\mathbf{v}\|_2 \geq t) \leq 2e^{1/12}s\exp\left(-\frac{t^2 m}{e\lambda^2 + 2\sqrt{e}\lambda t}\right).$$

Finally taking the union bound yields the desired result

$$\mathbb{P}(\max_{\ell\in\overline{S}}\|(\tilde{\mathbf{A}}_{\mathbf{P}})_\ell^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S\mathbf{v}\|_2 \geq t) \leq 2e^{1/12}sN\exp\left(-\frac{t^2 m}{e\lambda^2 + 2\sqrt{e}\lambda t}\right).$$

□

The next lemma yields a similar bound as in Lemma 3.26.

Lemma 3.36. *Assume the conditions of Lemma 3.35. Then for* $e\sqrt{300m}(1+\sqrt{\lambda s}) \geq t \geq 6e\sqrt{300}(1+\sqrt{\lambda s})/\sqrt{m}$

$$\mathbb{P}(\|((\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S)\mathbf{v}\|_2 \geq t\|\mathbf{v}\|_2) \leq 6\exp\left(-\frac{t\sqrt{m}}{e\sqrt{300}(1+\sqrt{\lambda s})}\right).$$

Proof. Assume without loss of generality that $\|\mathbf{v}\|_2 = 1$. We follow the notation in the proof of Theorem 3.33 and use the relation (3.93) to obtain

$$((\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S)\mathbf{v} = \frac{1}{m}\sum_{(i,j)\neq(k,\ell)} g_{ij}g_{k\ell}\delta_{ik}\mathbf{E}_{j\ell}(P_jP_\ell)\mathbf{v} + \frac{1}{m}\sum_{i,j}(g_{ij}^2-1)\mathbf{E}_{jj}(P_j)\mathbf{v},$$

where $i, k \in [m]$ and $j, \ell \in S$. Recall that the notation $\overrightarrow{\mathbf{E}}_j(x)$ corresponds to the block column vector of size s with vector x in its j-th entry and 0 elsewhere. Then $\mathbf{E}_{j\ell}(P_jP_\ell)\mathbf{v} = \overrightarrow{\mathbf{E}}_j(P_jP_\ell v_\ell)$. We introduce the notation $\tau = (i,j)$, $\xi = (k,\ell)$ and $H_{\tau,\xi} = \delta_{ik}\overrightarrow{\mathbf{E}}_j(P_jP_\ell v_\ell)$. Then we can write

$$\begin{aligned}((\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S)\mathbf{v} &= \frac{1}{m}\sum_{\tau\neq\xi} g_\tau g_\xi H_{\tau,\xi} + \frac{1}{m}\sum_{i,j}(g_{ij}^2-1)\mathbf{E}_{jj}(P_j)\mathbf{v} \\ &=: \mathcal{B} + \mathcal{C}. \end{aligned} \tag{3.103}$$

After following the step (3.94) in the proof of Theorem 3.33, we will use the noncommutative Khintchine inequality for decoupled Gaussian chaos as in (3.95) in order to bound the moments of $\|\mathcal{B}\|_2$. Observe that the ℓ_2-norm can be treated as spectral norm and therefore the following bound follows for an integer n

$$\mathbb{E}\|\mathcal{B}\|_2^{2n} \leq \mathbb{E}\|\mathcal{B}\|_{S_{2n}}^{2n} = \frac{1}{m^{2n}}\mathbb{E}\left\|\sum_{\tau\neq\xi} g_\tau g_\xi H_{\tau,\xi}\right\|_{S_{2n}}^{2n} \leq \frac{4^{2n}}{m^{2n}}\mathbb{E}\left\|\sum_{\tau\neq\xi} g_\tau g'_\xi H_{\tau,\xi}\right\|_{S_{2n}}^{2n}$$
$$\leq \frac{4^{2n}}{m^{2n}} \times 2C_n^2 \times \max\left\{\left\|\left(\sum_{\tau\neq\xi} H_{\tau,\xi}H_{\tau,\xi}^*\right)^{1/2}\right\|_{S_{2n}}^{2n}, \left\|\left(\sum_{\tau\neq\xi} H_{\tau,\xi}^*H_{\tau,\xi}\right)^{1/2}\right\|_{S_{2n}}^{2n}, \|F\|_{S_{2n}}^{2n}, \|\tilde{F}\|_{S_{2n}}^{2n}\right\}, \tag{3.104}$$

where $F, \tilde{F}$ are the block matrices $F = (H_{\tau,\xi})_{\tau,\xi}$, $\tilde{F} = (H_{\tau,\xi}^*)_{\tau,\xi}$ with the assumption that $H_{\tau,\tau} = 0$ and $C_n = \frac{(2n)!}{2^n n!}$. We denote the terms in the maximum argument above, from *(a)* to *(d)*.

a) We write the first term as

$$\sum_{\tau\neq\xi} H_{\tau,\xi}H_{\tau,\xi}^* = \sum_{(i,j)\neq(k,\ell)} \delta_{ik}\overrightarrow{\mathbf{E}}_j(P_jP_\ell v_\ell)\overrightarrow{\mathbf{E}}_j^*(P_jP_\ell v_\ell) = m\sum_{j\neq\ell}\mathbf{E}_{jj}[(P_jP_\ell v_\ell)(P_jP_\ell v_\ell)^*].$$

Then it follows that

$$\left\|\left(\sum_{\tau\neq\xi} H_{\tau,\xi}H_{\tau,\xi}^*\right)^{1/2}\right\|_{S_{2n}}^{2n} = \mathrm{Tr}\left(\sum_{\tau\neq\xi} H_{\tau,\xi}H_{\tau,\xi}^*\right)^n$$
$$= m^n\mathrm{Tr}\sum_{\substack{j_1,\ldots,j_n\\ \ell_1,\ldots,\ell_n\\ j_i\neq\ell_i\ \forall i}} \mathbf{E}_{j_1j_1}[(P_{j_1}P_{\ell_1}v_{\ell_1})(P_{j_1}P_{\ell_1}v_{\ell_1})^*]\mathbf{E}_{j_2j_2}[(P_{j_2}P_{\ell_2}v_{\ell_2})(P_{j_2}P_{\ell_2}v_{\ell_2})^*]\ldots$$
$$= m^n\mathrm{Tr}\sum_{\substack{j\\ \ell_1,\ldots,\ell_n\\ j\neq\ell_i\ \forall i}} \mathbf{E}_{jj}[(P_jP_{\ell_1}v_{\ell_1})(P_jP_{\ell_1}v_{\ell_1})^*(P_jP_{\ell_2}v_{\ell_2})(P_jP_{\ell_2}v_{\ell_2})^*\ldots]$$
$$= m^n\sum_j \mathrm{Tr}\left(\sum_{\ell\neq j}(P_jP_\ell v_\ell)(P_jP_\ell v_\ell)^*\right)^n = m^n\sum_j\left\|\left(\sum_{\ell\neq j}(P_jP_\ell v_\ell)(P_jP_\ell v_\ell)^*\right)^n\right\|_{S_1}$$
$$\leq m^n\sum_j \mathrm{rank}\left(\sum_{\ell\neq j}(P_jP_\ell v_\ell)(P_jP_\ell v_\ell)^*\right)^n\left\|\sum_{\ell\neq j}(P_jP_\ell v_\ell)(P_jP_\ell v_\ell)^*\right\|^n$$
$$\leq m^n\sum_j\left(\sum_{\ell\neq j}\mathrm{rank}(P_jP_\ell v_\ell)(P_jP_\ell v_\ell)^*\right)\left(\sum_{\ell\neq j}\|P_jP_\ell v_\ell\|_2^2\right)^n$$
$$\leq m^n\sum_j s\left(\sum_{\ell\neq j}\lambda^2\|v_\ell\|_2^2\right)^n \leq m^n\sum_j s\lambda^{2n} = m^n s^2\lambda^{2n}.$$

We used the relation $\|A\|_{S_1} \leq \mathrm{rank}(A)\|A\|$ due to (A.7) in the first inequality and $\|\mathbf{v}\|_2 = 1$ in the last inequality above.

b) We follow similar steps as above

$$\sum_{\tau\neq\xi} H^*_{\tau,\xi}H_{\tau,\xi} = \sum_{(i,j)\neq(k,\ell)} \delta_{ik}\overrightarrow{\mathbf{E}}^*_j(P_jP_\ell v_\ell)\overrightarrow{\mathbf{E}}_j(P_jP_\ell v_\ell)$$
$$= m\sum_{j\neq\ell}\|P_jP_\ell v_\ell\|_2^2 \leq m\sum_j\sum_{\ell\neq j}\lambda^2\|v_\ell\|_2^2 \leq ms\lambda^2.$$

Then we can easily obtain

$$\left\|\left(\sum_{\tau\neq\xi} H^*_{\tau,\xi}H_{\tau,\xi}\right)^{1/2}\right\|^{2n}_{S_{2n}} = m^n s^n \lambda^{2n}.$$

c) We have

$$\|F\|^{2n}_{S_{2n}} = \mathrm{Tr}(FF^*)^n = \mathrm{Tr}\sum_{\substack{\tau_1,\dots,\tau_n\\ \xi_1,\dots,\xi_n\\ \tau_i\neq\xi_i\ \forall i}} H^*_{\tau_1\xi_1}H_{\tau_1\xi_2}H^*_{\tau_2\xi_2}H_{\tau_2\xi_3}D^*_{\tau_3\xi_3}\dots H^*_{\tau_n\xi_n}H_{\tau_n\xi_1}$$
$$= \mathrm{Tr}\sum_{\substack{i_1,\dots,i_n\\ j_1,\dots,j_n\\ k_1,\dots,k_n\\ \ell_1,\dots,\ell_n}} \delta_{i_1k_1}\overrightarrow{\mathbf{E}}^*_{j_1}(P_{j_1}P_{\ell_1}v_{\ell_1})\delta_{i_1k_2}\mathbf{E}_{j_1}(P_{j_1}P_{\ell_2}v_{\ell_2})\delta_{i_2k_2}\mathbf{E}^*_{j_2}(P_{j_2}P_{\ell_2})\delta_{i_2k_3}\mathbf{E}_{j_2}(P_{j_2}P_{\ell_3}v_{\ell_3})\dots$$
$$= m\mathrm{Tr}\sum_{\substack{j_1,\dots,j_n\\ \ell_1,\dots,\ell_n\\ j_i\neq\ell_i\ \forall i}} \langle P_{j_1}P_{\ell_1}v_{\ell_1}, P_{j_1}P_{\ell_2}v_{\ell_2}\rangle\langle P_{j_2}P_{\ell_2}v_{\ell_2}, P_{j_2}P_{\ell_3}v_{\ell_3}\rangle\dots\langle P_{j_n}P_{\ell_n}v_{\ell_n}, P_{j_n}P_{\ell_1}v_{\ell_1}\rangle$$
$$\leq m\sum_{\substack{j_1,\dots,j_n\\ \ell_1,\dots,\ell_n\\ j_i\neq\ell_i\ \forall i}} \|P_{j_1}P_{\ell_1}\|\|v_{\ell_1}\|_2\|P_{j_1}P_{\ell_2}v_{\ell_2}\|_2\|P_{j_2}P_{\ell_2}\|\|v_{\ell_2}\|_2\dots\|P_{j_n}P_{\ell_n}\|\|v_{\ell_n}\|_2\|P_{j_n}P_{\ell_1}v_{\ell_1}\|$$
$$\leq m\sum_{\substack{j_1,\dots,j_n\\ \ell_1,\dots,\ell_n\\ j_i\neq\ell_i\ \forall i}} \lambda^n\|v_{\ell_1}\|_2^2\|v_{\ell_2}\|_2^2\dots\|v_{\ell_n}\|_2^2 \leq m\lambda^{2n}s^n\left(\sum_\ell\|v_\ell\|_2^2\right)^n \leq m\lambda^{2n}s^n.$$

d) Finally we estimate the last term in the maximum argument in (3.104),

$$\|F\|^{2n}_{S_{2n}} = \mathrm{Tr}(FF^*)^n$$
$$= \mathrm{Tr}\sum_{\substack{i_1,\dots,i_n\\ j_1,\dots,j_n\\ k_1,\dots,k_n\\ \ell_1,\dots,\ell_n}} \delta_{i_1k_1}\overrightarrow{\mathbf{E}}_{j_1}(P_{j_1}P_{\ell_1}v_{\ell_1})\delta_{i_1k_2}\mathbf{E}^*_{j_1}(P_{j_1}P_{\ell_2}v_{\ell_2})\delta_{i_2k_2}\mathbf{E}_{j_2}(P_{j_2}P_{\ell_2})\delta_{i_2k_3}\mathbf{E}^*_{j_2}(P_{j_2}P_{\ell_3}v_{\ell_3})\dots$$
$$= m\mathrm{Tr}\sum_{\substack{j_1,\dots,j_n\\ \ell_1,\dots,\ell_n\\ j_i\neq\ell_i\ \forall i}} \mathbf{E}_{j_1j_1}[(P_{j_1}P_{\ell_1}v_{\ell_1})(P_{j_1}P_{\ell_2}v_{\ell_2})^*]\mathbf{E}_{j_2j_2}[(P_{j_2}P_{\ell_2}v_{\ell_2})(P_{j_2}P_{\ell_3}v_{\ell_3})^*]\dots$$
$$= m\mathrm{Tr}\sum_{\substack{j\\ \ell_1,\dots,\ell_n\\ j\neq\ell_i\ \forall i}} (P_jP_{\ell_1}v_{\ell_1})(P_jP_{\ell_2}v_{\ell_2})^*(P_jP_{\ell_2}v_{\ell_2})(P_jP_{\ell_3}v_{\ell_3})^*\dots(P_jP_{\ell_n}v_{\ell_n})(P_{j_1}P_{\ell_1}v_{\ell_1})^*$$
$$= m\mathrm{Tr}\sum_{\substack{j\\ \ell_1,\dots,\ell_n\\ j\neq\ell_i\ \forall i}} (P_jP_{\ell_1}v_{\ell_1})\|P_jP_{\ell_2}v_{\ell_2}\|_2^2\dots\|P_jP_{\ell_n}v_{\ell_n}\|_2^2(P_jP_{\ell_1}v_{\ell_1})^*$$

$$\leq m \sum_{\substack{j \\ \ell_1,\dots,\ell_n \\ j\neq \ell_i\ \forall i}} \lambda^{2n-2}\mathrm{rank}[(P_jP_{\ell_1}v_{\ell_1})(P_jP_{\ell_1}v_{\ell_1})^*]\|P_jP_{\ell_1}v_{\ell_1}\|_2^2\|v_{\ell_2}\|_2^2\dots\|v_{\ell_n}\|_2^2$$

$$\leq ms\lambda^{2n}\left(\sum_{\ell}\|v_\ell\|_2^2\right)^n \leq ms\lambda^{2n}.$$

Then we have

$$\max\{a,b,c,d\} \leq m^n s^n \lambda^n.$$

Putting this estimate into (3.104) yields

$$\mathbb{E}\|\mathcal{B}\|_2^{2n} \leq \frac{4^{2n}}{m^{2n}}(2C_n^2)\ m^n s^n \lambda^n.$$

After an interpolation step similar to (3.97), we have the following result for general power p

$$(\mathbb{E}\|\mathcal{B}\|_2^p)^{1/p} \leq (4p/e)(2^{5/2})^{1/p}(\lambda s/m)^{1/2}. \tag{3.105}$$

Next task is to bound the ℓ_2-norm of the diagonal term $\mathcal{C}$ defined in (3.103). Since

$$\|\mathcal{C}\|_2^2 = \left\|\frac{1}{m}\sum_{i,j}(g_{ij}^2-1)\mathbf{E}_{jj}(P_j)\mathbf{v}\right\|_2^2 = \left\|\frac{1}{m}\sum_j \Phi_j^{(m)}\overrightarrow{\mathbf{E}}_j(v_j)\right\|_2^2 = \frac{1}{m^2}\sum_j(\Phi_j^{(m)})^2\|v_j\|_2^2$$

where $\Phi^{(m)}$ was defined in the proof of Theorem 3.33 and $\overrightarrow{\mathbf{E}}_j(x)$ denotes the block column vector of size s with vector x in its j-th entry and 0 elsewhere. Then we have for $p \geq 2$

$$\mathbb{E}\|\mathcal{C}\|_2^p \leq \frac{1}{m^p}\mathbb{E}\left(\sum_j(\Phi_j^{(m)})^2\|v_j\|_2^2\right)^{p/2},$$

$$(\mathbb{E}\|\mathcal{C}\|_2^p)^{1/p} \leq \frac{1}{m}\left(\mathbb{E}\left(\sum_j(\Phi_j^{(m)})^2\|v_j\|_2^2\right)^{p/2}\right)^{\frac{2}{p}\cdot\frac{1}{2}}.$$

We use Lemma A.12 due to Latała with the choice of $q = p/2$ in order to estimate the $p/2$-nd moment above. Then the first term in the maximum argument in (A.8) yields

$$\sum_j \mathbb{E}(\Phi_j^{(m)})^2\|v_j\|_2^2 = 2m,$$

where we used $\|\mathbf{v}\|_2 = 1$ and the simple calculation

$$\mathbb{E}(\Phi_j^{(m)})^2 = \mathbb{E}\sum_{k,\ell=1}^m(g_k^2-1)(g_\ell^2-1) = \mathbb{E}\sum_{k=1}^m(g_k^2-1)^2 = m\ \mathbb{E}(g_k^4+1-2g_k^2) = 2m.$$

For the second term in (A.8) recall from (3.100) that

$$\mathbb{E}(\Phi_j^{(m)}\|v_j\|_2)^p \leq 2\|v_j\|_2^p(2\sqrt{m}\sqrt{p}+8p)^p.$$

Hence,

$$\left(\sum_j \mathbb{E}(\Phi_j^{(m)}\|v_j\|_2)^p\right)^{2/p} = 2^{2/p}(2\sqrt{m}\sqrt{p}+8p)^2\left(\sum_j\|v_j\|_2^p\right)^{2/p}$$

$$= 2^{2/p}(2\sqrt{m}\sqrt{p}+8p)^2\|\mathbf{v}\|_{2,p}^2 \leq 2^{2/p}(2\sqrt{m}\sqrt{p}+8p)^2.$$

The last inequality is due to the ordering of the $\ell_{2,p}$-norms, i.e., $\|\mathbf{v}\|_{2,p} \leq \|\mathbf{v}\|_2 = 1$ for $p \geq 2$. Therefore Lemma A.12 gives us

$$\left(\mathbb{E} \left(\sum_j (\Phi_j^{(m)})^2 \|v_j\|_2^2 \right)^{p/2} \right)^{\frac{2}{p}} \leq K \frac{p/2}{\ln(p/2)} 2^{2/p} (2\sqrt{m}\sqrt{p} + 8p)^2,$$

where the universal constant $K < 6$. Simply estimating $\frac{p/2}{\ln(p/2)} \leq p/2$ for $p \geq 6$, we conclude that for $m \geq p \geq 6$

$$\begin{aligned}(\mathbb{E}\|\mathcal{C}\|_2^p)^{1/p} &\leq \frac{\sqrt{3}}{m} 2^{1/p} \sqrt{p}(2\sqrt{m}\sqrt{p} + 8p) \leq 2^{1/p}\sqrt{3}\frac{2p\sqrt{m} + 8p\sqrt{m}}{m} \\ &= p(10\sqrt{3})2^{1/p}(1/\sqrt{m}).\end{aligned}$$

Combining this with (3.105) yields

$$\begin{aligned}(\mathbb{E}\|((\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S)\mathbf{v}\|_2^p)^{1/p} &\leq (4p/e)(2^{5/2})^{1/p}(\lambda s/m)^{1/2} + p(10\sqrt{3})2^{1/p}(1/\sqrt{m}) \\ &\leq 6^{1/p} p\sqrt{300}\frac{1+\sqrt{\lambda s}}{\sqrt{m}}.\end{aligned}$$

for $m \geq p \geq 6$. For this range of p, we apply Proposition A.10 in order to have following estimate for the tail probability

$$\mathbb{P}(\|((\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S)\mathbf{v}\|_2 \geq t) \leq 6\exp\left(-\frac{t\sqrt{m}}{e\sqrt{300}(1+\sqrt{\lambda s})}\right),$$

where t is in the appropriate range as stated in the lemma. This ends the proof. □

• *Proof of Theorem 3.3 continued*

We mainly follow the ideas presented in the proof of the Bernoulli case in Section 3.3.3. It was based on using the *golfing scheme* in order to construct an inexact dual vector satisfying the conditions of Lemma 3.22 which are sufficient for nonuniform recovery. Recall that we partition m rows of $\tilde{\mathbf{A}}_{\mathbf{P}}$ into L disjoint of size m_j with $\sum_{j=1}^L m_j = m$. Then we recursively define the dual vector $\mathbf{u}$ as

$$\mathbf{u}^{(n)} = \frac{1}{m_n}(\mathbf{A}_{\mathbf{P}}{}^{(n)})^*(\mathbf{A}_{\mathbf{P}}{}^{(n)})_S(\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S^{(n-1)}) + \mathbf{u}^{(n-1)},$$

where $\mathbf{u}^{(0)} = 0$. We introduce the residual vector $\mathbf{w}^{(n)} = \mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S^{(n)}$ at every step n. Then we recall the following relation derived earlier

$$\mathbf{w}^{(n)} = \left[\mathbf{P}_S - \frac{1}{m_n}(\mathbf{A}_{\mathbf{P}}{}^{(n)})_S^*(\mathbf{A}_{\mathbf{P}}{}^{(n)})_S\right]\mathbf{w}^{(n-1)}.$$

Now for a moment, we assume that the following inequalities hold for each n with high probability,

$$\|\mathbf{w}^{(n)}\|_2 \leq r_n\|\mathbf{w}^{(n-1)}\|_2, \quad n \in [L], \tag{3.106}$$

$$\max_{i\in\overline{S}} \left\| \frac{1}{m_n}(\mathbf{A}_{\mathbf{P}}{}^{(n)})_i^*(\mathbf{A}_{\mathbf{P}}{}^{(n)})_S\mathbf{w}^{(n-1)} \right\|_2 \leq t_n\|\mathbf{w}^{(n-1)}\|_2, \quad n \in [L], \tag{3.107}$$

with the parameters r_n, t_n to be specified later. Recall that the followings also hold

$$\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 \leq \sqrt{s}\prod_{n=1}^L r_n.$$

$$\max_{i\in\overline{S}} \|u_i\|_2 \leq \sqrt{s} \sum_{n=1}^{L} t_n \prod_{j=1}^{n-1} r_j$$

with the understanding that $\prod_{j=1}^{n-1} r_j = 1$ if $n = 1$. Define the probabilities that (3.106) and (3.107) do not hold as $p_1(n)$ and $p_2(n)$ respectively. Then by Lemma 3.36 and the independence of the blocks,

$$p_1(n) \leq \varepsilon$$

provided

$$m_n \geq \frac{300e^2}{r_n^2}(1+\sqrt{\lambda s})^2 \ln^2(6/\varepsilon).$$

Recall that Lemma 3.36 holds only for an appropriate range of r_n, which we choose below. Similarly, due to Lemma 3.35 and the independence of the blocks

$$p_2(n) \leq \varepsilon$$

provided

$$m_n \geq \left(\frac{e\lambda^2}{t_n^2} + \frac{2\sqrt{e}\lambda}{t_n}\right) \ln(2e^{1/12}sN/\varepsilon).$$

We now choose the parameters L, m_n, t_n, r_n for $n \in [L]$ as follows

$$\begin{aligned}
&L = \lceil \ln(s)/2 \rceil + 2,\\
&m_1, m_2 \geq c(1+\sqrt{\lambda s})^2 \max\left\{\ln(s)\ln^2(6/\varepsilon), \ln(6sN/\varepsilon)\right\},\\
&m_n \geq c(1+\sqrt{\lambda s})^2 \max\left\{\ln^2(6L/\varepsilon), \frac{\ln(6sNL/\varepsilon)}{\ln^2(s)}\right\}\\
&r_1 = r_2 = \frac{1}{2\sqrt{\ln(s)}} \quad \text{and} \quad r_n = \frac{1}{2},\\
&t_1 = t_2 = \frac{1}{8\sqrt{s}}, \quad \text{and} \quad t_n = \frac{\ln(s)}{8\sqrt{s}}, \quad n = 3, \ldots, L,
\end{aligned}$$

where c is an absolute constant. First observe that the choices of r_n and m_n obey the condition stated in Lemma 3.36. Then as shown in (3.74) and (3.75), we have $\|\mathrm{sgn}(\mathbf{x}_S) - \mathbf{u}_S\|_2 \leq 1/4$ and $\max_{i\in\overline{S}} \|u_i\|_2 \leq 1/4$ as required by Lemma 3.22. Furthermore, by our choices of parameters above, we have, $p_1(1), p_1(2), p_2(1), p_2(2) \leq \varepsilon/2$ and $p_1(n), p_2(n) \leq \varepsilon/(2L)$. Then

$$\sum_{n=1}^{L} p_1(n) \leq 2\varepsilon \quad \text{and} \quad \sum_{n=1}^{L} p_2(n) \leq 2\varepsilon.$$

The overall number of samples obey

$$\begin{aligned}
m &= \sum_{n=1}^{L} m_n = m_1 + m_2 + \sum_{n=3}^{L} m_n\\
&\geq 2c(1+\sqrt{\lambda s})^2 \max\left\{\ln(s)\ln^2(6/\varepsilon), \ln(6sN/\varepsilon)\right\}\\
&\quad + c(1+\sqrt{\lambda s})^2 \max\left\{\lceil \ln(s)/2 \rceil \ln^2(6\lceil \ln(s)/2 \rceil/\varepsilon), \ln(6sN\lceil \ln(s)/2 \rceil/\varepsilon)\right\}.
\end{aligned}$$

We can further improve this condition with the refined trick presented at the end of Section 3.3.3. The trick was having more than L submatrices of $\mathbf{A_P}$ and then only using the ones which satisfies

desired properties. Necessary calculations yield that

$$m \geq \tilde{c}(1+\sqrt{\lambda s})^2 \ln(sN)\ln(\varepsilon^{-1}) \tag{3.108}$$

measurements are enough to satisfy Condition (3.46) of Lemma 3.22 with failure probability ε. Finally we deal with Condition (3.45). For the first part, Theorem 3.33 with parameter $\delta = 1/2$ implies that $\|[(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S]_{|\mathcal{H}}^{-1}\| \leq 2$ with probability at least ε provided

$$m \geq C(1+\sqrt{\lambda s})^2 \ln^2\left(6s^2 \sum_{j\in S} k_j(\varepsilon^{-1})\right), \tag{3.109}$$

where C is an absolute constant. The second part of Condition (3.45) follows from Corollary 3.34 with $\delta = 1$. Thus, $\max_{i\in\overline{S}} \|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_i\| \leq 1$ with probability at least ε provided

$$m \geq C(1+\sqrt{\lambda(s+1)})^2 \ln^2\left(6N(s+1)^2 \sum_{j\in N} k_j(\varepsilon^{-1})\right). \tag{3.110}$$

We have so far shown that the conditions of Lemma 3.22 hold with probability at least $1-3\varepsilon$ provided (3.108), (3.109) and (3.110) hold. Using $s < N$ and replacing ε by $\varepsilon/3$, we can see that Condition (3.6)

$$m \geq \tilde{C}(1+\lambda s)\ln^2\left(6N\sum_{j=1}^{N} k_j\right)\ln^2(\varepsilon^{-1})$$

implies all three conditions with an appropriate absolute constant $\tilde{C}$. This ends our proof. □

3.3.6. Alternative approach to the proof of Theorem 3.3

The main tools for the probabilistic estimates in Section 3.3.5 are the noncommutative Khintchine inequalities. These methods give estimates for the moments of matrix valued Gaussian sums and then allow us to yield tail bounds we need in order to prove our result. However, the proofs become quite long and technical due to the involved calculations of the Schatten norms. In this section we present a noncommutative Bernstein type inequality due to Koltchinskii [**73**] for the sum of matrices which are not uniformly bounded. In addition, using this result we provide another proof of Theorem 3.33 which is crucial towards the proof of our following alternative result to Theorem 3.3.

Theorem 3.37. *Let $\mathbf{x} \in \mathcal{H}$ be s-sparse. Let $A \in \mathbb{R}^{m\times N}$ be a Gaussian matrix and $(W_j)_{j=1}^N$ in $\mathbb{R}^d$ be given with parameter $\lambda \in [0,1]$. Assume that*

$$m \geq C(\ln(s)+\lambda s)\ln\left(\ln(s)+\lambda s\right)\ln(2Nsd/\varepsilon), \tag{3.111}$$

where $C > 0$ is a universal constant. Then with probability at least $1-\varepsilon$, $(L1)$ recovers $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}_{\mathbf{P}}\mathbf{x}$.

We do not give a complete proof of Theorem 3.37, and we believe that Theorem 3.39 below, which is the alternative to Theorem 3.33, gives a clear idea how to pursue the rest of the proof. Note that in (3.111) we obtain slightly worse log-factors in the number of required measurements than in Theorem 3.3, but with a significantly shorter and less technical proof. The following proposition is a modified version of [**73**, Proposition 2] where we optimize some parameters appearing in the original proof.

Proposition 3.38. *Assume $X, X_1, \ldots, X_n$ are i.i.d. random self-adjoint $d \times d$ matrices with $\mathbb{E}X = 0$, $\sigma_x^2 := \|\mathbb{E}X^2\|$ and $U_x := \left\|\|X\|\right\|_{\psi_1}$. Then, for all $t > 0$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| \geq t\right) \leq 2d \exp\left(-C\frac{t^2}{n\sigma_x^2 + U_x t + \frac{U_x}{2}\ln\left(\frac{2U_x n}{t}\right)t}\right), \tag{3.112}$$

for an absolute constant $C > 0$.

Note that this tail estimate is in essence similar to the noncommutative matrix Bernstein inequality, Theorem A.3, with the uniform bound on $\|X_i\|$ replaced by the weaker ψ_1-norm. In the original result [**73**, Proposition 2], the variance term σ_x^2 appears as a denominator in the tail bound which is not desirable in our situation since in the proof of Theorem 3.39, a trivial lower bound for σ_x^2 causes the number of measurements m to grow unboundedly. A generalized version of Proposition 3.38 for the sum of independent but not identically distributed matrices can be found in [**74**] and for rectangular random matrices in [**72**].

PROOF. Let $Y_n = X_1 + \ldots + X_n$. For a self adjoint matrix A, denote by $\lambda^+(A), \lambda^-(A)$ the largest and the smallest eigenvalue. Since Y_n is self-adjoint, $\|Y_n\| \geq t$ if and only if $\lambda^+(Y_n) \geq t$ or $\lambda^-(Y_n) \leq -t$, implying that

$$\mathbb{P}(\|Y_n\| \geq t) \leq \mathbb{P}(\lambda^+(Y_n) \geq t) + \mathbb{P}(\lambda^-(Y_n) \leq -t).$$

It is enough to control one of the probabilities on the right hand side since the other follows similarly from replacing Y_n by $-Y_n$. For all $\lambda > 0$, we have

$$\mathbb{P}(\lambda^+(Y_n) \geq t) \leq \mathbb{P}(\mathrm{Tr}(e^{\lambda Y_n}) \geq e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}\mathrm{Tr}(e^{\lambda Y_n}). \tag{3.113}$$

By the well-known *Golden-Thompson inequality* [**4**] which states that $\mathrm{Tr}(e^{A+B}) \leq \mathrm{Tr}(e^A e^B)$ for self-adjoint matrices A and B, one has

$$\begin{aligned}\mathbb{E}\mathrm{Tr}(e^{\lambda Y_n}) &= \mathbb{E}\mathrm{Tr}(e^{\lambda Y_{n-1} + \lambda X_n}) \leq \mathbb{E}\mathrm{Tr}(e^{\lambda Y_{n-1}} e^{\lambda X_n}) = \mathrm{Tr}\mathbb{E}(e^{\lambda Y_{n-1}} e^{\lambda X_n}) \\ &= \mathrm{Tr}(\mathbb{E}e^{\lambda Y_{n-1}}\mathbb{E}e^{\lambda X_n}) \leq \mathbb{E}\mathrm{Tr}(e^{\lambda Y_{n-1}})\left\|\mathbb{E}e^{\lambda X_n}\right\|.\end{aligned}$$

By induction, we conclude that

$$\mathbb{E}\mathrm{Tr}(e^{\lambda Y_n}) \leq \mathbb{E}\mathrm{Tr}(e^{\lambda X_1})\prod_{i=2}^n \left\|\mathbb{E}e^{\lambda X_i}\right\|.$$

Since $\mathbb{E}\mathrm{Tr}(e^{\lambda X_1}) = \mathrm{Tr}(\mathbb{E}e^{\lambda X_1}) \leq d\left\|\mathbb{E}e^{\lambda X}\right\|$, we obtain

$$\mathbb{E}\mathrm{Tr}(e^{\lambda Y_n}) \leq d\left\|\mathbb{E}e^{\lambda X}\right\|^n. \tag{3.114}$$

Koltchinskii [**73**, Proposition 2] uses the Taylor expansion for $e^{\lambda X}$ and the condition $\mathbb{E}X = 0$ to obtain

$$\begin{aligned}\mathbb{E}e^{\lambda X} &= \mathrm{Id} + \mathbb{E}\left[\lambda^2 X^2\left(\frac{1}{2!} + \frac{\lambda X}{3!} + \frac{\lambda^2 X^2}{4!} + \cdots\right)\right] \\ &\leq \mathrm{Id} + \lambda^2\mathbb{E}\left[X^2\left(\frac{1}{2!} + \frac{\lambda\|X\|}{3!} + \frac{\lambda^2\|X\|^2}{4!} + \cdots\right)\right] \\ &= \mathrm{Id} + \lambda^2\mathbb{E}\left[X^2\left(\frac{e^{\lambda\|X\|} - 1 - \lambda\|X\|}{\lambda^2\|X\|^2}\right)\right].\end{aligned}$$

We denote the function $\phi(u) := \frac{e^u-1-u}{u^2}$. Therefore, for all $\tau > 0$,

$$\begin{aligned}\left\|\mathbb{E}e^{\lambda X}\right\| &\le 1 + \lambda^2 \left\|\mathbb{E}[X^2\phi(\lambda\|X\|)]\right\| \\ &\le 1 + \lambda^2 \left\|\mathbb{E}X^2\right\| \phi(\lambda\tau) + \lambda^2 \mathbb{E}\left[\|X\|^2\phi(\lambda\|X\|)\mathbb{1}_{(\|X\|\ge\tau)}\right], \end{aligned} \tag{3.115}$$

where $\mathbb{1}$ denotes the characteristic function and we used the Jensen's inequality. Let $M := 2U_x$ and assume that $\lambda \le 1/M$. Then

$$\begin{aligned}\mathbb{E}\left[\|X\|^2\phi(\lambda\|X\|)\mathbb{1}_{(\|X\|\ge\tau)}\right] &\le \mathbb{E}\left[\|X\|^2\phi\left(\frac{\|X\|}{M}\right)\mathbb{1}_{(\|X\|\ge\tau)}\right] \le M^2\mathbb{E}\left[e^{\|X\|/M}\mathbb{1}_{(\|X\|\ge\tau)}\right] \\ &\le M^2\mathbb{E}^{1/2}(e^{2\|X\|/M})\mathbb{P}^{1/2}(\|X\|\ge\tau),\end{aligned}$$

where we used the monotonicity of ϕ in the first inequality and the Hölder's inequality in the last inequality. By the definition of the Orlicz norm $U_x = \left\|\|X_\ell\|\right\|_{\psi_1}$, we have $\mathbb{E}e^{2\|X\|/M} \le 2$ and also

$$\mathbb{P}(\|X\| \ge \tau) \le \exp\left(-\frac{2\tau}{M}\right).$$

Putting these estimates in (3.115), we obtain

$$\left\|\mathbb{E}e^{\lambda X}\right\| \le 1 + \lambda^2\sigma_x^2\phi(\lambda\tau) + \lambda^2M^2\sqrt{2}\exp\left(-\frac{\tau}{M}\right). \tag{3.116}$$

Now observe that $\phi(u) \ge 1/2$ for $u \ge 0$. Then we can reformulate (3.116) as

$$\left\|\mathbb{E}e^{\lambda X}\right\| \le 1 + \left(\lambda^2\sigma_x^2 + 3\lambda^2M^2\exp\left(-\frac{\tau}{M}\right)\right)\phi(\lambda\tau).$$

Now further assume that

$$\lambda \le \frac{1}{\tau} \le \frac{1}{M}. \tag{3.117}$$

Then $\lambda\tau \le 1$ which implies $\phi(\lambda\tau) \le 1$. Therefore,

$$\left\|\mathbb{E}e^{\lambda X}\right\| \le 1 + \left(\lambda^2\sigma_x^2 + 3\lambda^2M^2\exp\left(-\frac{\tau}{M}\right)\right) \le \exp\left(\lambda^2\sigma_x^2 + 3\lambda^2M^2\exp\left(-\frac{\tau}{M}\right)\right).$$

Combining this with (3.113) and (3.114), we arrive at

$$\mathbb{P}(\|Y_n\| \ge t) \le 2d\exp\left(-\lambda t + n\lambda^2\sigma_x^2 + 3n\lambda^2M^2\exp\left(-\frac{\tau}{M}\right)\right).$$

Next we minimize the convex function of λ in the exponential under the constraint $\frac{1}{\tau} \ge \lambda > 0$. A bit of analysis yields

$$\mathbb{P}(\|Y_n\| \ge t) \le 2d\exp\left(-\frac{t^2}{4n\left(\sigma_x^2 + 3M^2\exp\left(-\frac{\tau}{M}\right)\right) + 2\tau t}\right).$$

Next we optimize the denominator in the exponential above with respect to $\tau \ge M$ (recall the assumption (3.117)). It is a convex function of τ which takes the minimum at $\tau_0 = M\ln\left(\frac{6nM}{t}\right)$ which is greater than M as required. We plug this choice in

$$\mathbb{P}(\|Y_n\| \ge t) \le 2d\exp\left(-\frac{t^2}{4n\sigma_x^2 + 2Mt + 2M\ln\left(\frac{6nM}{t}\right)t}\right).$$

This gives the desired result. □

- *Conditioning of the submatrix* $(\mathbf{A_P})_S$

We recall the rescaled matrix $\tilde{\mathbf{A}}_\mathbf{P} = \frac{1}{\sqrt{m}}\mathbf{A_P}$. The following result is an alternative to Theorem 3.33.

Theorem 3.39. *Let $A \in \mathbb{R}^{m\times N}$ be a measurement matrix whose entries are i.i.d. Gaussian random variables and $(W_j)_{j=1}^N$ be a fusion frame in $\mathbb{R}^d$ given with parameter $\lambda \in [0,1]$. Then, the block matrix $\tilde{\mathbf{A}}_{\mathbf{P}}$ satisfies*

$$\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\| \le \delta$$

for $0 < \delta \le 1$ with probability at least $1-\varepsilon$, provided

$$m \ge C\delta^{-2}(\ln(s) + \lambda s)\ln\left(\frac{\ln(s)+\lambda s}{\delta}\right)\ln(2sd/\varepsilon),$$

where $C > 0$ is an absolute constant.

Proof. As in the proof of Theorem 3.24, denote $\mathbf{Y}_\ell = (g_{\ell j}P_j)_{j\in S}$ for $\ell \in [m]$ as the ℓ-th block column vector of $(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*$. We introduce the block matrices $\mathbf{X}_\ell := \frac{1}{m}(\mathbf{Y}_\ell\mathbf{Y}_\ell^* - \mathbb{E}\mathbf{Y}_\ell\mathbf{Y}_\ell^*)$ which have mean zero. Then we write

$$(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S = \sum_{\ell=1}^m \mathbf{X}_\ell.$$

We use Proposition 3.38 in order to estimate the tail of $\|\sum_{\ell=1}^m \mathbf{X}_\ell\|$. The variance term was estimated before for the Bernoulli case by

$$\sigma_x^2 \le \frac{\lambda^2 s}{m^2}.$$

This also holds for the Gaussian case since $\mathbb{E}g_i^2 = 1$ and $\mathbb{E}g_ig_j = 0,\ i \ne j$. It remains to estimate the Orlicz norm $U_x = \big\|\|X_\ell\|\big\|_{\psi_1}$. We have

$$\begin{aligned}\big\|\|X_\ell\|\big\|_{\psi_1} &= \frac{1}{m}\big\|\|\mathbf{Y}_\ell\mathbf{Y}_\ell^* - \mathbb{E}\mathbf{Y}_\ell\mathbf{Y}_\ell^*\|\big\|_{\psi_1} \le \frac{1}{m}\left(\big\|\|\mathbf{Y}_\ell\mathbf{Y}_\ell^*\|\big\|_{\psi_1} + \big\|\|\mathbb{E}\mathbf{Y}_\ell\mathbf{Y}_\ell^*\|\big\|_{\psi_1}\right)\\ &\le \frac{1}{m}\left(\big\|\|\mathbf{Y}_\ell\mathbf{Y}_\ell^*\|\big\|_{\psi_1} + \mathbb{E}\big\|\|\mathbf{Y}_\ell\mathbf{Y}_\ell^*\|\big\|_{\psi_1}\right) \le \frac{2}{m}\big\|\|\mathbf{Y}_\ell\mathbf{Y}_\ell^*\|\big\|_{\psi_1} = \frac{2}{m}\big\|\|\mathbf{Y}_\ell^*\|^2\big\|_{\psi_1}.\end{aligned}$$

The second inequality above is due to the Jensen's inequality [**54**, Theorem 7.10], since every norm is convex. We continue estimating

$$\begin{aligned}\|\mathbf{Y}_\ell^*\|^2 &= \max_{\substack{\|\mathbf{x}\|_2=1\\ \mathbf{x}\in\mathbb{R}^{S\cdot d}}}\left\|\sum_{j\in S} g_{\ell j}P_jx_j\right\|_2^2 = \max_{\|\mathbf{x}\|_2=1}\sum_{j,k} g_{\ell j}g_{\ell k}\langle P_jx_j, P_kx_k\rangle\\ &\le \max_{\|\mathbf{x}\|_2=1}\sum_{j\in S} g_{\ell j}^2\|x_j\|_2^2 + \max_{\|\mathbf{x}\|_2=1}\sum_{j\ne k}|g_{\ell j}g_{\ell k}|\|P_jP_k\|\|x_j\|_2\|x_k\|_2\\ &\le \max_{\substack{\|y\|_1=1\\ y\in\mathbb{R}^S}}\sum_{j\in S} g_{\ell j}^2\cdot y_j + \lambda\max_{\|\mathbf{x}\|_2=1}\left(\sum_{j\in S}|g_{\ell j}|\|x_j\|_2\right)^2\\ &= \max_{j\in S}|g_{\ell j}|^2 + \lambda\left(\max_{\substack{\|y\|_2=1\\ y\in\mathbb{R}^S}}\sum_{j\in S}|g_{\ell j}|y_j\right)^2 = \max_{j\in S}|g_{\ell j}|^2 + \lambda\sum_{j\in S} g_{\ell j}^2.\end{aligned}$$

The last line above follows from duals of the ℓ_2 and ℓ_∞-norms. Moreover,

$$\begin{aligned}\big\|\|X_\ell\|\big\|_{\psi_1} &\le \frac{2}{m}\big\|\|\mathbf{Y}_\ell^*\|^2\big\|_{\psi_1} \le \frac{2}{m}\big\|\max_{j\in S}|g_{\ell j}|^2\big\|_{\psi_1} + \frac{2\lambda}{m}\big\|\sum_{j\in S} g_{\ell j}^2\big\|_{\psi_1}\\ &\lesssim \frac{2}{m}\ln(s)\max_{j\in S}\big\|g_{\ell j}^2\big\|_{\psi_1} + \frac{2\lambda}{m}\sum_{j\in S}\big\|g_{\ell j}^2\big\|_{\psi_1} \qquad (3.118)\end{aligned}$$

$$\lesssim \frac{4\ln(s)}{m}\big\||g|\big\|_{\psi_2}^2 + \frac{4\lambda s}{m}\big\||g|\big\|_{\psi_2}^2, \tag{3.119}$$

where the g is a standard Gaussian random variable. Above, (3.118) follows from Lemma 3.21 and the triangle inequality; (3.119) is due to Lemma 3.20 and the equivalence of the $\|\cdot\|_{\psi_\alpha}$ and $\|\cdot\|'_{\psi_\alpha}$-norms. It is clear that $\big\||g|\big\|_{\psi_2} \leq c$ for some $c > 0$. Therefore

$$U_x = \big\|\|X_\ell\|\big\|_{\psi_1} \leq C\frac{\ln(s)+\lambda s}{m}.$$

for some C. Plugging the estimates for U_x and σ_x^2 into (3.112) in Proposition 3.38 we obtain

$$\mathbb{P}(\|(\tilde{\mathbf{A}}_{\mathbf{P}})_S^*(\tilde{\mathbf{A}}_{\mathbf{P}})_S - \mathbf{P}_S\| \geq \delta) \leq 2sd\exp\left(-C\frac{\delta^2 m}{\lambda^2 s + (\ln(s)+\lambda s)\delta + (\ln(s)+\lambda s)\ln\left(\frac{\ln(s)+\lambda s}{\delta}\right)\delta}\right).$$

Bounding this by ε yields the desired result. □

3.3.7. Proof of Theorem 3.4

The proof is analogous to the one of [**54**, Theorem 12.22]. It invokes Lemma 3.23 which gives sufficient conditions on the measurement matrices for robust and stable nonuniform recovery. Since Condition (3.49) requires normalization of the matrix, we will work with the matrix $\tilde{\mathbf{A}}_{\mathbf{P}} = \frac{1}{\sqrt{m}}\mathbf{A}_{\mathbf{P}}$. Then observe that the optimization problem $(L1)^\eta$ is equivalent to

$$\min_{\mathbf{z}\in\mathcal{H}} \|\mathbf{z}\|_{2,1} \quad \text{s.t.} \quad \left\|\tilde{\mathbf{A}}_{\mathbf{P}}\mathbf{z} - \frac{1}{\sqrt{m}}\mathbf{y}\right\|_2 \leq \eta.$$

We follow the golfing scheme as in the proof of Theorem 3.1, see Section 3.3.3. In particular, we make the same choices of the parameters L, r_n, t_n, q_n, m_n as before. We choose m_n as follows

$$\begin{aligned} m_1 &\geq c(1+\|\Lambda_S\|_\infty)\ln(N)L\ln(2L\varepsilon^{-1}),\\ m_n &\geq c(1+\|\Lambda_S\|_\infty)\ln(N)\ln(2L\varepsilon^{-1}), \quad n \geq 2. \end{aligned}$$

These choices change the number of overall samples m only up to a constant with respect to the choices in the proof of Theorem 3.1. Then Conditions (3.49), (3.50), (3.51), (3.52) are all satisfied for the normalized matrix $\tilde{\mathbf{A}}_{\mathbf{P}}$ with probability at least $1-\varepsilon$ with appropriate choices of the variables $\delta, \beta, \gamma, \theta$. It remains to verify that the vector $\mathbf{h} \in \mathbb{R}^{md}$ constructed in Section 3.3.3 satisfying $\mathbf{u} = \tilde{\mathbf{A}}_{\mathbf{P}}^*\mathbf{h}$ satisfies Condition (3.53). For simplicity, assume without loss of generality that the first L values of n are used in the construction of the dual vector in (3.67). Then recall that

$$\mathbf{u} = \sum_{n=1}^{L}\frac{1}{m_n}(\mathbf{A}_{\mathbf{P}}{}^{(n)})^*(\mathbf{A}_{\mathbf{P}}{}^{(n)})_S\mathbf{w}^{(n-1)} = \sum_{n=1}^{L}\frac{m}{m_n}(\tilde{\mathbf{A}}_{\mathbf{P}}^{(n)})^*(\tilde{\mathbf{A}}_{\mathbf{P}}^{(n)})_S\mathbf{w}^{(n-1)}.$$

Hence, $\mathbf{u} = \tilde{\mathbf{A}}_{\mathbf{P}}^*\mathbf{h}$ with $\mathbf{h}^* = ((\mathbf{h}^{(1)})^*, \dots, (\mathbf{h}^{(L)})^*, 0, \dots, 0)$ where

$$\mathbf{h}^{(n)} = \frac{m}{m_n}(\tilde{\mathbf{A}}_{\mathbf{P}}^{(n)})_S\mathbf{w}^{(n-1)} \in \mathbb{R}^{m_n d}, \quad n = 1, \dots, L'.$$

Then we have

$$\|\mathbf{h}\|_2^2 = \sum_{n=1}^{L}\|\mathbf{h}^{(n)}\|_2^2 = \sum_{n=1}^{L}\frac{m}{m_n}\left\|\sqrt{\frac{m}{m_n}}(\tilde{\mathbf{A}}_{\mathbf{P}}^{(n)})_S\mathbf{w}^{(n-1)}\right\|_2^2$$
$$\sum_{n=1}^{L}\frac{m}{m_n}\left\|\sqrt{\frac{1}{m_n}}(\mathbf{A}_{\mathbf{P}}{}^{(n)})_S\mathbf{w}^{(n-1)}\right\|_2^2.$$

We also recall the relation (3.66) of the vectors $\mathbf{w}^{(n)}$. This gives, for $n \geq 1$,

$$\begin{aligned}\left\|\sqrt{\frac{1}{m_n}}(\mathbf{A_P}^{(n)})_S\mathbf{w}^{(n-1)}\right\|_2^2 &= \left\langle \frac{1}{m_n}(\mathbf{A_P}^{(n)})_S^*(\mathbf{A_P}^{(n)})_S\mathbf{w}^{(n-1)}, \mathbf{w}^{(n-1)} \right\rangle \\ &= \left\langle \left(\frac{1}{m_n}(\mathbf{A_P}^{(n)})_S^*(\mathbf{A_P}^{(n)})_S - \mathbf{P}_S\right)\mathbf{w}^{(n-1)}, \mathbf{w}^{(n-1)} \right\rangle + \|\mathbf{w}^{(n-1)}\|_2^2 \\ &= \langle \mathbf{w}^{(n)}, \mathbf{w}^{(n-1)} \rangle + \|\mathbf{w}^{(n-1)}\|_2^2 \leq \|\mathbf{w}^{(n)}\|_2^2\|\mathbf{w}^{(n-1)}\|_2^2 + \|\mathbf{w}^{(n-1)}\|_2^2.\end{aligned}$$

Recall from the assumption (3.69) that $\|\mathbf{w}^{(n)}\|_2^2 \leq r_n'\|\mathbf{w}^{(n-1)}\|_2^2 \leq \|\mathbf{w}^{(n-1)}\|_2^2$. Then we obtain

$$\begin{aligned}\left\|\sqrt{\frac{1}{m_n}}(\mathbf{A_P}^{(n)})_S\mathbf{w}^{(n-1)}\right\|_2^2 &\leq 2\|\mathbf{w}^{(n-1)}\|_2^2 \leq \|\mathbf{w}^{(0)}\|_2^2\prod_{j=1}^{n-1}(r_j')^2 \\ &= 2\|\mathrm{sgn}(\mathbf{x})_S\|_2^2\prod_{j=1}^{n-1}(r_j')^2 = 2s\prod_{j=1}^{n-1}(r_j')^2.\end{aligned}$$

Assume that $m \leq C(1+\lambda s)\ln(N)\ln(sk)\ln(2\varepsilon^{-1})$ so that m is just large enough to satisfy (3.8). Recall the definition of $L = \lceil \ln(s)/\ln\ln(N)\rceil + 3$. Then by our choices of m_n, we have $\frac{m}{m_n} \leq L$ for $n \geq 2$ and $\frac{m}{m_1} \leq c$ for some $c > 0$. (If m is much larger, one can rescale m_n proportionally to achieve the same ratio.) This yields

$$\begin{aligned}\|\mathbf{h}\|_2^2 &\leq 2s\sum_{n=1}^{L}\frac{m}{m_n}\prod_{j=1}^{n-1}(r_j')^2 \leq 2s\left(\frac{c}{2\ln(N)} + \sum_{n=2}^{L}L\prod_{j=1}^{n-1}\frac{1}{2\ln N}\right) \\ &\leq 2C's\left(1 + \frac{L}{2\ln(N)}\frac{1}{[1-1/(2\ln(N))]}\right) \leq C''s,\end{aligned}$$

where we used the convention $\prod_{j=1}^{0}(r_j')^2 = 1$. Therefore, all conditions of Lemma 3.23 are satisfied for $\mathbf{x}$ and $\tilde{\mathbf{A}}_\mathbf{P}$ with probability at least $1-\varepsilon$. This completes the proof. □

3.4. Proofs of uniform results

3.4.1. Proof of Theorem 3.7

We know from Theorems 3.14 and 3.15 that if the FRIP constant δ_s of the matrix $\mathbf{A_P}$ is small enough then we have exact and robust uniform recovery via $(L1)$ and $(L1)^\eta$ programs respectively. Therefore our goal is to derive an estimate for the FRIP constant δ_s of the normalized matrix $\tilde{\mathbf{A}}_\mathbf{P} = \frac{1}{\sqrt{m}}\mathbf{A_P}$. The main tool we use for this estimate is provided by Krahmer, Mendelson and Rauhut in [**75**]. They derive a tail bound for a random variable X of the form

$$X = \sup_{A\in\mathcal{A}}\left|\|A\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|A\boldsymbol{\xi}\|_2^2\right|,$$

where $\mathcal{A}$ is a set of matrices and $\boldsymbol{\xi}$ is a subgaussian vector, i.e., has independent, mean-zero, variance 1, c-subgaussian entries. We will later see that the FRIP constant of a subgaussian matrix can be written as such a random variable. The tail bound for X is in terms of two types of complexity parameters of the set of matrices $\mathcal{A}$. The first one denoted by $d_F(\mathcal{A})$ and $d_{2\to2}(\mathcal{A})$, are the radius of $\mathcal{A}$ in the Frobenius norm and the operator norm respectively. In other words,

$$d_F(\mathcal{A}) = \sup_{A\in\mathcal{A}}\|A\|_F \quad \text{and} \quad d_{2\to2}(\mathcal{A}) = \sup_{A\in\mathcal{A}}\|A\|.$$

The second one is Talagrand's $\gamma_2(\mathcal{A}, \|\cdot\|_{2\to 2})$ functional, see [**100**] for a precise definition. With these notations, the result from [**75**] reads as follows.

Theorem 3.40. *Let $\mathcal{A} \subset \mathbb{C}^{m\times n}$ be a set of symmetric matrices, $\mathcal{A} = -\mathcal{A}$. Let $\boldsymbol{\xi}$ be a subgaussian vector of length n with parameter c. Set*

$$\begin{aligned} E &= d_F(\mathcal{A})\gamma_2(\mathcal{A}, \|\cdot\|) + \gamma_2(\mathcal{A}, \|\cdot\|)^2, \\ V &= d_{2\to 2}(\mathcal{A})[\gamma_2(\mathcal{A}, \|\cdot\|) + d_F(\mathcal{A})] \quad \text{and} \quad U = d_{2\to 2}^2(\mathcal{A}). \end{aligned} \tag{3.120}$$

Then, for $t > 0$,

$$\mathbb{P}\left(\sup_{A\in\mathcal{A}} \left|\|A\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|A\boldsymbol{\xi}\|_2^2\right| \geq C_1 E + t\right) \leq 2\exp\left(-C_2 \min\left\{\frac{t^2}{V^2}, \frac{t}{U}\right\}\right). \tag{3.121}$$

The constants $C_1, C_2 > 0$ depend only on c.

A similar result can also be found in [**37**] with a slight improvement of the parameter V. We recall that for a metric space (T, d) and $u > 0$ the covering number $\mathcal{N}(T, d, u)$ is defined as the smallest number of open balls of radius u in (T, d) needed to cover T. The well-known Dudley integral (see, e.g., [**100**]) relates the γ_a-functionals to such covering numbers. Specifically, the γ_2-functional of a set of matrices $\mathcal{A}$ equipped with the operator norm can be bounded as

$$\gamma_2(\mathcal{A}, \|\cdot\|) \leq c\int_0^{d_{2\to 2}(\mathcal{A})} \sqrt{\ln \mathcal{N}(\mathcal{A}, \|\cdot\|, u)}du. \tag{3.122}$$

Now we can pass to the proof of our result. The strategy is similar to the partial random Fourier case [**95**] by combining results from [**75**]. Recall from (3.31) that

$$\delta_s = \sup_{\mathbf{x}\in D_{s,N}} \left|\|\tilde{\mathbf{A}}_{\mathbf{P}}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| = \sup_{\mathbf{x}\in D_{s,N}} \left|\|\tilde{\mathbf{A}}_{\mathbf{P}}\mathbf{x}\|_2^2 - \mathbb{E}\|\tilde{\mathbf{A}}_{\mathbf{P}}\mathbf{x}\|_2^2\right|,$$

and that $D_{s,N} = \{\mathbf{x} \in \mathcal{H} : x_i \in W_i, \|\mathbf{x}\|_2 \leq 1, \|\mathbf{x}\|_0 \leq s\}$. The relation $\mathbb{E}\|\tilde{\mathbf{A}}_{\mathbf{P}}\mathbf{x}\|_2^2 = \|\mathbf{x}\|_2^2$ follows easily by noticing that each entry of A is an independent random variable with mean-zero and variance 1. Now observe that

$$\tilde{\mathbf{A}}_{\mathbf{P}} = \frac{1}{\sqrt{m}} \sum_{i\in[m], j\in[N]} \xi_{ij}\mathbf{Q}_{ij},$$

where $\mathbf{Q}_{ij} := \mathbf{E}_{ij}(P_j)$ and $\mathbf{E}_{ij}(A)$ denotes the block matrix with $m \times N$ blocks each of size $d \times d$ where $A \in \mathbb{R}^{d\times d}$ is at the intersection of i-th block row and j-th block column, and everywhere else is 0. Moreover,

$$\tilde{\mathbf{A}}_{\mathbf{P}}\mathbf{x} = \frac{1}{\sqrt{m}} \sum_{i\in[m], j\in[N]} \xi_{ij}\mathbf{Q}_{ij}\mathbf{x}.$$

Here, for $\mathbf{x} \in D_{s,N}$, we can see that if $j \in \text{supp}(\mathbf{x})$, the block column vector $\mathbf{Q}_{ij}\mathbf{x} \in \mathbb{R}^{m\cdot d}$ consists of the vector x_j in its i-th entry and 0 everywhere else. If $j \notin \text{supp}(\mathbf{x})$, then $\mathbf{Q}_{ij}\mathbf{x} = 0$. We define the matrix $V_{\mathbf{x}}$ whose columns are $\frac{1}{\sqrt{m}}\mathbf{Q}_{ij}\mathbf{x}$ for $i \in [m], j \in [N]$, i.e.,

$$V_{\mathbf{x}} = \frac{1}{\sqrt{m}}(\mathbf{Q}_{11}\mathbf{x}|\mathbf{Q}_{12}\mathbf{x}|\dots|\mathbf{Q}_{mN}\mathbf{x}).$$

Then $\tilde{\mathbf{A}}_{\mathbf{P}}\mathbf{x} = V_{\mathbf{x}}\boldsymbol{\xi}$ where $\boldsymbol{\xi}$ is a subgaussian vector of length mN. Denoting the set $\mathcal{A} = \{V_{\mathbf{x}} : \mathbf{x} \in D_{s,N}\}$, we can write

$$\delta_s = \sup_{A\in\mathcal{A}} \left| \|A\boldsymbol{\xi}\|_2^2 - \mathbb{E}\|A\boldsymbol{\xi}\|_2^2 \right| = \sup_{\mathbf{x}\in D_{s,N}} \left| \|V_{\mathbf{x}}\boldsymbol{\xi}\|_2^2 - \|\mathbf{x}\|_2^2 \right|.$$

Since $\mathbf{x} \in D_{s,N}$ implies $-\mathbf{x} \in D_{s,N}$, the set $\mathcal{A}$ is symmetric. Therefore we can use Theorem 3.40 in order to derive a tail estimate of the FRIP constant. The next lemma gives an estimate for the variable E defined in (3.120),

$$E = d_F(\mathcal{A})\gamma_2(\mathcal{A}, \|\cdot\|) + \gamma_2(\mathcal{A}, \|\cdot\|)^2.$$

Lemma 3.41. *Assume for some $\delta \in (0,1)$*

$$m \geq \tilde{C}\delta^{-2}(\lambda s + \sqrt{s})\ln(Nd)\ln^2(sk)\left(\ln(N) + k\ln\left(s\sqrt{k}\right)\right), \tag{3.123}$$

where $\tilde{C} > 0$ is a universal constant. Then $E \leq \delta$.

PROOF. For any $\mathbf{x} \in D_{s,N}$ it is easy to see that $\|V_{\mathbf{x}}\|_F = 1$. This gives $d_F(\mathcal{A}) = 1$. Next, to estimate the γ_2-functional in (3.120), recall from (3.122) that

$$\gamma_2(\mathcal{A}, \|\cdot\|) \lesssim \int_0^{d_{2\to 2}(\mathcal{A})} \sqrt{\ln \mathcal{N}(\mathcal{A}, \|\cdot\|, u)}du.$$

For a given $\mathbf{x} \in D_{s,N}$, let $\hat{\mathbf{x}} = (x_1|x_2|\dots|x_N)$ be the matrix with columns x_i. Then the diameter of $\mathcal{A}$ can be bounded by

$$d_{2\to 2}(\mathcal{A}) = \sup_{\mathbf{x}\in D_{s,N}} \|V_{\mathbf{x}}\| = \frac{1}{\sqrt{m}} \sup_{\mathbf{x}\in D_{s,N}} \|\hat{\mathbf{x}}\| \leq \frac{1}{\sqrt{m}} \sup_{\mathbf{x}\in D_{s,N}} \|\hat{\mathbf{x}}\|_F = \frac{1}{\sqrt{m}}. \tag{3.124}$$

The covering number will be estimated separately for small and large value of u. One estimate is good for small values of u. First we introduce the set $B_S^2 := \{\mathbf{x} : \mathrm{supp}(\mathbf{x}) \subset S \subset [N], \|\mathbf{x}\|_2 \leq 1\}$. Furthermore define the norm $|||\mathbf{x}||| := \|V_{\mathbf{x}}\|$. The $V_{\mathbf{x}}$ is linear in $\mathbf{x}$ so that $|||\mathbf{x} - \mathbf{y}||| = \|V_{\mathbf{x}} - V_{\mathbf{y}}\|$. Also, observe that

$$|||\mathbf{x}||| = \|V_{\mathbf{x}}\| = \frac{1}{\sqrt{m}}\|\hat{\mathbf{x}}\| \leq \frac{1}{\sqrt{m}}\|\mathbf{x}\|_2. \tag{3.125}$$

Then using subadditivity of covering numbers and a standard volumetric argument (see, e.g., [**94**, Chapter 8.4])

$$\begin{aligned}
\mathcal{N}(\mathcal{A}, \|\cdot\|, u) = \mathcal{N}(D_{s,N}, |||\cdot|||, u) &= \sum_{S\subset[N],|S|=s} \mathcal{N}\left(B_S^2, |||\cdot|||, u\right) \\
&\leq \sum_{S\subset[N],|S|=s} \mathcal{N}\left(B_S^2, \frac{\|\cdot\|_2}{\sqrt{m}}, u\right) = \sum_{S\subset[N],|S|=s} \mathcal{N}\left(B_S^2, \|\cdot\|_2, u\sqrt{m}\right) \\
&\leq \binom{N}{s}\left(1 + \frac{2}{u\sqrt{m}}\right)^{sk} \leq \left(\frac{eN}{s}\right)^s \left(1 + \frac{2}{u\sqrt{m}}\right)^{sk}.
\end{aligned}$$

Then for $u > 0$, we obtain

$$\ln \mathcal{N}(\mathcal{A}, \|\cdot\|, u) \leq s\ln(eN/s) + sk\ln\left(1 + \frac{2}{u\sqrt{m}}\right). \tag{3.126}$$

We also used that $\dim(B_S^2) = sk$. For large values of u, we first define the set

$$B_{2,1} := \left\{\mathbf{x} \in \mathcal{H} : \|\mathbf{x}\|_{2,1} \leq 1, \|\mathbf{x}\|_2 \leq \frac{1}{\sqrt{s}}\right\}.$$

Then it is evident that $D_{s,N} \subset \sqrt{s}B_{2,1}$. Therefore,

$$\mathcal{N}(\mathcal{A}, \|\cdot\|, u) = \mathcal{N}(D_{s,N}, |||\cdot|||, u) \leq \mathcal{N}(\sqrt{s}B_{2,1}, |||\cdot|||, u) = \mathcal{N}\left(B_{2,1}, |||\cdot|||, \frac{u}{\sqrt{s}}\right).$$

The next lemma provides an estimate of $\mathcal{N}(B_{2,1}, |||\cdot|||, u)$.

Lemma 3.42. *For $u > 0$,*

$$\sqrt{\ln \mathcal{N}(B_{2,1}, |||\cdot|||, u)} \leq \sqrt{\left(\frac{8\lambda + 16/\sqrt{s}}{mu^2} + \frac{6}{\sqrt{m}u}\right)} \sqrt{\ln(md + mN)\left[\ln(N) + k\ln\left(1 + \frac{4}{u\sqrt{m}}\right)\right]}.$$

Proof. Fix $u > 0$ and $\mathbf{x} \in B_{2,1}$. The idea is to approximate $\mathbf{x}$ by a finite set of very sparse vectors of ℓ_2-norm 1. To this end, we discretize the unit sphere of each fusion frame subspace W_j. Denote $S_j = \{\mathbf{y} \in \mathcal{H} : \|y_j\|_2 = 1; y_i = 0,\ i \neq j\}$. A standard volumetric argument implies that

$$\mathcal{N}(S_j, \|\cdot\|_2, \tilde{\varepsilon}) \leq \left(1 + \frac{2}{\tilde{\varepsilon}}\right)^k$$

for $\tilde{\varepsilon}$ to be specified later. For each j, let $T_j \subset S_j$ be the covering set of S_j with this cardinality. We use 1-sparse elements from the set $\mathcal{T} = \bigcup_{j\in[N]} T_j$ in order to find a vector $\mathbf{z}$ that is close to $\mathbf{x}$. The so-called empirical method of Maurey will be employed for that. To this end, we define a random vector $\tilde{\mathbf{Z}}$ as follows

$$\mathbb{P}\left(\tilde{\mathbf{Z}} = \overrightarrow{\mathbf{E}}_j\left(\frac{x_j}{\|x_j\|_2}\right)\right) = \|x_j\|_2$$

for j such that $\|x_j\|_2 \neq 0$ and $\tilde{\mathbf{Z}} = 0$ with probability $1 - \|\mathbf{x}\|_{2,1}$. Here $\overrightarrow{\mathbf{E}}_j(x) \in \mathbb{R}^{Nd}$ corresponds to the block column vector of size N with vector x in its j-th block and 0 elsewhere. Since $\|\mathbf{x}\|_{2,1} \leq 1$, this is a valid probability distribution. Observe that $\mathbb{E}\tilde{\mathbf{Z}} = \mathbf{x}$. For a number M to be determined later, let $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_M$ be independent copies of $\tilde{\mathbf{Z}}$, and put

$$\tilde{\mathbf{z}} = \frac{1}{M}\sum_{\ell=1}^{M} \tilde{\mathbf{Z}}_\ell.$$

We now denote $\mathbf{Z}_\ell \in \mathcal{T}$ as the closest vector to $\tilde{\mathbf{Z}}_\ell$ in the set $\mathcal{T}$ for all ℓ. Then we have $\|\tilde{\mathbf{Z}}_\ell - \mathbf{Z}_\ell\|_2 \leq \tilde{\varepsilon}$. The M-sparse vector

$$\mathbf{z} = \frac{1}{M}\sum_{\ell=1}^{M} \mathbf{Z}_\ell$$

is our candidate to approximate $\mathbf{x}$. We estimate the distance of $\mathbf{z}$ to $\mathbf{x}$ in $|||\cdot|||$ by first using the triangle inequality,

$$\begin{aligned}
|||\mathbf{z} - \mathbf{x}||| &\leq |||\mathbf{z} - \tilde{\mathbf{z}}||| + |||\tilde{\mathbf{z}} - \mathbf{x}||| \leq |||\frac{1}{M}\sum_{\ell=1}^{M}(\mathbf{Z}_\ell - \tilde{\mathbf{Z}}_\ell)||| + |||\tilde{\mathbf{z}} - \mathbf{x}||| \\
&\leq \frac{1}{M\sqrt{m}}\sum_{\ell=1}^{M}\|\tilde{\mathbf{Z}}_\ell - \mathbf{Z}_\ell\|_2 + |||\tilde{\mathbf{z}} - \mathbf{x}||| \leq \frac{\tilde{\varepsilon}}{\sqrt{m}} + |||\tilde{\mathbf{z}} - \mathbf{x}||| = u/2 + |||\tilde{\mathbf{z}} - \mathbf{x}|||, \qquad (3.127)
\end{aligned}$$

where we have chosen $\tilde{\varepsilon} = \frac{u\sqrt{m}}{2}$. The second inequality above follows from (3.125). Next we show that $|||\tilde{\mathbf{z}} - \mathbf{x}||| \leq u/2$ with nonzero probability for large enough M. Observe that

$$|||\tilde{\mathbf{z}} - \mathbf{x}||| = \|V_{\tilde{\mathbf{z}}} - V_{\mathbf{x}}\| = \left\|\frac{1}{M}\sum_{\ell=1}^{M}(V_{\tilde{\mathbf{Z}}_\ell} - V_{\mathbf{x}})\right\|.$$

Since this is a sum of centered random matrices, in order to estimate the tail probability of its norm, we employ Theorem A.4. Denote $B_\ell := \frac{1}{M}(V_{\tilde{\mathbf{Z}}_\ell} - V_{\mathbf{x}})$. We first need to bound the variance term

$$\sigma^2 = \max\left\{\left\|\sum_{\ell=1}^{M}\mathbb{E}(B_\ell B_\ell^*)\right\|, \left\|\sum_{\ell=1}^{M}\mathbb{E}(B_\ell^* B_\ell)\right\|\right\}. \tag{3.128}$$

For the first part,

$$\begin{aligned}\mathbb{E}(B_\ell B_\ell^*) &= \frac{1}{M^2}\mathbb{E}(V_{\tilde{\mathbf{Z}}_\ell} - V_{\mathbf{x}})(V_{\tilde{\mathbf{Z}}_\ell} - V_{\mathbf{x}})^* \\ &= \frac{1}{M^2}[\mathbb{E}(V_{\tilde{\mathbf{Z}}_\ell}V_{\tilde{\mathbf{Z}}_\ell}^*) - \mathbb{E}(V_{\tilde{\mathbf{Z}}_\ell})V_{\mathbf{x}}^* - V_{\mathbf{x}}\mathbb{E}(V_{\tilde{\mathbf{Z}}_\ell}^*) + V_{\mathbf{x}}V_{\mathbf{x}}^*] \\ &= \frac{1}{M^2}[\mathbb{E}(V_{\tilde{\mathbf{Z}}_\ell}V_{\tilde{\mathbf{Z}}_\ell}^*) - V_{\mathbf{x}}V_{\mathbf{x}}^*]\end{aligned}$$

where we used $\mathbb{E}(V_{\tilde{\mathbf{Z}}_\ell}) = V_{\mathbf{x}}$. Furthermore, by the triangle inequality,

$$\left\|\sum_{\ell=1}^{M}\mathbb{E}(B_\ell B_\ell^*)\right\| = \frac{1}{M}\|\mathbb{E}(V_{\tilde{\mathbf{Z}}}V_{\tilde{\mathbf{Z}}}^*) - V_{\mathbf{x}}V_{\mathbf{x}}^*\| \le \frac{1}{M}(\|\mathbb{E}(V_{\tilde{\mathbf{Z}}}V_{\tilde{\mathbf{Z}}}^*)\| + \|V_{\mathbf{x}}V_{\mathbf{x}}^*\|).$$

The expectation of the discrete distribution can be written as follows

$$\begin{aligned}\mathbb{E}(V_{\tilde{\mathbf{Z}}}V_{\tilde{\mathbf{Z}}}^*) &= \sum_{i=1}^{N}\|x_i\|_2 V_{\vec{\mathbf{E}}_i(x_i/\|x_i\|_2)}V_{\vec{\mathbf{E}}_i(x_i/\|x_i\|_2)}^* \\ &= \frac{1}{m}\sum_{i=1}^{N}\|x_i\|_2\sum_{j=1}^{m}\mathbf{E}_{jj}\left(\frac{x_i x_i^*}{\|x_i\|_2^2}\right) \\ &= \frac{1}{m}\sum_{j=1}^{m}\mathbf{E}_{jj}\left(\sum_{i=1}^{N}\frac{x_i x_i^*}{\|x_i\|_2}\right).\end{aligned} \tag{3.129}$$

Above $\mathbf{E}_{jj}(A)$ denotes the $m \times m$ diagonal block matrix with the matrix $A \in \mathbb{R}^{d\times d}$ in its j-th diagonal entry and 0 elsewhere. Here, we introduce the function

$$f_n(\mathbf{x}) := \left\|\sum_{i_1\neq i_2\neq\cdots\neq i_n}\frac{x_{i_1}x_{i_1}^*x_{i_2}x_{i_2}^*\cdots x_{i_n}x_{i_n}^*}{\|x_{i_1}\|_2\|x_{i_2}\|_2\cdots\|x_{i_n}\|_2}\right\|,$$

for all $\mathbf{x} \in B_{2,1}$ and $n \ge 2, n \in \mathbb{N}$. Then we have the following result.

Lemma 3.43. *For all $\mathbf{x} \in B_{2,1}$, it holds*

$$f_n(\mathbf{x})^2 \le \frac{\lambda^{2n-2}}{s} + f_{2n}(\mathbf{x}).$$

Proof. We have

$$f_n(\mathbf{x})^2 = \left\|\sum_{i_1\neq\cdots\neq i_n}\frac{x_{i_1}x_{i_1}^*\cdots x_{i_n}x_{i_n}^*}{\|x_{i_1}\|_2\cdots\|x_{i_n}\|_2}\right\|^2 = \left\|\sum_{\substack{i_1\neq\cdots\neq i_n\\ j_1\neq\cdots\neq j_n}}\frac{x_{i_1}x_{i_1}^*\cdots x_{i_n}x_{i_n}^*x_{j_1}x_{j_1}^*\cdots x_{j_n}x_{j_n}^*}{\|x_{i_1}\|_2\cdots\|x_{i_n}\|_2\|x_{j_1}\|_2\cdots\|x_{j_n}\|_2}\right\|$$

$$\le \left\|\sum_{\substack{i_1\neq\cdots\neq i_n=j_1\\ \neq j_2\neq\cdots\neq j_n}}\frac{x_{i_1}x_{i_1}^*\cdots x_{i_n}\|x_{i_n}\|_2^2x_{i_n}^*x_{j_2}x_{j_2}^*\cdots x_{j_n}x_{j_n}^*}{\|x_{i_1}\|_2\cdots\|x_{i_n}\|_2^2\|x_{j_2}\|_2\cdots\|x_{j_n}\|_2}\right\| + \left\|\sum_{\substack{i_1\neq\cdots\neq i_n\neq j_1\\ \neq j_2\neq\cdots\neq j_n}}\frac{x_{i_1}x_{i_1}^*\cdots x_{i_n}x_{i_n}^*x_{j_1}x_{j_1}^*\cdots x_{j_n}x_{j_n}^*}{\|x_{i_1}\|_2\cdots\|x_{i_n}\|_2\|x_{j_1}\|_2\cdots\|x_{j_n}\|_2}\right\|$$

$$\leq \sum_{\substack{i_1\neq\cdots\neq i_n\\ \neq j_2\neq\cdots\neq j_n}} \left\| \frac{x_{i_1}x_{i_1}^*\cdots x_{i_n}x_{i_n}^* x_{j_2}x_{j_2}^*\cdots x_{j_n}x_{j_n}^*}{\|x_{i_1}\|_2\cdots\|x_{i_{n-1}}\|_2\|x_{j_2}\|_2\cdots\|x_{j_n}\|_2}\right\| + f_{2n}(\mathbf{x})$$

$$\leq \sum_{\substack{i_1\neq\cdots\neq i_n\\ \neq j_2\neq\cdots\neq j_n}} \left(\frac{\lambda\|x_{i_1}\|_2\|x_{i_2}\|_2\cdots\lambda\|x_{i_n}\|_2\|x_{j_2}\|_2\cdots\lambda\|x_{j_{n-1}}\|_2\|x_{j_n}\|_2\|x_{i_1}x_{j_n}^*\|}{\|x_{i_1}\|_2\cdots\|x_{i_{n-1}}\|_2\|x_{j_2}\|_2\cdots\|x_{j_n}\|_2}\right) + f_{2n}(\mathbf{x})$$

$$\leq \lambda^{2n-2} \sum_{\substack{i_1,\cdots,i_n,\\ j_2,\cdots,j_n}} \left(\|x_{i_1}\|_2\cdots\|x_{i_{n-1}}\|_2\|x_{i_n}\|_2^2\|x_{j_2}\|_2\cdots\|x_{j_n}\|_2\right) + f_{2n}(\mathbf{x})$$

$$\leq \lambda^{2n-2}\sum_{i_n}\|x_{i_n}\|_2^2\left(\sum_j\|x_j\|_2\right)^{2n-2} + f_{2n}(\mathbf{x}) \leq \frac{\lambda^{2n-2}}{s} + f_{2n}(\mathbf{x}),$$

where we used the Cauchy-Schwarz inequality, the triangle inequality and $\mathbf{x}\in B_{2,1}$. □

Proof of Lemma 3.42 continued. We estimate (3.129) by using Lemma 3.43 as follows

$$\begin{aligned}
\|\mathbb{E}(V_{\tilde{\mathbf{Z}}}V_{\tilde{\mathbf{Z}}}^*)\| &= \frac{1}{m}\left\|\sum_{i=1}^N \frac{x_i x_i^*}{\|x_i\|_2}\right\| = \frac{1}{m}\left\|\sum_{i,j=1}^N \frac{x_i x_i^* x_j x_j^*}{\|x_i\|_2\|x_j\|_2}\right\|^{1/2} \\
&\leq \frac{1}{m}\left(\left\|\sum_{i=1}^N \frac{\|x_i\|_2^2 x_i x_i^*}{\|x_i\|_2^2}\right\| + f_2(\mathbf{x})\right)^{1/2} \\
&\leq \frac{1}{m}\left(\sum_{i=1}^N \|x_i x_i^*\| + \left(\frac{\lambda^2}{s} + f_4(\mathbf{x})\right)^{1/2}\right)^{1/2} \\
&\leq \frac{1}{m}\left(\sum_{i=1}^N \|x_i\|_2^2 + \left(\frac{\lambda^2}{s} + \left(\frac{\lambda^6}{s} + f_8(\mathbf{x})\right)^{1/2}\right)^{1/2}\right)^{1/2}.
\end{aligned}$$

Observe that the exponents of λ grow like 2^n-2. Iterating this process infinitely many times, we obtain

$$\|\mathbb{E}(V_{\tilde{\mathbf{Z}}}V_{\tilde{\mathbf{Z}}}^*)\| \leq \frac{1}{m}\sqrt{\frac{1}{s}+\sqrt{\frac{\lambda^2}{s}+\sqrt{\frac{\lambda^6}{s}+\cdots}}} = \frac{1}{m}\sqrt{\frac{1}{s}+\lambda\sqrt{\frac{1}{s}+\lambda\sqrt{\frac{1}{s}+\cdots}}} =: \frac{1}{m}A.$$

One can see the relation $\sqrt{\frac{1}{s}+\lambda A} = A$. Solving this quadratic equation yields $A \leq \frac{\lambda+\sqrt{\lambda^2+4/s}}{2} \leq \lambda + 1/\sqrt{s}$. Therefore

$$\|\mathbb{E}(V_{\tilde{\mathbf{Z}}}V_{\tilde{\mathbf{Z}}}^*)\| \leq \frac{\lambda+1/\sqrt{s}}{m}.$$

Also by (3.125), we have

$$\|V_{\mathbf{x}}^*V_{\mathbf{x}}\| = \|V_{\mathbf{x}}V_{\mathbf{x}}^*\| \leq \|V_{\mathbf{x}}\|^2 \leq \frac{1}{m}\|\mathbf{x}\|_2^2 \leq \frac{1}{ms}. \tag{3.130}$$

We conclude that

$$\left\|\sum_{\ell=1}^M \mathbb{E}(B_\ell B_\ell^*)\right\| \leq \frac{1/s+\lambda+1/\sqrt{s}}{Mm} \leq \frac{\lambda+2/\sqrt{s}}{Mm}. \tag{3.131}$$

We proceed by estimating the second part of the variance term (3.128). Similarly as above we have

$$\left\|\sum_{\ell=1}^M \mathbb{E}(B_\ell^* B_\ell)\right\| = \frac{1}{M}\|\mathbb{E}(V_{\tilde{\mathbf{Z}}}^*V_{\tilde{\mathbf{Z}}}) - V_{\mathbf{x}}^*V_{\mathbf{x}}\| \leq \frac{1}{M}(\|\mathbb{E}(V_{\tilde{\mathbf{Z}}}^*V_{\tilde{\mathbf{Z}}})\| + \|V_{\mathbf{x}}^*V_{\mathbf{x}}\|). \tag{3.132}$$

We now denote $H_{ii}(a)$ as the $N \times N$ diagonal matrix with the scalar a in its i-th diagonal entry and 0 elsewhere and $\mathbf{E}_{jj}(A)$ denotes the $m \times m$ diagonal block matrix with the matrix $A \in \mathbb{R}^{N\times N}$ in its j-th diagonal entry and 0 elsewhere. It follows that

$$\begin{aligned}
\|\mathbb{E}(V_{\tilde{\mathbf{Z}}}^* V_{\tilde{\mathbf{Z}}})\| &= \left\| \sum_{i=1}^{N} \|x_i\|_2 V_{\overrightarrow{\mathbf{E}}_i(x_i/\|x_i\|_2)}^* V_{\overrightarrow{\mathbf{E}}_i(x_i/\|x_i\|_2)} \right\| \\
&= \frac{1}{m} \left\| \sum_{i=1}^{N} \|x_i\|_2 \sum_{j=1}^{m} \mathbf{E}_{jj} \left[H_{ii} \left(\frac{x_i^* x_i}{\|x_i\|_2^2} \right) \right] \right\| \\
&= \frac{1}{m} \left\| \sum_{j=1}^{m} \mathbf{E}_{jj} \left(\sum_{i=1}^{N} H_{ii}(\|x_i\|_2) \right) \right\| = \frac{1}{m} \left\| \sum_{i=1}^{N} H_{ii}(\|x_i\|_2) \right\| \\
&= \frac{1}{m} \max_{i\in[N]} \|x_i\|_2 \leq \frac{1}{m\sqrt{s}}.
\end{aligned}$$

Combining this with (3.130) and (3.132), we obtain

$$\left\| \sum_{\ell=1}^{M} \mathbb{E}(B_\ell^* B_\ell) \right\| \leq \frac{2}{Mm\sqrt{s}}.$$

Then (3.131) dominates the right hand side of (3.128), which yields

$$\sigma^2 \leq \frac{\lambda + 2/\sqrt{s}}{Mm}. \tag{3.133}$$

Finally we estimate $\|B_\ell\|$ uniformly for all ℓ.

$$\|B_\ell\| \leq \frac{1}{M}(\|V_{\tilde{\mathbf{Z}}_\ell}\| + \|V_{\mathbf{x}}\|) \leq \frac{1}{M}\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{ms}} \right) \leq \frac{2}{M\sqrt{m}}. \tag{3.134}$$

Plugging the estimates (3.133) and (3.134) into Theorem A.4 gives

$$\mathbb{P}(|||\tilde{\mathbf{z}} - \mathbf{x}||| \geq u/2) \leq (md + mN) \exp\left(- \frac{u^2/8}{\frac{\lambda+2/\sqrt{s}}{Mm} + \frac{u}{3M\sqrt{m}}} \right).$$

Denote $D := md + mN$. A sufficient condition for $\mathbb{P}(|||\tilde{\mathbf{z}} - \mathbf{x}||| \leq u/2) > 0$ to hold is

$$\exp\left(\frac{u^2}{\frac{8\lambda+16/\sqrt{s}}{Mm} + \frac{8u}{3M\sqrt{m}}} \right) > D.$$

Taking the logarithm of both sides and rearranging the terms yield the condition

$$M > \ln(D) \left(\frac{8\lambda + 16/\sqrt{s}}{mu^2} + \frac{8}{3\sqrt{m}u} \right). \tag{3.135}$$

Under this condition, there exists a realization of the vector $\tilde{\mathbf{z}}$ for which $|||\tilde{\mathbf{z}} - \mathbf{x}||| \leq u/2$. This implies with (3.127) that there exists a vector of the form $\mathbf{z} = \frac{1}{M}\sum_{\ell=1}^{M} \mathbf{Z}_\ell$ with

$$|||\mathbf{z} - \mathbf{x}||| \leq u.$$

Note that $\mathbf{z}$ is at most M sparse and $\mathbf{Z}_\ell \in \mathcal{T}$. Since each $\mathbf{Z}_\ell$ takes

$$|\mathcal{T}| = \left| \bigcup_{j\in[N]} T_j \right| \leq N \left(1 + \frac{4}{u\sqrt{m}} \right)^k$$

values so that $\mathbf{z}$ can take at most $N^M \left(1+\frac{4}{u\sqrt{m}}\right)^{kM}$ values. The choice

$$M = \left\lfloor \ln(D) \left(\frac{8\lambda + 16/\sqrt{s}}{mu^2} + \frac{6}{\sqrt{m}u} \right) \right\rfloor$$

satisfies (3.135). Therefore we deduce that the covering numbers can be estimated by

$$\begin{aligned}\sqrt{\ln \mathcal{N}(B_{2,1}, |||\cdot|||, u)} &\leq \sqrt{\ln\left[N^M \left(1+\frac{4}{u\sqrt{m}}\right)^{kM} \right]} \\ &\leq \sqrt{\left(\frac{8\lambda + 16/\sqrt{s}}{mu^2} + \frac{6}{\sqrt{m}u} \right)} \sqrt{\ln(D)\left[\ln(N) + k\ln\left(1+\frac{4}{u\sqrt{m}}\right)\right]}.\end{aligned}$$

This completes the proof of Lemma 3.42. □

Proof of Lemma 3.41 continued. Noting that $D_{s,N} \subset \sqrt{s}B_{2,1}$, we have the two following bounds due to (3.126) and Lemma 3.42

$$\sqrt{\ln \mathcal{N}(\mathcal{A}, \|\cdot\|, u)} \leq \sqrt{s\ln(eN/s)} + \sqrt{sk\ln\left(1+\frac{2}{u\sqrt{m}}\right)} \tag{3.136}$$

$$\begin{aligned}\sqrt{\ln \mathcal{N}(\mathcal{A}, \|\cdot\|, u)} &= \sqrt{\ln \mathcal{N}\left(B_{2,1}, |||\cdot|||, \frac{u}{\sqrt{s}}\right)} \\ &\leq \sqrt{\left(\frac{8\lambda s + 16\sqrt{s}}{mu^2} + \frac{6\sqrt{s}}{\sqrt{m}u} \right)} \sqrt{\ln(D)\left[\ln(N) + k\ln\left(1+\frac{4\sqrt{s}}{u\sqrt{m}}\right)\right]},\end{aligned} \tag{3.137}$$

for $u > 0$. Next we combine these inequalities to estimate the Dudley integral in (3.122). We integrate (3.136) from 0 to some $\kappa \in (0, 1/\sqrt{m})$ and (3.137) from κ to $1/\sqrt{m}$.

$$\begin{aligned}I_1 &:= \int_0^\kappa \left(\sqrt{s\ln(eN/s)} + \sqrt{sk\ln\left(1+\frac{2}{u\sqrt{m}}\right)} \right) du \\ &\leq \kappa\sqrt{s\ln(eN/s)} + \kappa\sqrt{sk}\sqrt{\ln\left(1+\frac{2}{\kappa\sqrt{m}}\right)}.\end{aligned}$$

Hereby, we applied [**94**, Lemma 10.3]. The choice $\kappa = \frac{\sqrt{1+\lambda s}}{\sqrt{sk}\sqrt{m}}$ yields

$$I_1 \leq \frac{\sqrt{1+\lambda s}\sqrt{\ln(eN/s)}}{\sqrt{k}\sqrt{m}} + \frac{\sqrt{1+\lambda s}}{\sqrt{m}}\sqrt{\ln\left(1+\frac{\sqrt{sk}}{\sqrt{1+\lambda s}}\right)}.$$

Furthermore,

$$\begin{aligned}I_2 &:= \int_\kappa^{1/\sqrt{m}} \sqrt{\left(\frac{8\lambda s + 16\sqrt{s}}{mu^2} + \frac{6\sqrt{s}}{\sqrt{m}u} \right)} \sqrt{\ln(D)\left[\ln(N) + k\ln\left(1+\frac{4\sqrt{s}}{u\sqrt{m}}\right)\right]} du \\ &\leq \sqrt{\ln(D)\left[\ln(N) + k\ln\left(1+\frac{4s\sqrt{k}}{\sqrt{1+\lambda s}}\right)\right]} \int_\kappa^{1/\sqrt{m}} \left(\sqrt{\frac{8\lambda s + 16\sqrt{s}}{mu^2}} + \sqrt{\frac{6\sqrt{s}}{\sqrt{m}u}} \right) du \\ &\leq \sqrt{\ln(D)\left[\ln(N) + k\ln\left(1+\frac{4s\sqrt{k}}{\sqrt{1+\lambda s}}\right)\right]} \left(\sqrt{\frac{8\lambda s + 16\sqrt{s}}{m}} \int_\kappa^{1/\sqrt{m}} u^{-1} du + \sqrt{\frac{6\sqrt{s}}{\sqrt{m}}} \int_\kappa^{1/\sqrt{m}} u^{-1/2} du \right) \\ &\leq \sqrt{\ln(D)\left[\ln(N) + k\ln\left(1+\frac{4s\sqrt{k}}{\sqrt{1+\lambda s}}\right)\right]} \left(\sqrt{\frac{8\lambda s + 16\sqrt{s}}{m}} \ln\left(\frac{\sqrt{sk}}{\sqrt{1+\lambda s}}\right) + \sqrt{\frac{6\sqrt{s}}{\sqrt{m}}}(m^{-1/4} - \sqrt{\kappa}) \right).\end{aligned}$$

In the first inequality above, we used $u \geq \kappa$ and the triangle inequality. The term $\sqrt{\kappa}$ in the last line can be ignored. Then using the estimates we derived and (3.122), we arrive at

$$\gamma_2(\mathcal{A}, \|\cdot\|) \leq c \int_0^{d_{2\to 2}(\mathcal{A})} \sqrt{\ln \mathcal{N}(\mathcal{A}, \|\cdot\|, u)} du \leq I_1 + I_2 \leq \delta/2,$$

provided Condition (3.123) is satisfied with an appropriate absolute constant $\tilde{C}$. Then it follows from (3.120) that

$$E = \gamma_2(\mathcal{A}, \|\cdot\|) + \gamma_2(\mathcal{A}, \|\cdot\|)^2 \leq \delta.$$

This ends the proof of Lemma 3.41. □

Proof of Theorem 3.7 continued. So far, we have estimated E in (3.120). It remains to estimate the tail probability (3.121) given in Theorem 3.40. We choose the constant C in (3.10) sufficiently larger than $\tilde{C}$ in (3.123) so that

$$C_1 E \leq \delta/2,$$

where C_1 is the constant in (3.121). Next we bound the terms V and U in (3.121). We have

$$\begin{aligned} V &= d_{2\to 2}(\mathcal{A})[\gamma_2(\mathcal{A}, \|\cdot\|) + d_F(\mathcal{A})] \\ &\leq \frac{1}{\sqrt{m}}(\gamma_2(\mathcal{A}, \|\cdot\|) + 1) \leq 2/\sqrt{m}, \end{aligned}$$

since $\gamma_2(\mathcal{A}, \|\cdot\|) \leq 1$ by (3.123). Also we have

$$U = d_{2\to 2}^2(\mathcal{A}) = 1/m$$

by (3.124). Plugging these estimates into (3.121) and choosing $t = \delta/2$, we obtain

$$\mathbb{P}(\delta_s \geq \delta) \leq \mathbb{P}(\delta_s \geq C_1 E + \delta/2) \leq 2 \exp\left(-C_2 \frac{\delta^2 m}{16}\right) \leq \varepsilon$$

provided Condition (3.11) is satisfied with $C' = 16/C_2$. This finally completes the proof of Theorem 3.7. □

3.4.2. Proof of Theorem 3.8

Gaussian width. Let $g \in \mathbb{R}^m$ be a standard Gaussian random vector. Then for

$$E_m := \mathbb{E}\|g\|_2 = \sqrt{2}\frac{\Gamma((m+1)/2)}{\Gamma(m/2)}$$

we have

$$\frac{m}{\sqrt{m+1}} \leq E_m \leq \sqrt{m}, \tag{3.138}$$

see [**54**, **62**]. We introduce the *Gaussian width* of a set $T \subset \mathbb{R}^N$ by

$$\ell(T) := \mathbb{E} \sup_{x \in T} \langle g, x \rangle.$$

The following version of Gordon's escape through the mesh lemma [**62**] appears in [**54**, Theorem 9.21].

Lemma 3.44. *Let $A \in \mathbb{R}^{m \times N}$ be a Gaussian random matrix and T be a subset of the unit sphere $S^{N-1} = \{x \in \mathbb{R}^N, \|x\|_2 = 1\}$. Then*

$$\mathbb{E} \inf_{x \in T} \|Ax\|_2 \geq E_m - \ell(T).$$

Our strategy is to establish the ℓ_2-stable and robust null space property of order s for a Gaussian matrix A. Then Theorem 3.19 automatically gives us the desired result. Thus, we need to show that

$$\|\mathbf{x}_S\|_2 \leq \frac{\rho}{\sqrt{s}}\|\mathbf{x}_{\overline{S}}\|_{2,1} + \tau\|\mathbf{A}_\mathbf{P}\mathbf{x}\|_2 \tag{3.139}$$

for all $\mathbf{x} \in \mathcal{H}$ and $S \subset [N], |S| = s$. For $\rho \in (0, 1]$, we introduce the set

$$T_{\rho,s} := \left\{\mathbf{z} \in \mathcal{H} : \|\mathbf{z}_S\|_2 \geq \frac{\rho}{\sqrt{s}}\|\mathbf{z}_{\overline{S}}\|_{2,1} \text{ for some } S \subset [N], |S| = s\right\},$$

and the sphere $S_\mathcal{H} = \{\mathbf{y} \in \mathcal{H} : \|\mathbf{y}\|_2 = 1\}$. We will first show that with high probability

$$\min\{\|\mathbf{A}_\mathbf{P}\mathbf{z}\|_2 : \mathbf{z} \in T_{\rho,s} \cap S_\mathcal{H}\} > 0, \tag{3.140}$$

and this will be a step towards showing (3.139). We define the set $B_{2,1} := \{\mathbf{x} \in \mathcal{H} : \|\mathbf{x}\|_{2,1} \leq \sqrt{s}, \|\mathbf{x}\|_2 \leq 1\}$.

Lemma 3.45. *It holds*

$$T_{\rho,s} \cap S_\mathcal{H} \subset (1 + \rho^{-1})B_{2,1}.$$

PROOF. Take an arbitrary $\mathbf{x} \in T_{\rho,s}$ such that $\|\mathbf{x}\|_2 = 1$. Then there is an $S \subset [N], |S| = s$, such that $\|\mathbf{x}_S\|_2 \geq \frac{\rho}{\sqrt{s}}\|\mathbf{x}_{\overline{S}}\|_{2,1}$. Using this relation, it follows that

$$\|\mathbf{x}\|_{2,1} = \|\mathbf{x}_S\|_{2,1} + \|\mathbf{x}_{\overline{S}}\|_{2,1} \leq \sqrt{s}\|\mathbf{x}_S\|_2 + \frac{\sqrt{s}}{\rho}\|\mathbf{x}_S\|_2 \leq \sqrt{s}(1 + \rho^{-1}),$$

where we used that $\|\mathbf{x}\|_2 = 1$ and the Cauchy-Schwarz inequality. This proves our claim that $\mathbf{x} \in (1 + \rho^{-1})B_{2,1}$. □

Our goal is to show that (3.140) holds with high probability when the matrix A is Gaussian. We denote $T := T_{\rho,s} \cap S_\mathcal{H}$ for simplicity of notation. We notice that $\mathcal{F}(A) := \inf_{\mathbf{x}\in T}\|\mathbf{A}_\mathbf{P}\mathbf{x}\|_2$ defines a Lipschitz function with Lipschitz constant $L = 1$ with respect to the Frobenius norm (which corresponds to the ℓ_2-norm by identifying $\mathbb{R}^{m\times N}$ with $\mathbb{R}^{mN}$). Indeed, for two matrices $A, B \in \mathbb{R}^{m\times N}$,

$$\begin{aligned}\inf_{\mathbf{x}\in T}\|\mathbf{A}_\mathbf{P}\mathbf{x}\|_2 &\leq \inf_{\mathbf{x}\in T}(\|\mathbf{B}_\mathbf{P}\mathbf{x}\|_2 + \|(\mathbf{A}_\mathbf{P} - \mathbf{B}_\mathbf{P})\mathbf{x}\|_2) \leq \inf_{\mathbf{x}\in T}(\|\mathbf{B}_\mathbf{P}\mathbf{x}\|_2 + \|(\mathbf{A}_\mathbf{P} - \mathbf{B}_\mathbf{P})\|) \\ &\leq \inf_{\mathbf{x}\in T}\|\mathbf{B}_\mathbf{P}\mathbf{x}\|_2 + \|A - B\| \leq \inf_{\mathbf{x}\in T}\|\mathbf{B}_\mathbf{P}\mathbf{x}\|_2 + \|A - B\|_F.\end{aligned}$$

The second inequality follows from $T \subset S_\mathcal{H}$. Interchanging the roles of A and B yields

$$|\mathcal{F}(A) - \mathcal{F}(B)| \leq \|A - B\|_F.$$

Then the concentration of measure (A.11) yields

$$\mathbb{P}\left(\inf_{\mathbf{x}\in T}\|\mathbf{A}_\mathbf{P}\mathbf{x}\|_2 \leq \mathbb{E}\inf_{\mathbf{x}\in T}\|\mathbf{A}_\mathbf{P}\mathbf{x}\|_2 - t\right) \leq e^{-t^2/2}. \tag{3.141}$$

Our next goal is to derive a lower bound on $\mathbb{E}\inf_{\mathbf{x}\in T}\|\mathbf{A}_\mathbf{P}\mathbf{x}\|_2$. Actually a stronger concentration inequality holds for $X := \inf_{\mathbf{x}\in T}\|\mathbf{A}_\mathbf{P}\mathbf{x}\|_2$ according to (A.12), which satisfies the hypothesis of Lemma A.15 for $t \geq 1$ with constants $c = 2, \sigma_1^2 = \sigma_2^2 = 2$. Therefore we have

$$\sqrt{\mathbb{E}X^2} \leq \mathbb{E}X + C, \quad \text{for some } C > 0. \tag{3.142}$$

Before obtaining a lower bound for $\mathbb{E}X^2$, we introduce some notation. For a block vector $\mathbf{x} \in T$ with blocks $x_1, \dots, x_N$, let $D_{\mathbf{x}} \in \mathbb{R}^{N \times d}$ be the matrix with rows x_i. We define the sets

$$\begin{aligned}
\mathcal{U} &= \{D_{\mathbf{x}} : \mathbf{x} \in T\},\\
\mathcal{V} &= \{v \in \mathbb{R}^N : v = U_i \ \text{ for some } U \in \mathcal{U} \text{ and } i \in [d]\},\\
\widetilde{\mathcal{V}} &= \{v/\|v\|_2 : v \in \mathcal{V} \text{ and } \|v\|_2 \geq 1/\sqrt{d}\}.
\end{aligned}$$

Recall that U_i denotes the i-th column of U. Then we have

$$\begin{aligned}
X^2 = \inf_{\mathbf{x}\in T} \|\mathbf{A_P x}\|_2^2 \ &= \ \inf_{\mathbf{x}\in T} \|\mathbf{A_I x}\|_2^2 \ = \ \inf_{U\in\mathcal{U}} \sum_{i=1}^d \|AU_i\|_2^2\\
&= \inf_{\substack{c\in\mathbb{R}^d\\ \|c\|_2=1}} \inf_{\substack{U\in\mathcal{U}\\ \|U_i\|_2=|c_i|}} \sum_{i=1}^d \|AU_i\|_2^2 \ \geq \ \inf_{\substack{v\in\mathcal{V}\\ \|v\|_2\geq\frac{1}{\sqrt{d}}}} \|Av\|_2^2 \ \geq \ \frac{1}{d} \inf_{\substack{v\in\mathcal{V}\\ \|v\|_2\geq\frac{1}{\sqrt{d}}}} \|A(v/\|v\|_2)\|_2^2. \qquad (3.143)
\end{aligned}$$

Above we used that $\|U\|_F = 1$ for any $U \in \mathcal{U}$. The first inequality uses the observation that for a vector $a \in S^{d-1}$ there exist an index $i \in [d]$ such that $|a_i| \geq 1/\sqrt{d}$. Taking the expectation in (3.143) yields

$$\mathbb{E}X^2 \ \geq \ \frac{1}{d}\,\mathbb{E}\inf_{v\in\widetilde{\mathcal{V}}} \|Av\|_2^2 \ \geq \ \frac{1}{d}\left(\mathbb{E}\inf_{v\in\widetilde{\mathcal{V}}} \|Av\|_2\right)^2 \geq \ \frac{1}{d}\left(E_m - \mathbb{E}\sup_{v\in\widetilde{\mathcal{V}}} \langle g, v\rangle\right)^2, \qquad (3.144)$$

where $g \in \mathbb{R}^N$ is a standard Gaussian vector and $\widetilde{\mathcal{V}} \subset S^{N-1}$. Above, the Jensen's inequality and Lemma 3.44 are invoked in the second and the last inequalities respectively. This leads to bounding the Gaussian width of $\widetilde{\mathcal{V}}$ as follows

$$\begin{aligned}
\mathbb{E}\sup_{v\in\widetilde{\mathcal{V}}} \langle g, v\rangle \ &\leq \ \sqrt{d}\,\mathbb{E} \sup_{\substack{v\in\mathcal{V}\\ \|v\|_2\geq\frac{1}{\sqrt{d}}}} |\langle g, v\rangle| \ \leq \ \sqrt{d}\,\mathbb{E}\sup_{v\in\mathcal{V}} |\langle g, v\rangle| \ = \ \sqrt{d}\,\mathbb{E}\sup_{U\in\mathcal{U}} \|U^* g\|_\infty\\
&= \ \sqrt{d}\,\mathbb{E}\sup_{\mathbf{x}\in T} \|D_{\mathbf{x}}^* g\|_\infty \ \leq \ \sqrt{d}\,\mathbb{E}\sup_{\mathbf{x}\in T} \|D_{\mathbf{x}}^* g\|_2 \ \leq \ \sqrt{d}\,\sqrt{\mathbb{E}\sup_{\mathbf{x}\in(1+\rho^{-1})B_{2,1}} \|D_{\mathbf{x}}^* g\|_2^2}, \qquad (3.145)
\end{aligned}$$

where we used the Jensen's inequality once again and Lemma 3.45. We continue estimating

$$\begin{aligned}
&\mathbb{E}\sup_{\mathbf{x}\in(1+\rho^{-1})B_{2,1}} \|D_{\mathbf{x}}^* g\|_2^2 \ = \ \mathbb{E}\sup_{\mathbf{x}\in(1+\rho^{-1})B_{2,1}} \left(\sum_{i=1}^N g_i^2\|x_i\|_2^2 + \sum_{i\neq j} g_i g_j \langle x_i, x_j\rangle\right)\\
&\leq \ \mathbb{E}\left(\max_{i\in[N]} g_i^2\right) \sup_{\mathbf{x}\in(1+\rho^{-1})B_{2,1}} \left(\sum_{i=1}^N \|x_i\|_2^2\right) + \mathbb{E}\sup_{\mathbf{x}\in(1+\rho^{-1})B_{2,1}} \sum_{i\neq j} |g_i||g_j|\|P_i P_j\|\|x_i\|_2\|x_j\|_2\\
&\leq \ \mathbb{E}\left(\max_{i\in[N]} g_i^2\right)(1+\rho^{-1})^2 + \mathbb{E}\sup_{\mathbf{x}\in(1+\rho^{-1})B_{2,1}} \lambda\left(\sum_{i=1}^N |g_i|\|x_i\|_2\right)^2\\
&\leq \ \mathbb{E}\left(\max_{i\in[N]} g_i^2\right)(1+\rho^{-1})^2 + \lambda\,\mathbb{E}\left(\max_{i\in[N]} |g_i|\right)^2 (1+\rho^{-1})^2 s \ \leq \ c(1+\rho^{-1})^2(1+\lambda s)\ln(N),
\end{aligned}$$

for some constant $c > 0$. The last inequality above uses Lemma A.9. Plugging this estimate into (3.145) and then into (3.144) yields

$$\mathbb{E}X^2 \geq \frac{1}{d}(E_m - R)^2,$$

where $R := (1+\rho^{-1})\sqrt{c(1+\lambda s)\ d\ \ln(N)}$. Then using the lower bound (3.138) for E_m and (3.142) implies that

$$\mathbb{E}X \geq \sqrt{\mathbb{E}X^2} - C \geq \frac{1}{\sqrt{d}}\left(\frac{m}{\sqrt{m+1}} - R\right) - C.$$

Combining this with (3.141) gives

$$\mathbb{P}\left(\inf_{\mathbf{x}\in T} \|\mathbf{A_P x}\|_2 > \frac{1}{\sqrt{d}}\left(\frac{m}{\sqrt{m+1}} - R\right) - C - t\right) \geq 1 - e^{-t^2/2}.$$

Setting $t = \sqrt{2\ln(\varepsilon^{-1})}$ yields that

$$\mathbb{P}(\inf_{\mathbf{x}\in T} \|\mathbf{A_P x}\|_2 > 1/\tau) \geq 1 - \varepsilon,$$

provided

$$\frac{1}{\sqrt{d}}\left(\frac{m}{\sqrt{m+1}} - R\right) - C - \sqrt{2\ln(\varepsilon^{-1})} > \frac{1}{\tau}.$$

This means that under Condition (3.13), for any $\mathbf{x} \in \mathcal{H}$ such that $\|\mathbf{A_P x}\|_2 \leq \frac{1}{\tau}\|\mathbf{x}\|_2$ and for any set $S \subset [N], |S| = s$, it holds with probability at least $1-\varepsilon$ that

$$\|\mathbf{x}_S\|_2 \leq \frac{\rho}{\sqrt{s}}\|\mathbf{x}_{\overline{S}}\|_{2,1}.$$

For all other $\mathbf{x} \in \mathcal{H}$, clearly $\|\mathbf{x}\|_2 \leq \tau\|\mathbf{A_P x}\|_2$ which shows that for all $\mathbf{x} \in \mathcal{H}$, (3.139) holds. This ends the proof. □

3.5. Proof of the necessary condition

3.5.1. Covering numbers and coding theory

Let T be a subset of a metric space (X, d). For $t > 0$, the covering number $\mathcal{N}(T, d, t)$ is defined as the smallest integer $\mathcal{N}$ such that T can be covered with balls $B(x_\ell, t) = \{x \in X, d(x, x_\ell) \leq t\}, x_\ell \in T, \ell \in [\mathcal{N}]$, i.e.,

$$T \subset \bigcup_{\ell=1}^{\mathcal{N}} B(x_\ell, t).$$

The packing number $\mathcal{P}(T, d, t)$ is defined, for $t > 0$, as the maximal integer $\mathcal{P}$ such that there are points $x_\ell \in T$, $\ell \in [\mathcal{P}]$, which are t-separated, i.e., $d(x_\ell, x_k) > t$ for all $k, \ell \in [\mathcal{P}], k \neq \ell$. If $X = \mathbb{R}^n$ is a normed vector space and the metric d is induced by the norm via $d(u, v) = \|u - v\|$, we also write $\mathcal{N}(T, \|\cdot\|, t)$ and $\mathcal{P}(T, \|\cdot\|, t)$.

Proposition 3.46. *Let $\|\cdot\|$ be some norm on $\mathbb{R}^n$ and let $U = \{x \in \mathbb{R}^n : \frac{1}{2} \leq \|x\| \leq 2\}$ be a spherical shell. Then the packing and covering numbers satisfy, for $t > 0$,*

$$\left(\frac{2}{t}\right)^n - \left(\frac{1}{2t}\right)^n \leq \mathcal{N}(U, \|\cdot\|, t) \leq \mathcal{P}(U, \|\cdot\|, t).$$

Proof. For the first inequality, let $\{x_1, \ldots, x_{\mathcal{N}}\} \subset U$ be the minimal set covering U. This implies that

$$\mathcal{N}\mathrm{vol}(tB) \geq \mathrm{vol}\left(\bigcup_{\ell=1}^{\mathcal{N}} B(x_\ell, t)\right) \geq \mathrm{vol}(U) = \mathrm{vol}(2B) - \mathrm{vol}(B/2).$$

On $\mathbb{R}^n$ the volume satisfies the homogeneity relation $\mathrm{vol}(tB) = t^n\mathrm{vol}(B)$. Hence we have $\mathcal{N}t^n\mathrm{vol}(B) \geq 2^n\mathrm{vol}(B) - (1/2)^n\mathrm{vol}(B)$ which yields $\mathcal{N} \geq (2/t)^n - (1/2t)^n$ as desired. □

Here is another similar result on packing numbers, for a reference see [**54**].

Proposition 3.47. *Let $\|\cdot\|$ be some norm on $\mathbb{R}^n$ and let U be a subset of the unit ball $B = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$. Then the packing number satisfies, for $t > 0$,*

$$\mathcal{N}(U, \|\cdot\|, t) \leq \mathcal{P}(U, \|\cdot\|, t) \leq \left(1 + \frac{2}{t}\right)^n.$$

Next we present a lemma from coding theory which will be useful in our proof of Theorem 3.10. Let us first start with some notation from the literature. Let A be a finite set of q letters, also called *alphabet*. A *code* C is any nonempty subset of the set A^n of n-tuples of elements from A. The number n is the *length* of the code, and the set A^n is the *codespace*. The number of members in C is the *size* and is denoted $|C|$. The members of the codespace are referred as *words*, those belonging to C being *codewords*. We define the *Hamming* distance between any two words x, y of same length as

$$d_H(x, y) = \{i : x_i \neq y_i\}.$$

The minimum distance of a code C in A^n is defined as

$$d_{min}(C) = \min_{\substack{x,y \in C \\ x \neq y}} d_H(x, y).$$

In this context, the maximum possible size of a code C of length n with a fixed $d_{min}(C)$ is of particular interest to us. Such a bound can be seen as a packing problem with respect to Hamming distance and available in the literature as the *Gilbert-Varshamov* bound [**58**, **110**].

Lemma 3.48. *(Gilbert-Varshamov bound) Let A be a finite alphabet of q elements. There exists a code $C \subset A^n$ of length n with $d_{min}(C) \geq e$ such that*

$$|C| \geq \frac{q^n}{\sum_{i=0}^{e-1} \binom{n}{i}(q-1)^i}.$$

3.5.2. Proof of Theorem 3.10

The proof is based on the following combinatorial lemma [**14**, **53**, **63**, **83**, **87**] and the results in Section 3.5.1 .

Lemma 3.49. *Given integers $s < N$, there exist*

$$n \geq \left(\frac{N}{8s}\right)^s \tag{3.146}$$

subsets $I_1, \ldots, I_n$ of $[N]$ such that each I_i has cardinality $2s$ and

$$\text{card}(I_i \cap I_\ell) < s \quad \textit{whenever } i \neq \ell.$$

First recall that we restrict our attention to the set

$$\mathcal{H} = \{(x_i)_{i=1}^N : x_i \in W_i,\ \forall i \in [N]\} \subset \mathbb{R}^{d \times N}.$$

Let us consider the quotient space

$$X := \mathcal{H} \,/\, \ker \mathbf{A}_{\mathbf{P}|\mathcal{H}} = \{[\mathbf{x}] := \mathbf{x} + \ker \mathbf{A}_{\mathbf{P}|\mathcal{H}}, \mathbf{x} \in \mathcal{H}\},$$

which is equipped with the quotient norm

$$\|[\mathbf{x}]\| := \inf_{\mathbf{v} \in \ker \mathbf{A}_{\mathbf{P}|\mathcal{H}}} \|\mathbf{x} - \mathbf{v}\|_{2,1},\ \mathbf{x} \in \mathcal{H}.$$

Above $\mathbf{A}_{\mathbf{P}|\mathcal{H}}$ is the restriction of $\mathbf{A}_{\mathbf{P}}$ to $\mathcal{H}$. Let B_X be the unit ball of X with respect to this norm. Given a $4s$-sparse vector $\mathbf{x} \in \mathcal{H}$, we notice that every vector $\mathbf{z} = \mathbf{x} - \mathbf{v}$ with $\mathbf{v} \in \ker \mathbf{A}_{\mathbf{P}|\mathcal{H}}$ satisfies $\mathbf{A}_{\mathbf{P}}\mathbf{z} = \mathbf{A}_{\mathbf{P}}\mathbf{x}$. Thus, the assumption of the theorem gives $\|[\mathbf{x}]\| = \|\mathbf{x}\|_{2,1}$. Next we pack the spherical shells $S_i = \{y \in W_i : \frac{1}{2s} \leq \|y\|_2 \leq \frac{2}{s}\}$ of the subspaces W_i. Proposition 3.46 yields that we can find such a packing with packing distance of $1/s$ where

$$\mathcal{P}(S_i, \|\cdot\|_2, 1/s) \geq 2^k - (1/2)^k =: q,$$

Recall $\dim(W_i) = k$. For each i, let $T_i \subset S_i$ be the packing set of S_i with this cardinality. Now let $I_1, \ldots, I_n$ be the sets introduced in Lemma 3.49, and for each $j \in [n]$ let us define the set of $2s$-sparse vectors with support set from I_j

$$\mathcal{T}(I_j) := \{\mathbf{x} \in \mathcal{H} : \operatorname{supp}(\mathbf{x}) = I_j,\ x_i \in T_i \ \text{ for } i \in I_j\}.$$

It is clear that for each j, the total number of vectors in $\mathcal{T}(I_j)$ is at least q^{2s}. We wish to find a large enough subset of each $\mathcal{T}(I_j)$, say $\widehat{\mathcal{T}}(I_j)$, such that any two distinct vectors $\mathbf{x}, \mathbf{y} \in \widehat{\mathcal{T}}(I_j)$ satisfy $\operatorname{card}\{i : x_i \neq y_i\} \geq s$. This problem is exactly the same as the one in Lemma 3.48. We can think of q as the size of our alphabet and we have words of length $2s$ from this alphabet. Then we search for a lower bound on the number of words possible which are mutually distant to each other with at least a Hamming distance of s. Lemma 3.48 says that there exist such a subset $\widehat{\mathcal{T}}(I_j)$ satisfying

$$|\widehat{\mathcal{T}}(I_j)| \geq \frac{q^{2s}}{\sum_{i=0}^{s-1} \binom{2s}{i}(q-1)^i} \geq \frac{q^{2s}}{2^{2s}q^s} \geq (q/4)^s.$$

Let us define the set

$$\mathcal{V} = \bigcup_{j=1}^{n} \widehat{\mathcal{T}}(I_j).$$

We claim that the set $\{[\mathbf{x}] : \mathbf{x} \in \mathcal{V}\}$ is a 1-separating subset of the scaled unit ball $4B_X$. Observe that every $\mathbf{x} \in \mathcal{V}$ has a support set I_j for some $j \in [n]$ which has cardinality $2s$. Therefore, it holds that

$$\|[\mathbf{x}]\| = \|\mathbf{x}\|_{2,1} = \sum_{i \in I_j} \|x_i\|_2 \leq 2s \cdot \frac{2}{s} = 4,$$

since $x_i \in S_i$. Next we see that any two elements from the set $\mathcal{V}$ are at least 1-separated. We consider two cases in order to prove this argument. First assume that $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ both belong to the same set $\widehat{\mathcal{T}}(I_j)$ for some j. Then we know that $\operatorname{card}\{i : x_i \neq y_i\} \geq s$ holds. Whenever $x_i \neq y_i$, we have $\|x_i - y_i\|_2 \geq 1/s$ since $x_i, y_i \in T_i$. Hence

$$\|[\mathbf{x}] - [\mathbf{y}]\| = \|[\mathbf{x} - \mathbf{y}]\| = \|\mathbf{x} - \mathbf{y}\|_{2,1} = \sum_{i : x_i \neq y_i} \|x_i - y_i\|_2 \geq s \cdot \frac{1}{s} = 1.$$

The second equality holds since $\mathbf{x} - \mathbf{y}$ is $2s$-sparse. For the case that $\mathbf{x} \in \widehat{\mathcal{T}}(I_j)$ and $\mathbf{y} \in \widehat{\mathcal{T}}(I_\ell)$ with $j \neq \ell$, observe that $\mathbf{x} - \mathbf{y}$ is $4s$-sparse and the symmetric difference of the support sets satisfy $\operatorname{card}(I_j \triangle I_\ell) \geq 2s$. Then it follows that

$$\|[\mathbf{x}] - [\mathbf{y}]\| = \|[\mathbf{x} - \mathbf{y}]\| = \|\mathbf{x} - \mathbf{y}\|_{2,1} \geq \sum_{i : I_j \triangle I_\ell} (\|x_i\|_2 + \|y_i\|_2) \geq 2s \cdot \frac{1}{2s} = 1,$$

since $x_i, y_i \in S_i$. This proves our claim that $\mathcal{V}$ separates $4B_X$ by 1 and $|\mathcal{V}| \geq n(q/4)^s$. The ball $4B_X$ has dimension $r := \operatorname{rank}(\mathbf{A}_{\mathbf{P}}) \leq md$. According to Proposition 3.47, this implies that

$n(q/4)^s \leq 9^r \leq 9^{md}$. In view of (3.146), we obtain for $k \geq 1$

$$9^{md} \geq \left(\frac{N}{8s}\right)^s \left(\frac{2^k - (1/2)^k}{4}\right)^s \geq \left(\frac{N}{32s}\right)^s \left(\frac{3}{2}\right)^{ks},$$

since $2^k - (1/2)^k \geq (3/2)^k$ for $k \geq 1$. Taking the logarithm on both sides gives the desired result. □

3.6. Discussion

The fusion frame setup can be considered as a special case of the block sparsity setup where an additional subspace structure is assumed. In our work we have shown that if it is known a priori that the subspaces, where the signals lie in, are closer to being orthogonal, less random measurements become sufficient for successfully recovering those vectors. In order to achieve this, we use convex programming as our reconstruction method. The crucial difference of our setup to the usual block sparsity problems is using the information that $\mathbf{x} \in \mathcal{H}$ within our program, which in turn allows us to derive our novel results. As we noted earlier, it was shown in [**9**] that $m \gtrsim s\ln(N/s)$ is sufficient for uniform recovery with many random measurement ensembles. In the nonuniform setting, we improve this result since the number of measurements decreases with increasing incoherence of the subspaces. In the uniform setting, our result is an improvement when λ becomes small enough. To the best of our knowledge, these results are the first in the literature that study the effect of the incoherence of the subspaces on the sparse recovery of deterministic vectors. Indeed, [**9**] also contains results with this flavor but they are in terms of the coherence and formulated in an 'average case scenario'. The authors [**9**] study the recovery of random vectors chosen from a fusion frame according to a Gaussian distribution on the subspaces and study the effect of the dimension of the subspaces on recoverability. Their result shows that under the condition

$$\|A_S^\dagger A_\ell\|_2 \leq \alpha < 1 \quad \text{for all} \quad \ell \notin S, \tag{3.147}$$

where A is the measurement matrix, a vector $\mathbf{x}$ supported on S with $|S| = s$ can be recovered by $(L1)$ if

$$k \gtrsim \|\Lambda^S\|_\infty \max\{\ln(N-s), \ln(s)\}, \tag{3.148}$$

where k is the subspace dimension. We note that an additional condition on m is needed for 'suitable' measurement matrices in order to ensure that (3.147) holds. Interestingly, this result involves the incoherence matrix Λ from Section 3.1.5 similarly to our nonuniform result Theorem 3.1, but unlike our result, (3.148) improves with increasing k. In a preceding paper [**50**], Eldar and Rauhut consider a similar average case analysis for the joint sparsity case. They use the same probabilistic model for the sparse signal to be recovered and provide mild conditions that imply (3.147). Moreover, their results show that adding more channels to the joint sparsity problem, which can be interpreted as adding dimensions in the block sparsity problem, improves the recoverability of jointly sparse vectors.

Before comparing our results with the related literature, we give a brief account of some challenges we have faced towards proving our uniform results. In general, proving uniform recovery results tends to be harder than proving nonuniform recovery ones. In the scalar case, a uniform recovery result for Gaussian matrices based on the concentration of measure of Lipschitz functions

was given by one of the seminal papers by Candès and Tao [**20**]. However, in the fusion frame case, this approach does not work since the parameter λ does not appear as a factor in the number of measurements. Another approach is to bound the FRIP constant δ_s of the matrix $\mathbf{A_P}$. The techniques developed by Rudelson and Vershynin [**95**], where they invoke Dudley's inequality directly to bound the expectation of δ_s, also fail to yield λ. In order to achieve our results presented in this thesis, we employ a recent method to bound suprema of chaos processes by Krahmer, Mendelson and Rauhut [**75**]. Consequently, λ appears as a linear factor in the number of measurements, however not in the optimal order, see Theorem 3.7.

In the literature, there are other works where the recovery of block sparse vectors is considered. As mentioned earlier, those results do not assume any structure of the subspaces underlying the signals of interest. Our work as well as [**9**] also provides comparison with the case of the block sparsity problem. In the minimization program $(L1)$, if we omit the assumption $\mathbf{x} \in \mathcal{H}$ and replace $\mathbf{A_P}$ with $\mathbf{A_I}$, the problem reduces to the block sparsity setup. This is equivalent to assuming $W_j = \mathbb{R}^d$ which implies $\lambda = 1$. Then our nonuniform result implies that a block s-sparse vector $\mathbf{x}$ can be recovered from its m random measurements with high probability via ℓ_1/ℓ_2 minimization provided $m \gtrsim s\ln(N)$. However, since each of our measurements, $y_i \in \mathbb{R}^d$, is itself a vector, the number of measurements we are taking is actually md.

Eldar and Bölcskei in [**49**, **51**] consider the recovery of block sparse vectors via an extension of orthogonal matching pursuit for the block sparse case (BOMP) and provide a sufficient condition in terms of the sparsity level and 'block coherence' of the matrices which they define in their papers. Their condition is a deterministic one for the measurement matrices which gives nonuniform results and it is not answered whether certain type of random matrices satisfy this condition.

In [**23**], Candès and Recht provide a nonuniform result with Gaussian measurement matrices and the same recovery method $(L1)$ under the condition

$$md \gtrsim s(\sqrt{d} + \sqrt{2\ln(N)})^2 + sd. \tag{3.149}$$

Their method uses the duality based recovery results due to Fuchs [**55**]. Their measurement matrix $A \in \mathbb{R}^{md\times Nd}$ has all of its entries from the standard Gaussian distribution, unlike our measurement matrix $\mathbf{A_P}$. When we do not assume any structure as in the block sparsity setup, (3.149) provides a slightly better condition than our nonuniform one, $md \gtrsim sd\ln(N)$. However, when λ becomes smaller, our result scales better. Candès and Recht also provide an extension of their result to the subgaussian matrices with minor modifications. In [**91**], Rao et. al. provide an identical result to (3.149) by using convex geometry. This approach provides recovery from noisy measurements as well, however is restricted to only Gaussian matrices.

In earlier works [**97**, **99**], the authors look at the null space of the measurement matrix $A \in \mathbb{R}^{md\times Nd}$ rather than the RIP by using Gaussian widths. They provide uniform results in the asymptotic regime $N, d \to \infty$, in a similar spirit to the results discussed in Section 2.4. They consider the case that

$$m = \alpha N, \quad s \leq \beta N,$$

where α, β are some fixed values. They show that if A is a Gaussian matrix, then with high probability ℓ_1/ℓ_2 minimization recovers all block s-sparse vectors, provided $\alpha \geq 4\beta(1-\beta)$. This gives roughly the condition $m \geq Cs$, however is valid in the asymptotic regime. Another result

from [**89**] also uses Gaussian widths and provide the nonasymptotic condition $m \gtrsim s(d+\ln(N/s))$ for nonuniform recovery which is similar to (3.149) in essence.

CHAPTER 4

Conclusion

Chapter 2 was devoted to obtaining optimal constants for nonuniform sparse recovery with subgaussian matrices via ℓ_1-minimization. In particular, we provide roughly $m \geq 2s\ln(N)$ as a condition for exact recovery with Gaussian and Bernoulli matrices, which contains the optimal constant 2 and a near-optimal log-factor. Since the earliest papers in CS, the same order in the number of measurements has been shown in many different works, but the constants are generally worse than what we present. Our results also apply to the complex vectors and hold with high probability in finite dimensions, in contrast to some of the previous results operating in the asymptotic regime. We provide short and simple proofs based on elementary probabilistic tools. In addition to exact recovery, we show that nearly sparse signals could be accurately recovered from a small number of noisy measurements. Our stable results also give the optimal constant 2.

In Chapter 3, we studied the sparse recovery problem in fusion frames. In particular, we improved the familiar bound $m \gtrsim s\ln(N)$ on the necessary number of measurement by exploiting an incoherence property of the fusion frame subspaces. As in Chapter 2 we deduced our results by using subgaussian random matrices, including Gaussian and Bernoulli matrices as examples. We presented three different type of results regarding the sparse recovery with fusion frames: nonuniform guarantees, uniform guarantees and necessary conditions. In essence, nonuniform and uniform results state that as the fusion frame subspaces become closer to being orthogonal, less number of measurements suffice for successful recovery. We presented two uniform results namely Theorems 3.7 and 3.8 following different approaches, one of which, we think, is suboptimal in terms of the sparsity s and the other is suboptimal in terms of the ambient dimension d. An important open problem is to improve those results and possibly obtain the same order as the nonuniform condition (3.5). Our necessary condition (3.16) provides the minimal number of measurements required for sparse recovery and we show that it is quite close to the sufficient conditions. We also showed that the nonuniform and uniform results are stable and robust under noise and we presented numerical experiments in order to support our results. In the absence of the incoherence property, our results also compare to the block sparsity setup which is well studied in the literature and essentially gives similar results.

The use of 'completely' random matrices, where all entries are independent, is limited in some applications as noted earlier. Therefore, it is of interest to also consider structured random matrices such as randomly sampled bounded orthonormal systems, for sparse recovery problem in the fusion frames and obtain similar theoretical guarantees as in Chapter 3.

The incoherence parameter λ in Chapter 3 takes only the least principal angles between subspaces into account. This is somewhat pessimistic since the existence of any two subspaces within the fusion frame of co-dimension 1 is enough to yield $\lambda = 1$. One can also investigate the

possibility to generalize our results when other distance measures are considered, e.g., the chordal distance [**76**] which takes all principal angles into account.

As a final note, the necessary condition for sparse recovery with fusion frames in Theorem 3.10 currently applies only when the $(L1)$ program is used as the reconstruction method and we restrict ourselves to the exactly sparse case. We believe it is possible to extend this result to *all* types of reconstruction methods and to the stable and robust case by using Gelfand widths of ℓ_1/ℓ_2 balls associated with the fusion frame similarly to [**53**].

APPENDIX A

A.1. Bernstein inequalities

The following estimate for sums of subgaussian random variables appears for instance in [**111**].

Lemma A.1. *Let $X_1, \dots, X_M$ be a sequence of independent mean-zero subgaussian random variables with the same parameter c as in (2.5). Let $a \in \mathbb{R}^M$ be some vector. Then $Z := \sum_{j=1}^M a_j X_j$ is subgaussian, that is, for $t > 0$,*

$$\mathbb{P}(|\sum_{j=1}^M a_j X_j| \geq t) \leq 2\exp(-t^2/(4c\|a\|_2^2)).$$

Proof. By independence we have

$$\begin{aligned}\mathbb{E}\exp(\theta \sum_{j=1}^M a_j X_j) = \mathbb{E}\prod_{i=1}^M \exp(\theta a_j X_j) = \prod_{i=1}^M \mathbb{E}\exp(\theta a_j X_j) \leq \prod_{i=1}^M \exp(\theta a_j X_j) \\ = \exp(c\|a\|_2^2\theta^2).\end{aligned}$$

This shows that Z subgaussian with parameter $c\|a\|_2^2$ in (2.5). We apply Markov's inequality to obtain

$$\mathbb{P}(Z \geq t) = \mathbb{P}(\exp(\theta Z) \geq \exp(\theta t)) \leq \mathbb{E}[\exp(\theta Z)]e^{-\theta t} \leq e^{c\|a\|_2^2\theta^2 - \theta t}.$$

The optimal choice $\theta = t/(2c\|a\|_2^2)$ yields

$$\mathbb{P}(Z \geq t) \leq e^{-t^2/(4c\|a\|_2^2)}.$$

Repeating the above computation with $-Z$ instead of Z shows that

$$\mathbb{P}(-Z \geq t) \leq e^{-t^2/(4c\|a\|_2^2)},$$

and the union bound yields the desired estimate $\mathbb{P}(|Z| \geq t) \leq 2e^{-t^2/(4c\|a\|_2^2)}$. □

The following result from [**111**] is the Bernstein inequality for subexponential random variables.

Proposition A.2. *Let $X_1, \dots, X_N$ be independent centered subexponential random variables, and $K = \max_i \|X_i\|'_{\psi_1}$. Then for every $a = (a_1, \dots, a_N) \in \mathbb{R}^N$ and every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^N a_i X_i\right| \geq t\right) \leq 2\exp\left(-c\frac{t^2}{K^2\|a\|_2^2 + K\|a\|_\infty t}\right),$$

where $c > 0$ is an absolute constant.

The following theorem is the noncommutative Bernstein inequality due to Tropp, [**106**, Theorem 1.4].

Theorem A.3. *(Matrix Bernstein inequality) Let $\{X_\ell\}_{\ell=1}^M \in \mathbb{R}^{d\times d}$ be a sequence of independent random self-adjoint matrices. Suppose that $\mathbb{E}X_\ell = 0$ and $\|X_\ell\| \le K$ a.s. and put*

$$\sigma^2 := \left\| \sum_{\ell=1}^M \mathbb{E}X_\ell^2 \right\|.$$

Then for all $t \ge 0$,

$$\mathbb{P}\left(\left\| \sum_{\ell=1}^M X_\ell \right\| \ge t \right) \le 2d \exp\left(\frac{-t^2/2}{\sigma^2 + Kt/3} \right). \tag{A.1}$$

Remark. One can improve the tail bound (A.1) provided the $\{X_\ell\}$ are identically distributed and the $\mathbb{E}X_\ell^2$ are not full rank, say $\operatorname{rank}(\mathbb{E}X_\ell^2) = r < d$. Then (A.1) can be replaced by

$$\mathbb{P}\left(\left\| \sum_{\ell=1}^M X_\ell \right\| \ge t \right) \le 2r \exp\left(\frac{-t^2/2}{\sigma^2 + Kt/3} \right). \tag{A.2}$$

Sketch of the proof. We mainly improve [**106**, Corollary 3.7] under the assumptions above on X_ℓ. By [**106**, Theorem 3.6], for each $\theta > 0$, it holds that

$$\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^M X_\ell \right) \ge t \right) \le e^{-\theta t} \operatorname{Tr}\exp\left(\sum_{\ell=1}^M \ln\left(\mathbb{E}e^{\theta X_\ell} \right) \right),$$

where $\lambda_{\max}$ denotes the largest eigenvalue. Moreover, [**106**, Lemma 6.7] states that, for $\theta > 0$,

$$\mathbb{E}e^{\theta X} \preccurlyeq \exp\left(g(\theta)\mathbb{E}X^2 \right)$$

for a self-adjoint, centered random matrix X, where $g(\theta) = e^\theta - \theta - 1$. Using this result we obtain

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}\left(\sum_{\ell=1}^M X_\ell \right) \ge t \right) &\le e^{-\theta t} \operatorname{Tr}\exp\left(g(\theta) \sum_{\ell=1}^M \mathbb{E}X_\ell^2 \right) = e^{-\theta t} \operatorname{Tr}\exp\left(g(\theta) M \mathbb{E}X_1^2 \right) \\ &\le e^{-\theta t} r \ \lambda_{\max}\left[\exp\left(g(\theta) M \mathbb{E}X_1^2 \right) \right] = e^{-\theta t} r \ \exp\left(g(\theta) \lambda_{\max}(M\mathbb{E}X_1^2) \right).\end{aligned}$$

The first equality above uses that X_ℓ are identically distributed. The second inequality is valid because, for a positive definite matrix B with rank r, we have $\operatorname{Tr}B \le r\lambda_{\max}(B)$ and $\operatorname{rank}(c\,\mathbb{E}X_\ell^2) = \operatorname{rank}\left(\exp(c\,\mathbb{E}X_\ell^2)\right) = r$, for some $c > 0$. The rest of the proof proceeds in the same way as the proof of [**106**, Theorem 1.4].

We also give a rectangular version of the matrix Bernstein inequality as it appears in [**106**, Theorem 1.6].

Theorem A.4. *(Matrix Bernstein:rectangular) Let $\{Z_\ell\}_{\ell=1}^M \in \mathbb{R}^{d_1\times d_2}$ be a sequence of independent random matrices. Suppose that $\mathbb{E}Z_\ell = 0$ and $\|Z_\ell\| \le K$ a.s. and put*

$$\sigma^2 := \max\left\{ \left\| \sum_{\ell=1}^M \mathbb{E}(Z_\ell Z_\ell^*) \right\|, \left\| \sum_{\ell=1}^M \mathbb{E}(Z_\ell^* Z_\ell) \right\| \right\}.$$

Then for all $t \ge 0$,

$$\mathbb{P}\left(\left\| \sum_{\ell=1}^M Z_\ell \right\| \ge t \right) \le (d_1 + d_2) \exp\left(\frac{-t^2/2}{\sigma^2 + Kt/3} \right).$$

The next lemma is a deviation inequality for sums of independent random vectors which is a corollary of Bernstein inequalities for suprema of empirical processes [**54**, Corollary 8.45]. A similar result can be also found in [**65**, Theorem 12].

Lemma A.5. *(Vector Bernstein inequality) Let $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_M$ be independent copies of a random vector $\mathbf{Y}$ on $\mathbb{R}^n$ satisfying $\mathbb{E}\mathbf{Y} = 0$. Assume $\|\mathbf{Y}\|_2 \le K$. Let*

$$Z = \left\| \sum_{\ell=1}^{M} \mathbf{Y}_\ell \right\|_2, \quad \mathbb{E}Z^2 = M\mathbb{E}\|\mathbf{Y}\|_2^2,$$

and

$$\sigma^2 = \sup_{\|\mathbf{x}\|_2 \le 1} \mathbb{E}|\langle x, \mathbf{Y}\rangle|^2. \tag{A.3}$$

Then, for $t > 0$,

$$\mathbb{P}(Z \ge \sqrt{\mathbb{E}Z^2} + t) \le \exp\left(-\frac{t^2/2}{M\sigma^2 + 2K\sqrt{\mathbb{E}Z^2} + tK/3} \right). \tag{A.4}$$

Remark. The so-called weak variance σ^2 in (A.3) can be estimated by

$$\sigma^2 = \sup_{\|\mathbf{x}\|_2 \le 1} \mathbb{E}|\langle x, \mathbf{Y}\rangle|^2 \le \mathbb{E} \sup_{\|\mathbf{x}\|_2 \le 1} |\langle x, \mathbf{Y}\rangle|^2 = \mathbb{E}\|\mathbf{Y}\|_2^2. \tag{A.5}$$

A.2. Noncommutative Khintchine inequalities

For a matrix A, we define its Schatten p-norm as

$$\|A\|_{S_p} := \|\sigma(A)\|_p, \quad 1 \le p \le \infty,$$

where $\sigma(A) = (\sigma_1(A), \dots, \sigma_n(A))$ denotes the sequence of singular values of A. In particular, the following relations hold for Schatten norms, see [**94**] for a reference,

$$\|A\|_{S_{2n}}^{2n} = \mathrm{Tr}((AA^*)^n), \tag{A.6}$$

and, if A has rank r,

$$\|A\|_{S_p} \le r^{1/p}\|A\|. \tag{A.7}$$

Next, we state the noncommutative Khintchine inequality for matrix valued Gaussian sums as it appears in [**94**]. The same result holds for Bernoulli random variables which was first formulated by F. Lust-Piquard [**80**]. The optimal constants were provided by A. Buchholz in [**12**, **13**].

Lemma A.6. *Let $\mathbf{g} = (g_1, \dots, g_M)$ be a Gaussian sequence, and let B_j, $j = 1, \dots, M$ be complex matrices of the same dimension. Choose $n \in \mathbb{N}$. Then*

$$\mathbb{E}\| \sum_{j=1}^{M} g_j B_j \|_{S_{2n}}^{2n} \le C_n \max\left\{ \left\| \left(\sum_{j=1}^{M} B_j B_j^* \right)^{1/2} \right\|_{S_{2n}}^{2n}, \left\| \left(\sum_{j=1}^{M} B_j^* B_j \right)^{1/2} \right\|_{S_{2n}}^{2n} \right\},$$

where $C_n = \frac{(2n)!}{2^n n!}$.

The following lemma is taken from [**10**] and provides an example of "decoupling" the sums of chaos variables. Decoupling is a technique that reduces stochastic dependencies in certain sums of

random variables. A typical example of such a sum is $\sum_{j\neq k}\epsilon_j\epsilon_k\mathbf{x}_{jk}$ where $\mathbf{x}_{jk}$ are some vectors and (ϵ_j) is a Bernoulli series. Many other results about decoupling can be found in [**35**].

Lemma A.7. *Let $\boldsymbol{\xi}=(\xi_1,\dots,\xi_M)$ be a sequence of independent random variables with $\mathbb{E}\xi_j=0$ for all $j\in[M]$. Let $B_{j,k},\ j,k\in[M]$ be a double sequence of elements in a vector space with norm $\|\cdot\|$. Then for $1\leq p\leq\infty$*

$$\mathbb{E}\left\|\sum_{j\neq k}^{M}\xi_j\xi_k B_{j,k}\right\|^p\leq 4^p\mathbb{E}\left\|\sum_{j\neq k}^{M}\xi_j\xi'_k B_{j,k}\right\|^p,$$

where $\boldsymbol{\xi}'$ denotes an independent copy of $\boldsymbol{\xi}$.

The following result is the noncommutative Khintchine inequality for decoupled Gaussian chaos and the proof is same for Rademacher chaos which is given in [**93**]. A slightly general version without explicit constants appears in [**90**].

Lemma A.8. *Let $B_{j,k}\in\mathbb{C}^{r\times t}, j,k=1,\dots,M$, be complex matrices of the same dimension. Let $\mathbf{g},\mathbf{g}'$ be independent Gaussian sequences. Then, for $n\in\mathbb{N}$,*

$$\left[\mathbb{E}\left\|\sum_{j,k=1}^{M}g_jg'_kB_{j,k}\right\|_{S_{2n}}^{2n}\right]^{1/(2n)}\leq 2^{1/(2n)}\left(\frac{(2n)!}{2^n n!}\right)$$
$$\times\ \max\left\{\left\|\left(\sum_{j,k=1}^{M}B_{j,k}B_{j,k}^*\right)^{1/2}\right\|_{S_{2n}},\left\|\left(\sum_{j,k=1}^{M}B_{j,k}^*B_{j,k}\right)^{1/2}\right\|_{S_{2n}},\|F\|_{S_{2n}},\|\tilde{F}\|_{S_{2n}}\right\},$$

where $F,\tilde{F}$ are the block matrices $F=(B_{j,k})_{j,k=1}^M$ and $\tilde{F}=(B_{j,k}^*)_{j,k=1}^M$.

A.3. Moments and tails

The following lemma is a special case of [**54**, Proposition 8.2].

Lemma A.9. *Let $g_1,g_2,\dots,g_N\in\mathbb{R}$ be a sequence of (not necessarily independent) standard Gaussian random variables. Then,*

$$\mathbb{E}\max_{i\in[N]}|g_i|^2\leq c\ln(N)$$

for some $c>0$.

The next result [**94**, **105**] provides a relation between moments and tails of random variables.

Proposition A.10. *Suppose Z is a random variable satisfying*

$$(\mathbb{E}|Z|^p)^{1/p}\leq\alpha\beta^{1/p}p^{1/\gamma}\quad\text{for all } p_0\leq p\leq p_1$$

for some constants $\alpha,\beta,\gamma,p_1>p_0>0$. Then

$$\mathbb{P}(|Z|\geq e^{1/\gamma}\alpha u)\leq\beta e^{-u^\gamma/\gamma}$$

for all $u\in[p_0^{1/\gamma},p_1^{1/\gamma}]$.

Now we present a slight variation on the previous result.

Proposition A.11. *Suppose Z is a random variable satisfying*

$$(\mathbb{E}|Z|^p)^{1/p} \leq \beta^{1/p}(\alpha_1 p + \alpha_2\sqrt{p}) \quad \textit{for all } p \geq p_0.$$

Then, for $t \geq e\alpha_1 p_0 + e\alpha_2\sqrt{p_0}$,

$$\mathbb{P}(|Z| \geq t) \leq \beta \exp\left(-\frac{t^2}{e^2\alpha_2^2 + 2e\alpha_1 t}\right).$$

Proof. By Markov's inequality, we obtain

$$\mathbb{P}(|Z| \geq e(\alpha_1 u + \alpha_2\sqrt{u})) \leq \frac{\mathbb{E}|Z|^p}{(e(\alpha_1 u + \alpha_2\sqrt{u}))^p} \leq \beta\left(\frac{\alpha_1 p + \alpha_2\sqrt{p}}{e(\alpha_1 u + \alpha_2\sqrt{u})}\right)^p.$$

Choosing $p = u$ yields $\mathbb{P}(|Z| \geq e(\alpha_1 u + \alpha_2\sqrt{u})) \leq \beta e^{-u}$ for $u \geq p_0$. Taking the inverse of the function of u in the probability expression, we can equivalently write for $t \geq e\alpha_1 p_0 + e\alpha_2\sqrt{p_0}$

$$\mathbb{P}(|Z| \geq t) \leq \beta \exp\left(-\frac{v}{c^2} h_1\left(\frac{ct}{v}\right)\right)$$

where we set the function $h_1(x) = 1 + x - \sqrt{1+2x}$ and the variables $v = \frac{e^2\alpha_2^2}{2}$ and $c = e\alpha_1$. It follows from the elementary inequality $h_1(x) \geq \frac{x^2}{2(1+x)}$ that

$$\mathbb{P}(|Z| \geq t) \leq \beta \exp\left(-\frac{t^2}{2(v+ct)}\right).$$

Plugging v and c finishes the proof. □

The next lemma is due to Latała, [**77**, Corollary 3] and gives an estimate for the moments of sums of independent random variables.

Lemma A.12. *There exists a universal constant $K < 6$ such that if X_j are independent nonnegative random variables and $q \geq 1$, then*

$$\left(\mathbb{E}\left(\sum_j X_j\right)^q\right)^{1/q} \leq K\frac{q}{\ln(q)} \max\left\{\sum_j \mathbb{E}X_j, \left(\sum_j \mathbb{E}X_j^q\right)^{1/q}\right\}. \tag{A.8}$$

A.4. Concentration inequalities

The following lemma from [**32**] provides elegant estimates for the extremal singular values of a Gaussian matrix, see also [**54**, Chapter 9]. Its proof relies on the Slepian-Gordon Lemma [**60**, **61**] and the concentration of measure for Lipschitz functions [**78**].

Lemma A.13. *Let A be an $m \times s$ Gaussian matrix with $m > s$ and let $\sigma_{\min}$ and $\sigma_{\max}$ be the smallest and largest singular values of the normalized matrix $\frac{1}{\sqrt{m}}A$. Then for $r > 0$,*

$$\mathbb{P}(\sigma_{\min} \leq 1 - \sqrt{s/m} - r) \leq e^{-mr^2/2}, \tag{A.9}$$

$$\mathbb{P}(\sigma_{\max} \geq 1 + \sqrt{s/m} + r) \leq e^{-mr^2/2}. \tag{A.10}$$

The following result is the concentration of measure inequality for Lipschitz functions. The proofs and more details can be found in many references, for instance [**8**, **78**, **101**].

Lemma A.14. *(Concentration of measure) Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz function:*

$$|f(x) - f(y)| \leq L\|x - y\|_2, \textit{ for all } x, y \in \mathbb{R}^n.$$

Let $g = (g_1, \dots, g_n)$ be a vector of independent standard normal random variables. Then, for all $t > 0$,

$$\mathbb{P}(\mathbb{E}[f(g)] - f(g) \geq t) \leq e^{-\frac{t^2}{2L^2}}, \tag{A.11}$$

and consequently

$$\mathbb{P}(|f(g) - \mathbb{E}[f(g)]| \geq t) \leq 2e^{-\frac{t^2}{2L^2}}. \tag{A.12}$$

The following result appears in [**2**].

Lemma A.15. *Assume X is a random variable such that for all $t \geq 0$,*

$$\mathbb{P}(|X - \mathbb{E}X| > t) \leq c \exp(-\min\{t^2/\sigma_1^2, t/\sigma_2\}),$$

for some $c, \sigma_1, \sigma_2 \geq 0$. Then

$$\left|(\mathbb{E}X^2)^{1/2} - |\mathbb{E}X|\right| \leq C\sqrt{\sigma_1^2 + \sigma_2^2},$$

for some constant C that depends only on c.

A.5. Stirling's formula

The Gamma function is defined for $x > 0$ via

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt,$$

see [**54**] for a reference. For integer values n, it coincides with the factorial function, i.e.,

$$\Gamma(n) = (n-1)!. \tag{A.13}$$

Stirling's formula states that, for $x > 0$,

$$\Gamma(x) = \sqrt{2\pi}\ x^{x-\frac{1}{2}} e^{-x} \exp\left(\frac{\theta(x)}{12x}\right) \tag{A.14}$$

with $0 \leq \theta(x) \leq 1$. It simply follows that

$$\sqrt{2\pi}\ x^{x-\frac{1}{2}} e^{-x} \leq \Gamma(x) \leq \sqrt{2\pi}\ x^{x-\frac{1}{2}} e^{-x} e^{\frac{1}{12x}}. \tag{A.15}$$

Using (A.13) and (A.14) yields

$$n! = \sqrt{2\pi n}\ n^n e^{-n} e^{R_n}, \tag{A.16}$$

with $0 \leq R_n \leq 1/(12n)$.

Bibliography

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[2] N. Ailon and H. Rauhut. Fast and RIP-optimal transforms. *Preprint*, 2013.

[3] 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.

[4] R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1997.

[5] P. E. Bjørstad and J. Mandel. On the spectra of sums of orthogonal projections with applications to parallel computing. *BIT*, 31(1):76–88, Mar. 1991.

[6] B. G. Bodmann. Optimal linear transmission by loss-sensitive packet encoding. *Appl. Comput. Harmon. Anal.*, 22(3):274–285, 2007.

[7] B. G. Bodmann. Random fusion frames are nearly equiangular and tight. *Linear Algebra and its Applications*, 439(5):1401–1414, 2013.

[8] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[9] P. Boufounos, G. Kutyniok, and H. Rauhut. Sparse recovery from combined fusion frame measurements. *IEEE Trans. Inform. Theory*, 57(6):3864–3876, 2011.

[10] J. Bourgain and L. Tzafriri. Invertibility of 'large' submatrices with applications to the geometry of Banach spaces and harmonic analysis. *Israel J. Math.*, 57(2):137–224, 1987.

[11] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge Univ. Press, 2004.

[12] A. Buchholz. Operator Khintchine inequality in non-commutative probability. *Math. Ann.*, 319:1–16, 2001.

[13] A. Buchholz. Optimal constants in Khintchine type inequalities for fermions, Rademachers and q-Gaussian operators. *Bull. Pol. Acad. Sci. Math.*, 53(3):315–321, 2005.

[14] H. Buhrman, P. B. Miltersen, J. Radhakrishnan, and S. Venkatesh. Are bitvectors optimal? In *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing*, STOC '00, pages 449–458, New York, NY, USA, 2000. ACM.

[15] T. T. Cai and A. Zhang. Sharp RIP bound for sparse signal and low-rank matrix recovery. *Appl. Comput. Harmon. Anal.*, 35(1):74 – 93, 2013.

[16] T. T. Cai and A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. *IEEE Trans. Inf. Theory*, 60:122–132, 2014.

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

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

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

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

[21] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[22] E. J. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inf. Theory*, 57(11):7235 – 7254, 2011.

[23] E. J. Candès and B. Recht. Simple bounds for recovering low-complexity models. *Math. Program.*, 141(1-2):577–589, 2012.

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

[25] P. Casazza and G. Kutyniok. *Fusion Frames*. Applied and Numerical Harmonic Analysis. Boston, MA: Birkhäuser. xvi, 2013.

[26] P. G. Casazza and G. Kutyniok. Frames of subspaces. *Wavelets, Frames and Operator Theory*, pages 87–113, 2004.

[27] P. G. Casazza, G. Kutyniok, S. Li, and C. J. Rozell. Modeling sensor networks with fusion frames. In *Wavelets XII, Special Session on Finite-Dimensional Frames, Time-Frequency Analysis, and Applications*, volume 6701, page 11. Proc. SPIE, 2007.

[28] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6), 2012.

[29] O. Christensen. *An introduction to frames and Riesz bases*. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston, 2003.

[30] J. Conway, R. Hardin, and N. Sloane. Packing lines, planes, etc.: packings in Grassmannian spaces. *Exp. Math.*, 5(2):139–159, 1996.

[31] L. Daudet. Sparse and structured decompositions of signals with the molecular matching pursuit. *IEEE Transactions on Audio, Speech & Language Processing*, 14(5):1808–1816, 2006.

[32] K. Davidson and S. Szarek. Local operator theory, random matrices and Banach spaces. In W. B. Johnson and J. Lindenstrauss, editors, *Handbook of the geometry of Banach spaces I*. Elsevier, 2001.

[33] M. Davies and R. Gribonval. Restricted isometry constants where ℓ^p sparse recovery can fail for $0 < p \leq 1$. *IEEE Trans. Inform. Theory*, 55(5):2203–2214, 2009.

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

[35] V. de la Peña and E. Giné. *Decoupling. From Dependence to Independence*. Probability and its Applications (New York). Springer-Verlag, New York, 1999.

[36] I. S. Dhillon, R. W. Heath Jr., T. Strohmer, and J. A. Tropp. Constructing packings in Grassmannian manifolds via alternating projection. *Experiment. Math.*, 17(1):9–35, 2008.

[37] S. Dirksen. Tail bounds via generic chaining. ArXiv:1309.3522.

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

[39] D. Donoho and J. Tanner. Thresholds for the recovery of sparse solutions via l1 minimization. In *Conf. on Information Sciences and Systems*, 2006.

[40] D. Donoho and J. Tanner. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *J. Amer. Math. Soc.*, 22(1):1–53, 2009.

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

[42] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[43] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decompositions. *IEEE Trans. Inform. Theory*, 47(7):2845–2862, 2001.

[44] D. L. Donoho and J. Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 367(1906):4273–4293, 2009.

[45] C. Dossal, M.-L. Chabanol, G. Peyré, and J. Fadili. Sharp support recovery from noisy random measurements by ℓ_1-minimization. *Appl. Comput. Harmonic Anal.*, 33(1):24 – 43, 2012.

[46] M. Duarte, M. Davenport, D. Takhar, J. Laska, S. Ting, K. Kelly, and R. G. Baraniuk. Single-Pixel Imaging via Compressive Sampling. *Signal Processing Magazine, IEEE,*, 25(2):83–91, March , 2008.

[47] R. J. Duffin and A. C. Schaeffer. A class of nonharmonic Fourier series. *Trans. Amer. Math. Soc.*, 72:341–366, 1952.

[48] M. Elad and A. M. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. Inf. Theory*, 48(9):2558–2567, 2002.

[49] Y. Eldar, P. Kuppinger, and H. Bölcskei. Compressed sensing of block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans. Signal Process.*, 58:3042–3054, Jun. 2010.

[50] Y. Eldar and H. Rauhut. Average case analysis of multichannel sparse recovery using convex relaxation. *IEEE Trans. Inform. Theory*, 56(1):505–519, 2010.

[51] Y. C. Eldar and H. Bölcskei. Block-sparsity: Coherence and efficient recovery. In *ICASSP*, pages 2885–2888. IEEE, 2009.

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

[53] S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich. The Gelfand widths of l_p-balls for $0 < p \leq 1$. *J. Complexity*, 26:629–640, 2010.

[54] S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[55] J.-J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Trans. Inf. Th*, page 1344, 2004.

[56] A. Garnaev and E. Gluskin. On widths of the Euclidean ball. *Sov. Math., Dokl.*, 30:200–204, 1984.

[57] A. C. Gilbert and J. A. Tropp. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inform. Theory*, 53(12):4655–4666, 2007.

[58] E. Gilbert. A comparison of signalling alphabets. *Bell System Tech. Jour.*, 31:504–522, 1952.

[59] E. Gluskin. Norms of random matrices and widths of finite-dimensional sets. *Math. USSR, Sb.*, 48:173–182, 1984.

[60] Y. Gordon. Some inequalities for Gaussian processes and applications. *Israel J. Math.*, 50(4):265–289, 1985.

[61] Y. Gordon. Elliptically contoured distributions. *Probab. Theory Related Fields*, 76(4):429–438, 1987.

[62] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. Geometric aspects of functional analysis, Isr. Semin. 1986-87, Lect. Notes Math. 1317, 84-106., 1988.

[63] R. Graham and N. Sloane. Lower bounds for constant weight codes. *IEEE Trans. Inf. Theor.*, 26(1):37–43, Sept. 2006.

[64] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inform. Theory*, 49(12):3320–3325, 2003.

[65] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. on Inform. Theory*, 57(3):1548–1566, 2011.

[66] J. P. Haldar, D. Hernando, and Z.-P. Liang. Compressed-sensing MRI with random encoding. *IEEE Trans. Med. Imaging*, 30(4):893–903, 2011.

[67] M. Herman and T. Strohmer. High-resolution radar via compressed sensing. *IEEE Trans. Signal Process.*, 57(6):2275–2284, 2009.

[68] R. A. Horn. *Topics in matrix analysis*. Cambridge University Press, New York, NY, USA, 1986.

[69] M. Hügel, H. Rauhut, and T. Strohmer. Remote sensing via l_1-minimization. *Found. Comput. Math.*, 14:115–150, 2014.

[70] M. Kabanava and H. Rauhut. Analysis ℓ_1-recovery with frames and Gaussian measurements. *Preprint*, 2013.

[71] B. Kashin. Diameters of some finite-dimensional sets and classes of smooth functions. *Math. USSR, Izv.*, 11:317–333, 1977.

[72] O. Klopp. Noisy low-rank matrix completion with general sampling distribution. Working Papers 2012-06, Centre de Recherche en Economie et Statistique, Mar. 2012.

[73] V. Koltchinskii. Von Neumann entropy penalization and low-rank matrix estimation. *Ann. Stat.*, 39(6):2936–2973, 2011.

[74] V. Koltchinskii. A remark on low rank matrix recovery and noncommutative Bernstein type inequalities. *IMS Collections*, 9:213–226, 2012.

[75] F. Krahmer, S. Mendelson, and H. Rauhut. Suprema of chaos processes and the restricted isometry property. *Comm. Pure Appl. Math.*, to appear.

[76] G. Kutyniok, A. Pezeshki, R. Calderbank, and T. Liu. Robust dimension reduction, fusion frames, and Grassmannian packings. *Appl. Comput. Harmon. Anal.*, 26(1):64–76, 2009.

[77] R. Latała. Estimation of moments of sums of independent real random variables. *Annals of Probability*, 25(3):1502–1513, 1997.

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

[79] P. W. H. Lemmens and J. J. Seidel. Equi-isoclinic subspaces of Euclidean spaces. *Proc. Nederl. Akad. Wetensch. Series A*, pages 76:98–107, 1973.

[80] F. Lust-Piquard. Inégalites de Khintchine dans $c_p(1 < p < \infty)$. *C. R. Acad. Sci. Paris S'er. I Math.*, 303:289–292, 1986.

[81] M. Lustig, D. Donoho, and J. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.*, 58(6):1182–1195, 2007.

[82] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Process.*, 41(12):3397–3415, 1993.

[83] S. Mendelson, A. Pajor, and M. Rudelson. The geometry of random $\{-1, 1\}$-polytopes. *Discr. and Comp. Geometry*, 34(3):365–379, 2005.

[84] S. Mendelson, A. Pajor, and N. Tomczak Jaegermann. Uniform uncertainty principle for Bernoulli and subgaussian ensembles. *Constr. Approx.*, 28(3):277–289, 2009.

[85] D. Needell and R. Vershynin. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Found. Comput. Math.*, 9(3):317–334, 2009.

[86] D. Needell and R. Vershynin. Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit. *J. Sel. Topics Signal Processing*, 4(2):310–316, 2010.

[87] N. Nisan and A. Wigderson. Hardness vs randomness. *J. Comput. Syst. Sci.*, 49(2):149–167, Oct. 1994.

[88] P. Oswald. Frames and space splittings in Hilbert spaces. Technical report, 1997.

[89] S. Oymak, C. Thrampoulidis, and B. Hassibi. Simple bounds for noisy linear inverse problems with exact side information. ArXiv:1312.0641.

[90] G. Pisier. Non-commutative vector valued L_p-spaces and completely p-summing maps. *Astérisque*, (247), 1998.

[91] N. S. Rao, B. Recht, and R. D. Nowak. Universal measurement bounds for structured sparse signal recovery. *Journ. of Mach. Learn. Research. - Proc. Track*, 22:942–950, 2012.

[92] H. Rauhut. On the impossibility of uniform sparse reconstruction using greedy methods. *Sampl. Theory Signal Image Process.*, 7(2):197–215, 2008.

[93] H. Rauhut. Circulant and Toeplitz matrices in compressed sensing. In *Proc. SPARS'09*, Saint-Malo, France, 2009.

[94] H. Rauhut. Compressive sensing and structured random matrices. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, volume 9 of *Radon Series Comp. Appl. Math.*, pages 1–92. deGruyter, 2010.

[95] M. Rudelson and R. Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61:1025–1045, 2008.

[96] M. K. Simon. *Probability Distributions Involving Gaussian Random Variables: A Handbook for Engineers, Scientists and Mathematicians*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[97] M. Stojnic. Block-length dependent thresholds in block-sparse compressed sensing. *submitted to IEEE Trans. on Inform. Theory*, 2009. available at ArXiv:0907.3679.

[98] M. Stojnic. ℓ_1-optimization and its various thresholds in compressed sensing. In *ICASSP*, pages 3910–3913, 2010.

[99] M. Stojnic, F. Parvaresh, and B. Hassibi. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans. on Signal Proc.*, 57(8):3075–3085, 2009.

[100] M. Talagrand. *The Generic Chaining*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005. Upper and lower bounds of stochastic processes.

[101] M. Talagrand. *Mean Field Models for Spin Glasses. Volume I: Basic Examples*. Springer, 2010.

[102] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[103] J. Tropp. Recovery of short, complex linear combinations via l_1 minimization. *IEEE Trans. Inf. Theory*, 51(4):1568–1570, 2005.

[104] J. Tropp and D. Needell. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl. Comput. Harmon. Anal.*, 26(3):301–321, 2008.

[105] J. A. Tropp. On the conditioning of random subdictionaries. *Appl. Comput. Harmon. Anal.*, 25:1–24, 2008.

[106] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[107] E. van den Berg and M. P. Friedlander. SPGL1: A solver for large-scale sparse reconstruction, June 2007. http://www.cs.ubc.ca/labs/scl/spgl1.

[108] E. van den Berg and M. P. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2008.

[109] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer, 1996.

[110] R. Varshamov. Estimate of the number of signals in error correcting codes. *Dokl. Acad. Nauk SSSR*, 117:739–741, 1952.

[111] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge Univ Press, 2012.

[112] G. K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34:31–44, April 1991.