



# Preserving injectivity under subgaussian mappings and its application to compressed sensing



Peter G. Casazza<sup>a</sup>, Xuemei Chen<sup>b,\*</sup>, Richard G. Lynch<sup>c</sup>

<sup>a</sup> Department of Mathematics, University of Missouri, Columbia, MO, United States of America

<sup>b</sup> Department of Mathematical Sciences, New Mexico State University, Las Cruces, NM, United States of America

<sup>c</sup> Department of Mathematics, Texas A&M University, College Station, TX, United States of America

## ARTICLE INFO

### Article history:

Received 28 June 2019

Received in revised form 24 April 2020

Accepted 19 May 2020

Available online 26 May 2020

Communicated by Holger Rauhut

### Keywords:

Compressed sensing

Dictionary

Null space property

Sparse

Subgaussian

Frame

## ABSTRACT

The field of compressed sensing has become a major tool in high-dimensional analysis, with the realization that vectors can be recovered from relatively very few linear measurements as long as the vectors lie in a low-dimensional structure, typically the vectors that are zero in most coordinates with respect to a basis. However, there are many applications where we instead want to recover vectors that are sparse with respect to a dictionary rather than a basis. That is, we assume the vectors are linear combinations of at most  $s$  columns of a  $d \times n$  matrix  $\mathbf{D}$ , where  $s$  is very small relative to  $n$  and the columns of  $\mathbf{D}$  form a (typically overcomplete) spanning set. In this direction, we show that as a matrix  $\mathbf{D}$  stays bounded away from zero in norm on a set  $S$  and a provided map  $\Phi$  comprised of i.i.d. subgaussian rows has number of measurements at least proportional to the square of  $w(\mathbf{D}S)$ , the Gaussian width of the related set  $\mathbf{D}S$ , then with high probability the composition  $\Phi\mathbf{D}$  also stays bounded away from zero. As a specific application, we obtain that the null space property of order  $s$  is preserved under such subgaussian maps with high probability. Consequently, we obtain stable recovery guarantees for dictionary-sparse signals via the  $\ell_1$ -synthesis method with only  $O(s \log(n/s))$  random measurements and a minimal condition on  $\mathbf{D}$ , which complements the compressed sensing literature.

Published by Elsevier Inc.

## 1. Introduction

An important problem in high-dimensional analysis is to recover a signal from undersampled and corrupted measurements. This problem is ill-posed if no further assumptions are imposed on the signal class. With the breakthrough of compressive sensing (CS) (see [26]), we now know that it is possible to recover signals from very few (typically noisy) measurements, provided that the signals are sitting in a low-dimensional structure.

\* Corresponding author.

E-mail address: xchen@nmsu.edu (X. Chen).

To make this more concrete, we will use standard CS terminology. We wish to recover a signal  $\mathbf{z}_0 \in \mathbb{R}^d$  from its undersampled and corrupted linear measurements  $\mathbf{y} = \Phi \mathbf{z}_0 + \mathbf{e} \in \mathbb{R}^m$ , with the noise satisfying  $\|\mathbf{e}\|_2 \leq \varepsilon$ . The number of measurements  $m$  is assumed to be far less than the ambient dimension  $d$ , meaning the system has infinitely many solutions in general. To surpass this hurdle, we assume the signal  $\mathbf{z}_0$  has a sparse structure, that is, it can be written as the linear combination of only a few atoms from a dictionary. In other words, if  $\mathbf{D}$  is the matrix whose columns are the atoms, then  $\mathbf{z}_0 = \mathbf{D}\mathbf{x}_0$  for some sparse vector  $\mathbf{x}_0$ .

To recover dictionary-sparse signals, there are mainly two classes of algorithms: convex programming [13,41,8,25] and greedy algorithms [45,21,16,36]. This paper will focus on convex problems of the following form:

$$\hat{\mathbf{z}} = \arg \min_{\mathbf{z} \in \mathbb{R}^d} f_{\mathbf{D}}(\mathbf{z}), \quad \text{subject to } \|\mathbf{y} - \Phi \mathbf{z}\|_2 \leq \varepsilon, \quad (1)$$

where  $f_{\mathbf{D}}(\mathbf{z})$  is a convex function of  $\mathbf{z}$ . The unified form (1) is not new, see [15] for example. Two questions of interest when attacking this problem are the following:

- (Q1) What dictionaries ensure the existence of a sensing matrix  $\Phi$  to recover  $\mathbf{z}_0$  stably from (1)?
- (Q2) Given such a dictionary  $\mathbf{D}$ , how do we find suitable sensing matrices with number of rows as small as possible?

We begin by introducing in the next two subsections some currently known results related to answering these fundamental questions.

### 1.1. The basis case

Most CS literature focuses on the case when  $\mathbf{D}$  is the canonical basis, i.e.,  $\mathbf{z}_0 = \mathbf{x}_0$ . A suitable sensing matrix would mean that  $\Phi$  is able to extract the low-dimensional information off of  $\mathbf{z}_0$ . This is reflected in popular conditions like low coherence, the restricted isometry property, and the null space property, as well as their variations. In this case, the most popular method is the  $\ell_1$ -minimization, also known as the Basis Pursuit [8,25], where  $f_{\mathbf{D}}(\mathbf{z}) = \|\mathbf{z}\|_1$ :

$$\hat{\mathbf{z}} = \arg \min_{\mathbf{z} \in \mathbb{R}^d} \|\mathbf{z}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \Phi \mathbf{z}\|_2 \leq \varepsilon. \quad (2)$$

Proposed by Candes and Tao [9,10], the *restricted isometry property (RIP)* is sufficient to recover sparse signals via (2). A matrix  $\Phi \in \mathbb{R}^{m \times d}$  satisfies the RIP with constant  $0 \leq \delta < 1$  and sparsity  $s$  if

$$(1 - \delta)\|\mathbf{z}\|_2^2 \leq \|\Phi \mathbf{z}\|_2^2 \leq (1 + \delta)\|\mathbf{z}\|_2^2$$

for all  $s$ -sparse signals  $\mathbf{z} \in \mathbb{R}^d$ . The smallest  $\delta_s \geq 0$  for which the RIP holds is called the *restricted isometry constant*. This condition ensures that distinct sparse vectors have sufficiently far away measurements, providing explicit recovery guarantees.

While the RIP is a sufficient condition for recovery guarantees via (2), another property known as the *null space property (NSP)* is both necessary and sufficient. A sensing matrix  $\Phi \in \mathbb{R}^{m \times d}$  is said to satisfy the null space property of order  $s$  if for any index set  $T$  with  $|T| \leq s$  and any  $\mathbf{z} \in \ker(\Phi) \setminus \{0\}$ ,  $\|\mathbf{z}_T\|_1 < \|\mathbf{z}_{T^c}\|_1$  holds. Here  $\mathbf{z}_T$  denotes the vector having the same entries as  $\mathbf{z}$  on the support  $T$  and zero elsewhere. It is known that the successful recovery of sparse vectors from Basis Pursuit (2) when  $\varepsilon = 0$  occurs if and only if the NSP holds [20,27]. Moreover, it was shown in [3] that the NSP is necessary and sufficient for the stable recovery via Basis Pursuit. Other than the characterization of Basis Pursuit, another advantage of this property is that it only depends on the kernel of  $\Phi$ , which means that this property is invariant under

linear combinations of measurements (the rows of  $\Phi$ ). By a compactness argument, the NSP of order  $s$  is equivalent to the existence of  $0 < \gamma < 1$  such that

$$\|\mathbf{z}_T\|_1 < \gamma \|\mathbf{z}_{T^c}\|_1, \text{ for all } \mathbf{z} \in \ker(\Phi) \setminus \{0\}. \quad (3)$$

This is the so-called *stable null space property*.

The stable NSP is a matrix property. However, we will abuse the notation and say a vector  $\mathbf{z}$  has the *stable NSP* if  $\|\mathbf{z}_T\|_1 < \gamma \|\mathbf{z}_{T^c}\|_1$  for any index set  $T$  with cardinality at most  $s$ . We let  $S_\gamma^d$  be the set of vectors on the unit ball of  $\mathbb{R}^d$  that do not have the stable NSP. Explicitly,

$$S_\gamma^d := \{\mathbf{x} : \|\mathbf{x}_T\|_1 \geq \gamma \|\mathbf{x}_{T^c}\|_1 \text{ for some } |T| \leq s\} \cap \mathbb{S}_2^{d-1}. \quad (4)$$

Taking the intersection with the unit ball in the definition of  $S_\gamma^d$  is mainly for convenience the stable NSP for  $\Phi$  is equivalent to the existence of a positive lower bound of  $\|\Phi\mathbf{x}\|_2$  on  $S_\gamma^d$ . In summary,  $\Phi$  having the NSP is equivalent to the existence of  $0 < \gamma < 1$  and  $\eta > 0$  such that

$$\inf \{\|\Phi\mathbf{x}\|_2 : \mathbf{x} \in S_\gamma^d\} \geq \eta. \quad (5)$$

We will be using the notation  $S_\gamma^n$  frequently, so we conveniently denote it by  $S_\gamma$  instead since the dimension should be clear from context.

Next, we provide a recovery result using (5) that we are not able find in the literature.

**Theorem 1.1.** *If a sensing matrix  $\mathbf{A} \in \mathbb{R}^{m \times n}$  satisfies the stable NSP of order  $s$  with*

$$\inf \{\|\mathbf{A}\mathbf{v}\|_2 : \mathbf{v} \in S_\gamma\} \geq \eta \quad \text{for some } \eta > 0,$$

*then given  $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$  with  $\|\mathbf{e}\|_2 \leq \varepsilon$ , we have*

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_2 \leq \frac{2\gamma + 2}{1 - \gamma} \sigma_s(\mathbf{x}) + \frac{2\varepsilon}{\eta}, \quad (6)$$

where

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \varepsilon.$$

The condition (5) is similar to the so called robust null space property [26, Definition 4.17], and it resembles the recovery result of the robust NSP. We include its proof in Section 4.4 for completeness. The argument is fairly standard. In fact, a very similar argument can be found in the proof of [19, Theorem 3], which uses the robust null space property.

The best answer to (Q2) so far in the basis case is to use random matrices as the sensing matrix  $\Phi$ . It is well-known that random matrices whose entries are drawn from Gaussian or Bernoulli random variable satisfy the RIP with high probability, provided that  $m$  is only on the order of  $s \log(d/s)$  [9,39,18]. On the other hand, one needs at least  $O(s \log(d/s))$  number of measurements to ensure recovery, regardless of the decoder [24], and therefore random constructions achieve this minimum. There are many other types of random matrices that recover the signal effectively with (2), but do not necessarily have RIP, the Weibull matrices [22] for example. This is evidence that RIP is stronger than the NSP, which is explicitly proven in [5] using a semi-deterministic construction. The deterministic construction of suitable sensing matrices is significantly harder, and it requires many more measurements [4].

Much more can be said about both the history and the theory of CS in the basis case. For those interested in learning more, see the book [26] and the survey [24].

### 1.2. The general dictionary case

The general setting  $\mathbf{z}_0 = \mathbf{D}\mathbf{x}_0$  where  $\mathbf{D}$  is an arbitrary full rank  $d \times n$  matrix is much more challenging. When  $d = n$ , the columns of  $\mathbf{D}$  form a basis for  $\mathbb{R}^d$  and it is not hard to see that we can translate this to the canonical basis case as described before. However, the difficulty occurs when we assume that  $n > d$  so the dictionary  $\mathbf{D}$  is overcomplete. In this case,  $\mathbf{z}_0$  has infinitely many representations in  $\mathbf{D}$ , including possibly more than one sparse representation. There are many applications where the signals are seen through such a transformation and the need to understand when stable recovery is achievable is immense [38,7,29,2,17,14,23,34].

We note that such an overcomplete dictionary is also often called a *finite frame for  $\mathbb{R}^d$* . The field of finite frame theory is rich and has proven to be a powerful asset in many modern, real-world applications. We refer the inquisitive readers to [11,12] for a more thorough introduction to the elements of finite frame theory.

Perhaps the most reasonable recovery problem to consider in the dictionary setting, since it is the natural extension of (2), is the  $\ell_1$ -synthesis method:

$$\begin{aligned} \hat{\mathbf{x}} &= \arg \min \|\mathbf{x}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \Phi \mathbf{D} \mathbf{x}\|_2 \leq \varepsilon \\ \hat{\mathbf{z}} &= \mathbf{D} \hat{\mathbf{x}}. \end{aligned} \quad (7)$$

We note that defining  $\|\mathbf{z}\|_{K_{\mathbf{D}}} := \min\{\|\alpha\|_1 : \mathbf{D}\alpha = \mathbf{z}\}$  gives the following reformulation of the  $\ell_1$ -synthesis method (7):

$$\hat{\mathbf{z}} = \arg \min_{\mathbf{z} \in \mathbb{R}^d} \|\mathbf{z}\|_{K_{\mathbf{D}}} \quad \text{subject to} \quad \|\mathbf{y} - \Phi \mathbf{z}\|_2 \leq \varepsilon. \quad (8)$$

Specifically, for any convex set  $K$ , the *Minkowski functional* of  $K$  is defined as  $\|\mathbf{v}\|_K := \inf\{\lambda > 0 : \lambda^{-1}\mathbf{v} \in K\}$  so that in the dictionary setting where  $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_n]$  with  $K_{\mathbf{D}} := \text{conv}\{\pm \mathbf{d}_i\}_{i=1}^N$ , we have  $\|\mathbf{z}\|_{K_{\mathbf{D}}} = \min\{\|\alpha\|_1 : \mathbf{D}\alpha = \mathbf{z}\}$  [43]. The Minkowski functional is also known as the gauge of  $K$ , or the atomic norm associated to  $K$ .

One way to guarantee the successful recovery of (7) is to require  $\Phi \mathbf{D}$  to have the NSP or the RIP. The work by Rauhut et al. [38] showed that if  $\Phi \in \mathbb{R}^{m \times d}$  is a random matrix satisfying a concentration inequality with  $m = O(s \log \frac{n}{s})$  and  $\mathbf{D}$  satisfies the RIP, then the matrix  $\Phi \mathbf{D}$  also satisfies the RIP.

Once the composition  $\Phi \mathbf{D}$  satisfies the RIP, the program (7) (or (8)) will stably recover the sparse representation  $\mathbf{x}_0$ , and consequently the signal  $\mathbf{z}_0$ . However, we often only care about the recovery of  $\mathbf{z}_0$  in this dictionary based sparsity problem, in which case we allow  $\hat{\mathbf{x}}$  to be far away from  $\mathbf{x}_0$ .

To approach the problem in this new light, the work in [7] instead proposed the model where  $f_{\mathbf{D}}(\mathbf{z}) = \|\mathbf{D}^* \mathbf{z}\|_1$  in (1), called the  $\ell_1$ -analysis method:

$$\hat{\mathbf{z}} = \arg \min_{\mathbf{z} \in \mathbb{R}^d} \|\mathbf{D}^* \mathbf{z}\|_1 \quad \text{subject to} \quad \|\mathbf{y} - \Phi \mathbf{z}\|_2 \leq \varepsilon. \quad (9)$$

They showed that successful recovery via (9) is possible when  $\mathbf{D}$  is a Parseval frame, i.e.  $\mathbf{D}\mathbf{D}^* = \mathbf{I}_d$  and provided that  $\Phi$  satisfies a dictionary based RIP,  $\mathbf{D}$ -RIP. The definition of  $\mathbf{D}$ -RIP is similar to the usual RIP, but with  $\mathbf{D}\mathbf{x}$  in place of  $\mathbf{x}$ .

The  $\ell_1$ -analysis and  $\ell_1$ -synthesis models assume different sparsity to begin with. The analysis model assumes the sparsity of the analysis coefficient  $\mathbf{D}^* \mathbf{z}$ , which has applications in imaging where  $\mathbf{D}$  can be the finite difference operator, wavelets, shearlets, etc. [31,30]. The  $\ell_1$ -synthesis model assumes one of the infinitely many coefficients for  $\mathbf{z}$  in  $\mathbf{D}$  is sparse, as introduced at the beginning. This is more inclusive as the analysis coefficient is a particular case where the dual frame is the analysis operator ( $\mathbf{z} = \mathbf{D}\mathbf{D}^* \mathbf{z}$ ), see [29] for more details. On the technical side, the synthesis approach often imposes more challenges due to

its setting, and the fact that we do not know which dual frame of  $\mathbf{D}$  generates a sparse representation. The work by Chen et al. [14] tackled the  $\ell_1$ -synthesis problem and aimed to lay a framework for this method. They proposed a dictionary based NSP for the sensing matrix,  $\mathbf{D}$ -NSP for short, which we now define.

**Definition 1.2.** Let  $\mathbf{D} \in \mathbb{R}^{d \times n}$  be a dictionary. A matrix  $\Phi \in \mathbb{R}^{m \times d}$  is said to satisfy the  $\mathbf{D}$ -NSP of order  $s$  if for any index set  $T$  with  $|T| \leq s$  and any  $\mathbf{v}$  such that  $\mathbf{D}\mathbf{v} \in \ker \Phi \setminus \{\mathbf{0}\}$ , there is some  $\mathbf{u} \in \ker \mathbf{D}$  so that

$$\|\mathbf{v}_T + \mathbf{u}\|_1 < \|\mathbf{v}_{T^c}\|_1.$$

The  $\mathbf{D}$ -NSP is a characterization of exact recovery of dictionary sparse signals via (7) when  $\varepsilon = 0$ , and therefore is a generalization of the NSP. The following result helps emphasize the general direction of this paper.

**Theorem 1.3** ([14, Theorem 7.2]). *If  $\mathbf{D}$  is full spark, then  $\Phi$  has the  $\mathbf{D}$ -NSP with sparsity  $s$  if and only if  $\Phi\mathbf{D}$  has the NSP with the same sparsity.*

A frame  $\mathbf{D} \in \mathbb{R}^{d \times n}$  is *full spark* if every collection of  $d$  frame vectors is linearly independent. Full spark is not a strong assumption on dictionaries. In fact, it is quite obvious that if we randomly choose the entries of  $\mathbf{D}$  according to any continuous distribution, then  $\mathbf{D}$  will be full spark with probability one. More details can be found in [1].

As a (surprising) result of Theorem 1.3, if the  $\ell_1$ -synthesis method is successful at all, almost always, we will recover both  $\mathbf{x}_0$  and  $\mathbf{z}_0$ , and  $\mathbf{D}$  will satisfy the NSP since  $\ker(\mathbf{D}) \subset \ker(\Phi\mathbf{D})$ . In other words, if we are using  $\ell_1$ -synthesis to recover  $\mathbf{z}_0$ , then it is very reasonable to assume that  $\mathbf{D}$  has NSP and the coefficients  $\mathbf{x}_0$  will be recovered simultaneously. Therefore we will study the properties of the composition  $\Phi\mathbf{D}$  to ensure the success of  $\ell_1$ -synthesis.

Like the basis case, most work for the dictionary case often uses random measurements. The paper [30] uses Gaussian measurements for the  $\ell_1$ -analysis method, providing both nonuniform and uniform guarantees. The work [22] also considers the  $\ell_1$ -analysis approach, but instead uses Weibull measurements. As mentioned earlier, the work by Rauhut et al. [38] does analyze the  $\ell_1$ -synthesis method, however, it requires the dictionary  $\mathbf{D}$  to have the RIP. We again note that there is a gap between the RIP and the NSP [5], so we would like to reduce this assumption on  $\mathbf{D}$ . Additionally, it is known that a subgaussian matrix  $\Phi$  satisfies with high probability the  $\mathbf{D}$ -RIP [7], from which it is not hard to see that if  $\Phi$  is subgaussian, then essentially (with high probability)  $\Phi\mathbf{D}$  has the RIP if and only if  $\mathbf{D}$  has the RIP. This further solidifies the need to weaken the RIP assumption. Another notable work on random measurements in this setting is by Vershynin [43], which directly measures the recovery error in expectation. However, the error bound does not promote sparsity, and therefore will not provide exact reconstruction for  $s$ -sparse representations.

This paper will study the recovery prospects of (7) or (8) when the measurements are chosen at random. This kind of model is used in data acquisition when random measurements can be extracted, and is also applied in machine learning where data are assumed to have certain distributions. In this paper, we will assume the measurements are subgaussian, which we will review in Section 2.2. We wish to justify the effectiveness of random sensing matrices for recovering dictionary-sparse signals when  $\ell_1$ -synthesis (7) is used. As explained, this reduces to verifying the null space property of the composite  $\Phi\mathbf{D}$ . The biggest question is how small the number of measurements  $m$  can be given  $n, d, s$  fixed. If we treat  $\Phi\mathbf{D}$  as a whole and focus on recovering the coefficient  $\mathbf{x}_0$ , then this reduces to the basis case and the optimal number of measurements is  $m = O(s \log \frac{n}{s})$ , which is usually achieved by random construction as is the case in [22,30]. So we wish to answer the question:

*Given that  $\mathbf{D}$  has the NSP, find the smallest number of measurements such that  $\Phi\mathbf{D}$  has the NSP when the rows of  $\Phi$  are subgaussian.* (\*)

Notice that the NSP requirement on  $\mathbf{D}$  makes sense and is inevitable since  $\ker(\mathbf{D}) \subset \ker(\Phi\mathbf{D})$ .

### 1.3. Our contribution and organization

The contribution of this paper is twofold. The first main result, Theorem 3.1, states that a certain property of an operator/dictionary can be preserved under a subgaussian random map, given that this map is projecting to a dimension that is on the order of the square of the Gaussian width of certain set. It is not a coincidence that we study this preservability since the problem we wish to solve has this flavor. However, this could potentially be used to analyze other properties of compressed sensing matrices, or even beyond the scope of sparse analysis. Our second main result is the application of Theorem 3.1 to the null space property, thus solving (\*). Specifically, Theorem 3.3 states if  $\Phi \in \mathbb{R}^{m \times d}$  is a sensing matrix with independently drawn rows from a subgaussian distribution,  $\mathbf{D} \in \mathbb{R}^{d \times n}$  satisfies the NSP, and the number of measurements  $m$  is on the order of  $s \log(n/s)$ , then  $\Phi\mathbf{D}$  also satisfies the NSP with high probability. Consequently, we get a recovery result stated in Corollary 3.6, which is the first recovery result with only  $s \log(n/s)$  subgaussian measurements that only requires the dictionary to be NSP. As far as the authors can tell, this is the first work on the  $\ell_1$ -synthesis algorithm with random measurements that only requires a minimal condition on  $\mathbf{D}$ .

The road map is as follows. In Section 2, we provide the required preliminary material and notations that will be used throughout the paper. In Section 3, we introduce our main results as described above and argue how they are essentially optimal. Furthermore, we obtain as a corollary a more suitable recovery guarantee for the  $\ell_1$ -synthesis method. We then provide the theory behind our results in Section 4, as well as alternative estimates for the Gaussian width. The more specific cases that  $\Phi$  has rows drawn independent from a multivariate normal distribution is addressed in Section 5, wherein we provide better estimates. Finally, we end with some conclusions and discussions.

## 2. Preliminaries

We use  $\|\cdot\|_p$  for the standard  $\ell_p$  norm and we let

$$\mathbb{S}_p^{n-1} := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p = 1\} \text{ and } \mathbb{B}_p^{n-1} := \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_p \leq 1\}.$$

We also denote  $[n] := \{1, \dots, n\}$ . If  $S \subset \mathbb{R}^n$  and  $\mathbf{D} \in \mathbb{R}^{d \times n}$  is a dictionary, then we write  $\mathbf{D}S$  for  $\mathbf{D}S = \{\mathbf{D}\mathbf{x} : \mathbf{x} \in S\}$ . Also, we denote the columns of  $\mathbf{D}$  by  $\mathbf{d}_i$  so that  $\mathbf{D} = [\mathbf{d}_1, \dots, \mathbf{d}_n]$ . For a dictionary  $\mathbf{D}$ ,  $\|\mathbf{D}\|_2$  is the operator norm.

We will use the notation  $X \sim N(\mu, \sigma^2)$  to mean that a one dimensional random variable  $X$  follows a normal distribution with mean  $\mu$  and variance  $\sigma^2$ , and  $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  means a multidimensional random variable  $\mathbf{X}$  follows a multivariate normal distribution with mean vector  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ .

### 2.1. The Gaussian width

In the proof of our main result, we will need to bound  $w(\mathbf{D}S_\gamma)$ , where  $w$  denotes the Gaussian width defined as follows.

**Definition 2.1.** The *Gaussian width* of a set  $S \subset \mathbb{R}^n$  is defined as

$$w(S) := \mathbb{E} \sup_{\mathbf{x} \in S} \langle \mathbf{g}, \mathbf{x} \rangle,$$

where  $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_n)$  is a standard Gaussian random vector.

The Gaussian width plays a central role in asymptotic convex geometry. In particular, thinking of each inner product  $\langle \mathbf{g}, \mathbf{x} \rangle$  as a random projection, the Gaussian width measures how well, on average, the vectors in  $S$  can line up with a randomly chosen direction. For example, the Gaussian width of the unit ball is  $w(\mathbb{B}_2^{n-1}) = \mathbb{E} \|\mathbf{g}\|_2 = \sqrt{2} \Gamma(\frac{n+1}{2}) / \Gamma(\frac{n}{2})$ , which is on the order of  $\sqrt{n}$ . It is in this way that the Gaussian width can be thought of as a way to measure the “size” of a set [44]. In terms of CS, bounding the Gaussian width is how one obtains the important concentration equality used in the now standard CS proofs [26, Chapter 9]. Therefore, it is natural that our proof techniques will make use of it as well. Lastly, we note that it is often required that the set  $S$  be symmetric about the origin, which  $S_\gamma$  satisfies.

We will need the following result. The argument is given on Page 10 of [37], but we will provide it in Section 4.3 for the sake of completeness.

**Lemma 2.2.** For any map  $\mathbf{F} \in \mathbb{R}^{d \times n}$  and any  $S \subset \mathbb{R}^n$ , we have

$$w(\mathbf{F}S) \leq \|\mathbf{F}\|_2 w(S).$$

## 2.2. Subgaussian vectors

The measurement matrix  $\Phi \in \mathbb{R}^{m \times d}$  in our result will have rows drawn i.i.d. from a subgaussian distribution, which we now define following [42].

**Definition 2.3.** A random vector  $\varphi \in \mathbb{R}^d$  is called a *subgaussian vector* with parameters  $(\alpha, \sigma)$  if it satisfies the following.

- (1) It is centered, that is,  $\mathbb{E}[\varphi] = \mathbf{0}$ .
- (2) There exists a positive  $\alpha$  such that  $\mathbb{E}[|\langle \varphi, \mathbf{z} \rangle|] \geq \alpha$  for every  $\mathbf{z} \in \mathbb{S}_2^{d-1}$ .
- (3) There exists a positive  $\sigma$  such that  $\Pr(|\langle \varphi, \mathbf{z} \rangle| \geq t) \leq 2 \exp(-t^2/(2\sigma^2))$  for every  $\mathbf{z} \in \mathbb{S}_2^{d-1}$ .

There are many examples of subgaussian vectors, including vectors with independent Gaussian entries, or independent Bernoulli entries, as well as independent bounded entries. We list the following two cases in detail since they are used in corresponding theorems in the next section.

**Example 2.4 (Standard Gaussian vector).** Let  $\varphi \in \mathbb{R}^d \sim N(\mathbf{0}, \mathbf{I}_d)$  be a standard Gaussian vector. If  $\mathbf{z} \in \mathbb{S}_2^{d-1}$ , then  $Z := \langle \varphi, \mathbf{z} \rangle \sim N(0, 1)$  and it is well known that

$$\mathbb{E}[|Z|] = \sqrt{\frac{2}{\pi}} \quad \text{and} \quad \Pr(|Z| \geq t) \leq \exp\left(-\frac{t^2}{2}\right),$$

so the standard Gaussian vector is subgaussian with parameters  $\alpha = \sqrt{2/\pi}$  and  $\sigma = 1$ .

**Example 2.5 (Nonstandard Gaussian vector).** Suppose that  $\varphi \in \mathbb{R}^d \sim N(\mathbf{0}, \Sigma)$  where the covariance matrix  $\Sigma$  has smallest and largest singular values,  $\sigma_{\min}^2$  and  $\sigma_{\max}^2$ , respectively. Then  $\psi := \Sigma^{-1/2} \varphi \sim N(\mathbf{0}, \mathbf{I}_d)$  and we can compute for all  $\mathbf{z} \in \mathbb{S}_2^{d-1}$  that

$$\mathbb{E}[|\langle \varphi, \mathbf{z} \rangle|] = \mathbb{E}[|\langle \Sigma^{1/2} \psi, \mathbf{z} \rangle|] = \mathbb{E}[|\langle \psi, \Sigma^{1/2} \mathbf{z} \rangle|] = \|\Sigma^{1/2} \mathbf{z}\|_2 \mathbb{E}\left[\left|\left\langle \psi, \frac{\Sigma^{1/2} \mathbf{z}}{\|\Sigma^{1/2} \mathbf{z}\|} \right\rangle\right|\right] \geq \sigma_{\min} \sqrt{\frac{2}{\pi}}$$

and in a similar fashion

$$\Pr[|\langle \varphi, \mathbf{z} \rangle| \geq t] = \Pr\left[\left|\left\langle \psi, \frac{\Sigma^{1/2} \mathbf{z}}{\|\Sigma^{1/2} \mathbf{z}\|} \right\rangle\right| \geq \frac{t}{\|\Sigma^{1/2} \mathbf{z}\|_2}\right] \leq \exp\left(-\frac{t^2}{2\sigma_{\max}^2}\right),$$

so that  $\varphi$  is subgaussian with parameters  $\alpha = \sigma_{\min} \sqrt{2/\pi}$  and  $\sigma = \sigma_{\max}$ .

### 2.3. The mean empirical width

If  $\{\mathbf{f}_i\}_{i=1}^m$  are independent copies of the random distribution  $\mathbf{f} \in \mathbb{R}^n$ , then we can define the *mean empirical width* of a set  $S \subset \mathbb{R}^n$  as

$$W_m(S; \mathbf{f}) := \mathbb{E} \sup_{\mathbf{x} \in S} \left\langle \mathbf{x}, \frac{1}{\sqrt{m}} \sum_{i=1}^m \varepsilon_i \mathbf{f}_i \right\rangle,$$

where  $\{\varepsilon_i\}_{i=1}^m$  are independent random variables taking values uniformly over  $\{\pm 1\}$  and are independent from everything else.

The mean empirical width  $W_m(S; \mathbf{f})$  is a distribution-dependent measure of the size of the set  $S$ . Note that  $W_m(S; \mathbf{f})$  reduces to the usual Gaussian width  $w(S)$  when  $\mathbf{f}$  follows a standard Gaussian distribution, as shown in Remark 2.7. Estimation of  $W_m(S; \mathbf{f})$  for any subgaussian vector  $\mathbf{f}$  is made in [42], where  $S$  is required to be  $\mathbb{S}_2^{n-1} \cup K$  for some cone  $K$ . However, the bound can be relaxed to any subset  $S$  by the observation of the generic chaining bound and the majorizing measure theorem [40, Theorem 2.2.18 and Theorem 2.4.1]. We will state this as a lemma.

**Lemma 2.6.** *If  $\mathbf{f} \in \mathbb{R}^n$  is subgaussian with parameters  $(\alpha, \sigma)$  and  $S$  is any subset of  $\mathbb{R}^n$ , then*

$$W_m(S; \mathbf{f}) \leq C \sigma w(S) \quad (10)$$

for some universal constant  $C$ .

The constant  $C$  will appear in the main results. It is a universal constant that does not rely on the choice of subgaussian distribution. See [40] for precise computations of this constant. However, better estimates can be made using different techniques as is shown in the next example.

**Remark 2.7.** When  $\mathbf{f}$  follows a centered multivariate normal distribution  $N(\mathbf{0}, \Sigma)$ , we can take  $C = 1$ . Let  $\mathbf{g}_i := \Sigma^{-1/2} \mathbf{f}_i \sim N(\mathbf{0}, \mathbf{I}_n)$ , then  $\mathbf{g} = \sum_{i=1}^m \varepsilon_i \mathbf{g}_i \sim N(\mathbf{0}, \mathbf{I}_n)$  as well, so

$$\begin{aligned} W_m(S; \mathbf{f}) &= \mathbb{E} \sup_{\mathbf{z} \in S} \left\langle \mathbf{z}, \frac{1}{\sqrt{m}} \sum_{i=1}^m \varepsilon_i \mathbf{f}_i \right\rangle = \mathbb{E} \sup_{\mathbf{z} \in S} \left\langle \mathbf{z}, \frac{1}{\sqrt{m}} \sum_{i=1}^m \varepsilon_i \Sigma^{1/2} \mathbf{g}_i \right\rangle \\ &= \mathbb{E} \sup_{\mathbf{z} \in S} \left\langle \Sigma^{1/2} \mathbf{z}, \frac{1}{\sqrt{m}} \sum_{i=1}^m \varepsilon_i \mathbf{g}_i \right\rangle = \mathbb{E} \sup_{\mathbf{x} \in \Sigma^{1/2} S} \langle \mathbf{x}, \mathbf{g} \rangle \\ &= w(\Sigma^{1/2} S) \leq \sigma_{\max} w(S), \end{aligned}$$

where the last inequality is due to Lemma 2.2.

The mean empirical width appears in the following important result. This theorem was originally stated in [32] and coined *Mendelson's Small Ball Method* by J. Tropp [42]. This will be a primary tool in obtaining our main estimates.

**Theorem 2.8** ([42, Proposition 5.1], cf. [32, Theorem 2.1]). *Fix a set  $S \subset \mathbb{R}^n$ . Let  $\mathbf{f}$  be a random vector in  $\mathbb{R}^n$  and let  $\mathbf{F} \in \mathbb{R}^{m \times n}$  have rows  $\{\mathbf{f}_i^\top\}_{i=1}^m$  that are independent copies of  $\mathbf{f}^\top$ . Define*

$$Q_\xi(S; \mathbf{f}) := \inf_{\mathbf{x} \in S} \Pr \left( |\langle \mathbf{x}, \mathbf{f} \rangle| \geq \xi \right) \quad \text{and} \quad W_m(S; \mathbf{f}) := \mathbb{E} \sup_{\mathbf{x} \in S} \left\langle \mathbf{x}, \frac{1}{\sqrt{m}} \sum_{i=1}^m \varepsilon_i \mathbf{f}_i \right\rangle,$$

where  $\{\varepsilon_i\}_{i=1}^m$  are independent random variables taking values uniformly over  $\{\pm 1\}$  and are independent from everything else. Then for any  $\xi > 0$  and  $t > 0$ , we have

$$\inf_{\mathbf{x} \in S} \|\mathbf{F}\mathbf{x}\|_2 \geq \xi \sqrt{m} Q_{2\xi}(S; \mathbf{f}) - 2W_m(S; \mathbf{f}) - \xi t \quad (11)$$

with probability  $\geq 1 - e^{-t^2/2}$ .

### 3. Main results

#### 3.1. Preservability under subgaussian maps

**Theorem 3.1** (Preservability under random maps). Let  $\mathbf{D} \in \mathbb{R}^{d \times n}$  be arbitrary, let  $S \subset \mathbb{R}^n$  be such that  $\inf \{\|\mathbf{D}\mathbf{x}\|_2 : \mathbf{x} \in S\} \geq \eta$  for some constant  $\eta > 0$ , and assume  $\boldsymbol{\varphi} \in \mathbb{R}^d$  is a subgaussian vector with parameters  $(\alpha, \sigma)$ . If  $\Phi \in \mathbb{R}^{m \times d}$  is a measurement matrix with rows that are independent copies of  $\boldsymbol{\varphi}^\top$  and that the number of measurements satisfies

$$m \geq \frac{4^8 \sigma^6}{\eta^2 \alpha^6} C^2 w^2(\mathbf{D}S),$$

then with probability at least

$$1 - \exp\left(-m \frac{\alpha^4}{64^2 \sigma^4}\right),$$

we have

$$\inf_{\mathbf{x} \in S} \|\Phi \mathbf{D}\mathbf{x}\|_2 \geq C \sigma w(\mathbf{D}S). \quad (12)$$

Theorem 3.1 is beyond the null space property. It says that if an operator stays bounded away from  $\mathbf{0}$  on some set, then this operator under a random map also stays bounded away from  $\mathbf{0}$  on the same set, given that the dimension of the random map is at least proportional to the square of the Gaussian width of the related set. This could be potentially useful for other dimension reduction analysis. The proof of Theorem 3.1 can be found in Section 4.1. In equation (12), the constant  $\eta$  does not show explicitly, but it is reflected in the number of measurements  $m$  since  $\Phi$  is random.

**Remark 3.2.** Theorem 3.1 can be compared to [6, Proposition 18]. Both statements are about the minimal number of measurements related to Gaussian width. However, ours has a dictionary  $\mathbf{D}$  incorporated.

As an application of Theorem 3.1, we let  $S = S_\gamma$ , as defined in (4), and compute the Gaussian width of  $\mathbf{D}S_\gamma$ . This is a key step in this paper and is not a simple task. See Theorem 4.7 on bounding the Gaussian width. Recall that  $S_\gamma$  is the set of vectors that violates the stable NSP. Theorem 3.1 and Theorem 4.7 imply the following theorem on preserving the null space property.

**Theorem 3.3.** Assume  $\Phi \in \mathbb{R}^{m \times d}$  is a sensing matrix comprised of rows drawn i.i.d. from a subgaussian distribution with parameters  $(\alpha, \sigma)$ . Take  $\mathbf{D} \in \mathbb{R}^{d \times n}$  to be a dictionary satisfying the stable NSP of order  $s$  with  $\inf \{\|\mathbf{D}\mathbf{x}\|_2 : \mathbf{x} \in S_\gamma\} \geq \eta$  for some  $\eta > 0$  and satisfying  $\max \{\|\mathbf{d}_i\|_2^2 : i \in [n]\} \leq \rho$ . If the number of measurements satisfies

$$m \geq \frac{36 \cdot 4^8 \sigma^6}{\eta^2 \alpha^6} \frac{\rho}{\gamma^2} C^2 s \log\left(\frac{\sqrt{2n}}{s}\right), \quad (13)$$

then with probability at least

$$1 - \exp\left(-m \frac{\alpha^4}{64^2 \sigma^4}\right),$$

the composition  $\Phi \mathbf{D}$  also has the stable NSP of order  $s$  with  $\inf \{\|\Phi \mathbf{D} \mathbf{x}\|_2 : \mathbf{x} \in S_\gamma\} \geq C \sigma w(\mathbf{D} S_\gamma)$ .

**Remark 3.4.** Theorem 3.3 only requires a minimal condition on  $\mathbf{D}$ . If  $\Phi \mathbf{D}$  has the NSP, then  $\mathbf{D}$  must also have the NSP (hence some kind of stable NSP) since  $\ker(\mathbf{D})$  is a proper subspace of  $\ker(\Phi \mathbf{D})$ . Thus,  $\mathbf{D}$  having the NSP is a very reasonable condition if we want stable recovery through  $\ell_1$ -synthesis.

**Remark 3.5.** The constants in (13) are well controlled. The ratio  $\frac{\sigma}{\alpha}$  reflects how well the distributions behave and is the condition number of  $\Sigma$  if  $\varphi \sim N(\mathbf{0}, \Sigma)$ . To reiterate, the constant  $C$  that appears in the above results is a universal constant. A more precise estimate can be made when  $\varphi$  follows the multivariate normal distribution. Furthermore,  $\rho$  also equals one when each frame vector has unit norm, which is often the case. The constants  $\eta$  and  $\gamma$  reflect the null space property of  $D$ .

### 3.2. Sparse recovery via the $\ell_1$ -synthesis method

As a consequence of Theorem 3.3, we obtain a uniform recovery result using Theorem 1.1 with  $\mathbf{A}$  replaced by  $\Phi \mathbf{D}$ . We will use the standard notation

$$\sigma_s(\mathbf{x}) := \inf \{\|\mathbf{x} - \mathbf{v}\|_1 : \mathbf{v} \text{ is } s\text{-sparse}\},$$

to denote the  $\ell_1$ -error of best  $s$ -term approximation to a vector  $\mathbf{x}$ . This infimum is achieved by taking  $\mathbf{v} := \mathbf{x}_T$ , where  $T$  is the index set containing the indices where the  $s$ -largest absolute value entries of  $\mathbf{x}$  occur.

**Corollary 3.6.** Suppose  $\Phi \in \mathbb{R}^{m \times d}$  is a sensing matrix comprised of rows drawn i.i.d. from a subgaussian distribution with parameters  $(\alpha, \sigma)$  and suppose  $\mathbf{D} \in \mathbb{R}^{d \times n}$  has the stable NSP of order  $s$  with  $\inf \{\|\mathbf{D} \mathbf{x}\|_2 : \mathbf{x} \in S_\gamma\} \geq \eta$  for some  $\eta > 0$  and satisfies  $\max \{\|\mathbf{d}_i\|_2^2 : i \in [n]\} \leq \rho$ . Let  $\mathbf{z}_0 = \mathbf{D} \mathbf{x}_0$ , and the measurements  $\mathbf{y}$  satisfying  $\|\mathbf{y} - \Phi \mathbf{z}_0\|_2 \leq \varepsilon$ . If the number of measurements satisfies

$$m \geq \frac{36 \cdot 4^8}{\eta^2} \frac{\sigma^6}{\alpha^6} \frac{\rho}{\gamma^2} C^2 s \log\left(\frac{\sqrt{2}n}{s}\right),$$

then with probability at least  $1 - \exp\left(-m \frac{\alpha^4}{64^2 \sigma^4}\right)$ , the  $\ell_1$ -synthesis method (7) provides a stable recovery for both the coefficients  $\mathbf{x}_0$  and the signal  $\mathbf{z}_0$  as

$$\begin{aligned} \|\hat{\mathbf{x}} - \mathbf{x}_0\|_2 &\leq \frac{2\gamma + 2}{1 - \gamma} \sigma_s(\mathbf{x}_0) + \frac{2\varepsilon}{C\sigma\eta} \\ \|\hat{\mathbf{z}} - \mathbf{z}_0\|_2 &\leq \|\mathbf{D}\|_2 \frac{2\gamma + 2}{1 - \gamma} \sigma_s(\mathbf{x}_0) + \frac{2\varepsilon \|\mathbf{D}\|_2}{C\sigma\eta}. \end{aligned}$$

## 4. Proofs of the results

We will apply Mendelson's Small Ball Method in our setting. Most of the work will be in properly estimating two important quantities, which will further lead to the need to estimate the Gaussian width  $w(\mathbf{D} S_\gamma)$ . This is what makes up Sections 4.1 and 4.2. In doing so, we will obtain the bound on the number

of measurements in Theorem 3.3 and its corollaries that forces  $\Phi \mathbf{D}$  to have the NSP with high probability given that  $\Phi$  is a matrix made up of independent copies of a subgaussian vector and  $\mathbf{D}$  satisfies the NSP. In Section 4.3 we prove Lemma 2.2 and use it to obtain a different estimate for the Gaussian width than in Section 4.2. Lastly, in Section 4.4, we provide a stable recovery result with stable NSP for completeness.

#### 4.1. Preservability under a random map

Notice that if the rows of  $\Phi$  are independent copies of a random vector  $\varphi$ , then the rows of  $\Phi \mathbf{D}$  are independent copies of the random vector  $\mathbf{D}^\top \varphi$ . We will apply Theorem 2.8 with  $\Phi \mathbf{D}$  in place of  $\mathbf{F}$  and the random vector  $\mathbf{D}^\top \varphi$  in place of  $\mathbf{f}$ , which in turn will require us to estimate the quantities  $Q_{2\xi}(S; \mathbf{D}^\top \varphi)$  and  $W_m(S; \mathbf{D}^\top \varphi)$ .

In the proof of Theorem 4.2, we will use the following lemma to bound  $Q_{2\xi}(S; \mathbf{D}^\top \varphi)$ . The proof can be found in [42, Section 6.5].

**Lemma 4.1.** *If  $\mathbf{f} \in \mathbb{R}^n$  is a subgaussian vector with parameters  $(\alpha, \sigma)$ , then*

$$\Pr[|\langle \mathbf{x}, \mathbf{f} \rangle| \geq t] \geq \frac{(\alpha - t)^2}{4\sigma^2}$$

for any  $0 < t < \alpha$  and  $\mathbf{x} \in \mathbb{S}_2^{n-1}$ .

**Theorem 4.2.** *Let  $\mathbf{D} \in \mathbb{R}^{d \times n}$  be arbitrary, let  $S \subset \mathbb{R}^n$  be so that  $\inf\{\|\mathbf{D}\mathbf{x}\|_2 : \mathbf{x} \in S\} \geq \eta$  for some constant  $\eta > 0$ , and let  $\varphi \in \mathbb{R}^d$  be a subgaussian measurement with parameter  $(\alpha, \sigma)$ . If  $\Phi \in \mathbb{R}^{m \times d}$  has rows that are independent copies of  $\varphi^\top$ , then*

$$\inf_{\mathbf{x} \in S} \|\Phi \mathbf{D}\mathbf{x}\|_2 \geq \frac{\alpha\eta}{4^3} \left(\frac{\alpha}{\sigma}\right)^2 \sqrt{m} - 2C\sigma w(\mathbf{D}S) - \frac{\alpha\eta}{4}t \quad (14)$$

for any  $t > 0$  with probability at least  $1 - e^{-t^2/2}$ .

**Proof.** We first apply Mendelson's Small Ball Method, Theorem 2.8, with  $\mathbf{F}$  replaced by  $\Phi \mathbf{D}$  and therefore  $\mathbf{f}$  replaced by  $\mathbf{D}^\top \varphi$  to obtain the bound

$$\inf_{\mathbf{x} \in S} \|\Phi \mathbf{D}\mathbf{x}\|_2 \geq \xi \sqrt{m} Q_{2\xi}(S; \mathbf{D}^\top \varphi) - 2W_m(S; \mathbf{D}^\top \varphi) - \xi t. \quad (15)$$

By Lemma 4.1, provided we choose  $\xi$  to satisfy  $2\xi/\eta < \alpha$ , we obtain for any  $\mathbf{x} \in S$

$$\begin{aligned} \Pr(|\langle \mathbf{D}\mathbf{x}, \varphi \rangle| \geq 2\xi) &= \Pr\left(\left|\left\langle \frac{\mathbf{D}\mathbf{x}}{\|\mathbf{D}\mathbf{x}\|_2}, \varphi \right\rangle\right| \geq \frac{\xi}{\|\mathbf{D}\mathbf{x}\|_2}\right) \\ &\geq \Pr\left(\left|\left\langle \frac{\mathbf{D}\mathbf{x}}{\|\mathbf{D}\mathbf{x}\|_2}, \varphi \right\rangle\right| \geq \frac{\xi}{\eta}\right) \geq \frac{(\alpha - 2\xi/\eta)^2}{4\sigma^2} \end{aligned} \quad (16)$$

and therefore

$$Q_{2\xi}(S; \mathbf{D}^\top \varphi) = \inf_{\mathbf{x} \in S} \Pr(|\langle \mathbf{x}, \mathbf{D}^\top \varphi \rangle| \geq 2\xi) = \inf_{\mathbf{x} \in S} \Pr(|\langle \mathbf{D}\mathbf{x}, \varphi \rangle| \geq 2\xi) \geq \frac{(\alpha - 2\xi/\eta)^2}{4\sigma^2}.$$

Lemma 2.6 readily gives the estimate

$$W_m(S; \mathbf{D}^\top \varphi) = W_m(\mathbf{D}S; \varphi) \leq C\sigma w(\mathbf{D}S).$$

Placing these two bounds into (15) and choosing  $\xi$  to satisfy  $2\xi/\eta = \alpha/2$  gives the bound in (14).  $\square$

**Remark 4.3.** Notice that Theorem 4.2 is a generalization of [42, Theorem 6.3]. In this generalization, it is crucial that  $D$  has NSP as is evident in (16).

Finally, we can provide the proof of our first main result, Theorem 3.1.

**Proof of Theorem 3.1.** Theorem 4.2 implies that

$$\inf_{x \in S} \|\Phi \mathbf{D} \mathbf{x}\|_2 \geq \frac{\alpha \eta}{4^3} \left( \frac{\alpha}{\sigma} \right)^2 \sqrt{m} - 2C\sigma w(\mathbf{D}S) - \frac{\alpha \eta}{4} t := a - b - \frac{\alpha \eta}{4} t.$$

Picking  $m$  and  $t$  to satisfy  $a \geq 2b$  and  $(\alpha \eta / 4)t = (a - b)/2$  gives

$$\inf_{x \in S} \|\Phi \mathbf{D} \mathbf{x}\|_2 \geq a - b - (a - b)/2 = (a - b)/2 \geq b/2 = C\sigma w(\mathbf{D}S).$$

All that is left is to rewrite these conditions in terms of  $m$  and  $t$ . We have

$$a \geq 2b \Leftrightarrow \frac{\alpha \eta}{4^3} \left( \frac{\alpha}{\sigma} \right)^2 \sqrt{m} \geq 4C\sigma w(\mathbf{D}S) \Leftrightarrow m \geq \frac{4^8 \sigma^6}{\eta^2 \alpha^6} C^2 w^2(\mathbf{D}S)$$

and

$$\frac{\alpha \eta}{4} t = \frac{a - b}{2} \geq \frac{a}{4} = \frac{\alpha \eta}{4^4} \left( \frac{\alpha}{\sigma} \right)^2 \sqrt{m} \Leftrightarrow t \geq \frac{1}{64} \sqrt{m} \left( \frac{\alpha}{\sigma} \right)^2 \Leftrightarrow -\frac{t^2}{2} \leq -m \frac{\alpha^4}{64^2 \sigma^4},$$

proving the result.  $\square$

#### 4.2. Estimating the Gaussian width

In order to prove Theorem 3.3, we apply Theorem 3.1 to the null space property, i.e., let  $S = S_\gamma$ . Therefore the last ingredient of the proof of Theorem 3.3 is to suitably bound the Gaussian width  $w(\mathbf{D}S_\gamma)$ . Note that it is pivotal to have a relatively optimal upper bound on  $w(\mathbf{D}S_\gamma)$  since the number of measurements  $m$  is on the order of the square of this width.

Recall that

$$w(\mathbf{D}S_\gamma) := \mathbb{E} \sup_{\mathbf{z} \in \mathbf{D}S_\gamma} \langle \mathbf{g}, \mathbf{z} \rangle = \mathbb{E} \sup_{\mathbf{x} \in S_\gamma} \langle \mathbf{D}^\top \mathbf{g}, \mathbf{x} \rangle, \quad (17)$$

where  $\mathbf{g} \in \mathbb{R}^d$  is a standard Gaussian vector.

We again point out that bounding  $w(\mathbf{D}S_\gamma)$  is not an easy task. At first glance, one may naively estimate  $w(\mathbf{D}S_\gamma)$  using its geometric properties. We have

$$w(\mathbf{D}S_\gamma) = w(\text{conv}(\mathbf{D}S_\gamma)) = w(\mathbf{D} \text{conv}(S_\gamma)) \leq w(\mathbf{D}\mathbb{B}_2^{n-1}),$$

where  $\text{conv}(S)$  denotes the convex hull of  $S$ . By Lemma 2.2, we obtain the crude estimate

$$w(\mathbf{D}\mathbb{B}_2^{n-1}) \leq 2\|\mathbf{D}\|_2 w(\mathbb{B}_2^{n-1}) \approx 2\|D\|_2 \sqrt{n}.$$

This is far less than ideal with  $\sqrt{n}$  and the potential dimension dependency from  $\|D\|_2$ .

The approach we will use is inspired by [26, Section 9.4], which estimates  $w(S_\gamma)$ , the width in the basis case. It is worth noting that the generalization  $w(\mathbf{D}S_\gamma)$  is nontrivial, as will be explained in the proof of Theorem 4.7.

First, we introduce the convex cone

$$K_{\gamma,s} := \left\{ \mathbf{u} \in \mathbb{R}^n : \mathbf{u}_\ell \geq 0, \ell \in [n], \sum_{\ell=1}^s \mathbf{u}_\ell \geq \gamma \sum_{\ell=s+1}^n \mathbf{u}_\ell \right\}$$

and its dual cone

$$K_{\gamma,s}^* := \{ \mathbf{z} \in \mathbb{R}^n : \langle \mathbf{z}, \mathbf{u} \rangle \geq 0 \text{ for all } \mathbf{u} \in K_{\gamma,s} \}.$$

Also, recall the *nonincreasing rearrangement* of a vector  $\mathbf{x} \in \mathbb{R}^d$  is the vector  $\mathbf{x}^* \in \mathbb{R}^d$  for which

$$x_1^* \geq x_2^* \geq \cdots \geq x_d^* \geq 0$$

and  $x_i^* = |x_{\pi(i)}|$  for some permutation  $\pi$ .

We can now state the following lemma. It has a proof that is similar to that of [26, Proposition 9.31], but we provide it for the sake of completeness.

**Lemma 4.4.** *If  $\mathbf{D} \in \mathbb{R}^{d \times n}$  is an arbitrary matrix and  $\mathbf{g} \in \mathbb{R}^d \sim N(\mathbf{0}, \mathbf{I}_d)$ , then*

$$w(\mathbf{D}S_\gamma) \leq \mathbb{E} \min_{\mathbf{z} \in K_{\gamma,s}^*} \|(\mathbf{D}^\top \mathbf{g})^* + \mathbf{z}\|_2.$$

**Proof.** Elements in  $S_\gamma$  are invariant under permutation of indices and entrywise sign changes, so

$$\max_{\mathbf{x} \in S_\gamma} \langle \mathbf{D}^\top \mathbf{g}, \mathbf{x} \rangle = \max_{\mathbf{x} \in S_\gamma} \langle (\mathbf{D}^\top \mathbf{g})^*, \mathbf{x} \rangle = \max_{\mathbf{u} \in K_{\gamma,s} \cap \mathbb{S}^{n-1}} \langle (\mathbf{D}^\top \mathbf{g})^*, \mathbf{u} \rangle \leq \min_{\mathbf{z} \in K_{\gamma,s}^*} \|(\mathbf{D}^\top \mathbf{g})^* + \mathbf{z}\|_2, \quad (18)$$

where the last inequality follows by the duality

$$\max_{\mathbf{x} \in K, \|\mathbf{x}\|_2 \leq 1} \langle -\mathbf{g}, \mathbf{x} \rangle \leq \min_{\mathbf{z} \in K^*} \|\mathbf{g} - \mathbf{z}\|_2,$$

as given in [26, (B.39)]. Taking the expected value of (18) completes the proof.  $\square$

The following Lemma is similar to [26, Remark 9.25], but here we assume a general variance.

**Lemma 4.5.** *If  $a \sim N(0, \sigma^2)$ , then*

$$\mathbb{E} S_t^2(a) \leq \sigma^4 \sqrt{\frac{2}{\pi e}} t^{-2} e^{-t^2/(2\sigma^2)},$$

where  $S_t$  is the soft thresholding operator defined by

$$S_t(u) := \begin{cases} u - t & \text{if } u > t \\ 0 & \text{if } |u| \leq t \\ u + t & \text{if } u < -t \end{cases}.$$

**Proof.** We compute

$$\mathbb{E} S_t^2(a) = \frac{2}{\sqrt{2\pi\sigma^2}} \int_0^\infty S_t^2(u) e^{-u^2/(2\sigma^2)} du$$

$$\begin{aligned}
&= \frac{1}{\sigma} \sqrt{\frac{2}{\pi}} \int_t^\infty (u-t)^2 e^{-u^2/(2\sigma^2)} du \\
&= \frac{1}{\sigma} \sqrt{\frac{2}{\pi}} \int_0^\infty v^2 e^{-(v+t)^2/(2\sigma^2)} dv \\
&= e^{-t^2/(2\sigma^2)} \frac{1}{\sigma} \sqrt{\frac{2}{\pi}} \int_0^\infty v^2 e^{-v^2/(2\sigma^2)} e^{-vt/\sigma^2} dv \\
&= e^{-t^2/(2\sigma^2)} \sqrt{\frac{2}{\pi}} \int_0^\infty \left( \frac{v}{\sigma} e^{-(v/\sigma)^2/2} \right) v e^{-vt/\sigma^2} dv \\
&\leq e^{-t^2/(2\sigma^2)} \sqrt{\frac{2}{\pi}} e^{-1/2} \int_0^\infty v e^{-vt/\sigma^2} dv \\
&= e^{-t^2/(2\sigma^2)} \sqrt{\frac{2}{\pi}} e^{-1/2} \frac{\sigma^4}{t^2} \\
&= \sigma^4 \sqrt{\frac{2}{\pi e}} t^{-2} e^{-t^2/(2\sigma^2)}
\end{aligned}$$

as desired.  $\square$

The following Lemma is a key step in suitably bounding the Gaussian width. It is inspired by [35, Lemma 3.5].

**Lemma 4.6.** *If  $\mathbf{D} \in \mathbb{R}^{d \times n}$  is any matrix with  $\max \{ \|\mathbf{d}_i\|_2^2 : i \in [n] \} \leq \rho$  and  $\mathbf{g} \in \mathbb{R}^d \sim N(\mathbf{0}, \mathbf{I}_d)$ , then*

$$\mathbb{E} \sqrt{\frac{1}{s} \sum_{\ell=1}^s ((\mathbf{D}^\top \mathbf{g})_\ell^*)^2} \leq \sqrt{4\rho \log(\sqrt{2}n/s)}.$$

**Proof.** Setting  $a_i := (\mathbf{D}^\top \mathbf{g})_i$  gives  $a_i \sim N(0, \sigma_i^2)$  with  $\sigma_i = \|\mathbf{d}_i\|_2$ , so

$$\mathbb{E} \exp\left(\frac{a_i^2}{4\sigma_i^2}\right) = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{x^2}{2\sigma_i^2}} e^{\frac{x^2}{4\sigma_i^2}} dx = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{x^2}{4\sigma_i^2}} dx = \sqrt{2}.$$

It follows from Jensen's inequality that

$$\begin{aligned}
&\mathbb{E}^2 \sqrt{\frac{1}{s} \sum_{\ell=1}^s ((\mathbf{D}^\top \mathbf{g})_\ell^*)^2} \leq \mathbb{E} \frac{1}{s} \sum_{\ell=1}^s (a_\ell^*)^2 = \mathbb{E} \frac{1}{s} \sum_{\ell=1}^s 4(\sigma_\ell^*)^2 \log \left( \exp \frac{(a_\ell^*)^2}{4(\sigma_\ell^*)^2} \right) \\
&\leq 4\rho \mathbb{E} \frac{1}{s} \sum_{\ell=1}^s \log \left( \exp \frac{(a_\ell^*)^2}{4(\sigma_\ell^*)^2} \right) \leq 4\rho \mathbb{E} \log \left( \frac{1}{s} \sum_{\ell=1}^s \exp \frac{(a_\ell^*)^2}{4(\sigma_\ell^*)^2} \right) \\
&\leq 4\rho \log \left( \frac{1}{s} \sum_{\ell=1}^s \mathbb{E} \exp \frac{(a_\ell^*)^2}{4(\sigma_\ell^*)^2} \right) \leq 4\rho \log \left( \frac{1}{s} \sum_{\ell=1}^n \mathbb{E} \exp \frac{a_\ell^2}{4\sigma_\ell^2} \right) = 4\rho \log \left( \frac{1}{s} \sqrt{2}n \right).
\end{aligned}$$

Taking the squared root of this inequality yields the desired result.  $\square$

**Theorem 4.7.** *If  $\mathbf{D} \in \mathbb{R}^{d \times n}$  is any matrix with  $\max \{\|\mathbf{d}_i\|_2^2 : i \in [n]\} \leq \rho$ , then*

$$w(\mathbf{D}S_\gamma) \leq 6\gamma^{-1} \sqrt{s\rho \log(\sqrt{2}n/s)}.$$

**Proof.** Define

$$F_{\gamma,s} := \bigcup_{t \geq 0} \{\mathbf{z} \in \mathbb{R}^n : z_\ell = t, \ell \in [s], z_k \geq -\gamma t, k = s+1, \dots, n\}$$

which satisfies  $F_{\gamma,s} \subset K_{\gamma,s}^*$  by Lemma 9.32 of [26]. The minimum over a smaller set can only be larger so we obtain  $w(\mathbf{D}S_\gamma) \leq \mathbb{E} \min_{\mathbf{z} \in F_{\gamma,s}} \|(\mathbf{D}^\top \mathbf{g})^* + \mathbf{z}\|_2$  by Lemma 4.4. By the definition of  $F_{\gamma,s}$  and concavity of the square root function, we get

$$\begin{aligned} w(\mathbf{D}S_\gamma) &\leq \mathbb{E} \min_{\mathbf{z} \in F_{\gamma,s}} \|(\mathbf{D}^\top \mathbf{g})^* + \mathbf{z}\|_2 \\ &\leq \mathbb{E} \min_{\substack{t \geq 0 \\ z_\ell \geq -\gamma t, \ell \in [n] \setminus [s]}} \left( \sqrt{\sum_{\ell=1}^s ((\mathbf{D}^\top \mathbf{g})_\ell^* + t)^2} + \sqrt{\sum_{\ell=s+1}^n ((\mathbf{D}^\top \mathbf{g})_\ell^* + z_\ell)^2} \right) \end{aligned}$$

Now fix  $t \geq 0$  to be chosen later. Again by concavity, and since  $(\mathbf{D}^\top \mathbf{g})_\ell^*$  has mean zero we obtain

$$w(\mathbf{D}S_\gamma) \leq \mathbb{E} \sqrt{\sum_{\ell=1}^s ((\mathbf{D}^\top \mathbf{g})_\ell^* + t)^2} + \mathbb{E} \sqrt{\min_{z_\ell \geq -t} \sum_{\ell=s+1}^n ((\mathbf{D}^\top \mathbf{g})_\ell^* + z_\ell)^2} \quad (19)$$

$$\leq \mathbb{E} \sqrt{\sum_{\ell=1}^s ((\mathbf{D}^\top \mathbf{g})_\ell^*)^2} + \sqrt{2t \sum_{\ell=1}^s \mathbb{E}(\mathbf{D}^\top \mathbf{g})_\ell^* + t\sqrt{s}} + \mathbb{E} \sqrt{\min_{z_\ell \geq -t} \sum_{\ell=s+1}^n ((\mathbf{D}^\top \mathbf{g})_\ell^* + z_\ell)^2} \quad (20)$$

$$\leq \mathbb{E} \sqrt{\sum_{\ell=1}^s ((\mathbf{D}^\top \mathbf{g})_\ell^*)^2 + t\sqrt{s}} + \mathbb{E} \sqrt{\sum_{\ell=s+1}^n S_{\gamma t}^2((\mathbf{D}^\top \mathbf{g})_\ell^*)}, \quad (21)$$

where  $S_t$  is the soft thresholding operator defined earlier. Note that Inequality (21) holds because of how the  $z_\ell$  are defined.

We have so far exactly followed the proof of [26, Proposition 9.33], but if one tries to estimate the first term of (21) in the same fashion as in [26, Proposition 8.2], a  $\sqrt{d}$  factor would result, causing the estimate to be too large. Instead, we apply Lemma 4.6, and get

$$\mathbb{E} \sqrt{\sum_{\ell=1}^s ((\mathbf{D}^\top \mathbf{g})_\ell^*)^2} \leq \sqrt{4\rho s \log(\sqrt{2}n/s)}.$$

Next, we bound the last term in (21) by

$$\mathbb{E} \sqrt{\sum_{\ell=s+1}^n S_{\gamma t}^2((\mathbf{D}^\top \mathbf{g})_\ell^*)} \leq \sqrt{\mathbb{E} \sum_{\ell=s+1}^n S_{\gamma t}^2((\mathbf{D}^\top \mathbf{g})_\ell^*)} \leq \sqrt{\mathbb{E} \sum_{\ell=s+1}^n S_{\gamma t}^2((\mathbf{D}^\top \mathbf{g})_\ell)},$$

where we used the fact that  $\{(\mathbf{D}^\top \mathbf{g})_\ell^* : \ell = s+1, \dots, n\}$  are the  $n-s$  smallest entries in magnitude to obtain the second inequality.

To estimate the second moment of the soft thresholding operator, we again use the fact that  $a_i := (\mathbf{D}^\top \mathbf{g})_i$  gives  $a_i \sim N(0, \sigma_i^2)$  with  $\sigma_i = \|\mathbf{d}_i\|_2$  and so Lemma 4.5 implies

$$\mathbb{E} S_{\gamma t}^2(a_i) \leq \sigma_i^4 \sqrt{\frac{2}{\pi e}} (\gamma t)^{-2} \exp\left(\frac{-(\gamma t)^2}{2\sigma_i^2}\right) \leq \rho^2 \sqrt{\frac{2}{\pi e}} (\gamma t)^{-2} \exp\left(\frac{-(\gamma t)^2}{2\rho}\right)$$

Finally, combining all of these estimates of the quantities in (21) gives

$$w(\mathbf{D}S_\gamma) \leq \sqrt{4\rho s \log(\sqrt{2}n/s)} + t\sqrt{s} + \sqrt{(n-s)\rho^2 \sqrt{\frac{2}{\pi e}} (\gamma t)^{-2} \exp\left(\frac{-(\gamma t)^2}{2\rho}\right)}.$$

Choosing  $t = \gamma^{-1} \sqrt{4\rho \log(\sqrt{2}n/s)}$  and using the fact that  $\gamma^{-1} \geq 1$  gives

$$\begin{aligned} w(\mathbf{D}S_\gamma) &\leq \sqrt{4\rho s \log(\sqrt{2}n/s)} + \gamma^{-1} \sqrt{4s\rho \log(\sqrt{2}n/s)} + \sqrt{\rho \frac{(n-s)s^2}{2n^2} \sqrt{\frac{2}{\pi e \log(\sqrt{2}n/s)}}} \\ &\leq \sqrt{4\rho s \log(\sqrt{2}n/s)} (1 + \gamma^{-1}) + \sqrt{\rho s} \\ &\leq 3\gamma^{-1} \sqrt{4\rho s \log(\sqrt{2}n/s)}. \quad \square \end{aligned}$$

#### 4.3. Proof of Lemma 2.2

Recall that a *Gaussian process*  $\{\mathbf{X}_t\}_{t \in T}$  for some index set  $T$  (which can be uncountably infinite) is a sequence of random variables  $\mathbf{X}_t$  so that any finite linear combination follows a Gaussian distribution. Slepian's lemma [28] gives a way to compare such processes.

**Lemma 4.8** (*Slepian's Lemma*). *If  $\{\mathbf{X}_t\}_{t \in T}$  and  $\{\mathbf{Y}_t\}_{t \in T}$  are Gaussian processes so that for any  $s, t \in T$ ,*

$$\mathbb{E} |\mathbf{X}_s - \mathbf{X}_t|^2 \leq \mathbb{E} |\mathbf{Y}_s - \mathbf{Y}_t|^2$$

*holds, then*

$$\mathbb{E} \sup_{t \in T} \mathbf{X}_t \leq \mathbb{E} \sup_{t \in T} \mathbf{Y}_t.$$

**Proof of Lemma 2.2.** We define the Gaussian processes  $\{\mathbf{X}_{\mathbf{u}}\}_{\mathbf{u} \in S}$  and  $\{\mathbf{Y}_{\mathbf{u}}\}_{\mathbf{u} \in S}$

$$\mathbf{X}_{\mathbf{u}} := \langle \mathbf{F}\mathbf{u}, \mathbf{g} \rangle = \|\mathbf{F}\mathbf{u}\|_2 \left\langle \frac{\mathbf{F}\mathbf{u}}{\|\mathbf{F}\mathbf{u}\|_2}, \mathbf{g} \right\rangle \quad \text{and} \quad \mathbf{Y}_{\mathbf{u}} := \|\mathbf{F}\|_2 \langle \mathbf{u}, \mathbf{g} \rangle,$$

where  $g \sim N(\mathbf{0}, \mathbf{I}_d)$ . We notice

$$\mathbb{E} |\mathbf{X}_{\mathbf{u}} - \mathbf{X}_{\mathbf{v}}|^2 = \|\mathbf{F}(\mathbf{u} - \mathbf{v})\|_2^2 \leq \|\mathbf{F}\|_2^2 \|\mathbf{u} - \mathbf{v}\|_2^2 = \mathbb{E} |\mathbf{Y}_{\mathbf{u}} - \mathbf{Y}_{\mathbf{v}}|^2.$$

Thus, Lemma 4.8 gives

$$w(\mathbf{F}S) = \mathbb{E} \sup_{\mathbf{u} \in S} \mathbf{X}_{\mathbf{u}} \leq \mathbb{E} \sup_{\mathbf{u} \in S} \mathbf{Y}_{\mathbf{u}} = \|\mathbf{F}\|_2 w(S). \quad \square$$

#### 4.4. Sparse recovery with stable NSP

**Proof of Theorem 1.1.** Let  $\mathbf{h} := \hat{\mathbf{x}} - \mathbf{x}$  and  $T$  be the support of the biggest  $s$  entries of  $\mathbf{x}$  in magnitude. Then by a standard compressed sensing argument, we have

$$\|\mathbf{h}_{T^c}\|_1 \leq \|\mathbf{h}_T\|_1 + 2\sigma_s(\mathbf{x}), \quad (22)$$

where  $\sigma_s(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}_T\|_1$ .

If  $\mathbf{h}/\|\mathbf{h}\|_2 \in S_\gamma$ , then

$$\eta\|\mathbf{h}\|_2 \leq \|\mathbf{A}\mathbf{h}\|_2 \leq 2\varepsilon. \quad (23)$$

On the other hand, if  $\mathbf{h}/\|\mathbf{h}\|_2 \notin S_\gamma$ , then the vector  $\mathbf{h}$  itself has the stable NSP and therefore  $\|\mathbf{h}_T\|_1 < \gamma\|\mathbf{h}_{T^c}\|_1$ . Combined with (22), we have

$$\|\mathbf{h}_T\|_1 \leq \frac{2\gamma}{1-\gamma}\sigma_s(\mathbf{x}). \quad (24)$$

The equations (22), (24), and the fact that  $\|\mathbf{h}\|_1 = \|\mathbf{h}_T\|_1 + \|\mathbf{h}_{T^c}\|_1$ , we get

$$\|\mathbf{h}\|_2 \leq \|\mathbf{h}\|_1 \leq \frac{2\gamma+2}{1-\gamma}\sigma_s(\mathbf{x}). \quad (25)$$

Combining the two cases (23) and (25), we get the desired result (6).  $\square$

### 5. Estimates for Gaussian distributions

The Gaussian distributions are important special cases of subgaussian distributions, so we list this special cases of Theorem 3.3 below, using the estimates in Example 2.5, as well as Remark 2.7. In Corollary 5.1,  $\kappa := \sigma_{\max}^2/\sigma_{\min}^2$  is the condition number of  $\Sigma$ .

**Corollary 5.1.** Suppose  $\Phi \in \mathbb{R}^{m \times d}$  is a sensing matrix with rows drawn i.i.d. from a Gaussian distribution  $N(\mathbf{0}, \Sigma)$  in which the covariance matrix  $\Sigma$  has condition number  $\kappa$ , and suppose  $\mathbf{D} \in \mathbb{R}^{d \times n}$  has the stable NSP of order  $s$  with  $\inf\{\|\mathbf{D}\mathbf{x}\|_2 : \mathbf{x} \in S_\gamma\} \geq \eta$  for some  $\eta > 0$  and satisfies  $\max\{\|\mathbf{d}_i\|_2^2 : i \in [n]\} \leq \rho$ . If the number of measurements satisfies

$$m \geq \frac{9 \cdot 2^{15} \pi^3}{\eta^2} \frac{\rho \kappa^3}{\gamma^2} s \log\left(\frac{\sqrt{2}n}{s}\right),$$

then with probability at least

$$1 - \exp\left(-m \frac{\kappa^2}{4^5 \pi^2}\right),$$

the composition  $\Phi\mathbf{D}$  also has the stable NSP of order  $s$  with  $\inf\{\|\Phi\mathbf{D}\mathbf{x}\|_2 : \mathbf{x} \in S_\gamma\} \geq \sigma_{\max} w(\mathbf{D}S_\gamma)$ .

The following theorem is an improvement on Corollary 5.1 in terms of the condition number  $\kappa$ . The techniques are the same ones used in proving Theorem 3.3, but for the Gaussian distribution we can improve Lemma 4.1 and hence improve the estimate on the marginal tail  $Q_\xi$ . See (26) below.

**Theorem 5.2.** Suppose  $\Phi \in \mathbb{R}^{m \times d}$  is a sensing matrix with rows drawn i.i.d. from a Gaussian distribution  $N(\mathbf{0}, \Sigma)$  in which the covariance matrix  $\Sigma$  has condition number  $\kappa$ , and suppose  $\mathbf{D} \in \mathbb{R}^{d \times n}$  has the NSP of order  $s$  with  $\inf \{\|\mathbf{D}\mathbf{x}\|_2 : \mathbf{x} \in S_\gamma\} \geq \eta$  for some  $\eta > 0$  and satisfies  $\max \{\|\mathbf{d}_i\|_2^2 : i \in [n]\} \leq \rho$ . If the number of measurements satisfies

$$m \geq \frac{18 \cdot 2^9 \pi e}{\eta^2} \frac{\rho \kappa}{\gamma^2} s \log(2n),$$

then with probability at least

$$1 - \exp\left(-m \frac{1}{128e\pi}\right),$$

the composition  $\Phi\mathbf{D}$  also has the NSP of order  $s$ .

**Proof.** Let  $\varphi \in \mathbb{R}^d \sim N(\mathbf{0}, \Sigma)$  be the vector that generates  $\Phi$ . The random vector  $\varphi$  has the same distribution as  $\Sigma^{1/2}\mathbf{g}$ , where  $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_d)$ . Therefore,

$$\langle \mathbf{D}\mathbf{x}, \varphi \rangle = \langle \mathbf{D}\mathbf{x}, \Sigma^{1/2}\mathbf{g} \rangle = \langle \Sigma^{1/2}\mathbf{D}\mathbf{x}, \mathbf{g} \rangle = \|\Sigma^{1/2}\mathbf{D}\mathbf{x}\|_2 \left\langle \frac{\Sigma^{1/2}\mathbf{D}\mathbf{x}}{\|\Sigma^{1/2}\mathbf{D}\mathbf{x}\|_2}, \mathbf{g} \right\rangle = \|\Sigma^{1/2}\mathbf{D}\mathbf{x}\|_2 Z$$

where  $Z \sim N(0, 1)$ . Therefore, we obtain

$$\begin{aligned} Q_{2\xi}(S_\gamma; \mathbf{D}^\top \varphi) &= \inf_{\mathbf{x} \in S_\gamma} \Pr(|\langle \mathbf{D}\mathbf{x}, \varphi \rangle| \geq 2\xi) \\ &= \inf_{\mathbf{x} \in S_\gamma} \Pr\left(|Z| \geq \frac{2\xi}{\|\Sigma^{1/2}\mathbf{D}\mathbf{x}\|_2}\right) \\ &\geq \Pr\left(|Z| \geq \frac{2\xi}{\eta\sigma_{\min}}\right) \\ &\geq \frac{\eta\sigma_{\min}}{4\xi} \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{2\xi^2}{\eta^2\sigma_{\min}^2}\right) \end{aligned} \quad (26)$$

where we used the well-known lower bound

$$\Pr(Z > t) \geq \frac{1}{2t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}, \quad \text{for any } t \geq 1$$

with the assumption that  $2\xi/(\eta\sigma_{\min}) \geq 1$ . We also have

$$W_m(S_\gamma; \mathbf{D}^\top \varphi) = W_m(\mathbf{D}S_\gamma; \mathbf{D}^\top \varphi) \leq \sigma_{\max} w(\mathbf{D}S_\gamma)$$

by Remark 2.7. Theorem 2.8 hence implies that

$$\begin{aligned} \inf_{\mathbf{x} \in S_\gamma} \|\Phi\mathbf{D}\mathbf{x}\|_2 &\geq \xi\sqrt{m}Q_{2\xi}(S_\gamma; \mathbf{D}^\top \varphi) - 2W_m(S_\gamma; \mathbf{D}^\top \varphi) - \xi t \\ &\geq \sqrt{m} \frac{\eta\sigma_{\min}}{4\sqrt{2\pi}} \exp\left(-\frac{2\xi^2}{\eta^2\sigma_{\min}^2}\right) - 2\sigma_{\max} w(\mathbf{D}S_\gamma) - \xi t \\ &=: a - b - \xi t \end{aligned}$$

with probability  $\geq 1 - e^{-t^2/2}$ . Picking  $m, \xi, t$  such that  $a \geq 2b$  and  $\xi t = (a - b)/2$  gives

$$\inf_{\mathbf{x} \in S} \|\Phi \mathbf{D} \mathbf{x}\|_2 \geq a - b - (a - b)/2 = (a - b)/2 \geq b/2 > 0,$$

guaranteeing the null space property of  $\Phi \mathbf{D}$ .

Lastly, we rewrite these conditions choosing  $2\xi = \eta\sigma_{\min}$  and invoke Theorem 4.7 to get

$$a \geq 2b \Leftrightarrow \sqrt{m} \frac{\eta\sigma_{\min}}{4\sqrt{2\pi}} e^{-1/2} \geq 4\sigma_{\max} w(\mathbf{D}S_\gamma) \Leftrightarrow m \geq \frac{18 \cdot 2^9 \pi e \rho \kappa}{\eta^2 \gamma^2} s \log(2n)$$

and

$$\frac{\eta\sigma_{\min}}{2} t = \frac{a - b}{2} \geq \frac{a}{4} = \sqrt{m} \frac{\eta\sigma_{\min}}{16\sqrt{2\pi}} e^{-1/2} \Leftrightarrow t \geq \frac{\sqrt{m}}{8\sqrt{2\pi}e} \Leftrightarrow -\frac{t^2}{2} \leq -m \frac{1}{128\pi e}$$

as desired.  $\square$

## 6. Conclusions and discussions

This paper generalizes recovery guarantees for  $\ell_1$ -synthesis with dictionary sparsity, which is of interesting in the field of compressed sensing. In particular, the recovery result Corollary 3.6 states that  $O(s \log(n/s))$  subgaussian measurements are enough while only imposing minimal conditions on  $\mathbf{D}$ , if  $\mathbf{D}$  is full spark. As far as the authors can tell, such work is first of its kind. In proving the result, the first step is a rather classical application of the Mendelson's small ball method. The second step, the Gaussian width computation, is the main technical contribution.

Similar recovery results for the basis case  $\mathbf{D} = \mathbf{I}$  was generalized to random distributions far beyond subgaussian [19,33]. It would be interesting to consider more general random matrices such as subexponential distributions. Another related direction of future work in the non-standard Gaussian case is to explore redefining the inner product as  $\mathbf{x}^T \Sigma^{-1/2} \mathbf{y}$ . Similar results as obtained above might hold true in this non-isotropic setting without extensive adjustments to our arguments.

## Acknowledgments

We would like to thank Eric Pinkham for taking part in our initial discussions to get the project moving. We also wish to thank Simon Foucart for his extremely helpful insights and suggestions. We thank the anonymous referees for their extremely helpful comments which improved the presentation of the paper. Finally, we are very gracious for the grant support we received: ARO W911NF-16-1-0008; NSF ATD 1321779; NSF DMS 1307685; NSF DMS 1725455; NSF DMS 1908880; and NSF DMS 1906725.

## References

- [1] B. Alexeev, J. Cahill, D.G. Mixon, Full spark frames, *J. Fourier Anal. Appl.* 18 (6) (2012) 1167–1194.
- [2] A. Aldroubi, X. Chen, A. Powell, Perturbations of measurement matrices and dictionaries in compressed sensing, *Appl. Comput. Harmon. Anal.* 33 (2) (2012) 282–291.
- [3] A. Aldroubi, X. Chen, A. Powell, Stability and robustness of  $\ell_q$  minimization using null space property, in: *Proceedings of SampTA 2011*, 2011.
- [4] J. Bourgain, S.J. Dilworth, K. Ford, S. Konyagin, D. Kutzarova, Explicit constructions of RIP matrices and related problems, *Duke Math. J.* 159 (1) (2011) 145–185.
- [5] J. Cahill, X. Chen, R. Wang, The gap between the null space property and the restricted isometry property, *Linear Algebra Appl.* 501 (2016) 363–375.
- [6] J. Cahill, D.G. Mixon, Robust width: a characterization of uniformly stable and robust compressed sensing, *arXiv:1408.4409*.
- [7] E.J. Candès, Y. Eldar, D. Needell, P. Randall, Compressed sensing with coherent and redundant dictionaries, *Appl. Comput. Harmon. Anal.* 31 (1) (2011) 59–73.
- [8] E.J. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Commun. Pure Appl. Math.* 59 (8) (2006) 1207–1223.

- [9] E.J. Candès, T. Tao, Decoding by linear programming, *IEEE Trans. Inf. Theory* 51 (12) (2005) 4203–4215.
- [10] E.J. Candès, T. Tao, Near optimal signal recovery from random projections: universal encoding strategies?, *IEEE Trans. Inf. Theory* 52 (12) (2006) 5406–5425.
- [11] P.G. Casazza, G. Kutyniok, *Finite Frames: Theory and Applications*, Springer, 2013.
- [12] P.G. Casazza, R.G. Lynch, A Brief Introduction to Hilbert Space Frame Theory and Its Applications. Finite Frame Theory, in: *AMS Short Course Lecture Notes*, in: *Proc. Sympos. Appl. Math.*, vol. 73, Amer. Math. Soc., Providence, RI, 2016, pp. 1–51.
- [13] S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, *SIAM J. Sci. Comput.* 20 (1) (1998) 33–61. Reprinted in *SIAM Rev.* 43 (1) (2001) 129–159.
- [14] X. Chen, H. Wang, R. Wang, A null space analysis of the  $\ell_1$ -synthesis method in dictionary-based compressed sensing, *Appl. Comput. Harmon. Anal.* 37 (3) (2014) 492–515.
- [15] V. Chandrasekaran, B. Recht, P. Parrilo, A. Willsky, The convex geometry of linear inverse problems, *Found. Comput. Math.* 12 (6) (2012) 805–849.
- [16] A. Cohen, W. Dahmen, R. DeVore, Orthogonal matching pursuit under the restricted isometry property, *Constr. Approx.* 45 (1) (2017) 113–127.
- [17] M.A. Davenport, D. Needell, M.B. Wakin, Signal space CoSaMP for sparse recovery with redundant dictionaries, *IEEE Trans. Inf. Theory* 59 (10) (2013) 6820–6829.
- [18] R. DeVore, G. Petrova, P. Wojtaszczyk, Instance-optimality in probability with an  $\ell_1$ -minimization decoder, *Appl. Comput. Harmon. Anal.* 27 (3) (2009) 275–288.
- [19] S. Dirksen, G. Lecué, H. Rauhut, On the gap between restricted isometry properties and sparse recovery conditions, *IEEE Trans. Inf. Theory* 64 (8) (2016) 5478–5487.
- [20] D. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via  $\ell_1$  minimization, *Proc. Natl. Acad. Sci.* 100 (5) (2003) 2197–2202.
- [21] S. Foucart, Hard thresholding pursuit: an algorithm for compressive sensing, *SIAM J. Numer. Anal.* 49 (6) (2011) 2543–2563.
- [22] S. Foucart, Stability and robustness of  $\ell_1$ -minimizations with Weibull matrices and redundant dictionaries, *Linear Algebra Appl.* 441 (2014) 4–21.
- [23] S. Foucart, Dictionary-sparse recovery via thresholding-based algorithms, *J. Fourier Anal. Appl.* 22 (1) (2016) 6–19.
- [24] S. Foucart, Flavors of compressive sensing, in: G. Fasshauer, L. Schumaker (Eds.), *Approximation Theory XV: San Antonio 2016. AT 2016*, in: *Springer Proceedings in Mathematics & Statistics*, vol. 201, Springer, Cham, 2016.
- [25] S. Foucart, M. Lai, Sparsest solutions of underdetermined linear systems via  $\ell_q$ -minimization for  $0 < q \leq 1$ , *Appl. Comput. Harmon. Anal.* 26 (3) (2009) 395–407.
- [26] S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, Birkhäuser, Boston, 2013.
- [27] R. Gribonval, M. Nielsen, Sparse representations in unions of bases, *IEEE Trans. Inf. Theory* 49 (12) (2003) 3320–3325.
- [28] M. Ledoux, M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*, vol. 23, Springer, 1991.
- [29] S. Li, T. Mi, Y. Liu, Performance analysis of  $\ell_1$ -synthesis with coherent frames, in: *2012 IEEE International Symposium on Information Theory Proceedings*, 2012.
- [30] M. Kabanava, H. Rauhut, Analysis  $\ell_1$ -recovery with frames and Gaussian measurements, *Acta Appl. Math.* 140 (2015) 173–195.
- [31] E.J. King, G. Kutyniok, X. Zhuang, Analysis of inpainting via clustered sparsity and microlocal analysis, *J. Math. Imaging Vis.* 48 (2) (2014) 205–234.
- [32] V. Koltchinskii, S. Mendelson, Bounding the smallest singular value of a random matrix without concentration, *Int. Math. Res. Not.* 2015 (23) (2015) 12991–13008.
- [33] G. Lecué, S. Mendelson, Sparse recovery under weak moment assumptions, *J. Eur. Math. Soc.* 19 (3) (2017) 881–904.
- [34] F. Krahmer, D. Needell, R. Ward, Compressive sensing with redundant dictionaries and structured measurements, *SIAM J. Math. Anal.* 47 (6) (2015) 4606–4629.
- [35] S. Mendelson, H. Rauhut, R. Ward, Improved bounds for sparse recovery from subsampled random convolutions, *Ann. Appl. Probab.* 28 (6) (2018) 3491–3527.
- [36] D. Needell, J. Tropp, CoSaMP: iterative signal recovery from incomplete and inaccurate samples, *Appl. Comput. Harmon. Anal.* 26 (3) (2009) 301–321.
- [37] Y. Plan, R. Vershynin, E. Yudovina, High-dimensional estimation with geometric constraints, *Inf. Inference* 6 (1) (2017) 1–40.
- [38] H. Rauhut, K. Schnass, P. Vandergheynst, Compressed sensing and redundant dictionaries, *IEEE Trans. Inf. Theory* 54 (5) (2008) 2210–2219.
- [39] M. Rudelson, R. Vershynin, On sparse reconstruction from Fourier and Gaussian measurements, *Commun. Pure Appl. Math.* 61 (8) (2008) 1025–1045.
- [40] M. Talagrand, *Upper and Lower Bounds for Stochastic Processes: Modern Methods and Classical Problems*, Springer Science & Business Media, 2014.
- [41] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. Ser. B* 58 (1) (1996) 267–288.
- [42] J.A. Tropp, Convex recovery of a structured signal from independent random linear measurements, in: *Sampling Theory, a Renaissance*, in: *Appl. Numer. Harmon. Anal.*, Birkhäuser/Springer, Cham, 2015, pp. 67–101.
- [43] R. Vershynin, Estimation in high dimensions: a geometric perspective, in: *Sampling Theory, a Renaissance*, in: *Appl. Numer. Harmon. Anal.*, Birkhäuser/Springer, Cham, 2015, pp. 3–66.
- [44] M. Wainwright, Statistics meets optimization: randomization and approximation for high-dimensional problems, *Nachdiplom Lecture*, ETH Zurich, 2015, <https://www.stat.berkeley.edu/~wainwrig/nachdiplom/>, Scribed by Y. Wei.
- [45] T. Zhang, Sparse recovery with orthogonal matching pursuit under RIP, *IEEE Trans. Inf. Theory* 57 (9) (2011) 6215–6221.