Course Notes (Paris 2009)

Ronald DeVore

June 14, 2009

Abstract

1 Lecture 1-2: Foundations of Compressed Sensing

1.1 Sparsity

Let X be a Banach space. By a dictionary $\mathcal{D} \subset X$ we mean any set of norm one elements. Let us consider $\Sigma_k := \Sigma_k(\mathcal{D})$ the set of all $S \in X$ such that

$$S = \sum_{g \in \Lambda} c_g g, \quad \#(\Lambda) \leq k \tag{1.1}$$

We say the elements in Σ_k are *k-sparse*. Notice that Σ_k is generally not a linear space: $\Sigma_k + \Sigma_k \neq \Sigma_k$.

In applications, we cannot expect our target function (image/signal/ solution to PDE) to be sparse so we consider how well we can approximate it by k-sparse elements. This is measured by

$$\sigma_k(f) := \sigma_k(f)_X := \inf_{S \in \Sigma_k} \|f - S\|_X, \tag{1.2}$$

which is called *the error of k term approximation* in $\mathbb{R}$. To measure how fast $\sigma_k(f)$ tends to zero we introduce the *primary approximation spaces* $\mathcal{A}^r = \mathcal{A}^r(\mathcal{D}, X)$, $r > 0$, which consists of all f such that

$$\|f\|_{\mathcal{A}^r} := \sup_{k \geq 1} k^r \sigma_k(f) < \infty. \tag{1.3}$$

In some cases, one can characterize the space $\mathcal{A}^r$. To get a feeling for results of this type let us consider the following setting where $\mathcal{D} = (\phi_j)_{j \geq 1}$ is an orthonormal basis for a real Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\| \cdot \| := \| \cdot \|_{\mathcal{H}}$. Each $f \in \mathcal{H}$ has an expansion

$$f = \sum_{j=1}^{\infty} a_j \phi_j, \quad a_j := a_j(f) := \langle f, \phi_j \rangle. \tag{1.4}$$

We say a sequence (b_j) is in weak ℓ_q if

$$\|(b_j)\|_{w\ell_q}^q := \sup_{\epsilon > 0} \epsilon^q \#\{j : \ |b_j| > \epsilon\} < \infty. \tag{1.5}$$

An equivalent definition is that the sequence b_j^* of rearrangements of the absolute values of the b_j into non-increasing order satisfies

$$b_n^* \leq M n^{-1/q}, \quad n \geq 1, \tag{1.6}$$

with the smallest M being the norm in $w\ell_q$.

Theorem 1.1 *A function $f \in \mathcal{H}$ is in $\mathcal{A}^r$, $r > 0$, if and only if $(a_j(f)) \in w\ell_q$ with $1/q = r + 1/2$ with equivalent norms: there exists constants c_1, c_2 such that*

$$c_1 \|f\|_{\mathcal{A}^r} \leq \|(a_j(f))\| w\ell_q \leq c_2 \|f\|_{\mathcal{A}^r}. \tag{1.7}$$

Proof: Suppose $(a_j(f)) \in w\ell_q$, then taking $\epsilon = M^{-1}2^{-j/q}$ with $M := \|(a_j)\|_{w\ell_q}$, we find from the definition of $w\ell_q$ that $\Lambda_j := \Lambda_j(f) := \{i : |a_i| > M2^{-j/q}\}$, has cardinality $\#(\Lambda_j) \leq 2^j$ for each $j \in \mathbb{Z}$. Hence,

$$\sigma_{2^k}^2 \leq \sum_{j>k} \sum_{i \in \Lambda_j \setminus \Lambda_{j-1}} |a_i|^2 \leq \sum_{j>k} 2^j M^2 2^{-2(j-1)/q} \leq 2M^2 \sum_{j \geq k} 2^{-2jr} \leq \frac{2}{1-2^{-2r}} M^2 2^{-2kr},$$

from which the left inequality easily follows.

If $f \in \mathcal{A}^r$, then with (a_n^*) the decreasing rearrangement of $(|a_j|)$ we have

$$n[a_{2n}^*]^2 \leq \sum_{n+1}^{2n} [a_k^*]^2 \leq \sigma_n^2(f) \leq \|f\|_{\mathcal{A}^r}^2 n^{-2r}.$$

A similar inequality holds for a_{2n+1}^* and we derive the right inequality in (1.7). □

Frequently, we want to measure approximation error in non-Hilbertian norms. In this case, one needs further properties of the basis relative to that norm. In classical settings such as for wavelets or Fourier decompositions, this is provided by Littlewood-Paley theory and square functions. For our purposes, it will be enough to consider sequence norms $\ell_p(\Lambda)$ where Λ is a finite or countably infinite set.

If we fix the $\ell_p = \ell_p(\Lambda)$ norm in which approximation error is to be measured, then for any $x \in \mathbb{R}^N$, we have for $q := (r + 1/p)^{-1}$,

$$c_0 \|x\|_{w\ell_q} \leq \|x\|_{\mathcal{A}^r} \leq c_1 r^{-1/p} \|x\|_{w\ell_q}, \quad x \in \mathbb{R}^N, \tag{1.8}$$

for two absolute constants $c_0, c_1 > 0$. This is proved in a similar manner to Theorem 1.1 where the constants in these inequalities do not depend on N. Therefore, $x \in \mathcal{A}^r$ is equivalent to $x \in w\ell_q$ with equivalent norms.

Since the ℓ_q norm is larger than the weak ℓ_q norm, we can replace the weak ℓ_q norm by the ℓ_q norm in the right inequality of (1.8). However, the constant can be improved via a direct argument. Namely, if $1/q = r + 1/p$, then for any $x \in \ell_q$, $q < p$,

$$\sigma_k(x)_{\ell_p} \leq \|x\|_{\ell_q} k^{-r}, \quad k = 1, 2, \ldots. \tag{1.9}$$

To prove this, take Λ' as the set of indices corresponding to the k largest entries in x. If ϵ is the size of the smallest entry in Λ', then $\epsilon \leq \|x\|_{w\ell_q} k^{-1/q} \leq \|x\|_{\ell_q} k^{-1/q}$ and therefore

$$\sigma_k(x)_{\ell_p}^p = \sum_{i \notin \Lambda'} |x_i|^p \leq \epsilon^{p-q} \sum_{i \notin \Lambda'} |x_i|^q \leq k^{-\frac{p-q}{q}} \|x\|_{\ell_q}^{p-q} \|x\|_{\ell_q}^q, \tag{1.10}$$

so that (1.9) follows.

From this, we see that if we consider the class $K = U(\ell_q^N)$, we have

$$\sigma_k(K)_{\ell_p} \leq k^{-r}, \tag{1.11}$$

with $r = 1/q - 1/p$. On the other hand, taking $x \in K$ such that $x_i = (2k)^{-1/q}$ for $2k$ indices and 0 otherwise, we find that

$$\sigma_k(x)_{\ell^p} = [k(2k)^{-p/q}]^{1/p} = 2^{-1/q} k^{-r}, \tag{1.12}$$

so that $\sigma_k(K)_{\ell_p}$ can be framed by

$$2^{-1/q} k^{-r} \leq \sigma_k(K)_{\ell_p} \leq k^{-r}. \tag{1.13}$$

2 Compressed sensing

The typical paradigm for obtaining a compressed version of a discrete signal represented by a vector $x \in \mathbb{R}^N$ is to choose an appropriate basis, compute the coefficients of x in this basis, and then retain only the k largest of these with $k < N$. If we are interested in a bit stream representation, we also need in addition to quantize these k coefficients.

Assuming, without loss of generality, that x already represents the coefficients of the signal in the appropriate basis, this means that we pick an approximation to x from Σ_k. The best performance that we can achieve by such an approximation process in some given norm $\|\cdot\|_X$ of interest is described by $\sigma_k(x)_X$.

The above compression scheme requires us to know all the entries in x. Compressed sensing asks whether we can obtain the same performance with less information about x. To formulate the problem, we are given a budget of n questions we can ask about x. These questions are required to take the form of asking for the values $\lambda_1(x), \dots, \lambda_n(x)$ where the λ_j are fixed linear functionals. The information we gather about x can therefore by described by

$$y = \Phi x, \tag{2.1}$$

where Φ is an $n \times N$ matrix called the *encoder* and $y \in \mathbb{R}^n$ is the *information vector*. The rows of Φ are representations of the linear functionals λ_j, $j = 1, \dots, n$.

To extract the information that y holds about x, we use a *decoder* Δ which is a mapping from $\mathbb{R}^n \to \mathbb{R}^N$. We emphasize that Δ is not required to be linear. Thus, $\Delta(y) = \Delta(\Phi x)$ is our approximation to x from the information we have retained. We shall denote by $\mathcal{A}_{n,N}$ the set of all encoder-decoder pairs (Φ, Δ) with Φ an $n \times N$ matrix.

The most common way of evaluating the performance of an encoding-decoding pair $(\Phi, \Delta) \in \mathcal{A}_{n,N}$ is to ask for the largest value of k such that the encoding-decoding is exact for all k-sparse vectors, i.e.

$$x \in \Sigma_k \Rightarrow \Delta(\Phi x) = x. \tag{2.2}$$

This has an easy solution [10]

Lemma 2.1 *If Φ is any $n \times N$ matrix and k is a positive integer, then the following are equivalent:*
(i) *There is a decoder Δ such that $\Delta(\Phi x) = x$, for all $x \in \Sigma_k$,*
(ii) $\Sigma_{2k} \cap \mathcal{N} = \{0\}$,
(iii) *For any set T with $\#T = 2k$, the matrix Φ_T has rank $2k$.*
(iv) *For any set T with $\#(T) = 2k$, the columns indexed by T are linearly independent.* (v) *The symmetric non-negative matrix $\Phi_T^t \Phi_T$ is invertible, i.e. positive definite.*

Proof: The equivalence of (ii-v) is linear algebra.

(i)$\Rightarrow$(ii): Suppose (i) holds and $x \in \Sigma_{2k} \cap \mathcal{N}$. We can write $x = x_0 - x_1$ where both $x_0, x_1 \in \Sigma_k$. Since $\Phi x_0 = \Phi x_1$, we have, by (i), that $x_0 = x_1$ and hence $x = x_0 - x_1 = 0$.

(ii)$\Rightarrow$(i): Given any $y \in \mathbb{R}^n$, we define $\Delta(y)$ to be any element in $\mathcal{F}(y)$ with smallest support. Now, if $x_1, x_2 \in \Sigma_k$ with $\Phi u = \Phi d$, then $x_1 - x_2 \in \mathcal{N} \cap \Sigma_{2k}$. From (ii), this means that $x_1 = x_2$. Hence, if $x \in \Sigma_k$ then $\Delta(\Phi x) = x$ as desired. $\Box$

It is easy to construct examples of matrices of size $n \times N$ with $n = 2k$ which satisfy the requirements of the Lemma. For example, if $0 < x_1 < \cdots < x_N = 1$, then the matrix $\Phi = (x_i^j)_{0 \le i \le n; 1 \le j \le N}$ works. Thus $2k$ measurements suffice to recover every k sparse vectors.

2.1 Instance optimality

We would like to measure the performance of a compressed sensing scheme (Δ, Φ) in a more robust way so that it includes all vectors x. Accordingly, we given the following definition:
We say that (Φ, Δ) is instance optimal in $\|\cdot\|_X$ of order k with constant C_0 if

$$\|x - \Delta(\Phi x)\|_X \le C_0 \sigma_k(x)_X, \tag{2.3}$$

holds for all $x \in \mathbb{R}^N$.

Notice that if we have instance optimality of order k for some norm then any k sparse vector x is captured exactly since $\sigma_k(x)_X = 0$. We shall see that the range of k for which instance optimality holds strongly depends on the norm X under consideration.

We have already seen in Lemma 2.1 that the performance of a matrix Φ in compressed sensing is determined by the null space

$$\mathcal{N} = \mathcal{N}(\Phi) := \{x \in \mathbb{R}^N : \Phi x = 0\}. \tag{2.4}$$

The importance of $\mathcal{N}$ is that if we observe $y = \Phi x$ without any a-priori information on x, the set of z such that $\Phi z = y$ is given by the affine space

$$\mathcal{F}(y) := x + \mathcal{N}. \tag{2.5}$$

The following result from [10] shows how the null space determines whether or not we have instance optimality.

Theorem 2.2 *Given an $n \times N$ matrix Φ, a norm $\|\cdot\|_X$ and a value of k, then a sufficient condition that there exists a decoder Δ such that (2.3) holds with constant C_0 is that*

$$\|\eta\|_X \leq \frac{C_0}{2}\sigma_{2k}(\eta)_X, \quad \eta \in \mathcal{N}. \tag{2.6}$$

A necessary condition is that

$$\|\eta\|_X \leq C_0\sigma_{2k}(\eta)_X, \quad \eta \in \mathcal{N}. \tag{2.7}$$

Proof: To prove the sufficiency of (2.6), we will define a decoder Δ for Φ as follows. Given any $y \in \mathbb{R}^N$, we consider the set $\mathcal{F}(y)$ and choose

$$\Delta(y) := \operatorname*{argmin}_{z\in\mathcal{F}(y)} \sigma_k(z)_X. \tag{2.8}$$

We shall prove that for all $x \in \mathbb{R}^N$

$$\|x - \Delta(\Phi x)\|_X \leq C_0\sigma_k(x)_X. \tag{2.9}$$

Indeed, $\eta := x - \Delta(\Phi x)$ is in $\mathcal{N}$ and hence by (2.6), we have

$$\begin{aligned}\|x - \Delta(\Phi x)\|_X &\leq (C_0/2)\sigma_{2k}(x - \Delta(\Phi x))_X \\ &\leq (C_0/2)(\sigma_k(x)_X + \sigma_k(\Delta(\Phi x)_X) \\ &\leq C_0\sigma_k(x)_X,\end{aligned}$$

where the second inequality uses the fact that $\sigma_{2k}(x+z)_X \leq \sigma_k(x)_X + \sigma_k(z)_X$ and the last inequality uses the fact that $\Delta(\Phi x)$ minimizes $\sigma_k(z)$ over $\mathcal{F}(y)$.

To prove the necessity of (2.7), let Δ be any decoder for which (2.3) holds. Let η be any element in $\mathcal{N} = \mathcal{N}(\Phi)$ and let η_0 be the best $2k$-term approximation of η in X. Let $\eta_0 = \eta_1 + \eta_2$ be any splitting of η_0 into two vectors of support size k, we can write

$$\eta = \eta_1 + \eta_2 + \eta_3, \tag{2.10}$$

with $\eta_3 = \eta - \eta_0$. Since $-\eta_1 \in \Sigma_k$ we have by (2.3) that $-\eta_1 = \Delta(\Phi(-\eta_1))$, but since $\eta \in \mathcal{N}$, we also have $-\Phi\eta = \Phi(\eta_2 + \eta_3)$ so that $-\eta_1 = \Delta(\Phi(\eta_2 + \eta_3))$. Using again (2.3) we derive

$$\begin{aligned}\|\eta\|_X &= \|\eta_2 + \eta_3 - \Delta(\Phi(\eta_2 + \eta_3))\|_X \leq C_0\sigma_k(\eta_2 + \eta_3) \\ &\leq C_0\|\eta_3\|_X = C_0\sigma_{2k}(\eta),\end{aligned}$$

which is (2.7). □

When X is an ℓ_p space, the best k term approximation is obtained by leaving the k largest components of x unchanged and setting all the others to 0. Therefore the property

$$\|\eta\|_X \leq C\sigma_k(\eta)_X, \tag{2.11}$$

can be reformulated by saying that

$$\|\eta\|_X \leq C\|\eta_{T^c}\|_X, \tag{2.12}$$

holds for all $T \subset \{1, \cdots, N\}$ such that $\#T \leq k$, where T^c is the complement set of T in $\{1, \cdots, N\}$. In going further, we shall say that Φ has the *null space property* in X of order k with constant C if (2.12) holds for all $\eta \in \mathcal{N}$ and $\#T \leq k$. Thus, we have

Corollary 2.3 *Suppose that X is an ℓ_p^N space, $k > 0$ an integer and Φ an encoding matrix. If Φ has the null space property* (2.12) *in X of order $2k$ with constant $C_0/2$, then there exists a decoder Δ so that (Φ, Δ) satisfies* (2.3) *with constant C_0. Conversely, the validity of* (2.3) *for some decoder Δ implies that Φ has the null space property* (2.12) *in X of order $2k$ with constant C_0.*

3 Gelfand widths: bounds for the range of k

Given a norm $\|\cdot\|_X$ in which we wish to measure error, we would like to know the largest range of k for which we can obtain instance optimality and then understand which schemes (Φ, Δ) achieve this range. We shall bound k by considering the performance of compressed sensing systems on compact sets K and showing this is related to certain well-known n widths.

Given K and X, we define

$$E_n(K)_X := \inf_{(\Phi,\Delta)\in\mathcal{A}_{n,N}} \sup_{x\in K} \|x - \Delta(\Phi x)\|_X, \tag{3.1}$$

which is a measure of the performance of the best compressed sensing systems on the set K.

We shall show that $E_n(K)_X$ is equivalent to the following Gelfand width:

$$d^n(K)_X := \inf_Y \sup\{\|x\|_X \; ; \; x \in K \cap Y\}, \quad n = 1, 2, \ldots, \tag{3.2}$$

where the infimum is taken over all subspaces Y of X of codimension less or equal to n.

Lemma 3.1 *Let $K \subset \mathbb{R}^N$ be any set for which $K = -K$ and for which there is a $C_0 > 0$ such that $K + K \subset C_0 K$. If $X \subset \mathbb{R}^N$ is any normed space, then*

$$d^n(K)_X \leq E_n(K)_X \leq C_0 d^n(K)_X, \quad 1 \leq n \leq N. \tag{3.3}$$

Proof: The proof will again bring out the role of the null space of Φ in the performance of Φ. Indeed, this null space $Y = \mathcal{N}$ of Φ is of codimension less or equal to n. Conversely, given any space $Y \subset \mathbb{R}^N$ of codimension n, we can associate its orthogonal complement $Y^\perp$ which is of dimension n and the $n \times N$ matrix Φ whose rows are formed by any basis for $Y^\perp$. Through this identification, we see that

$$d^n(K)_X = \inf_\Phi \sup\{\|\eta\|_X : \eta \in \mathcal{N} \cap K\}, \tag{3.4}$$

where the infimum is taken over all $n \times N$ matrices Φ.

Now, if (Φ, Δ) is any encoder-decoder pair and $z = \Delta(0)$, then for any $\eta \in \mathcal{N}$, we also have $-\eta \in \mathcal{N}$. It follows that either $\|\eta - z\|_X \geq \|\eta\|_X$ or $\| - \eta - z\|_X \geq \|\eta\|_X$. Since $K = -K$ we conclude that

$$d^n(K)_X \leq \sup_{\eta \in \mathcal{N} \cap K} \|\eta - \Delta(\Phi\eta))\|_X. \tag{3.5}$$

Taking an infimum over all encoder-decoder pairs in $\mathcal{A}_{n,N}$, we obtain the left inequality in (3.3).

To prove the right inequality, we choose an optimal Y for $d^n(K)_X$ and use the matrix Φ associated to Y (i.e., the rows of Φ are a basis for $Y^\perp$). We define a decoder Δ for Φ as follows. Given y in the range of Φ, we recall that $\mathcal{F}(y)$ is the set of x such that $\Phi x = y$. If $\mathcal{F}(y) \cap K \neq \emptyset$, we take any $\bar{x}(y) \in \mathcal{F}(y) \cap K$ and define $\Delta(y) := \bar{x}(y)$. When $\mathcal{F}(y) \cap K = \emptyset$, we define $\Delta(y)$ as any element from $\mathcal{F}(y)$. This gives

$$E_n(K)_X \leq \sup_{x, x' \in \mathcal{F}(y) \cap K} \|x - x'\|_X \leq \sup_{\eta \in C_0[K \cap \mathcal{N}]} \|\eta\|_X \leq C_0 d^n(K)_X, \tag{3.6}$$

where we have used the fact that $x - x' \in \mathcal{N}$ and $x - x' \in C_0 K$ by our assumptions on K. This proves the right inequality in (3.3). $\square$

The orders of the Gelfand widths of ℓ_q balls in ℓ_p are known. Historically, the most famous of these results is the following

$$c_0 \min\left\{1, \sqrt{\frac{\log(N/n)}{n}}\right\} \leq d^n(U(\ell_1^N))_{\ell_2^N} = E_n(U(\ell_1^N))_{\ell_2^N} \leq c_1 \min\left\{1, \sqrt{\frac{\log(N/n)}{n}}\right\}. \tag{3.7}$$

The upper bound in (3.7) was first proved by Kashin [17] with a slightly worse power of the logarithm. The above form was given by Gluskin and Garneev [14]. These results could be thought of as the start of compressed sensing. We will have more to say on this in a moment. For now let us mention another result (which can be proved using the techniques in Chapter 13 of [19]). For any $0 < q < 1$,

$$c_0 \Big[\min\left\{1, \frac{\log(N/n)}{n}\right\}\Big]^{1/q-1} \leq E_n(U(\ell_q^N))_{\ell_1^N} \leq c_1 \Big[\min\left\{1, \frac{\log(N/n)}{n}\right\}\Big]^{1/q-1}. \tag{3.8}$$

Let us see how we can use this last result to give a bound on the optimal range of k for which instance optimality can hold. Suppose that we have an $n \times N$ matrix which gives ℓ_1^N instance optimality for some C_0 and k. For any vector in $U(\ell_q^N)$ we know from (1.13) that $\sigma_k(x) \leq \|x\|_{\ell_q^N} k^{-1/q-1}$. It follows that if we have instance optimality of order k for some sensing system of size $n \times N$, then $E_n(U(\ell_q^N))_{\ell_1^N} \leq C_0 k^{-1/q+1}$. Applying (3.8) gives

$$c_0 \Big[\frac{\log(N/n)}{n}\Big]^{1/q-1} \leq E_n(U(\ell_q^N))_{\ell_1^N} \leq C_0 k^{-1/q+1}. \tag{3.9}$$

This means that $k \leq \frac{Cn}{\log(N/n)}$ with $C = (\frac{C_0}{c_0})^{1/q-1}$. Thus, this is the largest range of k for which we can have ℓ_1 instance optimality. Similar bounds can be established for instance optimality in other space $X = \ell_p^N$ and will be discussed shortly. For now we set out to see if we can find matrices that give instance optimality for this range of k.

4 Constructing good matrices

Now that we know the largest range of k possible in various settings of compressed sensing, we set out to see if we can construct matrices with this range of performance. All constructions of good CS matrices Φ are probabilistic.

We shall limit ourselves to random matrices of the following form (other possibilities can also be treated). We suppose that $\Phi = \Phi(\omega)$, $\omega \in \Omega$, is a family of random $n \times N$ matrices whose entries are given by independent realizations of a fixed symmetric random variable η defined on a probability space (Ω, ρ) with expectation $\mathbb{E}\eta = 0$ and variance $\mathbb{E}\eta^2 = 1/n$. The columns Φ_j, $j = 1, \dots, N$, of Φ will be vectors in $\mathbb{R}^n$ with $\mathbb{E}\|\Phi_j\|_{\ell_2^n} = 1$.

We shall show that under rather mild conditions on η, the matrices $\Phi(\omega)$ will be good matrices with very high probability. Indeed, it will be enough to assume that $r := \sqrt{n}\eta$ is sub-Gaussian, i.e.

$$\Pr\{|\eta| > \delta\} \leq C_o e^{-c_0 \delta^2}, \quad \delta > 0. \tag{4.1}$$

Two simple instances of random matrices which are often considered in compressed sensing are

(i) **Gaussian matrices:** $\Phi_{i,j} = \mathcal{N}(0, \frac{1}{n})$ are i.i.d. Gaussian variables of variance $1/n$.

(ii) **Bernoulli matrices:** $\Phi_{i,j} = \frac{\pm 1}{\sqrt{n}}$ are i.i.d. Bernoulli variables of variance $1/n$.

To understand the performance of the random matrices $\Phi(\omega)$ generated by such a choice η, we first examine the mapping properties of Φ. From the sub-Gaussian property one deduces:

Concentration of Measure Property (CMP) : *For any $x \in \mathbb{R}^N$ and any $0 < \delta < 1$, there is a set $\Omega_0(x, \delta)$ with*

$$\rho(\Omega_0(x, \delta)^c) \leq C_0 e^{-n c_0(\delta)}, \tag{4.2}$$

such that for each $\omega \in \Omega_0(x, \delta)$ we have

$$(1 - \delta)\|x\|_{\ell_2^N}^2 \leq \|\Phi(\omega)x\|_{\ell_2^n}^2 \leq (1 + \delta)\|x\|_{\ell_2^N}^2. \tag{4.3}$$

Lemma 4.1 *Let r be a zero mean random variable that satisfies* (4.1). *Then, the $n \times N$ random family $\Phi(\omega)$, whose entries $\phi_{i,j}$ are independent realizations of $\eta = \frac{1}{\sqrt{n}} r$ satisfies the CMP for all n and N.*

Proof: For a not too difficult proof of this fact see [12]. $\square$

For specific random variables such as Gaussian or Bernoulli random variables, there are several proofs in the literature of CMP. For example, it is proved in [1] that CMP holds with $c_0(\delta) = \delta^2/4 - \delta^3/6$ and $C_0 = 2$ for Bernoulli random variables.

There are several important consequences that can be drawn from the CMP. For us, the most important example is the Restricted Isometry Property (RIP) as introduced by Candés, Romberg, and Tao [5] which examines the mapping properties of Φ on Σ_k.

Restricted Isometry Property (RIP): *An $n \times N$ matrix A is said to have RIP of order k with constant δ if*

$$(1 - \delta)\|z\|_{\ell_2^N} \leq \|Az\|_{\ell_2^n} \leq (1 + \delta)\|z\|_{\ell_2^N}, \quad \forall z \in \Sigma_k. \tag{4.4}$$

We shall now show that random matrices with CMP will satisfy RIP for the large range of k.

Theorem 4.2 *Any random family of $n \times N$ matrices which satisfies CMP will automatically satisfy the RIP of order k and constant δ for any $k \leq c(\delta)n/\log(N/n)$ with probability $\geq 1-e^{-c_2 n}$ where c and c_2 depend only on δ.*

For the proof of this theorem we follow [3]. For any index set $T \subset \{1,\ldots,N\}$, let X_T be the linear space of all vectors in $\mathbb{R}^N$ which are supported on T.

Lemma 4.3 *Let $\Phi(\omega)$, $\omega \in \Omega$, satisfies* **CMP**. *Then, for any set T with $\#(T) = k < n$ and any $0 < \delta < 1$, we have*

$$(1-\delta)\|x\|_{\ell_2^N} \leq \|\Phi(\omega)x\|_{\ell_2^n} \leq (1+\delta)\|x\|_{\ell_2^N}, \quad \textit{for all} \;\; x \in X_T, \tag{4.5}$$

with probability

$$\geq 1 - 2(12/\delta)^k\, e^{-c_0(\delta/2)n}. \tag{4.6}$$

Proof: First note that it is enough to prove (4.5) in the case $\|x\|_{\ell_2^N} = 1$, since Φ is linear. Next, we choose a finite set of points Q_T such that $Q_T \subseteq X_T$, $\|q\|_{\ell_2^N} \leq 1$ for all $q \in Q_T$, and for all $x \in X_T$ with $\|x\|_{\ell_2^N} \leq 1$ we have

$$\min_{q \in Q_T} \|x - q\|_{\ell_2^N} \leq \delta/4. \tag{4.7}$$

It is well known from covering numbers and easy to prove (see e.g. Chapter 13 of [19]) that we can choose such a set Q_T with $\#(Q_T) \leq (12/\delta)^k$. We next use **CMP** with $\delta/2$, with the result that, with probability exceeding the right side of (4.6), we have

$$(1-\delta/2)\|q\|_{\ell_2^N}^2 \leq \|\Phi q\|_{\ell_2^n}^2 \leq (1+\delta/2)\|q\|_{\ell_2^N}^2, \quad \text{for all } q \in Q_T, \tag{4.8}$$

which trivially gives us

$$(1-\delta/2)\|q\|_{\ell_2^N} \leq \|\Phi q\|_{\ell_2^n} \leq (1+\delta/2)\|q\|_{\ell_2^N}, \quad \text{for all } q \in Q_T. \tag{4.9}$$

We now define A as the smallest number such that

$$\|\Phi x\|_{\ell_2^n} \leq (1+A)\|x\|_{\ell_2^N}, \quad \text{for all } x \in X_T,\ \|x\|_{\ell_2^N} \leq 1. \tag{4.10}$$

Our goal is to show that $A \leq \delta$. For this, we recall that for any $x \in X_T$ with $\|x\|_{\ell_2^N} \leq 1$, we can pick a $q \in Q_T$ such that $\|x-q\|_{\ell_2^N} \leq \delta/4$. In this case we have

$$\|\Phi x\|_{\ell_2^n} \leq \|\Phi q\|_{\ell_2^n} + \|\Phi(x-q)\|_{\ell_2^n} \leq 1 + \delta/2 + (1+A)\delta/4. \tag{4.11}$$

Since by definition A is the smallest number for which (4.10) holds, we obtain $A \leq \delta/2 + (1+A)\delta/4$. Therefore $A \leq \frac{3\delta/4}{1-\delta/4} \leq \delta$, as desired. We have proved the upper inequality in (4.5). The lower inequality follows from this since

$$\|\Phi x\|_{\ell_2^n} \geq \|\Phi q\|_{\ell_2^n} - \|\Phi(x-q)\|_{\ell_2^n} \geq 1 - \delta/2 - (1+\delta)\delta/4 \geq 1 - \delta, \tag{4.12}$$

which completes the proof. $\square$

Proof of Theorem : We know that for each of the k dimensional spaces X_T, the matrix $\Phi(\omega)$ will fail to satisfy (4.5) with probability

$$\leq 2(12/\delta)^k e^{-c_0(\delta/2)n}. \tag{4.13}$$

There are $\binom{N}{k} \leq (eN/k)^k$ such subspaces. Hence, the RIP will fail to hold with probability

$$\leq 2(eN/k)^k(12/\delta)^k e^{-c_0(\delta/2)n} = e^{-c_0(\delta/2)n+k[\log(eN/k)+\log(12/\delta)]+\log(2)}. \tag{4.14}$$

Thus, for a fixed $c_1 > 0$, whenever $k \leq c_1 n/\log(N/k)$, we will have that the exponent in the exponential on the right side of (4.14) is $\leq -c_2 n$ provided that $c_2 > c_0(\delta/2) - c_1[1 + (1 + \log(12/\delta))/\log(N/k)]$. Hence, we can always choose $c_1 > 0$ sufficiently small to ensure that $c_2 > 0$. This proves the theorem. From the validity of the theorem for the range of $k \leq c_1 n/\log(N/k)$, one can easily deduce its validity for $k \leq c_1' n/[\log(N/n) + 1]$ for $c_1' > 0$ depending only on c_1. $\square$

Remarks: The above theorem holds for any random family satisfying **CMP** not necessarily generated by draws of a single random variable η. For the matrices generate by a single random variable, we have shown that if $r := \sqrt{n}\eta$ is subgaussian then it has the **CMP**. Therefore, **SG**$\rightarrow$**CMP** $\rightarrow$ **RIP**. Much more is known about RIP. Two papers to look at are Rudelson and Vershynin [21] which treats RIP for Fourier matrices where there are still fundamental open questions and Adamczak, Litvak, Pajor, Tomczack-Jaegermann [2] which shows that weaker assumptions than **SG** suffice for **RIP**

5 Verifying instance optimality

We have claimed that matrices which satisfy **CMP** are good matrices for compressed sensing. To illustrate this fact, we shall now show that they satisfy instance optimality in ℓ_1^N for the largest range of k. The following lemma is proved using the method of Candés and Tao[6].

Lemma 5.1 *Let $a = \ell/k$, $b = \ell'/k$ with $\ell, \ell' \geq k$ integers. If Φ is any matrix which satisfies the RIP of order $(a+b)k$ with $\delta = \delta_{(a+b)k} < 1$. Then Φ satisfies the null space property in ℓ_1 of order ak with constant $C_0 = 1 + \frac{\sqrt{a}(1+\delta)}{\sqrt{b}(1-\delta)}$.*

Proof: It is enough to prove (2.12) in the case when T is the set of indices of the largest ak entries of η. Let $T_0 = T$, T_1 denote the set of indices of the next bk largest entries of η, T_2 the next bk largest, and so on. The last set T_s defined this way may have less than bk elements.

We define $\eta_0 := \eta_{T_0} + \eta_{T_1}$. Since $\eta \in \mathcal{N}$, we have $\Phi\eta_0 = -\Phi(\eta_{T_2} + \ldots + \eta_{T_s})$, so that

$$\begin{aligned} \|\eta_T\|_{\ell_2} &\leq \|\eta_0\|_{\ell_2} \leq (1-\delta)^{-1}\|\Phi\eta_0\|_{\ell_2} = (1-\delta)^{-1}\|\Phi(\eta_{T_2} + \ldots + \eta_{T_s})\|_{\ell_2} \\ &\leq (1-\delta)^{-1}\sum_{j=2}^{s}\|\Phi\eta_{T_j}\|_{\ell_2} \leq (1+\delta)(1-\delta)^{-1}\sum_{j=2}^{s}\|\eta_{T_j}\|_{\ell_2}, \end{aligned}$$

where we have used the **RIP** repeatedly. Now for any $i \in T_{j+1}$ and $i' \in T_j$, we have $|\eta_i| \leq |\eta_{i'}|$ so that $|\eta_i| \leq (bk)^{-1}\|\eta_{T_j}\|_{\ell_1}$. It follows that

$$\|\eta_{T_{j+1}}\|_{\ell_2} \leq (bk)^{-1/2}\|\eta_{T_j}\|_{\ell_1}, \quad j = 1, 2, \ldots, s-1, \tag{5.1}$$

so that

$$\|\eta_T\|_{\ell_2} \leq (1+\delta)(1-\delta)^{-1}(bk)^{-1/2}\sum_{j=1}^{s-1}\|\eta_{T_j}\|_{\ell_1} \leq (1+\delta)(1-\delta)^{-1}(bk)^{-1/2}\|\eta_{T^c}\|_{\ell_1}. \tag{5.2}$$

By the Cauchy-Schwartz inequality $\|\eta_T\|_{\ell_1} \leq (ak)^{1/2}\|\eta_T\|_{\ell_2}$, and we therefore obtain

$$\|\eta\|_{\ell_1} = \|\eta_T\|_{\ell_1} + \|\eta_{T^c}\|_{\ell_1} \leq (1 + \frac{\sqrt{a}(1+\delta)}{\sqrt{b}(1-\delta)})\|\eta_{T^c}\|_{\ell_1} \tag{5.3}$$

which verifies the null space property with the constant C_0. □

Since we know the null space property is sufficient for instance optimality, we have proved the following.

Theorem 5.2 *Let Φ be any matrix which satisfies the RIP of order $3k$. Define the decoder Δ for Φ as in (8.18) for $X = \ell_1$. Then (2.3) holds in $X = \ell_1$ with constant $C_0 = 2(1 + \sqrt{2}\frac{1+\delta}{1-\delta})$.*

Remarks: Candés [4] has shown that $3k$ can be replaced by $2k$ in the above theorem. There are also some papers trying to understand the weakest assumption on δ.

Let us also note that the same arguments as given above give the following *mixed norm instance optimality*

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq Ck^{-1/2}\sigma_k(f)_{\ell_1^N}, \tag{5.4}$$

which holds for any matrix satisfying RIP of order $3k$ and an appropriate decoder Δ.

6 Instance optimality in ℓ_2

The reader may be curious as to why we concentrated on instance optimality in ℓ_1^N and not in the more natural space ℓ_2^N. The reason is that instance optimality fails miserably in ℓ_2^N. The reason for this is that any properly normalized $n \times N$ compressed sensing matrix Φ with $n << N$ will necessarily have large norm on ℓ_2^N. Here is one particular way to fetter this out [10].

Theorem 6.1 *Any $n \times N$ matrix Φ of which satisfies instance optimality with $k = 1$ necessarily has $N \leq C_0^2 n$.*

Proof: We know that a necessary and sufficient condition for instance optimality is the null space property. So for any vector η in the null space of Φ, we have

$$\|\eta\|_{\ell_2}^2 \leq C_0^2\|\eta_{T^c}\|_{\ell_2}^2, \quad \#T \leq 1, \tag{6.1}$$

or equivalently for all $j \in \{1, \cdots, N\}$,

$$\sum_{i=1}^{N}|\eta_i|^2 \leq C_0^2\sum_{i\neq j}|\eta_i|^2. \tag{6.2}$$

From this, we derive that for all $j \in \{1, \cdots, N\}$,

$$|\eta_j|^2 \leq (C_0^2 - 1) \sum_{i \neq j} |\eta_i|^2 = (C_0^2 - 1)(\|\eta\|_{\ell_2}^2 - |\eta_j|^2), \tag{6.3}$$

and therefore

$$|\eta_j|^2 \leq A\|\eta\|_{\ell_2}^2, \tag{6.4}$$

with $A = 1 - \frac{1}{C_0^2}$.

Let $(e_j)_{j=1,\cdots,N}$ be the canonical basis of $\mathbb{R}^N$ so that $\eta_j = \langle \eta, e_j \rangle$ and let $v_1, \ldots, v_{N-n}$ be an orthonormal basis for $\mathcal{N}$. Denoting by $P = P_{\mathcal{N}}$ the orthognal projection onto $\mathcal{N}$, we apply (6.4) to $\eta := P(e_j) \in \mathcal{N}$ and find that for any $j \in \{1, \ldots, N\}$

$$|\langle P(e_j), e_j \rangle|^2 \leq A. \tag{6.5}$$

This means

$$\sum_{i=1}^{N-n} |\langle e_j, v_i \rangle|^2 \leq A, \quad j = 1, \ldots, N. \tag{6.6}$$

We sum (6.6) over $j \in \{1, \ldots, N\}$ and find

$$N - n = \sum_{i=1}^{N-n} \|v_i\|_{\ell_2}^2 \leq AN. \tag{6.7}$$

It follows that $(1 - A)N \leq n$. That is, $N \leq nC_0^2$ as desired. $\square$

7 Instance optimality in probability

While it is disturbing that instance optimality does not hold in ℓ_2^N, the situation is not so bleak if we rethink what we are doing. To obtain instance optimality for the large range of k for ℓ_1, we need to use probabilistic constructions since there are no known deterministic constructions. On the other hand, even if we had one of the favorable random matrices we would not be able to verify it since the RIP property cannot be checked in any reasonable computational time. Hence ultimately we are in a situation where we draw a matrix at random and know only that it will work with high probability. Then why not evaluate performance in this probabilistic setting as well?

So let us embed ourselves into the following setting. We let Ω be a probability space with probability measure ρ and let $\Phi = \Phi(\omega)$, $\omega \in \Omega$ be an $n \times N$ random matrix. To keep matters simple, let us assume that the entries of Φ are generated by independent draws of a random variable as we have previously considered. We seek results of the following type:

Instance Optimality in Probability: *for any $x \in \mathbb{R}^N$, if we draw Φ at random with respect to P, then*

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq C_0 \sigma_k(x)_{\ell_2} \tag{7.1}$$

holds for this particular x with high probability for some decoder Δ (dependent on the draw Φ).

It should be understood that Φ is drawn independently for each x in contrast to building a Φ such that (7.1) holds simultaneously for all $x \in \mathbb{R}^N$ which was our original definition of instance optimality.

We now describe our process for decoding $y = \Phi x$, when $\Phi = \Phi(\omega)$ is our given realization of the random matrix. (This method is numerically impractical but will be sufficient for theoretical results. Later we shall turn to more practical decoders.) Let $T \subset \{1, \dots, N\}$ be any subset of column indices with $\#(T) = k$ and let X_T be the linear subspace of $\mathbb{R}^N$ which consists of all vectors supported on T. For this T, we define

$$x_T^* := \operatorname*{argmin}_{z \in X_T} \|\Phi z - y\|_{\ell_2}. \tag{7.2}$$

In other words, x_T^* is chosen as the least squares minimizer of the residual in approximation by elements of X_T. Notice that x_T^* is supported on T. If Φ satisfies RIP of order k then the matrix $\Phi_T^t \Phi_T$ is nonsingular and the nonzero entries of x_T^* are given by

$$(\Phi_T^t \Phi_T)^{-1} \Phi_T^t y. \tag{7.3}$$

To decode y, we search over all subsets T of cardinality k and choose

$$T^* := \operatorname*{argmin}_{\#(T)=k} \|y - \Phi x_T^*\|_{\ell_2^n}. \tag{7.4}$$

Our decoding of y is now given by

$$x^* = \Delta(y) := x_{T^*}^*. \tag{7.5}$$

Theorem 7.1 [10] *Assume that Φ is a random matrix which satisfies RIP of order $2k$ and also satisfies CMP each with probability $1-\epsilon$. Then, for each $x \in \mathbb{R}^N$, the estimate (7.1) holds with $C_0 = 1 + \frac{2C}{1-\delta}$ and probability $1 - 2\epsilon$.*

Proof: Let $x \in \mathbb{R}^N$ be arbitrary and let $\Phi = \Phi(\omega)$ be the draw of the matrix Φ from the random ensemble. We denote by T the set of indices corresponding to the k largest entries of x. Thus

$$\|x - x_T\|_{\ell_2} = \sigma_k(x)_{\ell_2}. \tag{7.6}$$

Then,

$$\|x - x^*\|_{\ell_2} \le \|x - x_T\|_{\ell_2} + \|x_T - x^*\|_{\ell_2} \le \sigma_k(x)_{\ell_2} + \|x_T - x^*\|_{\ell_2}. \tag{7.7}$$

We bound the second term by

$$\begin{aligned}
\|x_T - x^*\|_{\ell_T^N} &\le (1-\delta)^{-1} \|\Phi(x_T - x^*)\|_{\ell_2} \\
&\le (1-\delta)^{-1} (\|\Phi(x - x_T)\|_{\ell_2} + \|\Phi(x - x^*)\|_{\ell_2}) \\
&= (1-\delta)^{-1} (\|y - \Phi x_T\|_{\ell_2} + \|y - \Phi x^*\|_{\ell_2}) \\
&\le 2(1-\delta)^{-1} \|y - \Phi x_T T\|_{\ell_2} = 2(1-\delta)^{-1} \|\Phi(x - x_T)\|_{\ell_2} \\
&\le 2C(1-\delta)^{-1} \|x - x_T\|_{\ell_2} = 2C(1-\delta)^{-1} \sigma_k(x)_{\ell_2}.
\end{aligned}$$

where the first inequality uses the RIP and the fact that $x_T - x^*$ is a vector with support of size less than $2k$, the third inequality uses the minimality of T^* and the fourth inequality uses the boundedness property in probability for $x - x_T$. $\Box$

8 Decoding

Up to this point we have completely ignored the practicality of the decoders used in our compressed sensing results. We shall now remedy this situation. The two most common decoders are constructed by ℓ_1 minimization and greedy algorithms. Both of these are reasonable to implement numerically. We shall only have time to discuss ℓ_1 minimization but there are now nice results for greedy decoders (see [20], [9]). We concentrate on how this decoder performs in terms of instance optimality in ℓ_1^N and instance optimality with high probability in ℓ_2^N?

The decoder for ℓ_1 minimization is

$$\Delta(y) := \operatorname*{argmin}_{\Phi z = y} \|z\|_{\ell_1}, \quad y \in \mathbb{R}^n. \tag{8.1}$$

It can be implemented numerically with linear programming using the simplex algorithm or interior point methods. The fact that ℓ_1-minimization is a good decoder was one of the main contributions of Donoho [13] and Candés, Romberg, and Tao [5, 7] and their results were the beginning of the subject of compressed sensing as it is now called. The following theorem is contained in [10] but can also be derived from the techniques in [5].

Theorem 8.1 *Let Φ be any matrix which satisfies the RIP of order $3k$ with $\delta_{3k} \le \delta < (\sqrt{2}-1)^2/3$. Define the decoder Δ for Φ as in (8.2). Then, (Φ, Δ) satisfies (2.3) in $X = \ell_1$ with $C_0 = \frac{2\sqrt{2}+2-(2\sqrt{2}-2)\delta}{\sqrt{2}-1-(\sqrt{2}+1)\delta}$.*

Remark: *Again, Candés [4] shows that $3k$ can be replaced by $2k$ with a somewhat more involved argument.*

Proof: We apply Lemma 5.1 with $a = 1$, $b = 2$ to see that Φ satisfies the null space property in ℓ_1 of order k with constant $C = 1 + \frac{1+\delta}{\sqrt{2}(1-\delta)} < 2$. This means that for any $\eta \in \mathcal{N}$ and T such that $\#T \le k$, we have

$$\|\eta\|_{\ell_1} \le C\|\eta_{T^c}\|_{\ell_1}, \tag{8.2}$$

and therefore

$$\|\eta_T\|_{\ell_1} \le (C-1)\|\eta_{T^c}\|_{\ell_1}. \tag{8.3}$$

Let $x^* = \Delta(\Phi x)$ be the solution of (8.1) so that $\eta = x^* - x \in \mathcal{N}$ and

$$\|x^*\|_{\ell_1} \le \|x\|_{\ell_1}. \tag{8.4}$$

Denoting by T the set of indices of the largest k coefficients of x, we can write

$$\|x_T^*\|_{\ell_1} + \|x_{T^c}^*\|_{\ell_1} \le \|x_T\|_{\ell_1} + \|x_{T^c}\|_{\ell_1}. \tag{8.5}$$

It follows that

$$\|x_T\|_{\ell_1} - \|\eta_T\|_{\ell_1} + \|\eta_{T^c}\|_{\ell_1} - \|x_{T^c}\|_{\ell_1} \le \|x_T\|_{\ell_1} + \|x_{T^c}\|_{\ell_1}, \tag{8.6}$$

and therefore

$$\|\eta_{T^c}\|_{\ell_1} \le \|\eta_T\|_{\ell_1} + 2\|x_{T^c}\|_{\ell_1} = \|\eta_T\|_{\ell_1} + 2\sigma_k(x)_{\ell_1}. \tag{8.7}$$

Using (8.3) and the fact that $C < 2$ we thus obtain

$$\|\eta_{T^c}\|_{\ell_1} \leq \frac{2}{2-C}\sigma_k(x)_{\ell_1}. \tag{8.8}$$

We finally use again (8.2) to conclude that

$$\|x - x^*\|_{\ell_1} \leq \frac{2C}{2-C}\sigma_k(x)_{\ell_1}, \tag{8.9}$$

which is the announced result. □

Our next goal is to show that the ℓ_1 minimization decoder can be used together with general random matrices to give instance optimality in probability for the large range of k. To establish this fact we need another mapping property of random matrices.

Lemma 8.2 *Let $\Phi(\omega)$ be an $n \times N$ random matrix which satisfies CMP. For each $x \in \mathbb{R}^N$ there is a set $\Omega_1(x)$ with*

$$\rho(\Omega_1(x)^c) \leq Ce^{\frac{-n}{2L}} \tag{8.10}$$

such that for all $\omega \in \Omega_1(x)$,

$$\|\Phi x\|_{\ell_\infty^n} \leq \frac{1}{\sqrt{L}}\|x\|_{\ell_2^N}, \quad \text{where } L := \log N/n. \tag{8.11}$$

Proof: We shall prove this lemma in the case that $\eta = \frac{1}{\sqrt{n}}r$ where r is the Bernoulli random variable taking values ± 1. In the general **SG** case, one has to analyze moments (see [12]. Without loss of generality we can assume that $\|x\|_{\ell_2^N} = 1$. Fix such an x. We note that each entry y_i of y is of the form

$$y_i = \frac{1}{\sqrt{n}}\sum_{j=1}^{N} x_j r_{i,j}, \tag{8.12}$$

where the $r_{i,j}$ are independent random variables and $x = (x_1, \dots, x_N)$. We shall use Hoeffding's inequality (see page 596 of [15]) which says that for independent mean zero random variables ϵ_j taking values in $[a_j, b_j]$, $j = 1, \dots, N$, we have

$$\Pr\left(|\sum_{j=1}^{N}\epsilon_j| \geq \delta\right) \leq 2e^{\frac{-2\delta^2}{\sum_{j=1}^{N}(b_j-a_j)^2}}. \tag{8.13}$$

We apply this to the random variables $\epsilon_j := \frac{1}{\sqrt{n}}x_j r_{i,j}$, $j = 1, \dots, N$, which take values in $\frac{1}{\sqrt{n}}[-x_j, x_j]$. Since $\sum_{j=1}^{N}(2x_j)^2 = 4$, we deduce that

$$\Pr(|y_i| \geq \delta) \leq 2e^{\frac{-n\delta^2}{2}}. \tag{8.14}$$

Applying a union bound, we get

$$\Pr\left(\|y\|_{\ell_\infty^n} \geq \delta\right) \leq 2ne^{\frac{-n\delta^2}{2}}. \tag{8.15}$$

If we now take $\delta = 1/\sqrt{L}$ we arrive at the lemma. □

There is one additional mapping property of random matrices which is instrumental in showing that ℓ_1 minimization can be used as a decoder and attain instance optimality in probability.

Clipped Ball Mapping Property (CBMP): *Let $\Phi(\omega)$ be a random family of $n \times N$ matrices whose entries are given by random draws of the random variable $\eta = \frac{1}{\sqrt{n}} r$ with r a* **SG** *random variable. Let $L := \log(N/n)$ as before. Then, with probability $\geq 1 - Ce^{-c\sqrt{nN}}$ on the draw of Φ the following holds: for each vector $y \in \mathbb{R}^n$ with $\|y\|_{\ell_2^n}, L^{-1/2}\|y\|_{\ell_\infty^n} \leq 1$, there is a $z \in \mathbb{R}^N$ such that $\Phi(z) = y$ and $\|z\|_{\ell_1^N} \leq C' \sqrt{\frac{n}{L}}$. In other words, with high probability the unit ball in ℓ_1^N is mapped onto a clipped ball in $\mathbb{R}^n$.*

Remark: Using arguments similar to the proof of ℓ_1 instance optimality we can also require that the vector z in **CBMP** satisfies $\|z\|_{\ell_2^N} \leq C\|y\|_{\ell_2^n}$

This mapping property was proved by A. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann [18] and reproved in [12]. We now use this mapping property to prove ℓ_2 instance optimality in probability.

Theorem 8.3 *Let $\Phi(\omega)$ be a random family of $n \times N$ matrices whose entries are given by random draws of the random variable $\eta = \frac{1}{\sqrt{n}} r$ with r a* **SG** *random variable and let Δ be the ℓ_1-minimization decoder. Let $L := \log(N/n)$ as before. For each $x \in \mathbb{R}^N$ and each $k \leq \tilde{a}n/\log(N/n)$, $N \geq [\ln 6]^2 n$, there is a set $\Omega(x,k)$ with*

$$\rho(\Omega(x,k)^c) \leq C[e^{-\tilde{c}_1 n} + e^{-\sqrt{Nn}} + e^{-n/24} + ne^{\frac{-n}{2\log(N/n)}}], \tag{8.16}$$

such that for each $\omega \in \Omega(x,k)$, we have

$$\|x - \Delta(\Phi x)\|_{\ell_2^N} \leq C'\sigma_k(x)_{\ell_2^N}, \tag{8.17}$$

where C and C' are absolute constants.

Proof: We will prove the theorem for the largest k satisfying $k \leq \tilde{a}n/L$. The theorem follows for all other k from the monotonicity of σ_k. Let x_k be a best approximation to x from Σ_k, so $\|x - x_k\|_{\ell_2^N} = \sigma_k(x)_{\ell_2^N} =: \sigma_k(x)$, and let $y' = \Phi(x - x_k)$. From **CMP** and Lemma 8.2, we have with high probability

$$\|y'\|_{\ell_2^n} \leq \sqrt{\frac{3}{2}}\|x - x_k\|_{\ell_2^N} = \sqrt{\frac{3}{2}}\sigma_k(x),$$

and

$$\|y'\|_{\ell_\infty^n} \leq \frac{1}{\sqrt{L}}\|x - x_k\|_{\ell_2^N} = \frac{1}{\sqrt{L}}\sigma_k(x).$$

The **CBMP** and the Remark following its definition says that there is a vector $z' \in \mathbb{R}^N$, such that $\Phi(x - x_k) = y' = \Phi z'$ and

$$\|z'\|_{\ell_2^N} \leq C\sigma_k(x), \quad \text{and } \|z'\|_{\ell_1^N} \leq C\sqrt{\frac{n}{L}}\sigma_k(x). \tag{8.18}$$

Note that $\sigma_k(x_k + z')_{\ell_1^N} \leq \|z'\|_{\ell_1^N}$, and therefore using (8.18) it follows that

$$\sigma_k(x_k + z')_{\ell_1^N} \leq C\sqrt{\frac{n}{L}}\sigma_k(x). \tag{8.19}$$

Since $\Phi x = \Phi(x_k + z')$, we have that $\bar{x} := \Delta(\Phi(x_k + z')) = \Delta(\Phi x)$. We know that with high probability Φ satisfies RIP of order $2k$ and hence the mixed-norm instance optimality (5.4). This means that

$$\|x_k + z' - \bar{x}\|_{\ell_2^N} \leq \frac{C}{\sqrt{k}} \sigma_k(x_k + z')_{\ell_1^N} \leq C' \sigma_k(x).$$

where the last inequality uses the definition of k. Therefore, it follows from (8.18) that

$$\begin{aligned} \|x - \bar{x}\|_{\ell_2^N} &\leq \|x - x_k - z'\|_{\ell_2^N} + \|x_k + z' - \bar{x}\|_{\ell_2^N} \\ &\leq \|x - x_k\|_{\ell_2^N} + \|z'\|_{\ell_2^N} + \|x_k + z' - \bar{x}\|_{\ell_2^N} \\ &\leq C\sigma_k(x), \end{aligned} \tag{8.20}$$

which proves the theorem. □

References

[1] D. Achilioptas, Database-friendly random projections, Proc. ACM Symposium on the Principles of Database Systems, pp. 274-281, 2001

[2] R. Adamczak, A.E. Litvak, A. Pajor, N. Tomczak-Jaegermann, *Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling*, preprint

[3] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, A simple proof of the restricted isometry property for random matrices, Constructive Approximation, to appear.

[4] E. Candés, *The restricted isometry property and its implications for compressed sensing*, Compte Rendus de l'Academie des Sciences, Paris, Series I, **3**46(2008), 589–592.

[5] E. J. Candès, J. Romberg and T. Tao, Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information, *IEEE Trans. Inf. Theory*, **5**2(2006), 489–509.

[6] E. Candès, J. Romberg, and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Comm. Pure and Appl. Math.*, **5**9(2006), 1207–1223.

[7] E. Candès and T. Tao, Decoding by linear programming, *IEEE Trans. Inf. Theory* **5**1(2005), 4203–4215.

[8] E. Candès and T. Tao, Near optimal signal recovery from random projections:universal encoding strategies, *IEEE Trans. Inf. Theory* **5**2(2006), 5406–5425.

[9] A. Cohen, W. Dahmen, and R. DeVore, Near optimal approximation of arbitrary vectors from highly incomplete measurements, preprint.

[10] A. Cohen, W. Dahmen, and R. DeVore, *Compressed sensing and best k-term approximation*, JAMS, **22**(2009), 211–231.

[11] R. DeVore, Nonlinear approximation, *Acta Numer.* **7** (1998), 51–150.

[12] R. DeVore, G. Petrova, and P. Wojtaczszyk, *Instance-optimality in Probability with an ℓ_1-Minimization Decoder*, ACHA (to appear).

[13] D. Donoho, Compressed Sensing, EEE Trans. Information Theory, **52** (2006), 1289-1306.

[14] A.Garnaev, E.D. Gluskin, The widths of a Euclidean ball, Doklady AN SSSR, 277 (1984), 1048-1052.

[15] Györfy, L., M. Kohler, A. Krzyzak, A. and H. Walk (2002) *A distribution-free theory of nonparametric regression*, Springer Verlag, Berlin.

[16] E.D. Gluskin, Norms of random matrices and widths of finite-dimensional sets, Math. USSR Sbornik, 48(1984), 173–182.

[17] B. Kashin, The widths of certain finite dimensional sets and classes of smooth functions, *Izvestia* **4**1(1977), 334–351.

[18] A. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Advances in Math. **1**95 (2005), 491–523.

[19] G.G. Lorentz, M. von Golitschek and Yu. Makovoz, *Constructive Approximation:Advanced Problems*, Springer Grundlehren, vol. 304, Springer Berlin Heidelberg, 1996.

[20] D. Needel and J. Tropp, *CoSaMP: Iterative signal recovery from incomplete and inaccurate samples*, preprint 2008.

[21] M. Rudelson and R. Vershynin, *On sparse reconstruction from Gaussian and Fourier measurements*, CPAM **61**(2008), 1025–1045.

[22] A. Pajor and N. Tomczak-Jaegermann, *Subspaces of small codimension of finite dimensional Banach spaces*, Proc. Amer. Math. Soc., vol. 97, 1986, pp. 637–642.