A Taste of Compressed Sensing [*]

Albert Cohen, Wolfgang Dahmen, and Ronald DeVore

January 30, 2007

Abstract

The usual paradigm for signal processing is to model a signal as a bandlimited function and capture the signal by means of its time samples. The Shannon-Nyquist theory says that the sampling rate needs to be at least twice the bandwidth. For broadbanded signals, such high sampling rates may be impossible to implement in circuitry. Compressed Sensing is a new area of signal processing whose aim is to circumvent this dilemma by sampling signals closer to their information rate instead of their bandwidth. Rather than model the signal as bandlimited, Compressed Sensing, assumes the signal can be represented or approximated by a few suitably chosen terms from a basis expansion of the signal. It also enlarges the concept of sample to include the application of any linear functional applied to the signal. In this paper, we shall give a brief introduction to compressed sensing that centers on the effectiveness and implementation of random sampling.

Key Words: Compressed sensing, best k-term approximation, instance optimality in probability, efficient decoding, orthogonal matching pursuit.

AMS Subject Classification: 68P30, 94A20, 94A34, 94A12

1 Introduction

Compressed Sensing (CS) is a new area of signal processing whose goal is to capture a signal/image with as few measurements as possible. While this idea has a long history in mathematics dating back to seminal work in approximation theory and widths of the 1970's (see for example [15, 14]), its potential in signal processing was recently brought into focus by the the work of Candés, Romberg, and Tao [6, 7] and Donoho [10]. Related developments occured in the work on finite rates of innovation by Vetterli, Marziliano and Blue (see e.g. [16]) and the problem of data sketching in theoretical computer science (see for example [11, 12, 5] as representative papers).

[*] This research was supported by the Office of Naval Research Contracts ONR-N00014-03-1-0051, ONR/DEPSCoR N00014-03-1-0675, and ONR/DEPSCoR N00014-05-1-0715; AFOSR Contract UF/USAF F49620-03-1-0381; DARPA/NGA Contract HM1582-05-2-0001; ARO/DoD Contract W911NF-05-1-0227; National Science Foundation Grant DMS-354707; and the French-German PROCOPE contract 11418YB

The idea behind CS is to encode a workable approximation of the signal with as few measurements as possible. There are several reasons for wanting to do this. The most obvious is the challenge of encoding broadbanded signals as mentioned in the Abstract. Since conventional sampling based on the model of bandlimited signals requires an inordinate number of samples, the only alternatives seem to be to either change our notion of signal model from bandlimited to something more representative of the information rate in the signal or to change the notion of sample itself. CS does both.

The new model for signals utilized in CS is based on sparse approximation. It assumes that the signal can be represented or well approximated by a linear combination of a small number of waveforms taken from a basis or dictionary. We shall discuss only the discrete version of Compressed Sensing in which the signal is a vector $x \in I\!R^N$ with N typically very large. For example, in image processing N would be the number of pixels which may be one million or more.

The typical paradigm for obtaining a compressed version of a discrete signal represented by a vector $x \in I\!R^N$ is to choose an appropriate basis, compute all of the coefficients of x in this basis, and then retain only the k largest of these with $k < N$. Without loss of generality, we shall assume that the appropriate basis is the canonical Kroenecker delta basis. Any other basis can be treated by including a linear transformation to the canonical basis. The view, expressed by CS is that rather than sample every entry of x, it may be possible to actually compute only a few *non-adaptive* linear measurements and still retain the necessary information about x in order to build a good compressed representation. These measurements are represented by a vector

$$y = \Phi x, \tag{1.1}$$

of dimension $n < N$ where Φ is an $n \times N$ *measurement matrix* (called a CS matrix). The recovery of an approximation x^* of x from these measurements is performed by the application of an operator Δ which we refer to as the *decoder* to y:

$$x^* := \Delta(y) = \Delta(\Phi x). \tag{1.2}$$

In contrast to Φ, the operator Δ is allowed to be non-linear. We call such an encoder-decoder pair (Φ, Δ) a *Compressed Sensing System*.

The main questions in compressed sensing are: (i) what is the optimal or near optimal performance we can obtain for a CS system? (ii) what are good CS matrices Φ? (iii) what are the most efficient decoders Δ realizing near optimal performance? This paper will summarize part of our current understanding of these questions.

2 Instance optimality: a measure of performance in Compressed Sensing

To measure the performance of a CS system, we have two issues: (i) what is the distortion between the signal x and the decoded approximation x^*?, (ii) what is the computational effort in the decoding? Since we are dealing with discrete signals, the most natural measures of distortion are the ℓ_p norms defined by

$$\|x\|_{\ell_p} := \|x\|_{\ell_p^N} := \begin{cases} \left(\sum_{i=1}^N |x_j|^p\right)^{1/p}, & 0 < p < \infty, \\ \max_{j=1,\dots,N} |x_j|, & p = \infty. \end{cases} \tag{2.1}$$

Since our assumption about the signal x centers on how well it can be approximated by a linear combination of a few terms of the basis, it is natural to measure the quality of performance by comparing the distortion $\|x - x^*\|_{\ell_p}$ with how well the signal can be approximated by a given budget of k terms from the basis. The best we could do given such a budget is to choose the k largest entries of x as the approximation. If we denote by $S_k \subset \{1, \cdots, N\}$ the set of indices corresponding to these k largest entries, then the performance of such an approximation process in the ℓ_p norm is the *best k-term approximation* error

$$\sigma_k(x)_{\ell_p} := \|x - x_{S_k}\|_{\ell_p} = \|x_{S_k^c}\|_{\ell_p}, \tag{2.2}$$

where for any set of indices $S \subset \{1, \dots, N\}$, we denote by S^c its complement and by x_S the vector obtained from x by setting to 0 all its component with indices not in S. Note x_{S_k} is best approximation of x from the set of *k-sparse vectors*, i.e. vectors which have at most k non-zero coordinates. This approximation process should be considered as *adaptive* since the indices of those coefficients that are retained will vary from one signal to the next. So we are interested in comparing the performance of the non-adaptive CS system with the optimal adaptive strategy of choosing the best k coordinates to approximate x.

In [3], we adressed the performance of CS systems by considering the following question:

For a given space $X = \ell_p^N$ (i.e. $I\!R^N$ endowed with the ℓ_p-norm) and $k < N$, what is the minimal value of n for which there exists a CS system (Φ, Δ) such that

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

for all $x \in I\!R^N$, with C_0 a constant independent of k and N?

We say that a pair (Φ, Δ) which satisfies property (2.3) is *instance-optimal of order k* with constant C_0. Note that we could reverse the roles of n and k above by fixing n and asking for the largest value of k for which (2.3) holds.

It was shown that the answer to the above question heavily depends on the ℓ_p norm under consideration. Let us illustrate this by quoting two contrasting results from [3]:

(i) In the case $p = 1$, for any $n \ge ak\log(N/k)$, with a a sufficiently large fixed constant, it is possible to build encoding-decoding pairs (Φ, Δ), with Φ an $n \times N$ matrix, that are instance-optimal of order k with constant C_0 depending only on a. Moreover the decoder Δ can be taken as

$$\Delta(y) := \underset{\Phi z = y}{\operatorname{argmin}} \|z\|_{\ell_1}. \tag{2.4}$$

These facts can also be derived from the results of Candés, Romberg and Tao in [7]. Therefore, it is possible to obtain the accuracy of k-term approximation whenever the budget n for the number of non-adaptive measurements exceeds the number k of adaptive measurements by the small factor $a \log(N/k)$.

(ii) In the case $p = 2$, if (Φ, Δ) is any CS system which is instance-optimal of order $k = 1$, then the number of measurement n is always larger than aN where $a > 0$ depends only on C_0 in (2.3). Therefore, the number of non-adaptive measurements has to be very large in order to compete with even one single adaptive measurement.

While it appears from this last result that compressed sensing may not perform well in ℓ_2, it turns out that this is not exactly the case. For example, a more optimistic result was established by Candés, Romberg and Tao in [7]. They show that if $n \geq ak \log(N/k)$ it is possible to build CS systems (Φ, Δ) such that for all $x \in \mathrm{IR}^{\mathrm{N}}$,

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq C_0 \frac{\sigma_k(x)_{\ell_1}}{\sqrt{k}}, \tag{2.5}$$

with the decoder again defined by (2.4). Of course, (2.5) is not the same as instance-optimal since it involves approximation in ℓ_1 on the right side. However, this estimate does show that k-sparse signals are exactly reconstructed and signals x which are well approximated by k-sparse signals in the ℓ_1 sense will be approximated well in the ℓ_2 sense as well.

While instance-optimality cannot hold in ℓ_2, it turns out that instance-optimal pairs for ℓ_2 can be constructed if one accepts a *probabilistic* statement in which $\Phi = \Phi(\omega)$, $\omega \in \Omega$, is a random matrix-valued variable on a probability space (Ω, ρ). A first result in this direction, obtained by Cormode and Mutukrishnan in [5], shows how to construct a random matrix Φ with $n \sim k(\log N)^{5/2}$ measurements and a decoder $\Delta = \Delta(\omega)$, $\omega \in \Omega$, such that for any $x \in I\!\!R^N$,

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq C_0 \sigma_k(x)_{\ell_2} \tag{2.6}$$

holds with overwhelming probability (larger than $1 - \epsilon(n)$ where $\epsilon(n)$ tends rapidly to 0 as $n \to +\infty$). Note that this result says that given x, the set of $\omega \in \Omega$ for which (2.6) fails to hold has small measure. This set of failure will depend on x. We shall refer to results of the form (2.6) as *instance-optimality in probability.* The above results on instance-optimal in probability can be improved in several directions as will be discussed later in this paper.

3 Optimal choices for the matrix Φ

If we fix the budget n of samples to be used in a CS system then the best systems from the viewpoint above are those that give instance-optimality of order k for the largest values of k. We have noted above that one may take k of the order $n/\log(N/n)$ and achieve instance-optimality in ℓ_1. One can show that this range of k cannot be improved (see [3]).

How can one design CS systems that have this optimal order of instance-optimality? Candés and Tao [6] have introduced a property which can be used in order to derive sufficient conditions for instance-optimality of order k:

RIP: *A matrix* Φ *satisfies the Restricted Isometry Property* (RIP) *of order* k *with constant* $\delta_k \in (0,1)$, *if*

$$(1-\delta_k)\|x_T\|_{\ell_2} \le \|\Phi x_T\|_{\ell_2} \le (1+\delta_k)\|x_T\|_{\ell_2}, \quad x \in I\!R^N, \ |T| \le k. \tag{3.1}$$

The requirement (3.1) says that the mapping Φ acts like an isometry on k-sparse vectors.

The following theorem (proved in [3]) shows that any matrix Φ which satisfies the RIP of order $3k$ admits a decoder Δ so that the CS System (Φ, Δ) is instance-optimal in ℓ_1 of order k.

Theorem 3.1 *Let* Φ *be any matrix which satisfies RIP of order* $3k$ *with* $\delta_{3k} < 1$. *Then there is a decoder* Δ *such that the CS system* (Φ, Δ) *satisfies instance optimality of order* k *in* ℓ_1 *with constant* $C_0 = 2\sqrt{2}\frac{1+\delta}{1-\delta}$.

For this theorem to be of any use, we need to have a way of constructing matrices Φ which satisfy the RIP and we must be able to verify the RIP for these matrices. We have already remarked that such constructions are known when $k \le an/\log(N/n)$. What do these constructions look like?

All of the constructions which exhibit the largest range of k utilize randomness. The simplest of these to describe is the following. We want to fill in the nN entries of the matrix Φ. We flip a fair coin and if it lands heads, we will fill the $(1,1)$ entry with the value $1/\sqrt{n}$ and if it lands tails then we fill this entry with the value $-1/\sqrt{n}$. The remaining entries are filled in with the same procedure by using independent coin flips. We do not know if, for a given set of coin flips, the resulting matrix Φ satisfies the RIP of order $k \approx n/\log(N/n)$. It of course depends on the luck of the draw in our coin flips. But what we can show is that with high probability the resulting matrix will have the RIP for this range of k. In other words most of the time we have the RIP.

In [2], we have given a very simple proof that the above construction as well as other standard random constructions of matrices lead to the RIP. These constructions include the replacement of coin flips by independent draws of more general random variables such as Gaussian. The derivation of RIP applies to matrices Φ whose entries are independent realizations of a random variable $r(\omega)$, $\omega \in \Omega$, for which the matrices $\Phi(\omega)$, $\omega \in \Omega$, satisfy the following concentration of measure inequality:

P1 : For any $x \in I\!R^{\mathrm{N}}$ and $\delta \in (0,1]$

$$\left|\|\Phi x\|_{\ell_2}^2 - \|x\|_{\ell_2}^2\right| \le \delta \|x\|_{\ell_2}^2 \tag{3.2}$$

holds with probability $\ge 1 - b_1 e^{-c_1 n \delta^2}$ where b_1 and c_1 are absolute constants.
Notice that the verification of P1 is a much simpler task than verifying RIP since we only need verify P1 with high probability. It is shown in [2] that P1 implies the RIP as is formulated in the following theorem

Theorem 3.2 *Suppose that* n, N, *and* $0 < \delta < 1$ *are given. If the random* $n \times N$ *matrices* $\Phi(\omega)$ *satisfy the concentration inequality* P1, *then there exist constants* $c_2, c_3 > 0$ *depending only on* δ *such that the RIP holds for* $\Phi(\omega)$ *with the prescribed* δ *and any* $k \le c_2 n/\log(N/k)$ *with probability larger than* $1 - e^{-c_3 n}$.

It is easy to verify that standard random variables such as the Bernoulli random variable used with coin flips or the Gaussian random variable satisfy P1 (see [4]). Unfortunately, for any given realization of Φ, there is no simple way to check whether the resulting matrix satisfies the RIP. Because of this, for large values of n, N, we are still unable to directly put our hands on a matrix which satisfies the RIP of large order k and therefore gives our best instance-optimal performance. There are some deterministic constructions of matrices and verifications that they satisfy the RIP but the range of k for which this is true is much more modest: $k \leq c\sqrt{n}/\log(N/n)$. One such construction given in [8] is based on finite fields. Other constructions can be made using coding theory.

4 Instance-optimality in probability

While probability and randomness are used to prove the existence of matrices which satisfy the RIP (and hence instance-optimality in ℓ_1) for the biggest range of k, we should note that the CS system itself does not use randomness. Once we have a matrix which satisfies the RIP, the compressed sensing is completely deterministic. However, the weakness in the above results is that although we know such favorable matrices exist, it is hard to put our hands on one and verify that it is a favorable choice.

There is another thread of compressed sensing where we put randomness into the system itself. We begin with a family $\Phi(\omega)$, $\omega \in \Omega$, of random matrices. Given a signal $x \in I\!R^N$, we draw one of these random matrices $\Phi(\omega)$ and use it to encode x. This gives a vector $y(\omega) = \Phi(\omega)x$. We decode $y(\omega)$ by using a decoder $\Delta = \Delta(\omega)$ which depends on the matrix Φ (and therefore on our draw). This gives the vector $x(\omega) := \Delta(y(\omega))$ and we are interested in how well this vector approximates x.

It is interesting that such a random CS system can perform significantly better than fixing the matrix once and for all. To explain this, we first state a theorem which follows from [3].

Theorem 4.1 *Assume that $\Phi(\omega)$, $\omega \in \Omega$, is a random matrix whose entries are generated by independent draws of a random variable $r(\omega)$. If $\Phi(\omega)$ satisfies* P1*, then there is a family of decoders $\Delta(\omega)$, $\omega \in \Omega$, with the following property. If x is any vector in $I\!R^N$, then with high probability, we have*

$$\|x - x(\omega)\|_{\ell_2} \leq C_0 \sigma_k(x)_{\ell_2}, \tag{4.1}$$

for some $k \geq an/\log(N/n)$ with C_0 and a absolute constants. In other words, with high probability, the decoded $x(\omega)$ approximates x with instance-optimal accuracy in ℓ_2 for the large range of k.

We invite the reader to consult [3] for the precise formulation of high probability in which sense (4.1) holds.

5 Decoders

The main separation between the theoretical results of the 1970's on approximation and the current results on CS systems occurs with decoding. The early theoretical results did not consider the question of practical decoders Δ to team with a CS matrix Φ. The celebrated contribution of the early work on compressed sensing by Candés-Tao and Donoho was to show that for certain constructions of CS matrices Φ, the decoding can be done using ℓ_1 minimization as described in (2.4). This decoding is a problem in linear programming and off the shelf methods for solving (2.4) by, for example, the simplex algorithm or interior point methods can be applied.

While ℓ_1 minimization is an implementable decoder, there is still much interest in trying to reduce the computational time in decoding. This is a major research area of Theoretical Computer Science and the work in that area constructs some specific matrices Φ and ad hoc decoders which perform better than the theoretical bounds for linear programming. Moreover, in the case of random CS systems, the applicability of ℓ_1 minimization in an ℓ_2 probabilistic instance-optimal system is not proven.

An attractive alternative to ℓ_1 minimization is Orthogonal Matching Pursuit (OMP) which is a prominent method in signal processing. Gilbert and Tropp [13] have studied OMP as a decoder for compressed sensing and proved first results in the probabilistic setting. The application of OMP for CS decoding can be described as follows. We denote the columns of Φ by $(\phi_j)_{j=1,\cdots,N}$ and we first try to find a good approximation to y which is a sparse linear combination of these columns. As the first step of OMP, we define

$$j_1 := \operatorname*{argmax}_{j=1,\cdots,N} |\langle y, \phi_j \rangle|, \tag{5.1}$$

and approximate y by its projection $y^1 := z^1_{j_1}\phi_{j_1}$ with $z^1_{j_1} := \langle y, \phi_{j_1}\rangle / \|\phi_{j_1}\|^2$. At the step i of the algorithm, we have defined a set of indices $\{j_1, \cdots, j_i\}$ and the approximant $y^i = \sum_{l=1}^i z^i_{j_l}\phi_{j_l}$ which is by definition the orthogonal projection of y onto $\text{Span}\{\phi_{j_1}, \cdots, \phi_{j_i}\}$. The new index is defined by

$$j_{i+1} := \operatorname*{argmax}_{j=1,\cdots,N} |\langle r^i, \phi_j \rangle|, \tag{5.2}$$

where $r^i := y - y^i$ is the residual. The components $z^i_{j_1}, \cdots, z^i_{j_i}$, when augmented by zeros in the other coordinates, define a sparse approximation to x that we denote by x^i and which is supported on $\{j_1, \cdots, j_i\}$:

$$x^i_j = z^i_j, \ \text{ if } \ j \in \{j_1, \cdots, j_i\}, \ \ 0 \ \text{ otherwise.} \tag{5.3}$$

The following striking result was proved in [13] for fairly general classes of random matrices (such as Gaussian and Bernoulli): if $n \geq ak \log N$ with a sufficiently large, then for any k sparse vector x, the OMP algorithm returns x exactly after k iterations ($x^k = x$), with probability greater than $1 - N^{-b}$ where b can be made arbitrarily large by taking a large enough.

Note that OMP is computationally relatively inexpensive, even when compared with ℓ_1 minimization. The decoding in the above algorithm would require $O(Nkn)$ arithmetic operations which is less than the best known theoretical bounds for ℓ_1 minimization.

Recently, the authors have shown that OMP is not only a valid strategy for the recovery of k-sparse vectors, but also for *arbitrary* N-dimensional vectors. Namely, we prove that for any $x \in I\!R^N$, and any $n \geq ak \log N$, the recursive application of $2k$ steps of the

OMP algorithm on the data y gives a decoding $\Delta(y) := x^{2k}$ which satisfies probabilistic instance-optimality in ℓ_2 provided that the random matrices $\Phi(\omega)$, $\omega \in \Omega$, satisfy property P1 and the following additional properties

P0: the vectors $(\phi_i)_{i=1,\cdots,N}$ are statistically independent.

P2: for any $z \in \mathrm{I\!R}^{\mathrm{n}}$, $l \in \{1, \dots, N\}$, and $\delta \in (0, 1]$,

$$|\langle z, \phi_l \rangle| \leq \delta \|z\|_{\ell_2} \tag{5.4}$$

holds with probability $\geq 1 - b_2 e^{-c_2 n \delta^2}$, where b_2 and c_2 are absolute constants.

Similar to property P1, P2 can be easily established for standard constructions of random matrices such as the Bernouli and Gaussian families (see [4]).

The following theorem is proved in [4]

Theorem 5.1 *If the random matrix* $\Phi(\omega)$, $\omega \in \Omega$, *satisfies* P0, P1 *and* P2, *then, for* $C_0 := 5 + 146\sqrt{2}$, *the vector* $x^* = x^{2k}$ *obtained after* $2k$ *iterations of the OMP algorithm satisfies*

$$\|x - x^*\|_{\ell_2} \leq C_0 \sigma_k(x), \tag{5.5}$$

with probability larger than $1 - N^{-\beta}$ *provided that* $n \geq ak \log N$. *Here* β *can be made arbitrarily large by choosing the constant* a *sufficiently large.*

Note that, given n, N, the specific decoder OMP provides ℓ_2-instance optimality for a slightly smaller range of k than claimed in Theorem 4.1. However, the decoder upon which the result in Theorem 4.1 is based upon is not computationally feasible because it requires solving least squares problems for all subsets of k columns of Φ, see [3].

Unlike the case of sparse signals we cannot expect the greedy algorithm to identify exactly the set S_k of k largest coefficients in absolute value because there may be many coefficients outside S_k of nearly the same size. All we can hope for is that it does identify with high probability a set of indices whose coefficients carry enough of the total ℓ_2-norm to restore x to accuracy $C\sigma_k(x)$. The proof in [4] shows that with high probability, the OMP chooses indices from the following set set

$$T_k := \Big\{ j : \; |x_j| \geq \frac{\sigma_k(x)}{\sqrt{k}} \Big\}. \tag{5.6}$$

Notice that T_k depends only on x and not on $\omega \in \Omega$.

The fact that selecting indices from T_k is sufficient for instance-optimality follows from the observation that T_k is not too different from S_k, according to the following

Lemma 5.2 *For each* $k = 1, 2, \dots$, *we have*

(i) $|T_k| \leq 2k$;

(ii) $\|x_{T_k^c}\| \leq \sqrt{2}\sigma_k(x)$.

It is interesting to remark that for general dictionaries, the OMP algorithm is known to converge slowly: its approximation error $\|y - y^k\|_{\ell_2}$ can at best be bounded by $k^{-1/2}$

(see [9] and [1] for a general discussion on the rate of convergence). In the present setting, its improved convergence properties are strongly tied to the probabilistic properties of Φ.

In the case of signals x with relatively low rates of best k-term approximation, we can actually use the results obtained in [1] in order to prove that the OMP decoder performs well without requiring the additional assumptions P0 and P2. Namely, we have the following.

Theorem 5.3 *Assume that the $n \times N$ matrices $\Phi(\omega)$, $\omega \in \Omega$, satisfy the concentration property* P1 *and let $1 \leq p < 2$. Then there exist absolute constants b_3, c_3 such that for any $x \in I\!R^N$ the output x^k of the kth step of OMP satisfies with probability larger than $1 - b_3 e^{-c_3 n}$*

$$\|x - x^k\|_{\ell_2} \leq 10\|x\|_{\ell_p} k^{-s}, \qquad s = \frac{1}{p} - \frac{1}{2}, \tag{5.7}$$

provided that $n \geq (2/c_1) k \log(N/n)$.

Proof: Recalling that x_{S_k} denotes the best k-term approximation to x and setting $y^k := \Phi x^k$, we have

$$\begin{aligned} \|x - x^k\|_{\ell_2} &\leq \sigma_k(x)_{\ell_2} + \|x_{S_k} - x^k\|_{\ell_2} \leq \sigma_k(x)_{\ell_2} + 2\|\Phi(x_{S_k} - x^k)\|_{\ell_2} \\ &\leq \sigma_k(x)_{\ell_2} + 2\Big\{\|\Phi(x_{S_k} - x)\|_{\ell_2} + \|y - y^k\|_{\ell_2}\Big\} \\ &\leq 4\sigma_k(x)_{\ell_2} + 2\|y - y^k\|_{\ell_2}, \end{aligned} \tag{5.8}$$

where we have used RIP of order $2k$ with $\delta = 1/2$ in the second inequality and P1 with $\delta = 1/2$ in the last step.

So it remains to estimate the residual error $\|y - y^k\|_{\ell_2}$. To this end, we invoke Theorem 2.3 in [1], which implies that

$$\|y - y^k\|_{\ell_2} \leq 2 \inf_{x' \in I\!R^N} \{\|y - \Phi x'\|_{\ell_2} + k^{-1/2}\|x'\|_{\ell_1}\}. \tag{5.9}$$

Since $y = \Phi x$ and $x - \hat{x}$ is independent of Φ, we can again employ P1 for $\delta = 1/2$ to conclude that

$$\|y - y^k\|_{\ell_2} \leq 3 \inf_{x' \in I\!R^N} \{\|x - x'\|_{\ell_2} + k^{-1/2}\|x'\|_{\ell_1}\} =: 3K(x, k^{-1/2}, \ell_2^N, \ell_1^N), \tag{5.10}$$

where $K(x, t, X, Y) := \inf_{x' \in X} \{\|x - x'\|_X + t\|x'\|_Y\}$ denotes the K-functional of Lions and Peetre.

For $\frac{1}{p} = \frac{\theta}{2} + \frac{1}{2}$ (i.e. $\theta/2 = s$), the space ℓ_p^N is contained in the interpolation space $[\ell_2^N, \ell_1^N]_{\theta,\infty}$ in the sense of the inequality

$$\sup_{t>0} K(x, t, \ell_2^N, \ell_1^N) t^{-\theta} \leq \|x\|_{\ell_p}, \tag{5.11}$$

between their respective norm. Therefore, we infer that $K(x, k^{-1/2}, \ell_2^N, \ell_1^N) \leq \|x\|_{\ell_p} k^{-s}$. It is also not hard to show that $\sigma_k(x)_{\ell_2} \leq \|x\|_{\ell_p} k^{-s}$ for $s = \frac{1}{p} - \frac{1}{2}$ which, in view of (5.8), (5.9), and (5.10), yields

$$\|x - x^k\|_{\ell_2} \leq 4\sigma_k(x)_{\ell_2} + 6K(x, k^{-1/2}, \ell_2^N, \ell_1^N) \leq 4\sigma_k(x)_{\ell_2} + 6\|x\|_{\ell_p} k^{-s} \leq 10\|x\|_{\ell_p} k^{-s}. \tag{5.12}$$

This confirms the assertion. □

Note that aside from the better bounds for the probability of success, the range k for which the above result holds is somewhat larger than in Theorem 5.1. On the other hand, it is considerably weaker because it provides the best rate only for the whole unit ball of ℓ_p^N, $p \in [1, 2)$ instead of in the instance-optimal sense for any rate of best k-term approximation.

References

[1] A. Barron, A. Cohen, W. Dahmen and R. DeVore, *Approximation and learning by greedy algorithms*, submitted 2006.

[2] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A Simple Proof of the Restricted Isometry Property for Random Matrices*, submitted 2006.

[3] A. Cohen, W. Dahmen and R. DeVore, *Compressed sensing and best k-term approximation*, submitted 2006

[4] A. Cohen, W. Dahmen and R. DeVore, *Near optimal approximation of arbitrary vectors from highly incomplete measurements*, submitted 2007

[5] G. Cormode and S. Muthukrishnan, *Towards an algorithmic theory of compressed sensing*, Technical Report 2005-25, DIMACS, 2005.

[6] E. Candès and T. Tao, *Decoding by linear programming*, IEEE Trans. Inf. Theory, **51**(2005), 4203–4215.

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

[8] R. DeVore, *Deterministic Constructions of Compressed Sensing Matrices*, submitted 2006.

[9] R. DeVore and V. Temlyakov, *Some remarks on greedy algorithms*, Advances in Computational Mathematics **5**(1996), 173-187.

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

[11] A. Gilbert, S. Guha, Y. Kotidis, P. Indyk, S. Muthukrishnan and M. Strauss, Fast, small space algorithm for approximate histogram maintenance, in Proc. of the 2002 ACM Symposium on Theory of Computing STOC, pp. 389–398.

[12] A. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan and M. Strauss, Near-optimal sparse fourier estimation via sampling, Proc. of the 2002 ACM Symposium on Theory of Computing STOC, pp. 152–161.

[13] A. C. Gilbert and J. A. Tropp, *Signal recovery from partial information via Orthogonal Matching Pursuit*, submitted 2005.

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

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

[16] M. Vetterli, P. Marziliano, and T. Blue, IEEE Transactions on Signal Processing, *Sampling signals with finite rate of innovation*, **50**(2002), 1417–1428.

Albert Cohen, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie 175, rue du Chevaleret, 75013 Paris, France, cohen@ann.jussieu.fr

Wolfgang Dahmen, Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, D-52056 Aachen Germany, dahmen@igpm.rwth-aachen,de

Ronald DeVore, Industrial Mathematics Institute, University of South Carolina, Columbia, SC 29208, devore@math.sc.edu