Compressive sampling

Emamnuel J. Candès*

Abstract. Conventional wisdom and common practice in acquisition and reconstruction of images from frequency data follow the basic principle of the Nyquist density sampling theory. This principle states that to reconstruct an image, the number of Fourier samples we need to acquire must match the desired resolution of the image, i.e. the number of pixels in the image. This paper surveys an emerging theory which goes by the name of "compressive sampling" or "compressed sensing," and which says that this conventional wisdom is inaccurate. Perhaps surprisingly, it is possible to reconstruct images or signals of scientific interest accurately and sometimes even exactly from a number of samples which is far smaller than the desired resolution of the image/signal, e.g. the number of pixels in the image.

It is believed that compressive sampling has far reaching implications. For example, it suggests the possibility of new data acquisition protocols that translate analog information into digital form with fewer sensors than what was considered necessary. This new sampling theory may come to underlie procedures for sampling and compressing data simultaneously.

In this short survey, we provide some of the key mathematical insights underlying this new theory, and explain some of the interactions between compressive sampling and other fields such as statistics, information theory, coding theory, and theoretical computer science.

Mathematics Subject Classification (2000). Primary 00A69, 41-02, 68P30; Secondary 62C65.

Keywords. Compressive sampling, sparsity, uniform uncertainty principle, underdertermined systems of linear equations, ℓ_1-minimization, linear programming, signal recovery, error correction.

1. Introduction

One of the central tenets of signal processing is the Nyquist/Shannon sampling theory: the number of samples needed to reconstruct a signal without error is dictated by its bandwidth – the length of the shortest interval which contains the support of the spectrum of the signal under study. In the last two years or so, an alternative theory of "compressive sampling" has emerged which shows that super-resolved signals and images can be reconstructed from far fewer data/measurements than what is usually considered necessary. The purpose of this paper is to survey and provide some of the key mathematical insights underlying this new theory. An enchanting aspect of compressive sampling it that it has significant interactions and bearings on some fields in the applied sciences and engineering such as statistics, information theory, coding

*The author is partially supported by an NSF grant CCF–515362.

Proceedings of the International Congress
of Mathematicians, Madrid, Spain, 2006

theory, theoretical computer science, and others as well. We will try to explain these connections via a few selected examples.

From a general viewpoint, sparsity and, more generally, compressibility has played and continues to play a fundamental role in many fields of science. Sparsity leads to efficient estimations; for example, the quality of estimation by thresholding or shrinkage algorithms depends on the sparsity of the signal we wish to estimate. Sparsity leads to efficient compression; for example, the precision of a transform coder depends on the sparsity of the signal we wish to encode [24]. Sparsity leads to dimensionality reduction and efficient modeling. The novelty here is that *sparsity has bearings on the data acquisition process itself, and leads to efficient data acquisition protocols.*

In fact, compressive sampling suggests ways to economically translate analog data into already compressed digital form [20], [7]. The key word here is "economically." Everybody knows that because typical signals have some structure, they can be compressed efficiently without much perceptual loss. For instance, modern transform coders such as JPEG2000 exploit the fact that many signals have a sparse representation in a fixed basis, meaning that one can store or transmit only a small number of adaptively chosen transform coefficients rather than all the signal samples. The way this typically works is that one acquires the full signal, computes the complete set of transform coefficients, encode the largest coefficients and discard *all* the others. This process of massive data acquisition followed by compression is extremely wasteful (one can think about a digital camera which has millions of imaging sensors, the pixels, but eventually encodes the picture on a few hundred kilobytes). This raises a fundamental question: because most signals are compressible, why spend so much effort acquiring all the data when we know that most of it will be discarded? Wouldn't it be possible to acquire the data in already compressed form so that one does not need to throw away anything? "Compressive sampling" also known as "compressed sensing" [20] shows that this is indeed possible.

This paper is by no means an exhaustive survey of the literature on compressive sampling. Rather this is merely an account of the author's own work and thinking in this area which also includes a fairly large number of references to other people's work and occasionally discusses connections with these works. We have done our best to organize the ideas into a logical progression starting with the early papers which launched this subject. Before we begin, we would like to invite the interested reader to also check the article [17] by Ronald DeVore – also in these proceedings – for a complementary survey of the field (Section 5).

2. Undersampled measurements

Consider the general problem of reconstructing a vector $x \in \mathbb{R}^N$ from linear measurements y about x of the form

$$y_k = \langle x, \varphi_k \rangle, \ k = 1, \dots, K, \quad \text{or} \quad y = \Phi x. \tag{2.1}$$

That is, we acquire information about the unknown signal by sensing x against K vectors $\varphi_k \in \mathbb{R}^N$. We are interested in the "underdetermined" case $K \ll N$, where we have many fewer measurements than unknown signal values. Problems of this type arise in a countless number of applications. In radiology and biomedical imaging for instance, one is typically able to collect far fewer measurements about an image of interest than the number of unknown pixels. In wideband radio frequency signal analysis, one may only be able to acquire a signal at a rate which is far lower than the Nyquist rate because of current limitations in Analog-to-Digital Converter technology. Finally, gene expression studies also provide examples of this kind. Here, one would like to infer the gene expression level of thousands of genes – that is, the dimension N of the vector x is in the thousands – from a low number of observations, typically in the tens.

At first glance, solving the underdertermined system of equations appears hopeless, as it is easy to make up examples for which it clearly cannot be done. But suppose now that the signal x is *compressible*, meaning that it essentially depends on a number of degrees of freedom which is smaller than N. For instance, suppose our signal is sparse in the sense that it can be written either exactly or accurately as a superposition of a small number of vectors in some fixed basis. Then this premise radically changes the problem, making the search for solutions feasible. In fact, accurate and sometimes exact recovery is possible by solving a simple convex optimization problem.

2.1. A nonlinear sampling theorem. It might be best to consider a concrete example first. Suppose here that one collects an incomplete set of frequency samples of a discrete signal x of length N. (To ease the exposition, we consider a model problem in one dimension. The theory extends easily to higher dimensions. For instance, we could be equally interested in the reconstruction of 2- or 3-dimensional objects from undersampled Fourier data.) The goal is to reconstruct the full signal f given only K samples in the Fourier domain

$$y_k = \frac{1}{\sqrt{N}} \sum_{t=0}^{N-1} x_t \, e^{-i2\pi\omega_k t/N}, \tag{2.2}$$

where the 'visible' frequencies ω_k are a subset Ω (of size K) of the set of all frequencies $\{0, \ldots, N-1\}$. Sensing an object by measuring selected frequency coefficients is the principle underlying Magnetic Resonance Imaging, and is common in many fields of science, including Astrophysics. In the language of the general problem (2.1), the sensing matrix Φ is obtained by sampling K rows of the N by N discrete Fourier transform matrix.

We will say that a vector x is *S-sparse* if its support $\{i : x_i \neq 0\}$ is of cardinality less or equal to S. Then Candès, Romberg and Tao [6] showed that one could almost always recover the signal x exactly by solving the convex program[1] ($\|\tilde{x}\|_{\ell_1} := \sum_{i=1}^N |\tilde{x}_i|$)

$$(\mathrm{P}_1) \qquad \min_{\tilde{x} \in \mathbb{R}^N} \|\tilde{x}\|_{\ell_1} \quad \text{subject to} \quad \Phi\tilde{x} = y. \tag{2.3}$$

[1] (P_1) can even be recast as a linear program [3], [15].

Theorem 2.1 ([6]). *Assume that x is S-sparse and that we are given K Fourier coefficients with frequencies selected uniformly at random. Suppose that the number of observations obeys*

$$K \geq C \cdot S \cdot \log N. \tag{2.4}$$

Then minimizing ℓ_1 reconstructs x exactly with overwhelming probability. In details, if the constant C is of the form $22(\delta + 1)$ in (2.4), then the probability of success exceeds $1 - O(N^{-\delta})$.

The first conclusion is that one suffers no information loss by measuring just about any set of K frequency coefficients. The second is that the signal x can be exactly recovered by minimizing a convex functional which does not assume any knowledge about the number of nonzero coordinates of x, their locations, and their amplitudes which we assume are all completely unknown a priori.

While this seems to be a great feat, one could still ask whether this is optimal, or whether one could do with even fewer samples. The answer is that in general, we cannot reconstruct S-sparse signals with fewer samples. There are examples for which the minimum number of samples needed for exact reconstruction by any method, no matter how intractable, must be about $S \log N$. Hence, the theorem is tight and ℓ_1-minimization succeeds nearly as soon as there is any hope to succeed by any algorithm.

The reader is certainly familiar with the Nyquist/Shannon sampling theory and one can reformulate our result to establish simple connections. By reversing the roles of time and frequency in the above example, we can recast Theorem 1 as a new nonlinear sampling theorem. Suppose that a signal x has support Ω in the frequency domain with $B = |\Omega|$. If Ω is a connected set, we can think of B as the bandwidth of x. If in addition the set Ω is known, then the classical Nyquist/Shannon sampling theorem states that x can be reconstructed perfectly from B equally spaced samples in the time domain[2]. The reconstruction is simply a linear interpolation with a "sinc" kernel.

Now suppose that the set Ω, still of size B, is unknown and not necessarily connected. In this situation, the Nyquist/Shannon theory is unhelpful – we can only assume that the connected frequency support is the entire domain suggesting that *all* N time-domain samples are needed for exact reconstruction. However, Theorem 2.1 asserts that far fewer samples are necessary. Solving (P_1) will recover x perfectly from about $B \log N$ time samples. What is more, these samples do not have to be carefully chosen; almost any sample set of this size will work. Thus we have a nonlinear analog (described as such since the reconstruction procedure (P_1) is nonlinear) to Nyquist/Shannon: we can reconstruct a signal with *arbitrary and unknown* frequency support of size B from about $B \log N$ *arbitrarily chosen* samples in the time domain.

Finally, we would like to emphasize that our Fourier sampling theorem is only a special instance of much more general statements. As a matter of fact, the results

[2]For the sake of convenience, we make the assumption that the bandwidth B divides the signal length N evenly.

extend to a variety of other setups and higher dimensions. For instance, [6] shows how one can reconstruct a piecewise constant (one or two-dimensional) object from incomplete frequency samples provided that the number of jumps (discontinuities) obeys the condition above by minimizing other convex functionals such as the total variation.

2.2. Background. Now for some background. In the mid-eighties, Santosa and Symes [44] had suggested the minimization of ℓ_1-norms to recover sparse spike trains, see also [25], [22] for early results. In the last four years or so, a series of papers [26], [27], [28], [29], [33], [30] explained why ℓ_1 could recover sparse signals in some special setups. We note though that the results in this body of work are very different than the sampling theorem we just introduced. Finally, we would like to point out important connections with the literature of theoretical computer science. Inspired by [37], Gilbert and her colleagues have shown that one could recover an S-sparse signal with probability exceeding $1-\delta$ from $S \cdot \mathrm{poly}(\log N, \log \delta)$ frequency samples placed on special equispaced grids [32]. The algorithms they use are not based on optimization but rather on ideas from the theory of computer science such as isolation, and group testing. Other points of connection include situations in which the set of spikes are spread out in a somewhat even manner in the time domain [22], [51].

2.3. Undersampling structured signals. The previous example showed that the structural content of the signal allows a drastic "undersampling" of the Fourier transform while still retaining enough information for exact recovery. In other words, if one wanted to sense a sparse object by taking as few measurements as possible, then one would be well-advised to measure randomly selected frequency coefficients. In truth, this observation triggered a massive literature. To what extent can we recover a compressible signal from just a few measurements. What are good sensing mechanisms? Does all this extend to object that are perhaps not sparse but well-approximated by sparse signals? In the remainder of this paper, we will provide some answers to these fundamental questions.

3. The Mathematics of compressive sampling

3.1. Sparsity and incoherence. In all what follows, we will adopt an abstract and general point of view when discussing the recovery of a vector $x \in \mathbb{R}^N$. In practical instances, the vector x may be the coefficients of a signal $f \in \mathbb{R}^N$ in an orthonormal basis Ψ

$$f(t) = \sum_{i=1}^{N} x_i \psi_i(t), \quad t = 1, \ldots, N. \tag{3.1}$$

For example, we might choose to expand the signal as a superposition of spikes (the canonical basis of $\mathbb{R}^N$), sinusoids, B-splines, wavelets [36], and so on. As a side

note, it is not important to restrict attention to orthogonal expansions as the theory and practice of compressive sampling accommodates other types of expansions. For example, x might be the coefficients of a digital image in a tight-frame of curvelets [5]. To keep on using convenient matrix notations, one can write the decomposition (3.1) as $x = \Psi f$ where Ψ is the N by N matrix with the waveforms ψ_i as rows or equivalently, $f = \Psi^* x$.

We will say that a signal f is sparse in the Ψ-domain if the coefficient sequence is supported on a small set and compressible if the sequence is concentrated near a small set. Suppose we have available undersampled data about f of the same form as before

$$y = \Phi f.$$

Expressed in a different way, we collect partial information about x via $y = \Phi' x$ where $\Phi' = \Phi\Psi^*$. In this setup, one would recover f by finding – among all coefficient sequences consistent with the data – the decomposition with minimum ℓ_1-norm

$$\min \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \Phi'\tilde{x} = y.$$

Of course, this is the same problem as (2.3), which justifies our abstract and general treatment.

With this in mind, the key concept underlying the theory of compressive sampling is a kind of uncertainty relation, which we explain next.

3.2. Recovery of sparse signals. In [7], Candès and Tao introduced the notion of uniform uncertainty principle (UUP) which they refined in [8]. The UUP essentially states that the $K \times N$ sensing matrix Φ obeys a "restricted isometry hypothesis." Let $\Phi_T,\ T \subset \{1, \ldots, N\}$ be the $K \times |T|$ submatrix obtained by extracting the columns of Φ corresponding to the indices in T; then [8] defines the S-restricted isometry constant δ_S of Φ which is the smallest quantity such that

$$(1 - \delta_S)\,\|c\|_{\ell_2}^2 \le \|\Phi_T c\|_{\ell_2}^2 \le (1 + \delta_S)\,\|c\|_{\ell_2}^2 \tag{3.2}$$

for all subsets T with $|T| \le S$ and coefficient sequences $(c_j)_{j\in T}$. This property essentially requires that every set of columns with cardinality less than S approximately behaves like an orthonormal system.

An important result is that if the columns of the sensing matrix Φ are approximately orthogonal, then the exact recovery phenomenon occurs.

Theorem 3.1 ([8]). *Assume that x is S-sparse and suppose that $\delta_{2S} + \delta_{3S} < 1$ or, better, $\delta_{2S} + \theta_{S,2S} < 1$. Then the solution $x^\star$ to* (2.3) *is exact, i.e., $x^\star = x$.*

In short, if the UUP holds at about the level S, the minimum ℓ_1-norm reconstruction is provably exact. The first thing one should notice when comparing this result with the Fourier sampling theorem is that it is deterministic in the sense that it does not involve any probabilities. It is also universal in that *all* sufficiently sparse vectors

are exactly reconstructed from Φx. In Section 3.4, we shall give concrete examples of sensing matrices obeying the exact reconstruction property for large values of the sparsity level, e.g. for $S = O(K/\log(N/K))$.

Before we do so, however, we would like to comment on the slightly better version $\delta_{2S} + \theta_{S,2S} < 1$, which is established in [10]. The number $\theta_{S,S'}$ for $S + S' \leq N$ is called the S, S'-*restricted orthogonality constants* and is the smallest quantity such that

$$|\langle \Phi_T c, \Phi_{T'} c' \rangle| \leq \theta_{S,S'} \cdot \|c\|_{\ell_2} \|c'\|_{\ell_2} \tag{3.3}$$

holds for all *disjoint* sets $T, T' \subseteq \{1, \ldots, N\}$ of cardinality $|T| \leq S$ and $|T'| \leq S'$. Thus $\theta_{S,S'}$ is the cosine of the smallest angle between the two subspaces spanned by the columns in T and T'. Small values of restricted orthogonality constants indicate that disjoint subsets of covariates span nearly orthogonal subspaces. The condition $\delta_{2S} + \theta_{S,2S} < 1$ is better than $\delta_{2S} + \delta_{3S} < 1$ since it is not hard to see that $\delta_{S+S'} - \delta_{S'} \leq \theta_{S,S'} \leq \delta_{S+S'}$ for $S' \geq S$ [8, Lemma 1.1].

Finally, now that we have introduced all the quantities needed to state our recovery theorem, we would like to elaborate on the condition $\delta_{2S} + \theta_{S,2S} < 1$. Suppose that $\delta_{2S} = 1$ which may indicate that there is a matrix $\Phi_{T_1 \cup T_2}$ with $2S$ columns ($|T_1| = S, |T_2| = S$) that is rank-deficient. If this is the case, then there is a pair (x_1, x_2) of nonvanishing vectors with x_1 supported on T_1 and x_2 supported on T_2 obeying

$$\Phi(x_1 - x_2) = 0 \iff \Phi x_1 = \Phi x_2.$$

In other words, we have two very distinct S-sparse vectors which are indistinguishable. This is why any method whatsoever needs $\delta_{2S} < 1$. For, otherwise, the model is not identifiable to use a terminology borrowed from the statistics literature. With this in mind, one can see that the condition $\delta_{2S} + \theta_{S,2S} < 1$ is only slightly stronger than this identifiability condition.

3.3. Recovery of compressible signals. In general, signals of practical interest may not be supported in space or in a transform domain on a set of relatively small size. Instead, they may only be concentrated near a sparse set. For example, a commonly discussed model in mathematical image or signal processing assumes that the coefficients of elements taken from a signal class decay rapidly, typically like a power law. Smooth signals, piecewise signals, images with bounded variations or bounded Besov norms are all of this type [24].

A natural question is how well one can recover a signal that is just nearly sparse. For an arbitrary vector x in $\mathbb{R}^N$, denote by x_S its best S-sparse approximation; that is, x_S is the approximation obtained by keeping the S largest entries of x and setting the others to zero. It turns out that if the sensing matrix obeys the uniform uncertainty principle at level S, then the recovery error is not much worse than $\|x - x_S\|_{\ell_2}$.

Theorem 3.2 ([9]). *Assume that x is S-sparse and suppose that $\delta_{3S} + \delta_{4S} < 2$. Then the solution $x^\star$ to* (2.3) *obeys*

$$\|x^* - x\|_{\ell_2} \le C \cdot \frac{\|x - x_S\|_{\ell_1}}{\sqrt{S}}. \tag{3.4}$$

For reasonable values of δ_{4S}, the constant in (3.4) *is well behaved; e.g. $C \le 8.77$ for $\delta_{4S} = 1/5$. Suppose further that $\delta_S + 2\theta_{S,S} + \theta_{2S,S} < 1$, we also have*

$$\|x^* - x\|_{\ell_1} \le C \, \|x - x_S\|_{\ell_1}, \tag{3.5}$$

for some positive constant C. Again, the constant in (3.5) *is well behaved.*

Roughly speaking, the theorem says that minimizing ℓ_1 recovers the S-largest entries of an N-dimensional unknown vector x from K measurements only. As a side remark, the ℓ_2-stability result (3.4) appears explicitly in [9] while the 'ℓ_1 instance optimality' (3.5) is implicit in [7] although it is not stated explicitely. For example, it follows from Lemma 2.1 – whose hypothesis holds because of Lemma 2.2. in [8] – in that paper. Indeed, let T be the set where x takes on its S-largest values. Then Lemma 2.1 in [7] gives $\|x^* \cdot 1_{T^c}\|_{\ell_1} \le 4\|x - x_S\|_{\ell_1}$ and, therefore, $\|(x^* - x) \cdot 1_{T^c}\|_{\ell_1} \le 5\|x - x_S\|_{\ell_1}$. We conclude by observing that on T we have

$$\|(x^* - x) \cdot 1_T\|_{\ell_1} \le \sqrt{S} \, \|(x^* - x) \cdot 1_T\|_{\ell_2} \le C \, \|x - x_S\|_{\ell_1},$$

where the last inequality follows from (3.4). For information, a more direct argument yields better constants.

To appreciate the content of Theorem 3.2, suppose that x belongs to a weak-ℓ_p ball of radius R. This says that if we rearrange the entries of x in decreasing order of magnitude $|x|_{(1)} \ge |x|_{(2)} \ge \cdots \ge |x|_{(N)}$, the ith largest entry obeys

$$|x|_{(i)} \le R \cdot i^{-1/p}, \quad 1 \le i \le N. \tag{3.6}$$

More prosaically, the coefficient sequence decays like a power-law and the parameter p controls the speed of the decay: the smaller p, the faster the decay. Classical calculations then show that the best S-term approximation of an object $x \in w\ell_p(R)$ obeys

$$\|x - x_S\|_{\ell_2} \le C_2 \cdot R \cdot S^{1/2 - 1/p} \tag{3.7}$$

in the ℓ_2 norm (for some positive constant C_2), and

$$\|x - x_S\|_{\ell_1} \le C_1 \cdot R \cdot S^{1 - 1/p}$$

in the ℓ_1-norm. For generic elements obeying (3.6), there are no fundamentally better estimates available. Hence, Theorem 3.2 shows that with K measurements only, we can achieve an approximation error which is as good as that one would obtain by knowing everything about the signal and selecting its S-largest entries.

3.4. Random matrices. Presumably all of this would be interesting if one could design a sensing matrix which would allow us to recover as many entries of x as possible with as few as K measurements. In the language of Theorem 3.1, we would like the condition $\delta_{2S} + \theta_{S,2S} < 1$ to hold for large values of S, ideally of the order of K. This poses a design problem. How should one design a matrix Φ – that is to say, a collection of N vectors in K dimensions – so that any subset of columns of size about S be about orthogonal? And for what values of S is this possible?

While it might be difficult to exhibit a matrix which provably obeys the UUP for very large values of S, we know that trivial randomized constructions will do so with overwhelming probability. We give an example. Sample N vectors on the unit sphere of $\mathbb{R}^K$ independently and uniformly at random. Then the condition of Theorems 3.1 and 3.2 hold for $S = O(K/\log(N/K))$ with probability $1 - \pi_N$ where $\pi_N = O(e^{-\gamma N})$ for some $\gamma > 0$. The reason why this holds may be explained by some sort of "blessing of high-dimensionality." Because the high-dimensional sphere is mostly empty, it is possible to pack many vectors while maintaining approximate orthogonality.

- *Gaussian measurements.* Here we assume that the entries of the K by N sensing matrix Φ are independently sampled from the normal distribution with mean zero and variance $1/K$. Then if

$$S \leq C \cdot K/\log(N/K), \tag{3.8}$$

  S obeys the condition of Theorems 3.1 and 3.2 with probability $1 - O(e^{-\gamma N})$ for some $\gamma > 0$. The proof uses known concentration results about the singular values of Gaussian matrices [16], [45].

- *Binary measurements.* Suppose that the entries of the K by N sensing matrix Φ are independently sampled from the symmetric Bernoulli distribution $P(\Phi_{ki} = \pm 1/\sqrt{K}) = 1/2$. Then it is conjectured that the conditions of Theorems 3.1 and 3.2 are satisfied with probability $1 - O(e^{-\gamma N})$ for some $\gamma > 0$ provided that S obeys (3.8). The proof of this fact would probably follow from new concentration results about the smallest singular value of a subgaussian matrix [38]. Note that the exact reconstruction property for S-sparse signals and (3.7) with S obeying (3.8) are known to hold for binary measurements [7].

- *Fourier measurements.* Suppose now that Φ is a partial Fourier matrix obtained by selecting K rows uniformly at random as before, and renormalizing the columns so that they are unit-normed. Then Candès and Tao [7] showed that Theorem 3.1 holds with overwhelming probability if $S \leq C \cdot K/(\log N)^6$. Recently, Rudelson and Vershynin [43] improved this result and established $S \leq C \cdot K/(\log N)^4$. This result is nontrivial and use sophisticated techniques from geometric functional analysis and probability in Banach spaces. It is conjectured that $S \leq C \cdot K/\log N$ holds.

- *Incoherent measurements.* Suppose now that Φ is obtained by selecting K rows uniformly at random from an N by N orthonormal matrix U and renormalizing the columns so that they are unit-normed. As before, we could think of U as the matrix $\Phi\Psi^*$ which maps the object from the Ψ to the Φ-domain. Then the arguments used in [7], [43] to prove that the UUP holds for incomplete Fourier matrices extend to this more general situation. In particular, Theorem 3.1 holds with overwhelming probability provided that

$$S \le C \cdot \frac{1}{\mu^2} \cdot \frac{K}{(\log N)^4}, \tag{3.9}$$

  where $\mu := \sqrt{N}\,\max_{i,j}|U_{i,j}|$ (observe that for the Fourier matrix, $\mu = 1$ which gives the result in the special case of the Fourier ensemble above). With $U = \Phi\Psi^*$,

$$\mu := \sqrt{N}\,\max_{i,j}|\langle\varphi_i, \psi_j\rangle| \tag{3.10}$$

  which is referred to as the mutual coherence between the measurement basis Φ and the sparsity basis Ψ [27], [28]. The greater the incoherence of the measurement/sparsity pair (Φ, Ψ), the smaller the number of measurements needed.

In short, one can establish the UUP for a few interesting random ensembles and we expect that in the future, many more results of this type will become available.

3.5. Optimality. Before concluding this section, it is interesting to specialize our recovery theorems to selected measurement ensembles now that we have established the UUP for concrete values of S. Consider the Gaussian measurement ensemble in which the entries of Φ are i.i.d. $N(0, 1/K)$. Our results say that one can recover any S-sparse vector from a random projection of dimension about $O(S \cdot \log(N/S))$, see also [18]. Next, suppose that x is taken from a weak-ℓ_p ball of radius R for some $0 < p < 1$, or from the ℓ_1-ball of radius R for $p = 1$. Then we have shown that for all $x \in w\ell_p(R)$

$$\|x^\star - x\|_{\ell_2} \le C \cdot R \cdot (K/\log(N/K))^{-r}, \quad r = 1/p - 1/2, \tag{3.11}$$

which has also been proven in [20]. An important question is whether this is optimal. In other words, can we find a possibly adaptive set of measurements and a reconstruction algorithm that would yield a better bound than (3.11)? By adaptive, we mean that one could use a sequential measurement procedure where at each stage, one would have the option to decide which linear functional to use next based on the data collected up to that stage.

It proves to be the case that one cannot improve on (3.11), and we have thus identified the optimal performance. Fix a class of object $\mathcal{F}$ and let $E_K(\mathcal{F})$ be the best reconstruction error from K linear measurements

$$E_K(\mathcal{F}) = \inf \sup_{f\in\mathcal{F}} \|f - D(y)\|_{\ell_2}, \quad y = \Phi f, \tag{3.12}$$

where the infimum is taken over all set of K linear functionals and all reconstruction algorithms D. Then it turns out $E_K(\mathcal{F})$ nearly equals the *Gelfand* numbers of a class $\mathcal{F}$ defined as

$$d_K(\mathcal{F}) = \inf_V \{\sup_{f\in\mathcal{F}} \|P_V f\| : \mathrm{codim}(V) < K\}, \tag{3.13}$$

where P_V is the orthonormal projection on the subspace V. Gelfand numbers play an important role in approximation theory, see [40] for more information. If $\mathcal{F} = -\mathcal{F}$ and $\mathcal{F} = \mathcal{F}+\mathcal{F} \le c_{\mathcal{F}}\mathcal{F}$, then $d_K(\mathcal{F}) \le E_K(\mathcal{F}) \le c_{\mathcal{F}} d_K(\mathcal{F})$. Note that $c_{\mathcal{F}} = 2^{1/p}$ in the case where $\mathcal{F}$ is a weak-ℓ_p ball. The thing is that we know the approximate values of the Gelfand numbers for many classes of interest. Suppose for example that $\mathcal{F}$ is the ℓ_1-ball of radius R. A seminal result of Kashin [35] and improved by Garnaev and Gluskin [31] shows that for this ball, the Gelfand numbers obey

$$C_1 \cdot R \cdot \sqrt{\frac{\log(N/K)+1}{K}} \le d_k(\mathcal{F}) \le C_2 \cdot R \cdot \sqrt{\frac{\log(N/K)+1}{K}}, \tag{3.14}$$

where C_1, C_2 are universal constants. Gelfand numbers are also approximately known for weak-ℓ_p balls as well; the only difference is that $((\log(N/K)+1)/K)^r$ substitutes $((\log(N/K)+1)/K)^{1/2}$. Hence, Kashin, Garnaev and Gluskin assert that with K measurements, the minimal reconstruction error (3.12) one can hope for is bounded below by a constant times $(K/\log(N/K))^{-r}$. Kashin's arguments [35] also used probabilistic functionals which establish the existence of recovery procedures for which the reconstruction error is bounded above by the right-hand side of (3.14). Similar types of recovery have also been known to be possible in the literature of theoretical computer science, at least in principle, for certain types of random measurements [1].

In this sense, our results – specialized to Gaussian measurements – are optimal for weak-ℓ_p norms. The novelty is that the information about the object can be retrieved from random coefficients by minimizing a simple linear program (2.3), and that the decoding algorithm adapts automatically to the weak-ℓ_p signal class, without knowledge thereof. Minimizing the ℓ_1-norm is adaptive and nearly gives the best possible reconstruction error simultaneously over a wide range of sparse classes of signals; no information about p and the radius R are required.

4. Robust compressive sampling

In any realistic application, we cannot expect to measure Φx without any error, and we now turn our attention to the robustness of compressive sampling vis a vis measurement errors. This is a very important issue because any real-world sensor is subject to at least a small amount of noise. And one thus immediately understands that to be widely applicable, the methodology needs to be stable. Small perturbations in the

observed data should induce small perturbations in the reconstruction. Fortunately, the recovery procedures may be adapted to be surprisingly stable and robust vis a vis arbitrary perturbations.

Suppose our observations are inaccurate and consider the model

$$y = \Phi x + e, \tag{4.1}$$

where e is a stochastic or deterministic error term with bounded energy $\|e\|_{\ell_2} \le \varepsilon$. Because we have inaccurate measurements, we now use a noise-aware variant of (2.3) which relaxes the data fidelity term. We propose a reconstruction program of the form

$$(\mathrm{P}_2) \qquad \min \ \|\tilde{x}\|_{\ell_1} \quad \text{such that} \quad \|\Phi\tilde{x} - y\|_{\ell_2} \le \varepsilon. \tag{4.2}$$

The difference with (P_1) is that we only ask the reconstruction be consistent with the data in the sense that $y - \Phi x^\star$ be within the noise level. The program (P_2) has a unique solution, is again convex, and is a special instance of a second order cone program (SOCP) [4].

Theorem 4.1 ([9]). *Suppose that x is an arbitrary vector in $\mathbb{R}^N$. Under the hypothesis of Theorem* 3.2*, the solution $x^\star$ to* (P_2) *obeys*

$$\|x^\star - x\|_{\ell_2} \le C_{1,S} \cdot \varepsilon + C_{2,S} \cdot \frac{\|x_0 - x_{0,S}\|_{\ell_1}}{\sqrt{S}}. \tag{4.3}$$

For reasonable values of δ_{4S} the constants in (4.3) *are well behaved, see* [9].

We would like to offer two comments. The first is that the reconstruction error is finite. This quiet observation is noteworthy because we recall that the matrix Φ is rectangular with many more columns than rows – thus having a fraction of vanishing singular values. Having said that, the mere fact that the severely ill-posed matrix inversion keeps the perturbation from "blowing up" may seem a little unexpected. Next and upon closer inspection, one sees that the reconstruction error is the sum of two terms: the first is simply proportional to the size of the measurement error while the second is the approximation error one would obtain in the noiseless case. In other words, the performance of the reconstruction degrades gracefully as the measurement noise increases. This brings us to our second point. In fact, it is not difficult to see that no recovery method can perform fundamentally better for arbitrary perturbations of size ε [9]. For related results for Gaussian sensing matrices, see [19].

5. Connections with statistical estimation

In the remainder of this paper, we shall briefly explore some connections with other fields, and we begin with statistics. Suppose now that the measurement errors in (4.1) are stochastic. More explicitly, suppose that the model is of the form

$$y = \Phi x + z, \tag{5.1}$$

where $z_1, \ldots, z_k$ are i.i.d. with mean zero and variance σ^2. In this section, we will assume that the z_k's are Gaussian although nothing in our arguments heavily relies upon this assumption. The problem is again to recover x from y which is a central problem in statistics since this is just the classical multivariate linear regression problem. Because the practical environment has changed dramatically over the last two decades or so, applications have emerged in which the number of observations is small compared to the dimension of the object we wish to estimate – here, $K \leq N$. This new paradigm sometimes referred to as "high-dimensional data" is currently receiving much attention and, clearly, the emerging theory of compressive sampling might prove very relevant.

The results from the previous sections are directly applicable. Suppose that x is S-sparse to simplify our exposition. Because $\|z\|_{\ell_2}^2$ is distributed as a chi-squared with K degrees of freedom, the reconstruction (4.2) would obey

$$\|x^* - x\|_{\ell_2}^2 \leq C \cdot K\sigma^2 \tag{5.2}$$

with high probability. While this may seem acceptable to the nonspecialist, modern results in the literature suggest that one might be able to get a better accuracy. In particular, one would like an adaptive error bound which depends upon the complexity of the true unknown parameter vector $x \in \mathbb{R}^N$. For example, if x only has S significant coefficients, we would desire an error bound of size about $S\sigma^2$; the less complex the estimand, the smaller the squared-error loss. This poses an important question: can we design an estimator whose accuracy depends upon the information content of the object we wish to recover?

5.1. Ideal model selection. To get a sense of what is possible, consider regressing the data y onto an arbitrary subset T by the method of least squares. Define $\hat{x}[T]$ to be the least squares estimate whose restriction to the set T is given by

$$\hat{x}_T[T] = (\Phi_T^T \Phi_T)^{-1} \Phi_T^T y, \tag{5.3}$$

and which vanishes outside T. Above, $\hat{x}_T[T]$ is the restriction of $\hat{x}[T]$ to T and similarly for x_T. Since $\hat{x}[T]$ vanishes outside T, we have

$$\mathbf{E}\|x - \hat{x}[T]\|^2 = \|x_T - \hat{x}_T[T]\|^2 + \sum_{i \notin T} |x_i|^2,$$

Consider the first term. We have

$$x_T - \hat{x}_T[T] = (\Phi_T^T \Phi_T)^{-1} \Phi_T^T (s + z),$$

where $s = \Phi_{T^c} x_{T^c}$. It follows that

$$\mathbf{E}\|x_T - \hat{x}_T[T]\|^2 = \|(\Phi_T^T \Phi_T)^{-1} \Phi_T^T s\|^2 + \sigma^2 \operatorname{Tr}((\Phi_T^T \Phi_T)^{-1}).$$

However, since all the eigenvalues of $\Phi_T^T \Phi_T$ belong to the interval $[1-\delta_{|T|}, 1+\delta_{|T|}]$, we have

$$\mathbf{E}\|x_T - \hat{x}_T[T]\|^2 \geq \frac{1}{1+\delta_{|T|}} \cdot |T| \cdot \sigma^2.$$

For each set T with $|T| \leq S$ and $\delta_S < 1$, we then have

$$\mathbf{E}\|x - \hat{x}[T]\|^2 \geq \sum_{i \in T^c c} x_i^2 + \frac{1}{2}\,|T| \cdot \sigma^2.$$

We now search for an *ideal estimator* which selects that estimator $\hat{x}[T^*]$ from the family $(\hat{x}[T])_{T \subset \{1,\ldots,N\}}$ with minimal Mean-Squared Error (MSE):

$$\hat{x}[T^*] = \mathrm{argmin}_{T \subset \{1,\ldots,N\}}\ \mathbf{E}\|x - \hat{x}[T]\|^2.$$

This estimator is ideal because we would of course not know which estimator $\hat{x}_T$ is best; that is, to achieve the ideal MSE, one would need an oracle that would tell us which model T to choose.

We will consider this ideal estimator nevertheless and take its MSE as a benchmark. The ideal MSE is bounded below by

$$\begin{aligned}\mathbf{E}\|x - \hat{x}[T^*]\|^2 &\geq \frac{1}{2}\ \min_T \left(\|x - \hat{x}[T]\|^2 + |T| \cdot \sigma^2\right) \\ &= \frac{1}{2} \sum_i \min(x_i^2, \sigma^2). \end{aligned} \tag{5.4}$$

Letting x_S be the best S-sparse approximation to x, another way to express the right-hand side (5.4) is in term of the classical trade-off between the approximation error and the number of terms being estimated times the noise level

$$\mathbf{E}\|x - \hat{x}_{T^*}\|^2 \geq \frac{1}{2} \inf_{S \geq 0} \left(\|x - x_S\|^2 + S\sigma^2\right).$$

Our question is of course whether there is a computationally efficient estimator which can mimic the ideal MSE.

5.2. The Dantzig selector. Assume for simplicity that the columns of Φ are normalized (there are straightforward variations to handle the general case). Then the Dantzig selector estimates x by solving the convex program

$$\text{(DS)} \qquad \min_{\tilde{x} \in \mathbb{R}^N} \|\tilde{x}\|_{\ell_1} \quad \text{subject to} \quad \sup_{1 \leq i \leq N} |(\Phi^T r)_i| \leq \lambda \cdot \sigma \tag{5.5}$$

for some $\lambda > 0$, and where r is the vector of residuals

$$r = y - \Phi\tilde{x}. \tag{5.6}$$

The solution to this optimization problem is the minimum ℓ_1-vector which is consistent with the observations. The constraints impose that the residual vector is within the noise level and does not correlate too well with the columns of Φ. For information, there exist related, yet different proposals in the literature, and most notably the lasso introduced by [47], see also [15]. Again, the program (DS) is convex and can be recast as a linear program (LP).

The main result in this line of research is that the Dantzig selector is not only computationally tractable, it is also accurate.

Theorem 5.1 ([10]). *Set* $\lambda := (1 + t^{-1})\sqrt{2\log p}$ *in* (5.5) *and suppose that* x *is* S*-sparse with* $\delta_{2S} + \theta_{S,2S} < 1 - t$. *Then with very large probability, the Dantzig selector* $\hat{x}$ *solution to* (5.5) *obeys*

$$\|\hat{x} - x\|^2 \leq O(\log p) \cdot \Big(\sigma^2 + \sum_i \min(x_i^2, \sigma^2)\Big). \tag{5.7}$$

Our result says that the Dantzig selector achieves a loss within a logarithmic factor of the ideal mean squared error one would achieve with an *oracle* which would supply perfect information about which coordinates are nonzero, and which were above the noise level. To be complete, it is possible to obtain similar bounds on the MSE.

There are extensions of this result to signals which are not sparse but compressible, e.g. for signals which belong to weak-ℓ_p balls. What is interesting here is that in some instances, even though the number of measurements is much smaller than the dimension of the parameter vector x, the Dantzig selector recovers the minimax rate that one would get if we were able to measure all the coordinates of x *directly* via $\tilde{y} = x + \sigma z$ where z is i.i.d. $N(0, 1)$.

6. Connections with error correction

Compressive sampling also interacts with the agenda of coding theory. Imagine we wish to transmit a vector x of length M to a remote receiver reliably. A frequently discussed approach consists in encoding the information x with an N by M coding matrix C with $N > M$. Assume that gross errors occur upon transmission and that a fraction of the entries of Cx are corrupted in a completely arbitrary fashion. We do not know which entries are affected nor do we know how they are affected. Is it possible to recover the information x exactly from the corrupted N-dimensional vector y?

To decode, [8] proposes solving the minimum ℓ_1-approximation problem

$$(\mathrm{D}_1) \qquad \min_{\tilde{x} \in \mathbb{R}^M} \|y - C\tilde{x}\|_{\ell_1}, \tag{6.1}$$

which can also be obviously recast as an LP. The result is that if C is carefully chosen, then (6.1) will correctly retrieve the information x with no error provided that the

fraction ρ of errors is not too large, $\rho \leq \rho^*$. This phenomenon holds for all x's and all corruption patterns.

To see why this phenomenon occurs, consider a matrix B which annihilates the $N \times M$ coding matrix C on the left, i.e. such that $BC = 0$; B is called a parity-check matrix and is any $(N - M) \times N$ matrix whose kernel is the range of C in $\mathbb{R}^N$. The transmitted information is of the form $y = Cx + e$, where e is a sparse vector of possibly gross errors, and apply B on both sides of this equation. This gives

$$\tilde{y} = B(Cx + e) = Be \tag{6.2}$$

since $BC = 0$. Therefore, the decoding problem is reduced to that of recovering the error vector e from the observations Be. Once e is known, Cx is known and, therefore, x is also known since we may just assume that C has full rank.

Now the reader knows that we could solve the underdetermined system (6.2) by ℓ_1-minimization. He also knows that if the UUP holds, the recovery is exact. Now (D_1) and (P_1) are equivalent programs. Indeed, it follows from the decomposition $\tilde{x} = x + h$ that

$$(\mathrm{D}_1) \iff \min_{h \in \mathbb{R}^M} \|e - Ch\|_{\ell_1}.$$

Now the constraint $Bd = Be$ means that $d = e - Ah$ for some $h \in \mathbb{R}^M$ and, therefore,

$$\begin{aligned} \min \|d\|_{\ell_1}, \quad Bd = Be &\iff \min_{h \in \mathbb{R}^n} \|d\|_{\ell_1}, \quad d = e - Ah \\ &\iff \min_{h \in \mathbb{R}^n} \|e - Ah\|_{\ell_1}, \end{aligned}$$

which proves the claim.

Hence, if one uses a random coding matrix which is a popular choice, we have the following result, see also [42]:

Theorem 6.1 ([8]). *Suppose the coding matrix C has i.i.d. $N(0, 1)$ entries. Then with probability exceeding $1 - O(e^{-\gamma M})$ for some $\gamma > 0$, (D_1) exactly decodes all $x \in \mathbb{R}^M$ provided that the fraction ρ of arbitrary errors obeys $\rho \leq \rho^*(M, N)$.*

In conclusion, one can correct a constant fraction of errors with arbitrary magnitudes by solving a convenient LP. In [8], the authors reported on numerical results showing that in practice (D_1) works extremely well and recovers the vector x exactly provided that the fraction of the corrupted entries be less than about 17% in the case where $N = 2M$ and less than about 34% in the case where $N = 4M$.

7. Further topics

Our intention in this short survey was merely to introduce the new compressive sampling concepts. We presented an approach based on the notion of uncertainty principle

which gives a powerful and unified treatment of some of the main results underlying this theory. As we have seen, the UUP gives conditions for exact, approximate, and stable recoveries which are almost necessary. Another advantage that one can hardly neglect is that this makes the exposition fairly simple. Having said that, the early papers on compressive sampling – e.g. [6], [7], [20] – have spurred a large and fascinating literature in which other approaches and ideas have been proposed. Rudelson and Vershynin have used tools from modern Banach space theory to derive powerful results for Gaussian ensembles [42], [14], [43]. In this area, Pajor and his colleagues have established the existence of abstract reconstruction procedures from subgaussian measurements (including random binary sensing matrices) with powerful reconstruction properties. In a different direction, Donoho and Tanner have leveraged results from polytope geometry to obtain very precise estimates about the minimal number of Gaussian measurements needed to reconstruct S-sparse signals [21], [23], see also [43]. Tropp and Gilbert reported results about the performance of greedy methods for compressive sampling [49]. Haupt and Nowak have quantified the performance of combinatorial optimization procedures for estimating a signal from undersampled random projections in noisy environments [34]. Finally, Rauhut has worked out variations on the Fourier sampling theorem in which a sparse continuous-time trigonometric polynomials is randomly sampled in time [41]. Because of space limitations, we are unfortunately unable to do complete justice to this rapidly growing literature.

We would like to emphasize that there are many aspects of compressive sampling that we have not touched. For example, we have not discussed the practical performance of this new theory. In fact, numerical experiments have shown that compressive sampling behaves extremely well in practice. For example, it has been shown that from $3S-4S$ nonadaptive measurements, one can reconstruct an approximation of an image in a fixed basis which is more precise than that one would get by measuring all the coefficients of the object in that basis and selecting the S largest [13], [50]. Further, numerical simulations with noisy data show that compressive sampling is very stable and performs well in noisy environments. In practice, the constants appearing in Theorems 4.1 and 5.1 are very small, see [9] and [10] for empirical results.

We would like to close this article by returning to the main theme of this paper, which is that compressive sampling invites to rethink sensing mechanisms. Because if one were to collect a comparably small number of general linear measurements rather than the usual pixels, one could in principle reconstruct an image with essentially the same resolution as that one would obtain by measuring all the pixels. Therefore, if one could design incoherent sensors (i.e. measuring incoherent linear functionals), the payoff could be extremely large. Several teams have already reported progress in this direction. For example, a team led by Baraniuk and Kelly have proposed a new camera architecture that employs a digital micromirror array to perform optical calculations of linear projections of an image onto pseudorandom binary patterns [46], [52]. Compressive sampling may also address challenges in the processing of wideband radio frequency signals since high-speed analog-to-digital convertor

technology indicates that current capabilities fall well short of needs, and that hardware implementations of high precision Shannon-based conversion seem out of sight for decades to come. Finally, compressive sampling has already found applications in wireless sensor networks [2]. Here, compressive sampling allows of *energy efficient* estimation of sensor data with comparably few sensor nodes. The power of these estimation schemes is that they require no prior information about the sensed data. All these applications are novel and exciting. Others might just be around the corner.

References

[1] Alon, N., Matias, Y., Szegedy, B., The space complexity of approximating the frequency moments. *J. Comput. System Sci.* **58** (1999), 137–147.

[2] Bajwa, W. U., Haupt, J., Sayeed, A. M., Nowak, R., Compressive wireless sensing. In *Proc. 5th Intl. Conf. on Information Processing in Sensor Networks* (IPSN '06), Nashville, TN, 2006, 134–142.

[3] Bloomfield, P., Steiger, W., *Least Absolute Deviations: Theory, Applications, and Algorithms*. Progr. Probab. Statist. 6, Birkhäuser, Boston, MA, 1983.

[4] Boyd, S., Vandenberghe, L., *Convex Optimization*. Cambridge University Press, Cambridge 2004.

[5] Candès, E. J., Donoho, D. L. New tight frames of curvelets and optimal Representations of objects with piecewise C^2 singularities. *Comm. Pure Appl. Math.* **57** (2004), 219–266.

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

[7] Candès, E. J., Tao, T., Near-optimal signal recovery from random projections and universal encoding strategies. *IEEE Trans. Inform. Theory*, 2004, submitted.

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

[9] Candès, E. J., Romberg, J., Tao, T., Signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.* **59** (8) (2005), 1207–1223.

[10] Candès, E. J., Tao, T., The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, to appear.

[11] Candès, E. J., Romberg, J., The role of sparsity and incoherence for exactly reconstructing a signal from limited measurements. Technical Report, California Institute of Technology, 2004.

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

[13] Candès, E. J., Romberg, J., Practical signal recovery from random projections. In *SPIE International Symposium on Electronic Imaging: Computational Imaging III*, San Jose, California, January 2005.

[14] Candès, E. J., Rudelson, M., Vershynin, R. and Tao, T. Error correction via linear programming. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science (FOCS)* (2005), IEEE Comput. Soc. Press, LosAlamitos, CA,295–308.

[15] Chen, S. S., Donoho, D. L., Saunders, M. A, Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** (1999), 33–61.

[16] Davidson, K. R., Szarek, S. J., Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces* (ed. by W. B. Johnson, J. Lindenstrauss), Vol. I, North-Holland, Amsterdam 2001, 317–366; Corrigendum, Vol. 2, 2003, 1819–1820.

[17] DeVore, R. A., Optimal computation. In *Proceedings of the International Congress of Mathematicians* (Madrid, 2006), Volume 1, EMS Publishing House, Zürich 2006.

[18] Donoho, D. L., For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest Solution. *Comm. Pure Appl. Math.* **59** (2006), 797–829.

[19] Donoho, D. L., For most large underdetermined systems of equations, the minimal ℓ_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.* **59** (2006), 907–934.

[20] Donoho, D. L., Compressed sensing. Technical Report, Stanford University, 2004.

[21] Donoho, D. L., Neighborly polytopes and sparse solutions of underdetermined linear equations. Technical Report, Stanford University, 2005.

[22] Donoho, D. L., Logan, B. F., Signal recovery and the large sieve, *SIAM J. Appl. Math.* **52** (1992), 577–591.

[23] Donoho, D. L., Tanner, J., Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** (2005), 9452–9457.

[24] Donoho, D. L., Vetterli, M., DeVore, R. A., Daubechies, I., Data compression and harmonic analysis. *IEEE Trans. Inform. Theory* **44** (1998), 2435–2476.

[25] Donoho, D. L., Stark, P. B., Uncertainty principles and signal recovery. *SIAM J. Appl. Math.* **49** (1989), 906–931.

[26] Donoho, D. L., Huo, X., Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory* **47** (2001), 2845–2862.

[27] Donoho, D. L., Elad, M., Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Natl. Acad. Sci. USA* **100** (2003), 2197–2202.

[28] Elad, M., Bruckstein, A. M., A generalized uncertainty principle and sparse representation in pairs of $\mathbb{R}^N$ bases. *IEEE Trans. Inform. Theory* **48** (2002), 2558–2567.

[29] Feuer, A., Nemirovski, A., On sparse representation in pairs of bases. *IEEE Trans. Inform. Theory* **49** (2003), 1579–1581.

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

[31] Garnaev, A., Gluskin, E., The widths of a Euclidean ball. *Dokl. Akad. Nauk. USSR* **277** (1984), 1048–1052; English transl. *Soviet Math. Dokl.* **30** (1984), 200–204.

[32] Gilbert, A. C., Muthukrishnan, S., Strauss, M., Improved time bounds for near-optimal sparse Fourier representation. In *Proceedings of SPIE 5914* (Wavelets XI), ed. by M. Papadakis, A. F. Laine, M. A. Unser, 2005.

[33] Gribonval, R., Nielsen, M., Sparse representations in unions of bases. *IEEE Trans. Inform. Theory* **49** (2003), 3320–3325.

[34] Haupt, J., Nowak, R., Signal reconstruction from noisy random projections. *IEEE Trans. Inform. Theory*, submitted.

[35] Kashin, B., The widths of certain finite dimensional sets and classes of smooth functions, *Izvestia* **41** (1977), 334–351.

[36] Mallat, S., *A Wavelet Tour of Signal Processing*. Academic Press, San Diego, CA, 1998.

[37] Mansour, Y., Randomized interpolation and approximation of sparse polynomials. *SIAM J. Comput.* **24** (1995), 357–368.

[38] Litvak, A. E., Pajor, A., Rudelson, M., Tomczak-Jaegermann, N., Smallest singular value of random matrices and geometry of random polytopes. Manuscript, 2004.

[39] Mendelson, S., Pajor, A., Tomczak-Jaegermann, N., Reconstruction and subgaussian processes. *C. R. Math. Acad. Sci. Paris* **340** (2005), 885–888.

[40] Pinkus, A., *N-Widths in Approximation Theory*. Ergeb. Math. Grenzgeb. (3) 7, Springer-Verlag, Berlin 1985.

[41] Rauhut, H., Random sampling of sparse trigonometric polynomials. Preprint, 2005.

[42] Rudelson, M., Vershynin, R., Geometric approach to error-correcting codes and reconstruction of signals. *Internat. Math. Res. Notices* **2005** (64) (2005), 4019–4041.

[43] Rudelson, M., Vershynin, R., Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements. Preprint, 2006.

[44] Santosa, F., Symes, W. W., Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Statist. Comput.* **7** (1986), 1307–1330.

[45] Szarek, S. J., Condition numbers of random matrices. *J. Complexity* **7** (1991), 131–149.

[46] Takhar, D., Laska, J. N., Wakin, M., Duarte, M. F., Baron, D., Sarvotham, S., Kelly, K. F., Baraniuk, R. G., A new compressive imaging camera architecture using optical-domain compression. *IS&T/SPIE Computational Imaging IV*, San Jose, January 2006.

[47] Tibshirani, R., Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** (1996), 267–288.

[48] Tropp, J. A., Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory* **52** (2006), 1030–1051.

[49] Tropp, J. A., Gilbert, A. C., Signal recovery from partial information via orthogonal matching pursuit. Preprint, University of Michigan, 2005.

[50] Tsaig, Y., Donoho, D. L., Extensions of compressed sensing. Technical report, Department of Statistics, Stanford University, 2004.

[51] Vetterli, M., Marziliano, P., Blu, T., Sampling signals with finite rate of innovation. *IEEE Trans. Signal Process.* **50** (2002), 1417–1428.

[52] Wakin, M., Laska, J. N., Duarte, M. F., Baron, D., Sarvotham, S., Takhar, D., Kelly, K. F., Baraniuk, R. G., Compressive imaging for video representation and coding. *Picture Coding Symposium*, special session on Next Generation Video Representation, Beijing, April 2006.

Applied and Computational Mathematics, California Institute of Technology, Pasadena, CA 91125, U.S.A.

E-mail: emmanuel@acm.caltech.edu