An Introduction To Compressive Sampling

[A sensing/sampling paradigm that goes against the common knowledge in data acquisition]

[Emmanuel J. Candès and Michael B. Wakin]

Conventional approaches to sampling signals or images follow Shannon's celebrated theorem: the sampling rate must be at least twice the maximum frequency present in the signal (the so-called Nyquist rate). In fact, this principle underlies nearly all signal acquisition protocols used in consumer audio and visual electronics, medical imaging devices, radio receivers, and so on. (For some signals, such as images that are not naturally bandlimited, the sampling rate is dictated not by the Shannon theorem but by the desired temporal or spatial resolution. However, it is common in such systems to use an antialiasing low-pass filter to bandlimit the signal before sampling, and so the Shannon theorem plays an implicit role.) In the field of data conversion, for example, standard analog-to-digital converter (ADC) technology implements the usual quantized Shannon representation: the signal is uniformly sampled at or above the Nyquist rate.

Digital Object Identifier 10.1109/MSP.2007.914731

This article surveys the theory of compressive sampling, also known as compressed sensing or CS, a novel sensing/sampling paradigm that goes against the common wisdom in data acquisition. CS theory asserts that one can recover certain signals and images from far fewer samples or measurements than traditional methods use. To make this possible, CS relies on two principles: *sparsity*, which pertains to the signals of interest, and *incoherence*, which pertains to the sensing modality.

- Sparsity expresses the idea that the "information rate" of a continuous time signal may be much smaller than suggested by its bandwidth, or that a discrete-time signal depends on a number of degrees of freedom which is comparably much smaller than its (finite) length. More precisely, CS exploits the fact that many natural signals are sparse or compressible in the sense that they have concise representations when expressed in the proper basis Ψ.
- Incoherence extends the duality between time and frequency and expresses the idea that objects having a sparse representation in Ψ must be spread out in the domain in which they are acquired, just as a Dirac or a spike in the time domain is spread out in the frequency domain. Put differently, incoherence says that unlike the signal of interest, the sampling/sensing waveforms have an extremely dense representation in Ψ.

CS THEORY ASSERTS THAT ONE CAN RECOVER CERTAIN SIGNALS AND IMAGES FROM FAR FEWER SAMPLES OR MEASUREMENTS THAN TRADITIONAL METHODS USE.

The crucial observation is that one can design efficient sensing or sampling protocols that capture the useful information content embedded in a sparse signal and condense it into a small amount of data. These protocols are nonadaptive and simply require correlating the signal with a small number of fixed waveforms that are incoherent with the sparsifying basis. What is most remarkable about these sampling protocols is that they allow a sensor to very efficiently capture the information in a sparse signal without trying to comprehend that signal. Further, there is a way to use numerical optimization to reconstruct the full-length signal from the small amount of collected data. In other words, CS is a very simple and efficient signal acquisition protocol which samples—in a signal independent fashion—at a low rate and later uses computational power for reconstruction from what appears to be an incomplete set of measurements.

Our intent in this article is to overview the basic CS theory that emerged in the works [1]–[3], present the key mathematical ideas underlying this theory, and survey a couple of important results in the field. Our goal is to explain CS as plainly as possible, and so our article is mainly of a tutorial nature. One of the charms of this theory is that it draws from various subdisciplines within the applied mathematical sciences, most notably probability theory. In this review, we have decided to highlight this aspect and especially the fact that randomness can—perhaps surprisingly—lead to very effective sensing mechanisms. We will also discuss significant implications, explain why CS is a concrete protocol for sensing and compressing data simultaneously (thus the name), and conclude our tour by reviewing important applications.

THE SENSING PROBLEM

In this article, we discuss sensing mechanisms in which information about a signal $f(t)$ is obtained by linear functionals recording the values

$$y_k = \langle f, \varphi_k \rangle, \qquad k = 1, \ldots, m. \tag{1}$$

That is, we simply correlate the object we wish to acquire with the waveforms $\varphi_k(t)$. This is a standard setup. If the sensing waveforms are Dirac delta functions (spikes), for example, then y is a vector of sampled values of f in the time or space domain. If the sensing waveforms are indicator functions of pixels, then y is the image data typically collected by sensors in a digital camera. If the sensing waveforms are sinusoids, then y is a vector of Fourier coefficients; this is the sensing modality used in magnetic resonance imaging (MRI). Other examples abound.

Although one could develop a CS theory of continuous time/space signals, we restrict our attention to discrete signals $f \in \mathbb{R}^n$. The reason is essentially twofold: first, this is conceptually simpler and second, the available discrete CS theory is far more developed (yet clearly paves the way for a continuous theory—see also "Applications"). Having said this, we are then interested in *undersampled* situations in which the number m of available measurements is much smaller than the dimension n of the signal f. Such problems are extremely common for a variety of reasons. For instance, the number of sensors may be limited. Or the measurements may be extremely expensive as in certain imaging processes via neutron scattering. Or the sensing process may be slow so that one can only measure the object a few times as in MRI. And so on.

These circumstances raise important questions. Is accurate reconstruction possible from $m \ll n$ measurements only? Is it possible to design $m \ll n$ sensing waveforms to capture almost all the information about f? And how can one approximate f from this information? Admittedly, this state of affairs looks rather daunting, as one would need to solve an underdetermined linear system of equations. Letting A denote the $m \times n$ sensing matrix with the vectors $\varphi_1^*, \ldots, \varphi_m^*$ as rows (a^* is the complex transpose of a), the process of recovering $f \in \mathbb{R}^n$ from $y = Af \in \mathbb{R}^m$ is ill-posed in general when $m < n$: there are infinitely many candidate signals $\tilde{f}$ for which $A\tilde{f} = y$. But one could perhaps imagine a way out by relying on realistic models of objects f which naturally exist. The Shannon

theory tells us that, if $f(t)$ actually has very low bandwidth, then a small number of (uniform) samples will suffice for recovery. As we will see in the remainder of this article, signal recovery can actually be made possible for a much broader class of signal models.

INCOHERENCE AND THE SENSING OF SPARSE SIGNALS

This section presents the two fundamental premises underlying CS: sparsity and incoherence.

SPARSITY

Many natural signals have concise representations when expressed in a convenient basis. Consider, for example, the image in Figure 1(a) and its wavelet transform in (b). Although nearly all the image pixels have nonzero values, the wavelet coefficients offer a concise summary: most coefficients are small, and the relatively few large coefficients capture most of the information.

Mathematically speaking, we have a vector $f \in \mathbb{R}^n$ (such as the n-pixel image in Figure 1) which we expand in an orthonormal basis (such as a wavelet basis) $\Psi = [\psi_1 \psi_2 \cdots \psi_n]$ as follows:

$$f(t) = \sum_{i=1}^{n} x_i \, \psi_i(t), \tag{2}$$

where x is the coefficient sequence of f, $x_i = \langle f, \psi_i \rangle$. It will be convenient to express f as Ψx (where Ψ is the $n \times n$ matrix with $\psi_1, \ldots, \psi_n$ as columns). The implication of sparsity is now clear: when a signal has a sparse expansion, one can discard the small coefficients without much perceptual loss. Formally, consider $f_S(t)$ obtained by keeping only the terms corresponding to the S largest values of (x_i) in the expansion (2). By definition, $f_S := \Psi x_S$, where here and below, x_S is the vector of coefficients (x_i) with all but the largest S set to zero. This vector is sparse in a strict sense since all but a few of its entries are zero; we will call *S-sparse* such objects with at most S nonzero entries. Since Ψ is an orthonormal basis (or "orthobasis"), we have $\|f - f_S\|_{\ell_2} = \|x - x_S\|_{\ell_2}$, and if x is sparse or *compressible* in the sense that the sorted magnitudes of the (x_i) decay quickly, then x is well approximated by x_S and, therefore, the error $\|f - f_S\|_{\ell_2}$ is small. In plain terms, one can "throw away" a large fraction of the coefficients without much loss. Figure 1(c) shows an example where the perceptual loss is hardly noticeable from a megapixel image to its approximation obtained by throwing away 97.5% of the coefficients.

This principle is, of course, what underlies most modern lossy coders such as JPEG-2000 [4] and many others, since a simple method for data compression would be to compute x from f and then (adaptively) encode the locations and values of the S significant coefficients. Such a process requires knowledge of all the n coefficients x, as the locations of the significant pieces of information may not be known in advance (they are signal dependent); in our example, they tend to be clustered around edges in the image. More generally, sparsity is a fundamental modeling tool which permits efficient fundamental signal processing; e.g., accurate statistical estimation and classification, efficient data compression, and so on. This article is about a more surprising and far-reaching implication, however, which is that sparsity has significant bearings on the acquisition process itself. Sparsity determines how efficiently one can acquire signals *nonadaptively*.

INCOHERENT SAMPLING

Suppose we are given a pair (Φ, Ψ) of orthobases of $\mathbb{R}^n$. The first basis Φ is used for sensing the object f as in (1) and the second is used to represent f. The restriction to pairs of orthobases is not essential and will merely simplify our treatment.

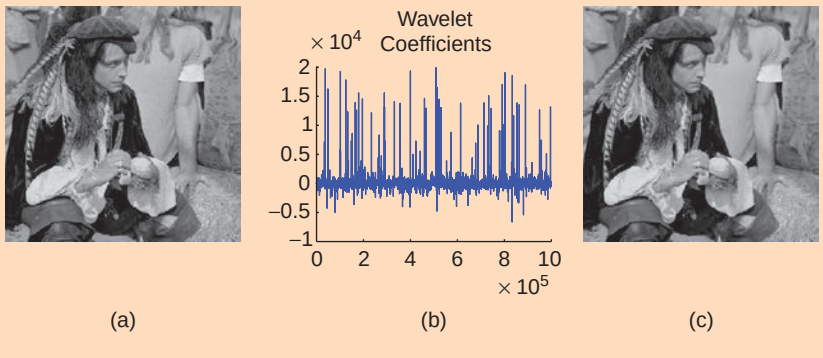
DEFINITION 1

The coherence between the sensing basis Φ and the representation basis Ψ is

$$\mu(\Phi, \Psi) = \sqrt{n} \cdot \max_{1 \le k, j \le n} |\langle \varphi_k, \psi_j \rangle|. \tag{3}$$

In plain English, the coherence measures the largest correlation between any two elements of Φ and Ψ; see also [5]. If Φ and Ψ contain correlated elements, the coherence is large. Otherwise, it is small. As for how large and how small, it follows from linear algebra that $\mu(\Phi, \Psi) \in [1, \sqrt{n}]$.

Compressive sampling is mainly concerned with low coherence pairs, and we now give examples of such pairs. In our first example, Φ is the canonical or spike basis $\varphi_k(t) = \delta(t - k)$ and



[FIG1] (a) Original megapixel image with pixel values in the range [0,255] and (b) its wavelet transform coefficients (arranged in random order for enhanced visibility). Relatively few wavelet coefficients capture most of the signal energy; many such images are highly compressible. (c) The reconstruction obtained by zeroing out all the coefficients in the wavelet expansion but the 25,000 largest (pixel values are thresholded to the range [0,255]). The difference with the original picture is hardly noticeable. As we describe in "Undersampling and Sparse Signal Recovery," this image can be perfectly recovered from just 96,000 incoherent measurements.

Ψ is the Fourier basis, $\psi_j(t) = n^{-1/2} e^{i\,2\pi\,jt/n}$. Since Φ is the sensing matrix, this corresponds to the classical sampling scheme in time or space. The time-frequency pair obeys $\mu(\Phi, \Psi) = 1$ and, therefore, we have *maximal incoherence*. Further, spikes and sinusoids are maximally incoherent not just in one dimension but in any dimension, (in two dimensions, three dimensions, etc.)

Our second example takes wavelets bases for Ψ and noiselets [6] for Φ. The coherence between noiselets and Haar wavelets is $\sqrt{2}$ and that between noiselets and Daubechies D4 and D8 wavelets is, respectively, about 2.2 and 2.9 across a wide range of sample sizes n. This extends to higher dimensions as well. (Noiselets are also maximally incoherent with spikes and incoherent with the Fourier basis.) Our interest in noiselets comes from the fact that 1) they are incoherent with systems providing sparse representations of image data and other types of data, and 2) they come with very fast algorithms; the noiselet transform runs in $O(n)$ time, and just like the Fourier transform, the noiselet matrix does not need to be stored to be applied to a vector. This is of crucial practical importance for numerically efficient CS implementations.

Finally, random matrices are largely incoherent with any fixed basis Ψ. Select an orthobasis Φ uniformly at random, which can be done by orthonormalizing n vectors sampled independently and uniformly on the unit sphere. Then with high probability, the coherence between Φ and Ψ is about $\sqrt{2\log n}$. By extension, random waveforms $(\varphi_k(t))$ with independent identically distributed (i.i.d.) entries, e.g., Gaussian or ± 1 binary entries, will also exhibit a very low coherence with any fixed representation Ψ. Note the rather strange implication here; if sensing with incoherent systems is good, then efficient mechanisms ought to acquire correlations with random waveforms, e.g., white noise!

MANY NATURAL SIGNALS ARE SPARSE OR COMPRESSIBLE IN THE SENSE THAT THEY HAVE CONCISE REPRESENTATIONS WHEN EXPRESSED IN THE PROPER BASIS.

UNDERSAMPLING AND SPARSE SIGNAL RECOVERY

Ideally, we would like to measure all the n coefficients of f, but we only get to observe a subset of these and collect the data

$$y_k = \langle f, \varphi_k \rangle, \quad k \in M, \tag{4}$$

where $M \subset \{1, \ldots, n\}$ is a subset of cardinality $m < n$. With this information, we decide to recover the signal by ℓ_1-norm minimization; the proposed reconstruction $f^\star$ is given by $f^\star = \Psi x^\star$, where $x^\star$ is the solution to the convex optimization program ($\|x\|_{\ell_1} := \sum_i |x_i|$)

$$\min_{\tilde{x} \in \mathbb{R}^n} \|\tilde{x}\|_{\ell_1} \quad \text{subject to} \quad y_k = \langle \varphi_k, \Psi\tilde{x} \rangle, \quad \forall k \in M. \tag{5}$$

That is, among all objects $\tilde{f} = \Psi\tilde{x}$ consistent with the data, we pick that whose coefficient sequence has minimal ℓ_1 norm. (As is well known, minimizing ℓ_1 subject to linear equality constraints can easily be recast as a linear program making available a host of ever more efficient solution algorithms.)

The use of the ℓ_1 norm as a sparsity-promoting function traces back several decades. A leading early application was reflection seismology, in which a sparse reflection function (indicating meaningful changes between subsurface layers) was sought from bandlimited data [7], [8]. However, ℓ_1-minimization is not the only way to recover sparse solutions; other methods, such as greedy algorithms [9], have also been proposed.

Our first result asserts that when f is sufficiently sparse, the recovery via ℓ_1-minimization is provably exact.

THEOREM 1 [10]

Fix $f \in \mathbb{R}^n$ and suppose that the coefficient sequence x of f in the basis Ψ is S-sparse. Select m measurements in the Φ domain uniformly at random. Then if

$$m \geq C \cdot \mu^2(\Phi, \Psi) \cdot S \cdot \log n \tag{6}$$

for some positive constant C, the solution to (5) is exact with overwhelming probability. (It is shown that the probability of success exceeds $1 - \delta$ if $m \geq C \cdot \mu^2(\Phi, \Psi) \cdot S \cdot \log(n/\delta)$. In addition, the result is only guaranteed for nearly all sign sequences x with a fixed support, see [10] for details.)

We wish to make three comments:

1) The role of the coherence is completely transparent; the smaller the coherence, the fewer samples are needed, hence our emphasis on low coherence systems in the previous section.

2) One suffers no information loss by measuring just about any set of m coefficients which may be far less than the signal size apparently demands. If $\mu(\Phi, \Psi)$ is equal or close to one, then on the order of $S \log n$ samples suffice instead of n.

3) The signal f can be exactly recovered from our condensed data set by minimizing a convex functional which does not assume any knowledge about the number of nonzero coordinates of x, their locations, or their amplitudes which we assume are all completely unknown a priori. We just run the algorithm and if the signal happens to be sufficiently sparse, exact recovery occurs.

The theorem indeed suggests a very concrete acquisition protocol: sample nonadaptively in an incoherent domain and invoke linear programming after the acquisition step. Following this protocol would essentially acquire the signal in a compressed form. All that is needed is a decoder to "decompress" this data; this is the role of ℓ_1 minimization.

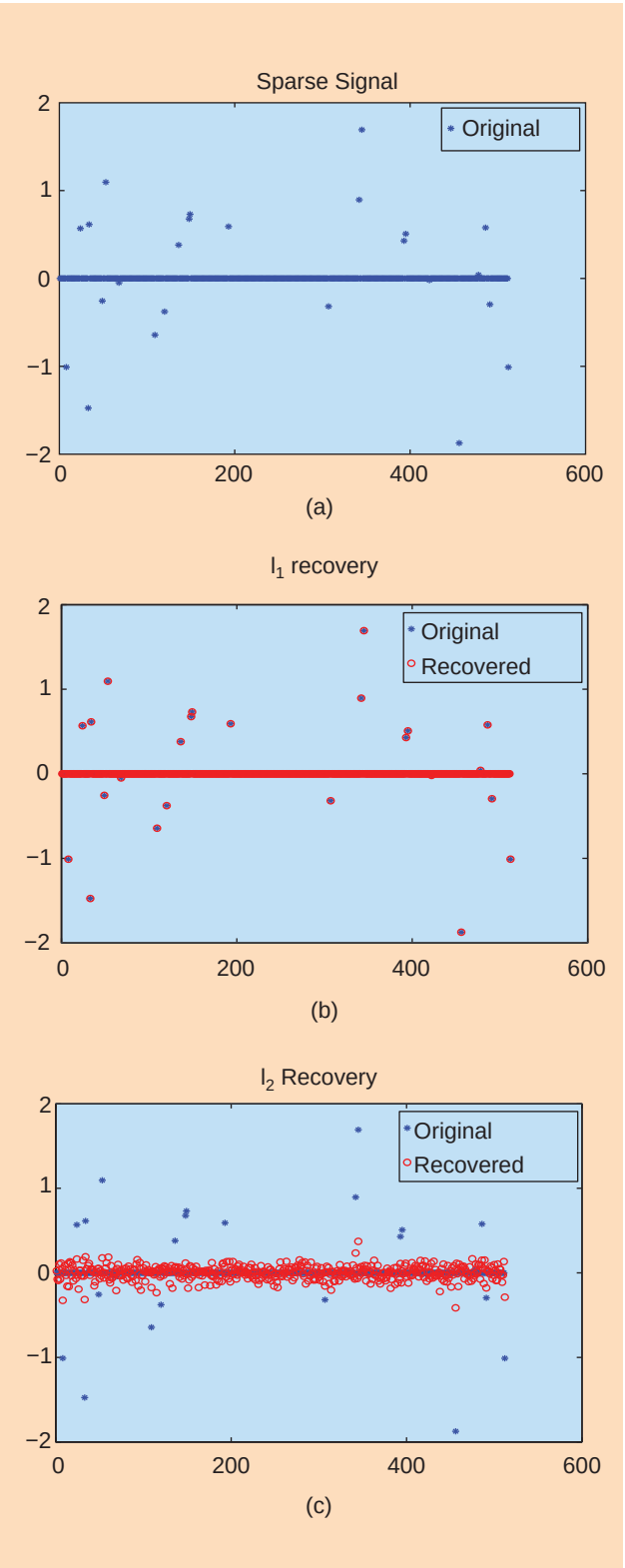
In truth, this random incoherent sampling theorem extends an earlier result about the sampling of spectrally sparse signals [1], which showed that randomness 1) can be a very effective

sensing mechanism and 2) is amenable to rigorous proofs, and thus perhaps triggered the many CS developments we have witnessed and continue to witness today. Suppose that we are interested in sampling ultra-wideband but spectrally sparse signals of the form $f(t) = \sum_{j=0}^{n-1} x_j e^{i2\pi jt/n}$, $t = 0, \ldots, n-1$, where n is very large but where the number of nonzero components x_j is less than or equal to S (which we should think of as comparably small). We do not know which frequencies are active nor do we know the amplitudes on this active set. Because the active set is not necessarily a subset of consecutive integers, the Nyquist/Shannon theory is mostly unhelpful (since one cannot restrict the bandwidth a priori, one may be led to believe that all n time samples are needed). In this special instance, Theorem 1 claims that one can reconstruct a signal with arbitrary and unknown frequency support of size S from on the order of $S \log n$ time samples, see [1]. What is more, these samples do not have to be carefully chosen; almost any sample set of this size will work. An illustrative example is provided in Figure 2. For other types of theoretical results in this direction using completely different ideas see [11]–[13].

It is now time to discuss the role played by probability in all of this. The key point is that to get useful and powerful results, one needs to resort to a probabilistic statement since one cannot hope for comparable results holding for all measurement sets of size m. Here is why. There are special sparse signals that vanish nearly everywhere in the Φ domain. In other words, one can find sparse signals f and very large subsets of size almost n (e.g., $n - S$) for which $y_k = \langle f, \varphi_k \rangle = 0$ for all $k \in M$. The interested reader may want to check the example of the Dirac comb discussed in [14] and [1]. On the one hand, given such subsets, one would get to see a stream of zeros and no algorithm whatsoever would of course be able reconstruct the signal. On the other hand, the theorem guarantees that the fraction of sets for which exact recovery does not occur is truly negligible (a large negative power of n). Thus, we only have to tolerate a probability of failure that is extremely small. For practical purposes, the probability of failure is zero provided that the sampling size is sufficiently large.

Interestingly, the study of special sparse signals discussed above also shows that one needs at least on the order of $\mu^2 \cdot S \cdot \log n$ samples as well. (We are well aware that there exist subsets of cardinality $2S$ in the time domain which can reconstruct any s-sparse signal in the frequency domain. Simply take $2s$ consecutive time points, see "What Is Comprehensive Sampling?" and [11] and [12], for example. But this is not what our claim is about. We want that most sets of a certain size provide exact reconstruction.) With fewer samples, the probability that information may be lost is just too high and reconstruction by any method, no matter how intractable, is impossible. In summary, when the coherence is one, say, we do not need more than $S \log n$ samples but we cannot do with fewer either.

We conclude this section with an incoherent sampling example, and consider the sparse image in Figure 1(c), which as we recall has only 25,000 nonzero wavelet coefficients. We then



[FIG2] (a) A sparse real valued signal and (b) its reconstruction from 60 (complex valued) Fourier coefficients by ℓ_1 minimization. The reconstruction is exact. (c) The minimum energy reconstruction obtained by substituting the ℓ_1 norm with the ℓ_2 norm; ℓ_1 and ℓ_2 give wildly different answers. The ℓ_2 solution does not provide a reasonable approximation to the original signal.

acquire information by taking 96,000 incoherent measurements (see [10] for the particulars of these measurements) and solve (5). The minimum-ℓ_1 recovery is *perfect*; that is, $f^\star = f$. This example shows that a number of samples just about $4\times$ the sparsity level suffices. Many researchers have reported on similar empirical successes. There is de facto a known four-to-one practical rule which says that for exact recovery, one needs about four incoherent samples per unknown nonzero term.

WHAT IS MOST REMARKABLE ABOUT THESE SAMPLING PROTOCOLS IS THAT THEY ALLOW A SENSOR TO VERY EFFICIENTLY CAPTURE THE INFORMATION IN A SPARSE SIGNAL WITHOUT TRYING TO COMPREHEND THAT SIGNAL.

ROBUST COMPRESSIVE SAMPLING

We have shown that one could recover sparse signals from just a few measurements but in order to be really powerful, CS needs to be able to deal with both nearly sparse signals and with noise. First, general objects of interest are not exactly sparse but approximately sparse. The issue here is whether or not it is possible to obtain accurate reconstructions of such objects from highly undersampled measurements. Second, in any real application measured data will invariably be corrupted by at least a small amount of noise as sensing devices do not have infinite precision. It is therefore imperative that CS be robust vis a vis such nonidealities. At the very least, small perturbations in the data should cause small perturbations in the reconstruction.

This section examines these two issues simultaneously. Before we begin, however, it will ease the exposition to consider the abstract problem of recovering a vector $x \in \mathbb{R}^n$ from data

$$y = Ax + z, \tag{7}$$

where A is an $m \times n$ "sensing matrix" giving us information about x, and z is a stochastic or deterministic unknown error term. The setup of the last section is of this form since with $f = \Psi x$ and $y = R\Phi f$ (R is the $m \times n$ matrix extracting the sampled coordinates in M), one can write $y = Ax$, where $A = R\,\Phi\,\Psi$. Hence, one can work with the abstract model (7) bearing in mind that x may be the coefficient sequence of the object in a proper basis.

RESTRICTED ISOMETRIES

In this section, we introduce a key notion that has proved to be very useful to study the general robustness of CS; the so-called *restricted isometry property* (RIP) [15].

DEFINITION 2

For each integer $S = 1, 2, \ldots$, define the isometry constant δ_S of a matrix A as the smallest number such that

$$(1 - \delta_S)\|x\|_{\ell_2}^2 \le \|Ax\|_{\ell_2}^2 \le (1 + \delta_S)\|x\|_{\ell_2}^2 \tag{8}$$

holds for all S-sparse vectors x.

We will loosely say that a matrix A obeys the RIP of order S if δ_S is not too close to one. When this property holds, A approximately preserves the Euclidean length of S-sparse signals, which in turn implies that S-sparse vectors cannot be in the null space of A. (This is useful as otherwise there would be no hope of reconstructing these vectors.) An equivalent description of the RIP is to say that all subsets of S columns taken from A are in fact nearly orthogonal (the columns of A cannot be exactly orthogonal since we have more columns than rows).

To see the connection between the RIP and CS, imagine we wish to acquire S-sparse signals with A. Suppose that δ_{2S} is sufficiently less than one. This implies that all pairwise distances between S-sparse signals must be well preserved in the measurement space. That is, $(1 - \delta_{2S})\|x_1 - x_2\|_{\ell_2}^2 \le \|Ax_1 - Ax_2\|_{\ell_2}^2 \le (1 + \delta_{2S})\|x_1 - x_2\|_{\ell_2}^2$ holds for all S-sparse vectors x_1, x_2. As demonstrated in the next section, this encouraging fact guarantees the existence of efficient and robust algorithms for discriminating S-sparse signals based on their compressive measurements.

GENERAL SIGNAL RECOVERY FROM UNDERSAMPLED DATA

If the RIP holds, then the following linear program gives an accurate reconstruction:

$$\min_{\tilde{x} \in \mathbb{R}^n} \|\tilde{x}\|_{\ell_1} \qquad \text{subject to} \qquad A\tilde{x} = y\ (= Ax). \tag{9}$$

THEOREM 2 [16]

Assume that $\delta_{2S} < \sqrt{2} - 1$. Then the solution $x^\star$ to (9) obeys

$$\|x^\star - x\|_{\ell_2} \le C_0 \cdot \|x - x_S\|_{\ell_1}/\sqrt{S} \quad \text{and}$$
$$\|x^\star - x\|_{\ell_1} \le C_0 \cdot \|x - x_S\|_{\ell_1} \tag{10}$$

for some constant C_0, where x_S is the vector x with all but the largest S components set to 0. (As stated, this result is due to the first author [17] and yet unpublished, see also [16] and [18].)

The conclusions of Theorem 2 are stronger than those of Theorem 1. If x is S-sparse, then $x = x_S$ and, thus, the recovery is exact. But this new theorem deals with all signals. If x is not S-sparse, then (10) asserts that the quality of the recovered signal is as good as if one knew ahead of time the location of the S largest values of x and decided to measure those directly. In other words, the reconstruction is nearly as good as that provided by an oracle which, with full and perfect knowledge about x, extracts the S most significant pieces of information for us.

Another striking difference with our earlier result is that it is deterministic; it involves no probability. If we are fortunate enough to hold a sensing matrix A obeying the hypothesis of

the theorem, we may apply it, and we are then guaranteed to recover *all* sparse S-vectors exactly, and essentially the S-largest entries of *all* vectors otherwise; i.e., there is no probability of failure.

What is missing at this point is the relationship between S (the number of components one can effectively recover) obeying the hypothesis and m the number of measurements or rows of the matrix. To derive powerful results, we would like to find matrices obeying the RIP with values of S close to m. Can one design such matrices? In the next section, we will show that this is possible, but first we examine the robustness of CS vis a vis data corruption.

ROBUST SIGNAL RECOVERY FROM NOISY DATA

We are given noisy data as in (7) and use ℓ_1 minimization with relaxed constraints for reconstruction:

$$\min \|\tilde{x}\|_{\ell_1} \qquad \text{subject to} \qquad \|A\tilde{x} - y\|_{\ell_2} \le \epsilon, \tag{11}$$

where ϵ bounds the amount of noise in the data. (One could also consider recovery programs such as the Dantzig selector [19] or a combinatorial optimization program proposed by Haupt and Nowak [20]; both algorithms have provable results in the case where the noise is Gaussian with bounded variance.) Problem (11) is often called the LASSO after [21]; see also [22]. To the best of our knowledge, it was first proposed in [8]. This is again a convex problem (a second-order cone program) and can be solved efficiently.

THEOREM 3 [16]

Assume that $\delta_{2S} < \sqrt{2} - 1$. Then the solution $x^\star$ to (11) obeys

$$\|x^\star - x\|_{\ell_2} \le C_0 \cdot \|x - x_S\|_{\ell_1}/\sqrt{S} + C_1 \cdot \epsilon \tag{12}$$

for some constants C_0 and C_1. (Again, this theorem is unpublished as stated and is a variation on the result found in [16].)

This can hardly be simpler. The reconstruction error is bounded by the sum of two terms. The first is the error which would occur if one had noiseless data. The second is just proportional to the noise level. Further, the constants C_0 and C_1 are typically small. With $\delta_{2S} = 1/4$ for example, $C_0 \le 5.5$ and $C_1 \le 6$. Figure 3 shows a reconstruction from noisy data.

This last result establishes CS as a practical and robust sensing mechanism. It works with all kinds of not necessarily sparse signals, and it handles noise gracefully. What remains to be done is to design efficient sensing matrices obeying the RIP. This is the subject of the next section.

RANDOM SENSING

Returning to the RIP, we would like to find sensing matrices with the property that column vectors taken from arbitrary subsets are nearly orthogonal. The larger these subsets, the better.

This is where randomness re-enters the picture. Consider the following sensing matrices: i) form A by sampling n column vectors uniformly at random on the unit sphere of $\mathbb{R}^m$; ii) form A by sampling i.i.d. entries from the normal distribution with mean 0 and variance $1/m$; iii) form A by sampling a random projection P as in "Incoherent Sampling" and normalize: $A = \sqrt{n/m}\,P$; and iv) form A by sampling i.i.d. entries from a symmetric Bernoulli distribution ($\mathbf{P}(A_{i,j} = \pm 1/\sqrt{m}) = 1/2$) or other sub-gaussian distribution. With overwhelming probability, all these matrices obey the RIP (i.e. the condition of our theorem) provided that

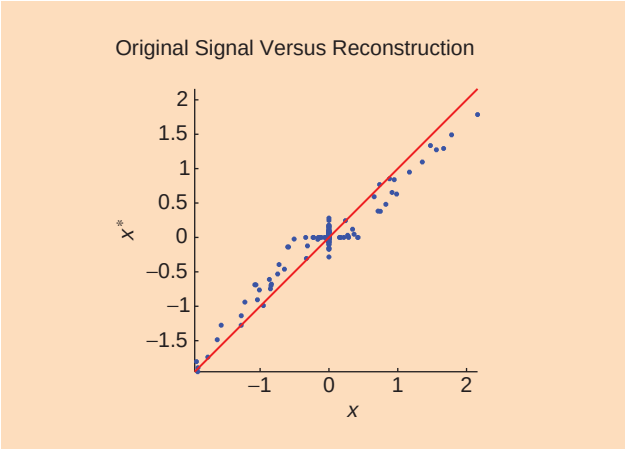
$$m \ge C \cdot S \log(n/S), \tag{13}$$

where C is some constant depending on each instance. The claims for i)–iii) use fairly standard results in probability theory; arguments for iv) are more subtle; see [23] and the work of Pajor and his coworkers, e.g., [24]. In all these examples, the probability of sampling a matrix not obeying the RIP when (13) holds is exponentially small in m. Interestingly, there are no measurement matrices and no reconstruction algorithm whatsoever which can give the conclusions of Theorem 2 with substantially fewer samples than the left-hand side of (13) [2], [3]. In that sense, using randomized matrices together with ℓ_1 minimization is a near-optimal sensing strategy.

One can also establish the RIP for pairs of orthobases as in "Incoherence and the Sensing of Sparse Signals." With $A = R\,\Phi\Psi$ where R extracts m coordinates uniformly at random, it is sufficient to have

$$m \ge C \cdot S (\log n)^4, \tag{14}$$

for the property to hold with large probability; see [25] and [2]. If one wants a probability of failure no larger than $O(n^{-\beta})$ for some $\beta > 0$, then the best known exponent in (14) is five instead of four (it is believed that (14) holds with just $\log n$).



[FIG3] A signal x (horizontal axis) and its reconstruction $x^\star$ (vertical axis) obtained via (11). In this example, $n = 512$ and $m = 256$. The signal is 64-sparse. In the model (7), the sensing matrix has i.i.d. $N(0, 1/m)$ entries and z is a Gaussian white noise vector adjusted so that $\|Ax\|_{\ell_2}/\|z\|_{\ell_2} = 5$. Here, $\|x^\star - x\|_{\ell_2} \approx 1.3 \cdot \epsilon$.

This proves that one can stably and accurately reconstruct nearly sparse signals from dramatically undersampled data in an incoherent domain.

Finally, the RIP can also hold for sensing matrices $A = \Phi\Psi$, where Ψ is an arbitrary orthobasis and Φ is an $m \times n$ measurement matrix drawn randomly from a suitable distribution. If one fixes Ψ and populates Φ as in i)–iv), then with overwhelming probability, the matrix $A = \Phi\Psi$ obeys the RIP provided that (13) is satisfied, where again C is some constant depending on each instance. These random measurement matrices Φ are in a sense *universal* [23]; the sparsity basis need not even be known when designing the measurement system!

MATHEMATICAL AND COMPUTATIONAL METHODS COULD HAVE AN ENORMOUS IMPACT IN AREAS WHERE CONVENTIONAL HARDWARE DESIGN HAS SIGNIFICANT LIMITATIONS.

WHAT IS COMPRESSIVE SAMPLING?

Data acquisition typically works as follows: massive amounts of data are collected only to be—in large part—discarded at the compression stage to facilitate storage and transmission. In the language of this article, one acquires a high-resolution pixel array f, computes the complete set of transform coefficients, encode the largest coefficients and discard all the others, essentially ending up with f_S. 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 in just a few hundred kilobytes).

CS operates very differently, and performs as "if it were possible to directly acquire just the important information about the object of interest." By taking about $O(S\log(n/S))$ random projections as in "Random Sensing," one has enough information to reconstruct the signal with accuracy at least as good as that provided by f_S, the best S-term approximation—the best compressed representation—of the object. In other words, CS measurement protocols essentially translate analog data into an already compressed digital form so that one can—at least in principle—obtain super-resolved signals from just a few sensors. All that is needed after the acquisition step is to "decompress" the measured data.

There are some superficial similarities between CS and ideas in coding theory and more precisely with the theory and practice of Reed-Solomon (RS) codes [26]. In a nutshell and in the context of this article, it is well known that one can adapt ideas from coding theory to establish the following: one can uniquely reconstruct any S-sparse signal from the data of its first $2S$ Fourier coefficients, $y_k = \sum_{t=0}^{n-1} x_t\, e^{-i\,2\pi\,kt/n}$, $k = 0, 1, 2, \ldots, 2S-1$, or from any set of $2S$ consecutive frequencies for that matter (the computational cost for the recovery is essentially that of solving an $S \times S$ Toeplitz system and of taking an n-point fast Fourier transform). Does this mean that one can use this technique to sense compressible signals? The answer is negative and there are two main reasons for this. First, the problem is that RS decoding is an algebraic technique, which cannot deal with nonsparse signals (the decoding finds the support by rooting a polynomial); second, the problem of finding the support of a signal—even when the signal is exactly sparse—from its first $2S$ Fourier coefficients is extraordinarily ill posed (the problem is the same as that of extrapolating a high degree polynomial from a small number of highly clustered values). Tiny perturbations of these coefficients will give completely different answers so that with finite precision data, reliable estimation of the support is practically impossible. Whereas purely algebraic methods ignore the conditioning of information operators, having well-conditioned matrices, which are crucial for accurate estimation, is a central concern in CS as evidenced by the role played by the RIP.

APPLICATIONS

The fact that a compressible signal can be captured efficiently using a number of *incoherent* measurements that is proportional to its information level $S \ll n$ has implications that are far reaching and concern a number of possible applications:

- Data compression. In some situations, the sparse basis Ψ may be unknown at the encoder or impractical to implement for data compression. As we discussed in "Random Sensing," however, a randomly designed Φ can be considered a universal encoding strategy, as it need not be designed with regards to the structure of Ψ. (The knowledge and ability to implement Ψ are required only for the decoding or recovery of f.) This universality may be particularly helpful for distributed source coding in multi-signal settings such as sensor networks [27]. We refer the reader to articles by Haupt et al. and Goyal et al. elsewhere in this issue for related discussions.
- Channel coding. As explained in [15], CS principles (sparsity, randomness, and convex optimization) can be turned around and applied to design fast error correcting codes over the reals to protect from errors during transmission.
- Inverse problems. In still other situations, the only way to acquire f may be to use a measurement system Φ of a certain modality. However, assuming a sparse basis Ψ exists for f that is also incoherent with Φ, then efficient sensing will be possible. One such application involves MR angiography [1] and other types of MR setups [28], where Φ records a subset of the Fourier transform, and the desired image f is sparse in the time or wavelet domains. Elsewhere in this issue, Lustig et al. discuss this application in more depth.
- Data acquisition. Finally, in some important situations the full collection of n discrete-time samples of an analog signal may be difficult to obtain (and possibly difficult to subsequently compress). Here, it could be helpful to design physical sampling devices that directly record discrete, low-rate incoherent measurements of the incident analog signal.

The last of these applications suggests that mathematical and computational methods could have an enormous impact

in areas where conventional hardware design has significant limitations. For example, conventional imaging devices that use CCD or CMOS technology are limited essentially to the visible spectrum. However, a CS camera that collects incoherent measurements using a digital micromirror array (and requires just one photosensitive element instead of millions) could significantly expand these capabilities. (See [29] and an article by Duarte et al. in this issue.)

Along these same lines, part of our research has focused on advancing devices for "analog-to-information" (A/I) conversion of high-bandwidth signals (see also the article by Healy et al. in this issue). Our goal is to help alleviate the pressure on conventional ADC technology, which is currently limited to sample rates on the order of 1 GHz. As an alternative, we have proposed two specific architectures for A/I in which a discrete, low-rate sequence of incoherent measurements can be acquired from a high-bandwidth analog signal. To a high degree of approximation, each measurement y_k can be interpreted as the inner product $\langle f, \varphi_k \rangle$ of the incident analog signal f against an analog measurement waveform φ_k. As in the discrete CS framework, our preliminary results suggest that analog signals obeying a sparse or compressible model (in some analog dictionary Ψ) can be captured efficiently using these devices at a rate proportional to their information level instead of their Nyquist rate. Of course, there are challenges one must address when applying the discrete CS methodology to the recovery of sparse analog signals. A thorough treatment of these issues would be beyond the scope of this short article and as a first cut, one might simply accept the idea that in many cases, discretizing/sampling the sparse dictionary allows for suitable recovery. Our two architectures are as follows:

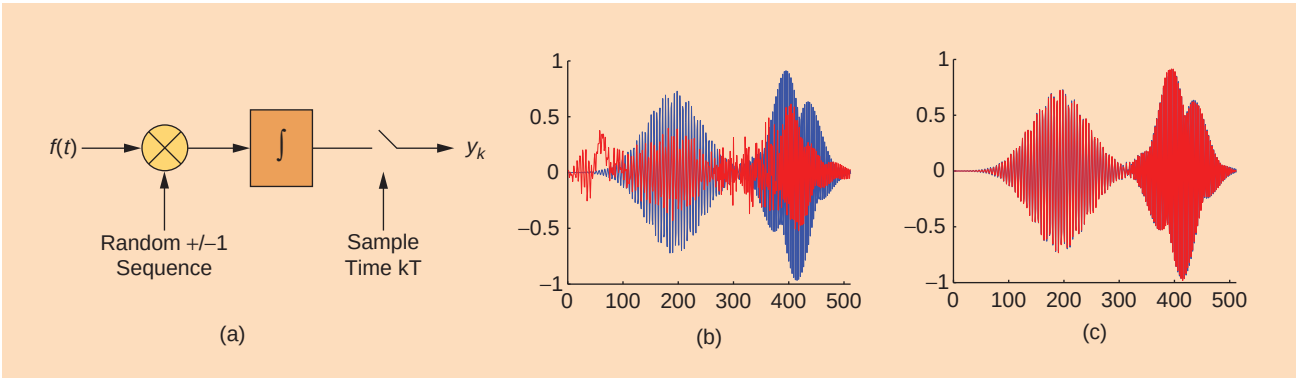
1) **Nonuniform Sampler (NUS).** Our first architecture simply digitizes the signal at randomly or pseudo-randomly sampled time points. That is, $y_k = f(t_k) = \langle f, \delta_{t_k} \rangle$. In effect, these time points are obtained by jittering nominal (low-rate) sample points located on a regular lattice. Due to the incoherence between spikes and sines, this architecture can be used to sample signals having sparse frequency spectra far below their Nyquist rate. There are of course tremendous benefits associated with a reduced sampling rate, as this provides added circuit settling time and has the effect of reducing the noise level.

PART OF OUR RESEARCH HAS FOCUSED ON ADVANCING DEVICES FOR "ANALOG-TO-INFORMATION" (A/I) CONVERSION OF HIGH-BANDWIDTH SIGNALS.

2) **Random Modulation Preintegration (RMPI).** Our second architecture is applicable to a wider variety of sparsity domains, most notably those signals having a sparse signature in the time-frequency plane. Whereas it may not be possible to digitize an analog signal at a very high rate rate, it may be quite possible to change its polarity at a high rate. The idea of the RMPI architecture [see Figure 4(a)] is then to multiply the signal by a pseudo-random sequence of ± 1s, integrate the product over time windows, and digitize the integral at the end of each time interval. This is a parallel architecture and one has several of these random multiplier-integrator pairs running in parallel using distinct sign sequences. In effect, the RMPI architecture correlates the signal with a bank of sequences of ± 1, one of the random CS measurement processes known to be universal, and therefore the RMPI measurements will be incoherent with any fixed time-frequency dictionary such as the Gabor dictionary described below.

For each of the above architectures, we have confirmed numerically (and in some cases physically) that the system is robust to circuit nonidealities such as thermal noise, clock timing errors, interference, and amplifier nonlinearities.

The application of A/I architectures to realistic acquisition scenarios will require continued development of CS algorithms and theory. To highlight some promising recent directions, we conclude with a final discrete example. We take f to be a one-dimensional signal of length $n = 512$ that contains two modulated pulses [see the blue curve in Figure 4(b)] From this signal, we



[FIG4] Analog-to-information conversion. (a) Random modulation preintegration (RMPI) system. (b) Original two-pulse signal (blue) and reconstruction (red) via $\ell_1$synthesis from random ± 1 measurements. (c) Two-pulse signal and reconstruction via reweighted ℓ_1 analysis.

collect $m = 30$ measurements using an $m \times n$ measurement matrix Φ populated with i.i.d. Bernoulli ± 1 entries. This is an unreasonably small amount of data corresponding to an undersampling factor of over 17. For reconstruction we consider a Gabor dictionary Ψ that consists of a variety of sine waves time limited by Gaussian windows, with different locations and scales. Overall the dictionary is approximately $43\times$ overcomplete and does not contain the two pulses that comprise f. The red curve in Figure 4(b) shows the result of minimizing $\|x\|_{\ell_1}$ such that $y = \Phi\Psi x$. The reconstruction shows pronounced artifacts, and we see $\|f - f^\star\|_{\ell_2}/\|f\|_{\ell_2} \approx 0.67$. However, we can virtually eliminate these artifacts by making two changes to the ℓ_1 recovery program. First, we instead minimize $\|\Psi^* \tilde{f}\|_{\ell_1}$ subject to $y = \Phi\tilde{f}$. (This variation has no effect when Ψ is an orthobasis.) Second, after obtaining an estimate $f^\star$, we reweight the ℓ_1 norm and repeat the reconstruction, with a lower penalty applied to those coefficients we anticipate to be large. Figure 4(c) shows the result after four iterations of reweighting; we see $\|f - f^\star\|_{\ell_2}/\|f\|_{\ell_2} \approx 0.022$. We refer the reader to [30] for more information on these directions. The point here is that even though the amount of data is ridiculously small, one has nevertheless captured most of the information contained in the signal. This, in a nutshell, is why CS holds such great promise.

AUTHORS

Emmanuel J. Candès (emmanuel@acm.caltech.edu) received his B. Sc. degree from the École Polytechnique, France, in 1993 and the Ph.D. degree in statistics from Stanford University in 1998. He is the Ronald and Maxine Linde Professor of Applied and Computational Mathematics at the California Institute of Technology. His research interests are in computational harmonic analysis, statistical estimation and detection, signal processing, scientific computing, inverse problems, and mathematical optimization. He received the Third Popov Prize in Approximation Theory in 2001, and the DOE Young Investigator Award in 2002. He was an Alfred P. Sloan Research Fellow in 2001. He has given plenary addresses at major international conferences, including ICIAM 2007 and ICIP 2007. In 2005, he was awarded the James H. Wilkinson Prize in Numerical Analysis and Scientific Computing by SIAM. He received the NSF 2006 Alan T. Waterman Medal.

Michael B. Wakin (wakin@umich.edu) received the B.S. degree in electrical engineering and the B.A. degree in mathematics in 2000 (summa cum laude), the M.S. degree in electrical engineering in 2002, and the Ph.D. degree in electrical engineering in 2007, all from Rice University. From 2006–2007, he was an NSF Mathematical Sciences postdoctoral research fellow in the Department of Applied and Computational Mathematics at the California Institute of Technology, and he is currently an assistant professor in the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor. His research interests include sparse, geometric, and manifold-based models for signal and image processing, approximation, compression, compressive sampling, and dimensionality reduction.

REFERENCES

[1] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[2] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[3] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289-1306, Apr. 2006.

[4] D.S. Taubman and M.W. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Norwell, MA: Kluwer, 2001.

[5] D.L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.

[6] R. Coifman, F. Geshwind, and Y. Meyer, "Noiselets," *Appl. Comput. Harmon. Anal.*, vol. 10, no. 1, pp. 27–44, 2001.

[7] J.F. Claerbout and F. Muir, "Robust modeling with erratic data," *Geophys. Mag.*, vol. 38, no. 5, pp. 826-844, Oct. 1973.

[8] F. Santosa and W.W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM J. Sci. Statist. Comput.*, vol. 7, no. 4, pp. 1307–1330, 1986.

[9] J. Tropp and A.C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Trans. Inform. Theory,* vol. 53, no. 12, pp. 4655-4666, 2007.

[10] E. Candès and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Prob.*, vol. 23, no. 3, pp. 969–985, 2007.

[11] P. Feng and Y. Bresler, "Spectrum-blind minimum-rate sampling and reconstruction of multiband signals," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, Atlanta, GA, vol. 2, 1996, pp. 1689–1692.

[12] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1417-1428, June 2002.

[13] A. Gilbert, S. Muthukrishnan, and M. Strauss, "Improved time bounds for near-optimal sparse Fourier representation," in *Proc. Wavelets XI SPIE Optics Photonics*, San Diego, CA, 2005.

[14] D.L. Donoho and P.B. Stark, "Uncertainty principles and signal recovery," *SIAM J. Appl. Math.*, vol. 49, no. 3, pp. 906–931, 1989.

[15] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inform. Theory*, vol. 51, no. 12, pp. 4203-4215, Dec. 2005.

[16] E. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. 59, no. 8, pp. 1207–1223, Aug. 2006.

[17] E.J. Candès, "Lectures on compressive sampling and frontiers in signal processing," *The Institute for Mathematics and its Applications*. University of Minnesota, June 2007 [Online]. Available: http://www.ima.umn.edu/2006-2007/ND6.4-15.07/abstracts.html

[18] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," 2006, Preprint.

[19] E. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," submitted for publication.

[20] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4036–4048, 2006.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267-288, 1996.

[22] S.S. Chen, D.L. Donoho, and M.A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[23] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," submitted for publication.

[24] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Uniform uncertainty principle for Bernoulli and sub-gaussian ensembles," 2006, Preprint.

[25] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," submitted for publication.

[26] R.E. Blahut, *Algebraic Codes for Data Transmission*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[27] D. Baron, M.B. Wakin, M.F. Duarte, S. Sarvotham, and R.G. Baraniuk, "Distributed compressed sensing," 2005, Preprint.

[28] M. Lustig, D.L. Donoho, and J.M. Pauly, "Rapid MR imaging with compressed sensing and randomly under-sampled 3DFT trajectories," in *Proc. 14th Ann. Meeting ISMRM*, Seattle, WA, May 2006.

[29] D. Takhar, V. Bansal, M. Wakin, M. Duarte, D. Baron, K.F. Kelly, and R.G. Baraniuk, "A compressed sensing camera: New theory and an implementation using digital micromirrors," in *Proc. Comp. Imaging IV SPIE Electronic Imaging*, San Jose, CA, 2006.

[30] E.J. Candès, M.B. Wakin, and S.P. Boyd, "Enhancing sparsity by reweighting ℓ_1," Tech. Rep., California Institute of Technol., 2007 [Online]. Available: http://www.acm.caltech.edu/~emmanuel/publications.html

SP