Extensions of Compressed Sensing

Yaakov Tsaig
David L. Donoho

October 20, 2004
Revised April 26, 2005

Abstract

We study the notion of Compressed Sensing (CS) as put forward in [13] and related work [18, 3, 4]. The notion proposes a signal or image, unknown but supposed to be compressible by a known transform, (eg. wavelet or Fourier), can be subjected to fewer measurements than the nominal number of data points, and yet be accurately reconstructed. The samples are nonadaptive and measure 'random' linear combinations of the transform coefficients. Approximate reconstruction is obtained by solving for the transform coefficients consistent with measured data and having the smallest possible ℓ^1 norm.

We present initial 'proof-of-concept' examples in the favorable case where the vast majority of the transform coefficients are zero. We continue with a series of numerical experiments, for the setting of ℓ^p-sparsity, in which the object has all coefficients nonzero, but the coefficients obey an ℓ^p bound, for some $p \in (0, 1]$. The reconstruction errors obey the inequalities paralleling the theory, seemingly with well-behaved constants.

We report that several workable families of 'random' linear combinations all behave equivalently, including random spherical, random signs, partial Fourier and partial Hadamard.

We next consider how these ideas can be used to model problems in spectroscopy and image processing, and in synthetic examples see that the reconstructions from CS are often visually "noisy". To suppress this noise we post-process using translation-invariant de-noising, and find the visual appearance considerably improved.

We also consider a multiscale deployment of compressed sensing, in which various scales are segregated and CS applied separately to each; this gives much better quality reconstructions than a literal deployment of the CS methodology.

These results show that, when appropriately deployed in a favorable setting, the CS framework is able to save significantly over traditional sampling, and there are many useful extensions of the basic idea.

Key Words and Phrases: Basis Pursuit. Underdetermined Systems of Linear Equations. Linear Programming. Random Matrix Theory.

Acknowledgements. Partial support from NSF DMS 00-77261, and 01-40698 (FRG) and ONR-MURI. Thanks to Michael Saunders for optimization advice, and Emmanuel Candès for discussions of his own related work with J. Romberg and T. Tao. Raphy Coifman asked us insistently about the question of the constants in (1.2) and inspired this report.

1 Introduction

In the modern multimedia-saturated world, 'everyone' knows that all humanly-intelligible data are highly compressible. In exploiting this fact, the dominant approach is to first sample the data, and then eliminate redundancy using various compression schemes. This raises the question:

why is it necessary to sample the data in a pedantic way and then later to compress it? Can't one directly acquire a compressed representation? Clearly, if this were possible, there would be implications in a range of different fields, extending from faster data acquisition, to higher effective sampling rates, and lower communications burden.

Several recent papers [18, 3, 13, 4], have shown that, under various assumptions, it may be possible to directly acquire a form of compressed representation. In this paper, we put such ideas to the test by making a series of empirical studies of the effectiveness of compressed sensing schemes.

1.1 The Approach

We adopt language and notation from [13]; the approach there is rather abstract, but has the advantage of factoring across many potential applications areas.

We assume the object of interest is a vector $x_0 \in \mathbf{R}^m$ – this could represent the m sampled values of a digital signal or image. We assume the object is *a priori* compressible by transform coding of the type used in e.g. JPEG or JPEG-2000. Mathematically, we let Ψ denote the matrix of the orthogonal transform in question, having columns ψ_i, $i = 1, \ldots, m$. By 'compressible' we mean that for some $p \leq 1$ and moderate $R > 0$, $\|\Psi^T x_0\|_p \leq R$; see [13] for further explanation of this condition, which imposes sparsity on the transform coefficients $\Psi^T x_0$.

To make our compressed measurements, we start from an n by m matrix Φ with $n < m$ satisfying certain conditions called CS1-CS3 in [13]. We form the matrix $\Xi = \Phi\Psi^T$, also $n \times m$. We take the n-pieces of measured information $y = (y(i) : 1 \leq i \leq n)$ according to the linear system $y = \Xi x_0$. Since $n < m$ this amounts to having fewer measurements than degrees of freedom for the signal x_0.

The individual measured values are of the form $y(i) = \langle \xi_i, x_0 \rangle$, i.e. each can be obtained by integrating the object x_0 against a measurement kernel ξ_i. Here ξ_i denotes the i-th row of Ξ; from the formula $\Xi = \Phi\Psi^T$ and the known properties of CS-matrices Φ, we may view each measurement kernel as a kind of 'random' linear combination of the basis elements ψ_j. The needed matrices Φ satisfying CS1-CS3 were shown to be constructible by random sampling from a uniform distribution on the columns of Φ. Following [13], the collection of matrices constructed in this manner shall be referred to as the uniform spherical ensemble.

To reconstruct an approximation to x_0 we solve

$$(L_1) \quad \min_x \|\Psi^T x\|_1 \text{ subject to } y = \Xi x. \tag{1.1}$$

Call the result $\hat{x}_{1,n}$. In words, $\hat{x}_{1,n}$ is, among all objects generating the same measured data, the one having transform coefficients with the smallest ℓ^1 norm. In [13] it was mentioned that this reconstruction procedure can be implemented by linear programming, and so may be considered computationally tractable. In fact, in the experiments presented throughout this paper, we use a primal-dual barrier method for the solution of this linear program, as proposed in [5].

In short, the approach involves linear, nonadaptive measurement, followed by nonlinear approximate reconstruction.

The paper [13] proved error bounds showing that despite the apparent undersampling ($n < m$), good accuracy reconstruction is possible for compressible objects. Such bounds take the form

$$\|\hat{x}_{1,n} - x_0\|_2 \leq C_p \cdot R \cdot (n/\log(m))^{1/2-1/p}, \qquad n, m > n_0. \tag{1.2}$$

As $p < 2$, this bound guarantees that reconstruction is accurate for large n, with a very weak dependence on m. Such bounds were interpreted in [13] to say that n CS measurements with

$n = O(N \log(m))$ are just as good as knowing the N biggest transform coefficients. Examples were sketched for model problems caricaturing imaging and spectroscopy.

In related prior work, classical literature in approximation theory [21, 17, 26] (developing the theory of Gel'fand n-widths) deals with closely related problems from an even more abstract viewpoint; see the discussion in [13]. In the literature of Information-Based Complexity such work had previously been shown to imply that certain classes of objects were best sampled not using point samples but instead using quasi-random linear combinations – see for example the discussion of recovery of monotone functions in [24]. More recently, Gilbert et al. [18] considered n-by-m matrices Φ made of n special rows out of the m-by-m Fourier matrix, while Candès, Romberg and Tao considered matrices Φ made of n randomly chosen rows from the Fourier matrix. Candès, Romberg, and Tao [3] considered also the use of ℓ^1 minimization, just as here while Gilbert et al. [18] considered a different nonlinear procedure. In addition, we note that at the time of revision, Candès, Romberg, and Tao [4, 30] obtained results generalizing the work in [13].

Finally, we comment that, at a higher level of abstraction, we are discussing here the idea of getting sparse solutions to underdetermined systems of equations using ℓ^1 methods, which forms part of a now-extensive body of work: [5, 8, 10, 11, 12, 14, 15, 19, 20, 23, 27, 28].

1.2 Questions ...

Readers may want to ask numerous questions about the result (1.2), starting with:

- *How large do n and m have to be?* Is (1.2) meaningless for practical problem sizes?
- *How large is the constant C_p?* Even if (1.2) applies, perhaps the constant is miserable?
- *When the object is not perfectly reconstructed, what sorts of errors occur?* Perhaps the error term, though controlled in size by (1.2), has an objectionable structure?

They might continue with

- *How should the CS framework be applied to realistic signals?* In [13], stylized models of spectroscopy and imaging were considered; for such models, it was proposed to deploy CS in a hybrid strategy, with the relatively small number of measurements at coarse scales obtained by classical linear sampling, and the bulk of measurements at fine scales obtained by the CS strategy. Are such ideas, originally derived for the purpose of enabling mathematical analysis, actually helpful in a concrete setting?
- *Are there artifacts caused by CS which should be suppressed?* Is any post-processing of CS needed to 'clean up' the reconstructions?
- *What happens if there is noise in the measurements?* Perhaps the framework falls apart if there's any noise in the observations – even just the small errors of floating point representation.

Supposing that such questions do not have devastating answers, there are also natural questions about extending the method by extending the kind of matrices Φ which are in use:

- *More Concrete CS matrices Φ.* It would be useful to have CS-matrices which can be obtained constructively/explicitly.

- *CS matrices* Φ *which can rapidly applied.* This is connected with the previous question. For applying the linear programming formulation of (1.1) it is very convenient [5] that Ψ and Φ each can be rapidly applied – along with their transposes.

We agree that these are all important questions to answer.

1.3 'Proofs-of-Concept' and Numerical Experiments

In this paper we approach the above questions through either proofs-of-concept or through computational experiments. We consider each question, describe a small-scale simulation to study it, generate synthetic datasets, and interpret the results of our simulation. While this simulation method alone cannot definitively answer some of our questions, it does provide valuable evidence which may inspire future theoretical studies, and suggest directions for future experimental and applications work.

In Section 2, as a 'warm-up', we give examples applying CS to signals which have only a small number of truly nonzero coefficients. Section 3 considers signals which have all coefficients nonzero, but which still have sparse coefficients as measured by ℓ^p norms, $0 < p \leq 1$. We analyze the empirical results, obtain numerical evidence for the constant C_p, and compare the empirical errors with the theoretical bound (1.2).

In Section 4, we explore the freedom available in the choice of CS-matrices, describing several different matrix ensembles which seem to give equivalent results. While [13] focused on matrices with iid columns, we have found that several other ensembles work well, including the partial Fourier ensembles [18, 3].

We observe that the results in the ℓ^p setting are in some sense 'noisy' when the number of sensed samples is small, even when the data measurements are noiseless. To alleviate that, in Section 5 we consider an extension of CS by post-processing to remove the 'noise' in CS reconstructions. Section 6 considers noisy observations, and develops a noise-aware reconstruction method.

In Section 7, we consider an extension of CS mentioned in [13]: a hybrid scheme, using conventional linear sampling and reconstruction for coarse-scale information and compressed sampling on fine scale information. The 'noise' in reconstructions can be dramatically lower for a given number of samples. Section 8 pushes this approach farther, deploying CS in a fully multiscale fashion. Different scales are segregated and CS is applied separately to each one. Again, the 'noise' can be dramatically lower.

In a final section we summarize our results, and identify issues for future work.

1.4 Reproducible Research

Upon publication, we will make the software used to generate the experiments in this paper available to the research community, as part of the *SparseLab* software package. The package, a collection of Matlab functions and scripts designed to solve sparse approximation problems, will be made available at http://sparselab.stanford.edu/, in the spirit of reproducible research [1]. In particular, all the figures appearing in this paper can be reproduced by code which is part of SparseLab. This will allow readers to study variations of our experiments and proofs-of-concept.

2 ℓ^0 Sparsity

We start with a warm-up question. Suppose we have an object x_0 with a very high degree of transform sparsity – only k nonzeros out of m coefficients; how big does n have to be for the CS

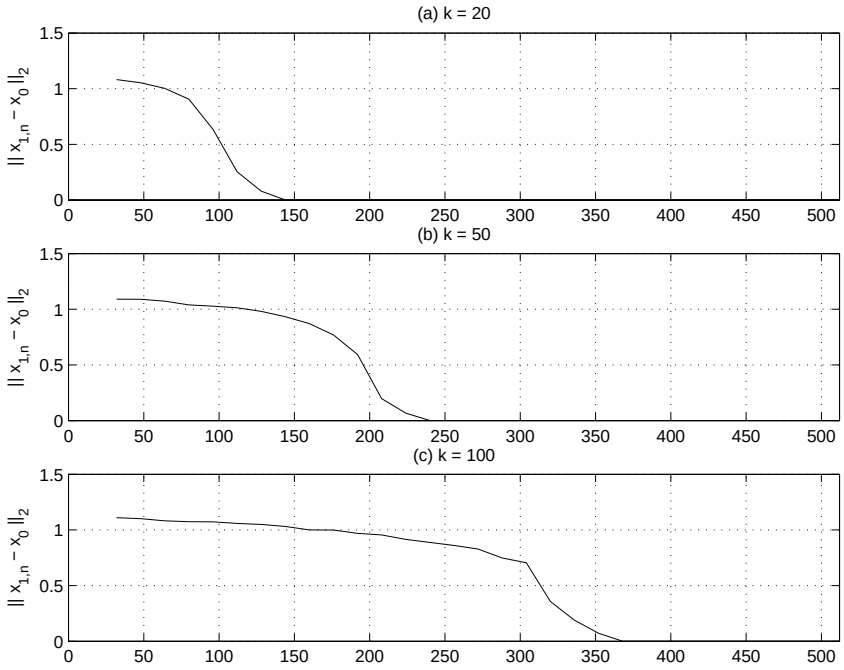

Figure 1: Error of reconstruction versus number of samples for (a) k = 20 nonzeros; (b) k = 50; (c) k = 100.

scheme to work well?

To study this question, we conduct a small-scale experiment, which may help many readers begin to develop an intuitive feel for this problem. For fixed m and a series of different values for k, we construct signals with k nonzeros, located in random positions in the transform domain, all of equal amplitude. We then apply the CS framework, while varying n, the number of samples used in the sensing scheme. The CS matrix Φ is drawn from the *uniform spherical ensemble*, i.e. the columns are drawn independently at random from a uniform distribution on the unit sphere $\mathbf{S}^{n-1}$ in Euclidean n-space. In addition, we set $\Psi = I$.

We repeat this experiment 20 times (independently), and record the worst-case behavior for each (n, m, k) instance, i.e. the maximal reconstruction error $\|\hat{x}_{1,n} - x_0\|_2$. Figure 1, panels (a)-(c), display the ℓ^2 reconstruction error as a function of n, with $m = 1024$, $k = 20, 50$ and 100. In each case we see that as n increases beyond about $2k$, the error starts to drop, and eventually becomes negligible at some multiple of k.

As a more 'signals-oriented' instance of this phenomenon, we considered the object *Blocks* from the Wavelab package [1]. As Figure 2 shows, the object is piecewise constant, and its Haar wavelet transform has relatively few nonzero coefficients. In fact, *Blocks* has $k = 77$ nonzero coefficients in a signal length $m = 2048$. Figures 3(a),(b) show the reconstruction results with $n = 340$, and 256 compressed samples, respectively. Clearly, the results are better for larger n, and somewhere about $n = 340 \approx 4k$ the method works well, while for $n = 256 \approx 3k$ the results are somewhat 'noisy'.

The results suggest a rule of thumb for these combinations of m, n and k: if an object has a representation using only k nonzeros at randomly-chosen sites, something like $n \approx 4k$ measurements would typically be needed.

(A referee asked whether the phenomenon observed in Figure 1 'held up' under variations in the relative sizes of different nonzeros in x_0. This is partially addressed by the *Blocks* example,

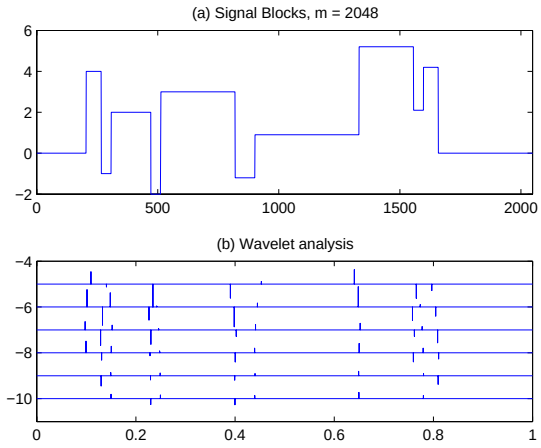

Figure 2: (a) Signal *Blocks*; (b) its expansion in a Haar wavelet basis. There are 77 nonzero coefficients.

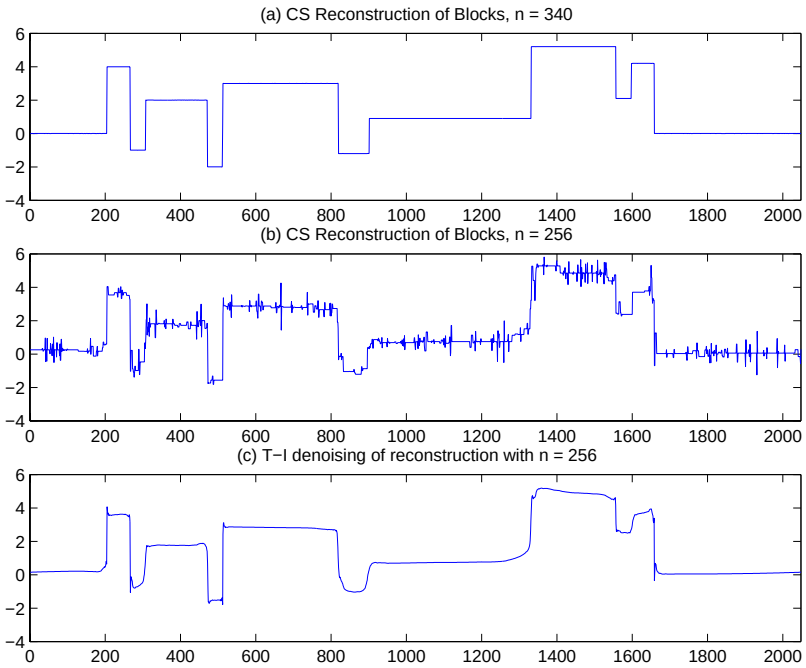

Figure 3: CS reconstructions of *Blocks* from (a) $n = 340$ and (b) $n = 256$ measurements. (a) shows perfect reconstruction (modulo numerical effects), while (b) is visibly noisy. Panel (c) shows Translation-Invariant denoising of (b).

which has wavelet coefficients exhibiting a wide range of amplitudes. More generally there is an underlying theory showing that (numerical effects aside), the amplitudes of the nonzeros of sparse signals don't matter, provided the number of nonzeros is below a threshold, which depends on n and m [11]. This theory concerns the equivalence between the sparsest solution and the minimum ℓ^1 solution. Define the ℓ^0 norm by the number of nonzeros: $\|x\|_0 = \#\{i : x(i) \neq 0\}$. Define the *equivalence breakdown point* EBP as a constant depending only on Φ, such that if y permits a representation $y = \Xi x_0$ with $\|\Psi^T x_0\|_0 < EBP$, then x_0 is the unique solution to (L_1). In such cases, if x_0 has $k < EBP$ nonzeros, (L_1) recovers x_0 perfectly. The amplitude and arrangement of the nonzeros simply don't matter. Actually, the breakdown point can be large. Theory shows that, for large random matrices Φ with n and m of comparable size (say within a constant factor of each other), drawn from the uniform spherical ensemble; then typically EBP is *proportional to* n [11]. The empirical finding that the error is minimal as soon as $n \approx 4k$ seems consistent with a (hoped-for) theoretical statement to the effect that $EBP \approx n/4$ at the (n, m) combinations used for Figure 1. The companion paper [29] explores breakdown phenomena more closely. Another referee has observed that the constant '4' in this relationship seems to be dropping as n increases. Our main goal here has been to introduce the reader to this phenomenon rather than carefully demarcate it. Far more extensive studies in [29] support our interpretation; these have also been supported by recent theoretical work showing the existence of a 'phase transition' boundary and calculating its position precisely, [32, 33]).

Of course the notion of ℓ^0 sparsity – while powerful – is inherently limited in its applicability. It provides an easily-understandable way to introduce the idea of sparsity to 'newcomers'; but in practice, real signals will not typically have exact zeros anywhere in the transform. Hence, examples of the type just shown serve merely as a warm-up.

3 ℓ^p Sparsity

We now consider a more widely applicable notion of sparsity, based on controlling the ℓ^p norm $\|x\|_p = (\sum_i |x(i)|^p)^{1/p}$ with $0 < p \leq 1$. Equating 'sparsity' with 'small ℓ^p norm' greatly expands the class of signals we may consider sparse, since objects can have all entries nonzero, but still have small ℓ^p norm; at the same time, signals sparse in the ℓ^0 case are (in an appropriate sense) also sparse in the ℓ^p sense. There is an intimate connection between ℓ^p norms, $0 < p \leq 1$ and the number of nonzeros (a.k.a. the ℓ^0 norm); for example, as $p \to 0$ the $\|x\|_p^p \to \|x\|_0$. See [13] and citations there for examples of ℓ^p constraints obeyed for natural classes of signals.

For the examples in this section, we generated random signals with controlled ℓ^p norm in the following manner. A signal θ is created with coefficients having random signs, and the amplitude, $|\theta|_{(k)}$, of the k-th largest-sized coefficient obeys

$$|\theta|_{(k)} = (k \cdot \log m)^{-1/p}, \quad k = 1, 2, \ldots. \tag{3.1}$$

Signals constructed in this manner have the property that

$$\|\theta\|_p^p = \sum_{k=1}^{m} (k \cdot \log m)^{-1} = \frac{1}{\log m} \sum_{k=1}^{m} \frac{1}{k} = 1 + o(1), \qquad m \to \infty.$$

Clearly, all of the coefficients $\theta_{(k)}$ are nonzero, i.e. $\|\theta\|_0 = m$.

Figure 4 panel (a) shows such a spiky object with $m = 1000, p = 3/4$, and panel (b) shows the reconstruction with $n = 100$ sensed samples. Note that the CS scheme recovers most of the 'important' coefficients well, but fails to recover some of the very small ones.

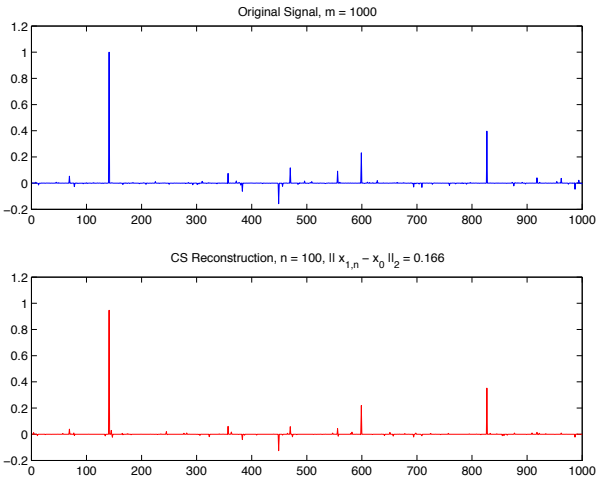

Figure 4: CS reconstruction of a signal with controlled ℓ^p norm, $p = 3/4$. (a) Original signal, $m = 1000$; (b) CS reconstruction with $n = 100$ samples, $\|x_{1,n} - x_0\| = 0.166$. The most substantial coefficients are recovered accurately.

With that, we turn to questions posed in Section 1.2 regarding the applicability of the bound (1.2). Namely, for this class of 'ℓ^p-sparse signals', how does the measured ℓ^2 error in the CS reconstruction compare to the theoretical predictions? How large is C_p in (1.2)?

This question is theoretical in nature, as it refers to an inequality covering all possible x-vectors with $\|x\|_p \leq R$, rather than typical ones seen in practice. Nevertheless experiments with typical signals can be informative. As a starter, we would like to have empirical evidence about the dependence of reconstruction error on n, m, and p. This should reflect the behavior of the constant C_p. To that end, consider the following simulation study. For a range of n, m, p, generate a matrix $\Phi_{n\times m}$ from the uniform spherical ensemble, set $\Psi = I$, and create a random signal x_0 of length m with a controlled ℓ^p norm in the manner described above. Compute the sensed measurement vector $y = \Phi\Psi^T x_0$ and solve (1.1). Finally, measure the reconstruction error $\|\hat{x}_{1,n} - x_0\|_2$. Repeat this basic procedure several times for each (n, m, p) triplet and recorded the largest error. We ran this study with p in the range $[0.25, 1]$, m in the range $[1000, 4000]$, and n in the range $[0.05m, 0.75m]$. We ran 20 instances of the experiment for each (n, m, p) triplet.

To summarize our results, we developed an empirical substitute $\tilde{C}_p$ for the theoretical constant C_p in (1.2). For each p, we maximized the error over all (n, m) pairs, and compared to the right-side of (1.2), obtaining:

$$\tilde{C}_p = \max_{n,m} \max_{x_0} \frac{\|\hat{x}_{1,n} - x_0\|_2}{\|x_0\|_p \cdot (\log(m)/n)^{1/p-1/2}}.$$

Here the maxima are taken over the full range of (n, m) and over the full collection of signals x_0 generated in our experiments. The results are given in Table 1.

Examining the table, we observe that very modest values of $\tilde{C}_p$ arise for the range of n, m we considered. This suggests that the theoretical value C_p may indeed not be of catastrophic proportions. Clearly, our estimate $\tilde{C}_p$ is at best a lower bound for C_p, since it simply indicates the worst values encountered in our simulations rather than the worst values possible. Still, it gives us practical information reflecting actual behavior at plausible 'ℓ^p sparse' signals, such as the one portrayed in Figure 4(a).

p	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6
$\tilde{C}_p$	0.127	0.133	0.122	0.0755	0.0525	0.0402	0.0253	0.0467
p	0.65	0.7	0.75	0.8	0.85	0.9	0.95	1
$\tilde{C}_p$	0.0619	0.0821	0.0905	0.113	0.159	0.177	0.191	0.216

Table 1: Empirically-derived constant $\tilde{C}_p$.

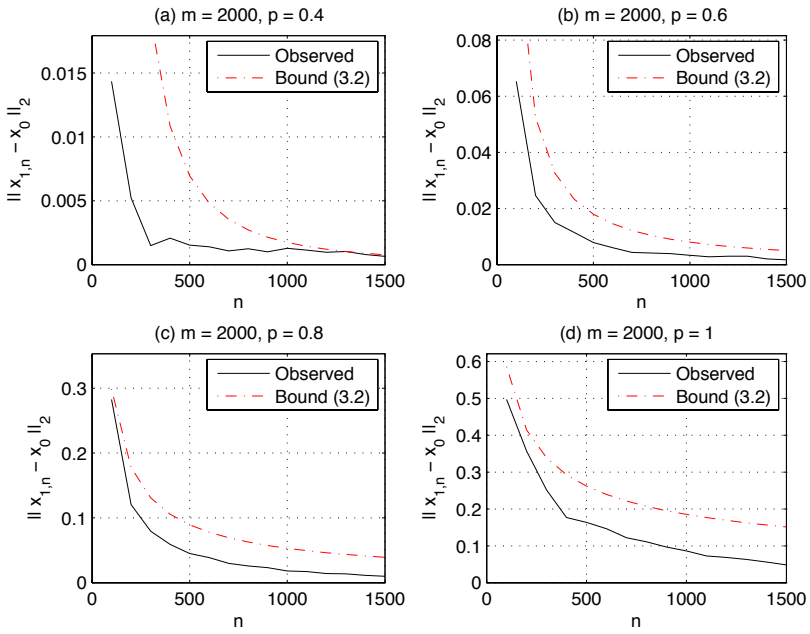

Figure 5: Reconstruction error versus number of measurements n, for fixed signal length $m = 2000$: (a) $p = 0.4$, (b) $p = 0.6$, (c) $p = 0.8$, (d) $p = 1$. Solid curves show maximum error in 20 pseudo-random replications at each (n, m) combination. Dashed curves show quasi-bound (3.2) using constants $\tilde{C}_p$ from Table 1.

With this estimate in hand, we wish to see if the empirical reconstruction error behaves in the manner (1.2) predicts, as n varies. For this purpose, we conducted a simulation study, this time keeping m, p constant as n grows. Results, in the form of error plots versus the number of samples n, are shown in Figure 5, panels (a)-(d), for $m = 2000$ and $p = 0.4, 0.6, 0.8, 1$, respectively. Each plot shows the measured empirical error, alongside a plot of the quasibound derived from the putative relationship

$$\|\hat{x}_{1,n} - x_0\|_2 \lesssim \tilde{C}_p \cdot \|x_0\|_p \cdot (n/\log(m))^{1/2-1/p}; \tag{3.2}$$

this is similar to (1.2) with C_p replaced by our empirical quantity $\tilde{C}_p$. We employ the quasi-inequality symbol $\lesssim$ to remind us that this is a tool for summarizing observed error behavior rather than a mathematical formula like (1.2).

Figure 5 shows how (3.2) compares to the observed error for practical signal lengths. As n increases, the error decreases like a power law depending on p. Moreover, we observe that for smaller p, the error tends to drop much more rapidly, as the bound predicts.

At the suggestion of the referees, we conducted another experiment, this time keeping n, the number of sensed samples, constant, and varying m, the signal length. Results are displayed in

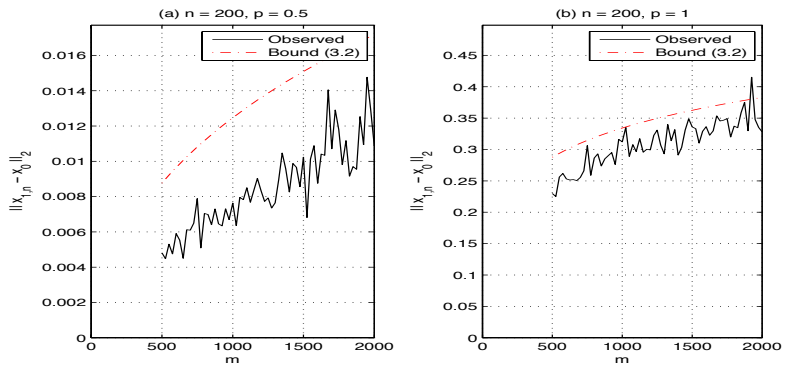

Figure 6: Reconstruction error versus signal length m, for fixed number of measurements $n = 200$: (a) $p = 0.5$, (b) $p = 1$. Solid curves show error in 20 pseudo-random replications. Dashed curves show quasi-bound (3.2) using constants $\tilde{C}_p$ from Table 1.

Figure 6, panels (a),(b), for $n = 200$ random measurements and $p = 0.5, 1$. Again, it appears that (3.2) reflects the behavior of the reconstruction error $\|\hat{x}_{1,n} - x_0\|_2$ as m grows.

A referee asked us to compare the bounds (3.2) with observed error behavior for 'ℓ^0 sparse' signals, discussed in the previous section. For signals of length m with k nonzeros of equal amplitude A, the ℓ^p norm is $A \cdot k^{1/p}$. This value may be used in the quasibound (3.2) along with the estimates for $\tilde{C}_p$ in Table 1, and contrasted with the empirical reconstruction error displayed in Figure 1. Figure 7 displays results for several values of p. Clearly, ℓ^0-sparse signals yield better error behavior than any of the bounds (3.2).

4 Different Ensembles of CS Matrices

In the examples shown so far, we have constructed CS matrices Φ from the uniform spherical ensemble described in [13]. Numerous other possibilities exist for construction of CS matrices. To clarify what we mean, we define four specific ensembles of random matrices

- *Random Signs Ensemble.* Here Φ_{ij} has entries $\pm 1/\sqrt{n}$ with signs chosen independently and both signs equally likely.
- *Uniform Spherical Ensemble.* The columns of Φ are iid random uniform on the sphere S^{n-1}.
- *Partial Fourier Ensemble.* We select at random n rows out of the $m \times m$ Fourier matrix, getting an n-by-m Partial Fourier matrix.
- *Partial Hadamard Ensemble.* We select at random n rows out of the $m \times m$ Hadamard matrix, getting an n-by-m Partial Hadamard matrix. (For this purpose we consider only $m = 2^k$.)

These choices are inspired by the following earlier work:

- The work of Kashin [21], followed by Garnaev and Gluskin [17], implicitly considered the random signs ensemble in the dual problem of Kolmogorov n-widths. Owing to a duality relationship between Gel'fand and Kolmogorov n-widths ([25]), and a relationship between Gel'fand n-widths and compressed sensing [13, 26] these matrices are suitable for use in the case $p = 1$.

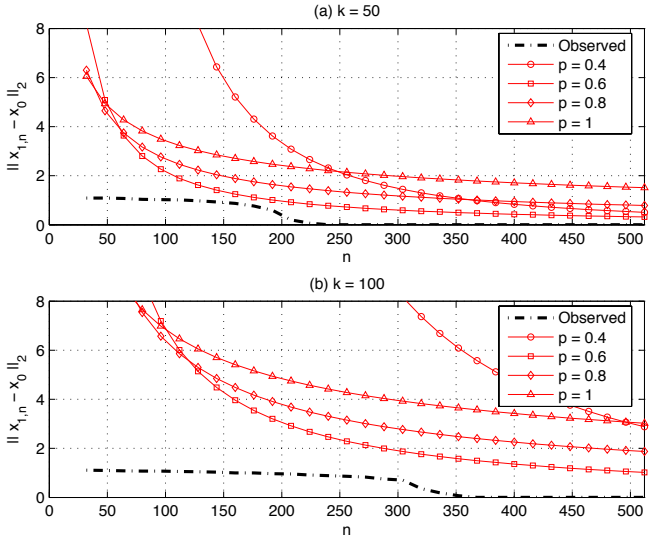

Figure 7: Reconstruction error versus number of measurements n, for fixed signal length $m = 1024$ and (a) k = 50 nonzeros, (b) k = 100 nonzeros. Dashed curve: observed error. Solid curve: error bounds (3.2) for various p.

- Donoho [11, 12, 13] considered the uniform Spherical ensemble.
- Candès, Romberg, and Tao [3] recently have generated a great deal of excitement by showing several interesting properties of Random Partial Fourier matrices and making claims about their possible use in Compressed Sensing [4].
- Partial Hadamard matrices are known to generate near-optimal subspaces in certain special cases for the related problem of determining Kolmogorov n-widths of the octahedron $b_{1,m}$ with respect to ℓ_m^∞ norm; see Pinkus' book [25].

We also remark that there are numerous very interesting practical applications where Partial Fourier and Partial Hadamard matrices are of direct interest, for example in Fourier transform imaging and Hadamard transform spectroscopy. Moreover, due to the special structure of the transforms underlying these matrices, the use of such matrices greatly expands the applicability of the CS scheme in cases where the data size is large, as would be the case for 2-D or 3-D data.

In Figure 8, we compare the quasibound (3.2) with actual errors in the different matrix ensembles just defined. In doing so, we follow the procedure of the previous section. Specifically, for each ensemble, we consider an object defined as in Section 3, i.e. an m-vector whose k-th largest amplitude coefficient $|\theta|_{(k)}$ obeys (3.1). A typical example is shown in Figure 4(a). We set $p = 3/4, m = 2048$, and consider families of experiments where n, the number of measurements, varies. For each n, we apply the CS framework and measure the ℓ^2 reconstruction error. We repeat this experiment 20 times for each (n, m, p) triplet, and record the maximal error. Figure 8 depicts error versus sampling n for the four different ensembles described above. We also display the error quasi-bound (3.2). Here the the constant $\tilde{C}_p$ was taken from Table 1, i.e. $\tilde{C}_{3/4} \approx 0.09$.

Figure 8 prompts several observations. First, the simulation results for the different ensembles are all qualitatively in agreement with the theoretical form of the error behavior in (1.2). Moreover, the relationship (3.2) gives a fairly good description of the true behavior observed in practice. And perhaps most importantly, we observe that the different ensembles show similar

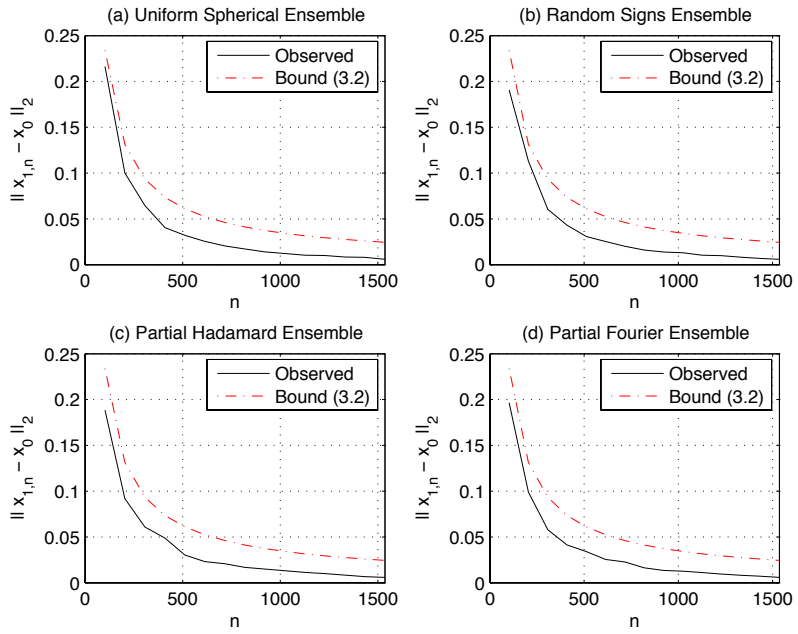

Figure 8: Error versus number of measurements n for $p = 3/4, m = 2048$: (a) Uniform Spherical Ensemble (b) Random Signs Ensemble (c) Partial Hadamard Ensemble; (d) Partial Fourier Ensemble.

behavior. This suggests that all such ensembles are equally good in practice. In the rest of the paper we continue to use the spherical uniform ensemble.

5 Denoising CS Results

When the object is undersampled, CS reconstructions are typically noisy. We considered the object *Bumps* from the Wavelab package [1], rendered with $m = 2048$. As Figure 9 panel (a) shows, the object is a superposition of smooth bumps. Panel (b) shows that the large coefficients at each scale happen near the bump locations, and panel (c) shows the decreasing rearrangement of the wavelet coefficients on a log scale. The linear appearance of this display is indicative of power-law behavior. We applied the CS framework, with Φ drawn, as usual, from a uniform spherical distribution, and Ψ an orthogonal wavelet basis, with Daubechies 'symmlet8' filters. Panel (a) of Figure 10 shows the results of the reconstruction with $n = 256$ measurements. Panel (b) shows the result with $n = 512$. Clearly both results are 'noisy', with, understandably, the 'noisier' one coming at the lower sampling rate.

Indeed, the results in Figures 3 and 10 show that 'noise' sometimes appears in reconstructions as the number of measurements decreases (even though the data are not noisy in these examples). To alleviate this phenomenon, we considered the test cases shown earlier, namely *Blocks* and *Bumps*, and applied translation-invariant wavelet de-noising [6] to the reconstructed 'noisy' signals. Results are shown in panel (c) of Figure 3 and panels (b) and (d) of Figure 10. At least visually, there is a great deal of improvement.

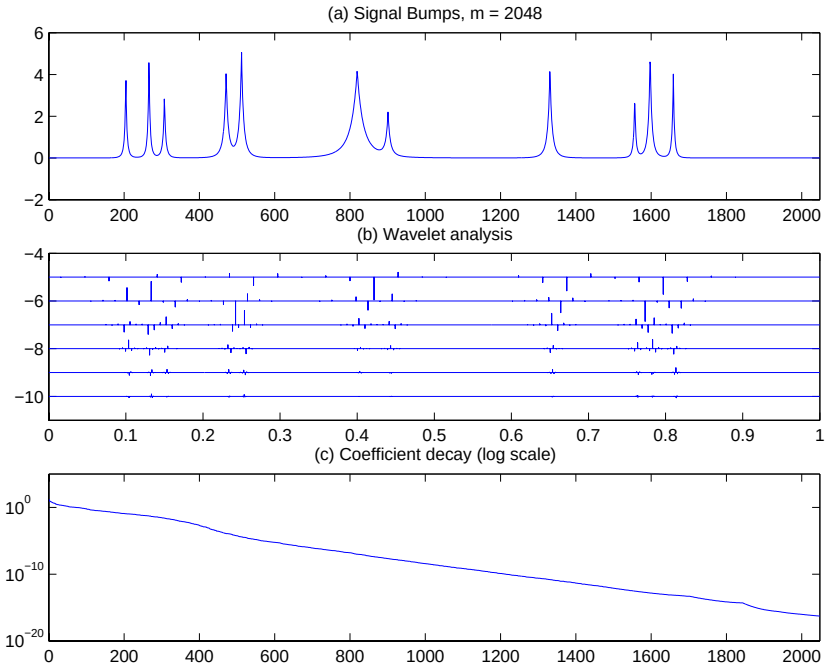

Figure 9: (a) Signal *Bumps*, $m = 2048$; (b) its wavelet coefficients; (c) Decay of coefficient magnitudes on a log scale. Rearranged coefficients exhibit roughly exponential decay.

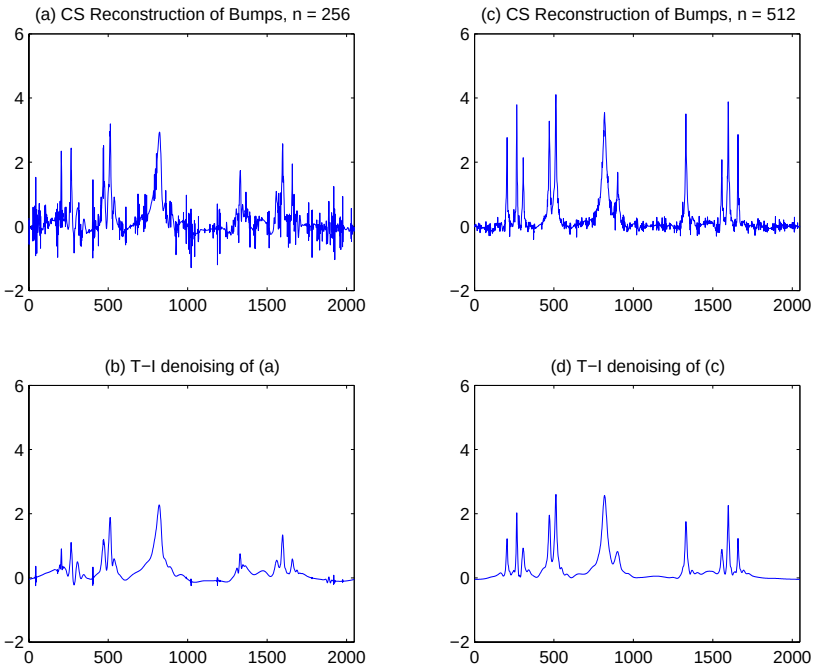

Figure 10: CS reconstruction of *Bumps* from (a) $n = 256$ and (c) $n = 512$ measurements. Translation-invariant denoising in (b) and (d).

6 Noise-Aware Reconstruction

So far we have not allowed for the possibility of measurement noise, digitization errors, etc. That is, we have assumed that the raw measurements y are perfectly-observed linear combinations of the underlying object – even in that case, the reconstructions can appear noisy, but not because of any 'noise' in the system. Now we consider the case where the data actually are noisy.

We remark that the theory allows for accommodating a small amount of noise already, through the ℓ^1-stability property proved in [13]. To go further, assume that rather than measuring $y = \Phi\Psi^T x$, we measure $y_n = \Phi\Psi^T x + z$, where z obeys $\|z\|_2 \le \epsilon$. To accommodate this noise, our primary adjustment would be to use *Basis Pursuit de-noising* (BPDN) rather than Basis Pursuit. For a given noise level $\epsilon > 0$, define the optimization problem

$$(L_{1,\epsilon}) \qquad \min \|\Psi^T x\|_1 \text{ subject to } \|y_n - \Phi\Psi^T x\|_2 \le \epsilon.$$

This can be written as a linearly constrained convex quadratic program, and is considered practical to solve. In [5] BPDN was successfully used in cases where $n < m$ and both are quite large: $n = 8192$ and $m = 262144$. Our proposal for dealing with noisy data is simply to measure $y_n = \Phi\Psi^T x + noise$ and then use $(L_{1,\epsilon})$ with an appropriate noise tolerance $\epsilon > 0$.

(A referee asked us to clarify the theoretical support for this suggestion. We discuss the implications of [10, 12], mentioning that work of Jean-Jacques Fuchs and Joel Tropp [16, 28] is also relevant. The results in [10, 12] suggest to expect stable reconstructions from the solution to $(L_{1,\epsilon})$. They suppose that we observe

$$y = \Phi\Psi^T x_0 + z \tag{6.1}$$

where z is an arbitrary disturbance. Under sufficient sparsity of the representation of x_0, they show that, if $\epsilon > \|z\|_2$, $(L_{1,\epsilon})$ gives stable recovery of the sparsest representation: for a constant C depending on n, m, and the sparsity of x_0,

$$\|\hat{x}_{1,\epsilon} - x_0\|_2 \le C\|z\|_2.$$

This motivates the use of $(L_{1,\epsilon})$ for compressed sensing of objects with ℓ^0 sparsity. How about ℓ^p sparsity? We note that for an object with coefficients $\Psi^T x \in \ell^p$, its best N-term approximation x_N obeys

$$\|x_N - x\|_2 \le \zeta_p \cdot \|\Psi^T x\|_p \cdot N^{1/2-1/p}, \qquad N \ge 0, \tag{6.2}$$

with $\zeta_p > 0$ a universal constant. To apply this, suppose we observe

$$y = \Phi\Psi^T x + z_0$$

where $\Psi^T x \in \ell^p$, we may rewrite this as

$$y = \Phi\Psi^T x_N + z_0 + z_1$$

where $z_1 = \Phi\Psi^T(x - x_N)$. Letting $\epsilon_1 = \|z_1\|_2$ (which is controlled by (6.2)) and $\epsilon_0 = \|z_0\|_2$, we get an instance of the model (6.1) with $z = z_0 + z_1$ and $\|z\|_2 \le \epsilon_0 + \epsilon_1$. More recent work [31, 34] also supports stability of this approach; see also Section 9.4.)

To see the performance of $(L_{1,\epsilon})$, we offer proof-of-concept results. We took the test signals *Blocks* and *Bumps*, shown in Figures 2 and 9, respectively, and added zero-mean white gaussian noise to them. The noise was rescaled to enforce a specific noise level $\|z\|_2 = 0.2$. We applied the compressed sensing scheme with denoising (CSDN) to the noisy wavelet expansions. For

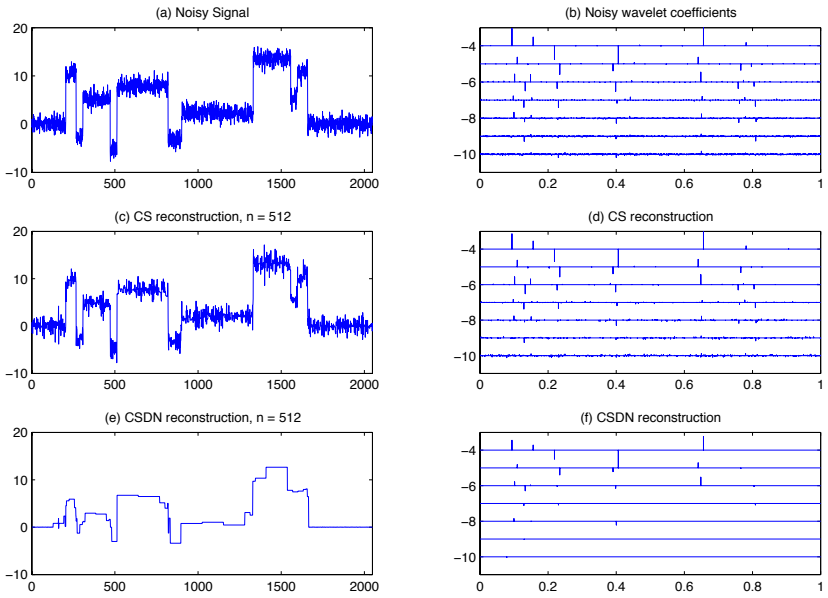

Figure 11: CSDN reconstruction of *Blocks*, $m = 2048, n = 512$. Signal and reconstructions are shown on the left panel, with corresponding wavelet expansions on the right panel.

comparison, we also attempted regular CS reconstruction. Results are shown in Figures 11 and 12. We used signal length $m = 2048$, and attempted reconstructions with $n = 512$. Indeed, the reconstruction achieved with CSDN is far superior to the CS reconstruction in both cases.

A referee has remarked (and we agree) that, in order to apply this method, it is important to know the noise level so that ϵ can be specified appropriately.

Another referee has remarked (and we agree) that for some compression purposes, measurements y will be quantized, inducing quantization noise in the observations and prompting the use of CSDN. Luckily, in this setting the noise level is known.

7 Two-Gender Hybrid CS

In [13] model spectroscopy and imaging problems were considered from a theoretical perspective. CS was deployed differently there than so far in this paper – in particular, it was not proposed that CS alone 'carry all the load'. In that deployment, CS was applied to measuring only *fine-scale* properties of the signal, while ordinary linear measurement and reconstruction was used to obtain the coarse-scale properties of the signal.

In more detail, the proposal was as follows; we spell out the ideas for dimension 1 only. Expand the object x_0 in the wavelet basis

$$x_0 = \sum_k \beta_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{j_1}\sum_k \alpha_{j,k}\psi_{j,k}$$

where j_0 is some specified coarse scale, j_1 is the finest scale, $\phi_{j_0,k}$ are male wavelets at coarse scale and $\psi_{j,k}$ are fine scale female wavelets. Let $\alpha = (\alpha_{j,k} : j_0 \le j \le j_1, 0 \le k < 2^j)$ denote the grouping together of all wavelet coefficients, and let $\beta = (\beta_{j_0,k} : 0 \le k < 2^{j_0})$ denote the male

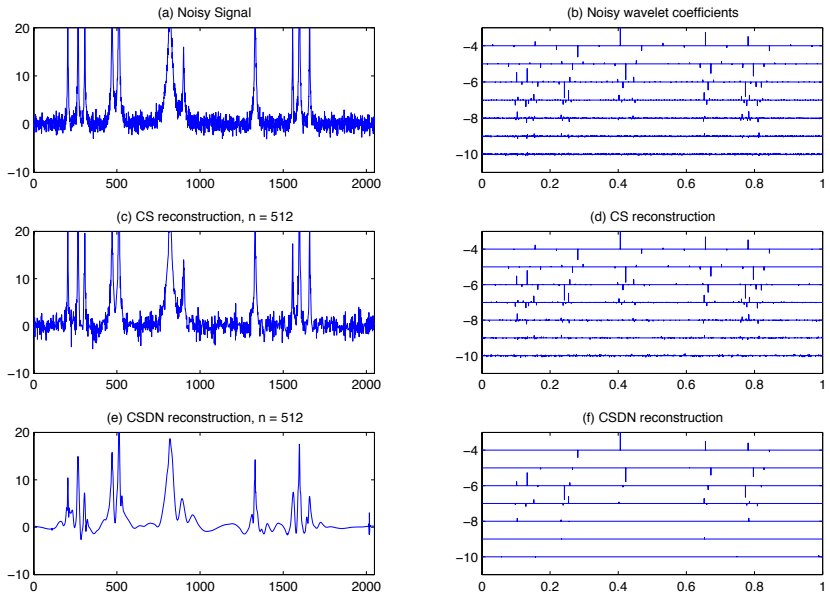

Figure 12: CSDN reconstruction of *Bumps*, $m = 2048, n = 512$. Signal and reconstructions are shown on the left panel, with corresponding wavelet expansions on the right panel.

coefficients. Now consider a scheme where different strategies are used for the two genders α and β. For the male coarse-scale coefficients, we simply take direct measurements

$$\hat{\beta} = (\langle \phi_{j_0,k}, x_0 \rangle : 0 \le k < 2^{j_0}).$$

For the female fine-scale coefficients, we apply the CS scheme. Let $m = 2^{j_1} - 2^{j_0}$, and let the $2^{j_1} \times m$ matrix Ψ have, for columns, the vectors $\psi_{j,k}$ in some standard order. Given an n by m CS matrix Φ, define $\Xi = \Phi\Psi^T$, so that, in some sense, the columns of of Ξ are 'noisy' linear combinations of columns of Ψ. Now make n measurements

$$y = \Xi x_0.$$

To reconstruct from these observations, define $\Omega = \Xi\Psi$ and consider the basis-pursuit optimization problem [5]:

$$(BP) \quad \min_a \|a\|_1 \text{ subject to } y_n = \Omega a; \tag{7.1}$$

a minor relabelling of the (L_1) problem. Call the answer $\hat{\alpha}$. The overall reconstruction is

$$\hat{x}_{hy} = \sum_k \hat{\beta}_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{j_1} \sum_k \hat{\alpha}_{j,k}\psi_{j,k},$$

a combination of linear reconstruction at coarse scales and nonlinear reconstruction based on undersampling at fine scales. When we speak of this scheme, of course, the total number of samples $n_{hy} = 2^{j_0} + n$ is the number of linear samples plus the number of compressed samples. The theory derived in [13] shows that this hybrid scheme, when applied to objects obeying certain constraints (e.g. bounded variation) gets accuracy comparable to linear sampling at scale 2^{-j_1}, only using many fewer total samples.

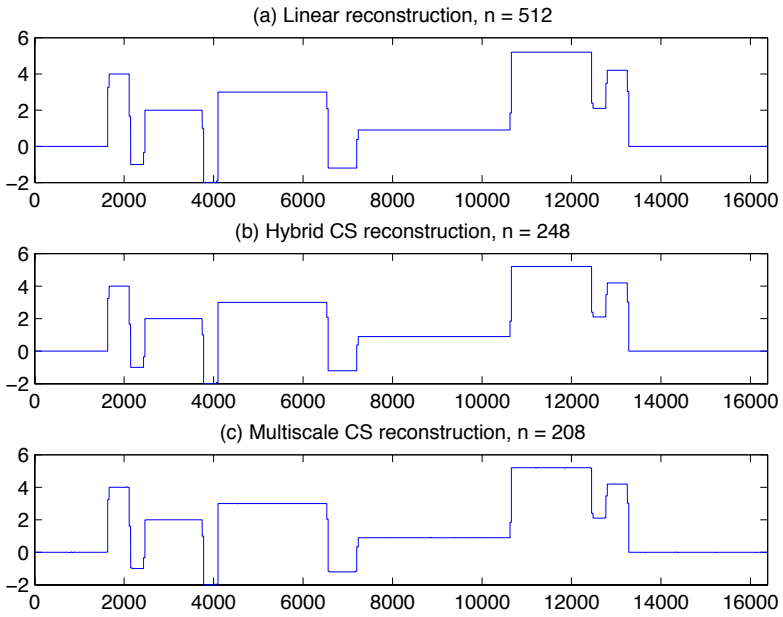

Figure 13: Reconstruction of Signal *Blocks*, $m = 16384$. (a) Linear reconstruction from 512 samples, $\|\hat{x}_{lin} - x_0\|_2 = 0.091$; (b) Hybrid CS reconstruction from 248 samples (Gender-Segregated), $\|\hat{x}_{hy} - x_0\|_2 = 0.091$; (c) Multiscale CS reconstruction from 208 samples, $\|\hat{x}_{ms} - x_0\|_2 = 0.091$.

The gender-segregated deployment of CS was suggested in [13] for mathematical convenience, but in experiments, it substantially outperforms straightforward gender-blind deployment. To see this, consider Figure 13. Panel (a) shows a blocky signal of original length $m = 16384$ reconstructed from $n = 512$ linear samples, where we set a coarsest scale $j_0 = 5$, and a finest scale $j_1 = 9$. Panel (b) shows reconstruction with $n_{hy} = 248$ hybrid compressed samples (32 male samples, 216 compressed female samples). As before, we used a sampling matrix Φ drawn from a uniform spherical distribution, and Ψ a Haar wavelet basis. The accuracy is evidently comparable. It is far better than panel (b) of Figure 3, which shows the result reconstruction from 256 standard CS samples, for $m = 2048$.

Now consider Figure 14. Panel (a) shows a bumpy signal of original length $m = 16384$ reconstructed from $n = 1024$ linear samples, where we set a coarsest scale $j_0 = 5$, and a finest scale $j_1 = 10$. We applied hybrid reconstruction, again using a sampling matrix Φ drawn from a uniform spherical distribution, and Ψ an orthogonal wavelet basis, with Daubechies' 'symmlet8' filters. Panel (b) shows the reconstruction result, with $n_{hy} = 640$ hybrid compressed samples (32 male samples, 608 compressed female samples). Again the reconstruction accuracy is comparable. It is far better than panel (b) of Figure 10, which shows the result of reconstruction from 512 standard CS samples, for $m = 2048$.

In [13] the idea of gender-segregated sampling was extended to higher dimensions in considering the class of images of bounded variation. The ideas are a straightforward extension of the 1-D case, and we shall not repeat them here. We investigate the performance of hybrid CS sampling applied to image data in the following experiment. Figure 15 shows the reconstruction results for a *Mondrian* image of size 1024×1024, so that $m = 2^{20}$. We again used the Haar wavelet expansion, which is naturally suited to images of this type, with a coarsest scale $j_0 = 3$, and a finest scale $j_1 = 6$. Panel (a) shows the result of linear reconstruction with $n = 4096$ sam-

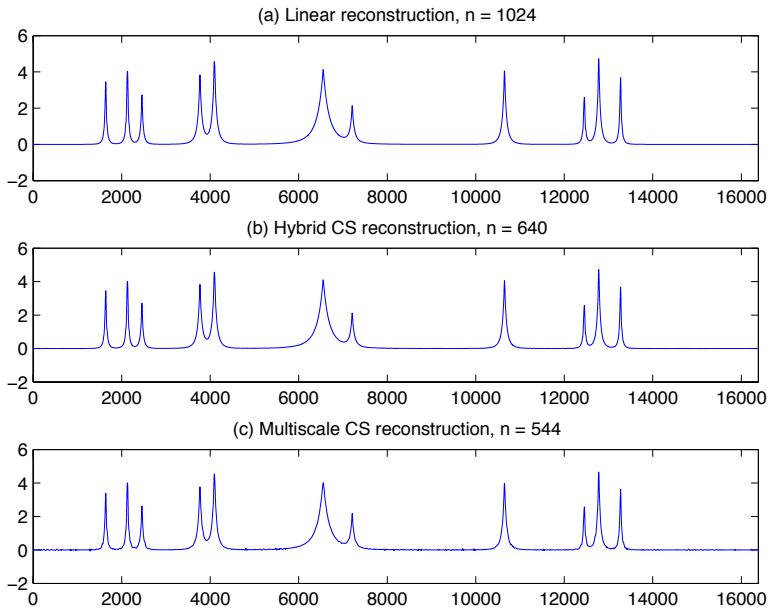

Figure 14: Reconstruction of Signal *Bumps*, $m = 16384$. (a) Linear reconstruction from 1024 samples, $\|\hat{x}_{lin} - x_0\|_2 = 0.0404$; (b) Hybrid CS reconstruction from 640 samples (Gender-Segregated), $\|\hat{x}_{hy} - x_0\|_2 = 0.0411$; (c) Multiscale CS reconstruction from 544 samples, $\|\hat{x}_{ms} - x_0\|_2 = 0.0425$.

ples, whereas Panel (b) has results for the hybrid CS scheme with $n_{hy} = 1152$ hybrid compressed samples (128 male samples, 1024 compressed female samples). The reconstruction accuracy is evidently comparable.

8 Multiscale Compressed Sensing

Encouraged by the apparent usefulness of Hybrid CS, we now consider a fully multiscale deployment of CS. The simplest way to explain the concept is to use for our multiscale system a standard 1-D orthogonal wavelet system. The same idea can be applied with other multiscale systems and in higher dimensions.

Expand the object x_0 in the wavelet basis

$$x_0 = \sum_k \beta_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{j_1}\sum_k \alpha_{j,k}\psi_{j,k}$$

Consider now a multilevel stratification of the object in question, partitioning the coefficient vector as

$$[(\beta_{j_0,\cdot}), (\alpha_{j_0,\cdot}), (\alpha_{j_0+1,\cdot}), \ldots, (\alpha_{j_1-1,\cdot})]$$

We then apply ordinary linear sampling to measure the coefficients $(\beta_{j_0,\cdot})$ directly, and then separately apply compressed sensing scale-by-scale, sampling data y_j about the coefficients $(\alpha_{j,\cdot})$ at level j using an $n_j \times 2^j$ CS matrix Φ_j. We obtain thereby a total of

$$n_{ms} = 2^{j_0} + n_{j_0} + n_{j_0+1} + \cdot + n_{j_1-1}$$

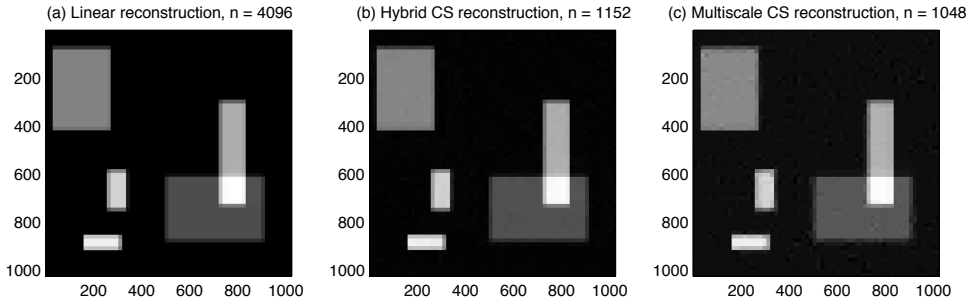

Figure 15: Reconstruction of *Mondrian* image, $m = 2^{20}$. (a) Linear reconstruction from 4096 samples, $\|\hat{x}_{lin} - x_0\|_2 = 0.227$; (b) Hybrid CS reconstruction from 1152 samples (Gender-Segregated), $\|\hat{x}_{hy} - x_0\|_2 = 0.228$; (c) Multiscale CS reconstruction from 1048 samples, $\|\hat{x}_{ms} - x_0\|_2 = 0.236$.

samples, compared to

$$m = 2^{j_0} + 2^{j_0} + 2^{j_0+1} + \cdot + 2^{j_1-1} = 2^{j_1}$$

coefficients in total. To obtain a reconstruction, we then solve the sequence of problems

$$(BP_j) \qquad \min_a \|a\|_1 \text{ subject to } y_j = \Phi_j a, \qquad j = j_0, \ldots, j_1 - 1,$$

calling the obtained solutions $\hat{\alpha}^{(j)}$; to reconstruct, we set

$$\hat{x}_{ms} = \sum_k \beta_{j_0,k}\phi_{j_0,k} + \sum_{j=j_0}^{j_1-1}\sum_k \hat{\alpha}_k^{(j)}\psi_{j,k}.$$

(Of course, variations are possible; we might group together several coarse scales $j_0, j_0 + 1, \ldots, j_0 + \ell$ to get a larger value of m.)

For an example of results obtained using Multiscale CS, consider Figure 13(c). It shows the signal *Blocks* reconstructed from $n_{ms} = 208$ compressed samples ($n_{j_0} = 32$ coarse-scale samples, $n_j = \min\{2^{j-1}, 48\}$ compressed samples at each detail scale j, $j_0 < j \le j_1$). Indeed, the reconstruction accuracy is comparable to the linear and gender-segregated results. Similarly, Figure 14(c) has multiscale reconstruction of *Bumps* from $n_{ms} = 544$ compressed samples ($n_{j_0} = 32$ coarse-scale samples, $n_j = \min\{2^{j-1}, 144\}$ compressed samples at each detail scale j, $j_0 < j \le j_1$). Again the reconstruction accuracy is comparable to that achieved by the other methods. Finally, Figure 15(c) has multiscale CS reconstruction of the *Mondrian* image from $n_{ms} = 1048$ compressed samples.

Consider now an example working with a frame rather than an ortho-basis, in this case the Curvelets frame [2]. Theory supporting the possible benefits of using this frame for cartoon-like images was developed in [13].

Like the wavelet basis, there is a scale parameter j which specifies the size of the curvelet frame element; we considered a deployment of multiscale compressed sensing which used different n_j at each level. In Figure 16 we give the results of this scheme applied to the familiar *Shepp-Logan Phantom* image.

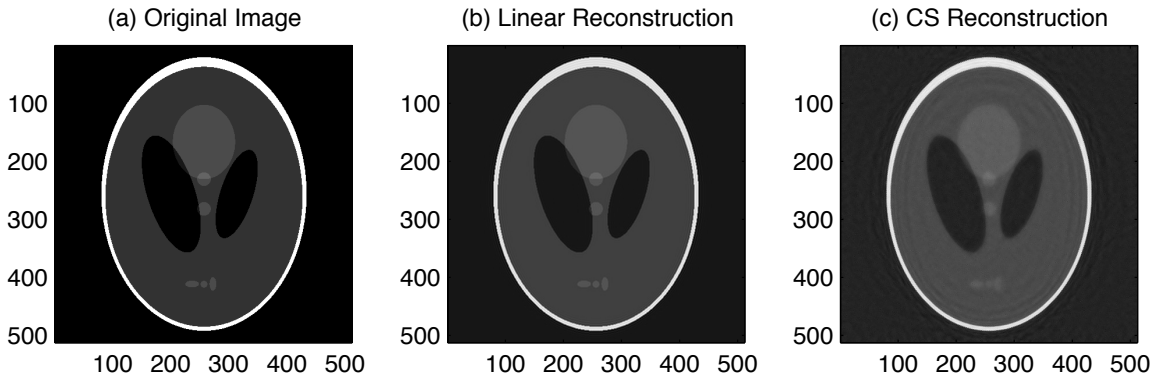

Figure 16: (a) *Shepp-Logan Phantom* image, 742400 Curvelet coefficients; (b) Reconstruction from 480256 Linear Samples, $\|\hat{x}_{lin} - x_0\|_2 = 0.142$; (c) Reconstruction from 103218 Multiscale Compressed Samples using the Curvelets Frame, $\|\hat{x}_{ms} - x_0\|_2 = 0.134$.

To help the reader gain more insight into the level-by-level performance, we compare in Figure 17 the image coefficients and the reconstructed ones. Clearly, there is additional noise in the reconstruction. Nonetheless, this noise is not visually evident in the overall CS reconstruction.

9 Conclusions

In this paper, we have reviewed the basic Compressed Sensing framework, in which we gather n pieces of information about an object which, nominally, has m degrees of freedom, $n < m$. When the object is compressible in the sense that the ℓ^p norm of its transform coefficients is well-controlled, then by measuring n essentially 'random' linear functionals of the object and reconstructing by ℓ^1 minimization on the transform coefficients, we get, in theory, an accurate reconstruction; see (1.2).

9.1 Initial Observations

We raised a number of applications-oriented questions. The first concerns the strength of the theoretical bound (1.2). This bound contains unspecified constants, and so the practical relevance of the framework depends heavily on the precise values of those constants. We conducted a number of simulations to compare observed behavior with the theoretical inequality (1.2). We obtained constants in the empirical relationship (3.2), interested to know if these are small enough to have practical implications, even at moderate n and m, m in the low thousands, n in the few hundreds. We concluded that this is so.

We note however, that the method comes off much more impressively when the underlying object has relatively few nonzeros than it does when the object has a large number of small nonzeros (as allowed by the ℓ^p model). At the problem sizes considered here, when the number of samples exceeds the number of nonzeros by a factor of 4 or so, the method often performs exceptionally well. Unfortunately, signals made of only a few nonzero terms are rather special.

We conducted numerical experiments on objects *Blocks* and *Bumps*, caricaturing spectra and scan lines in images; for such objects, whose visual appearance can be gauged, we found that below a certain threshold on n, the CS reconstructions look visually noisy – even though there is no noise in the observations.

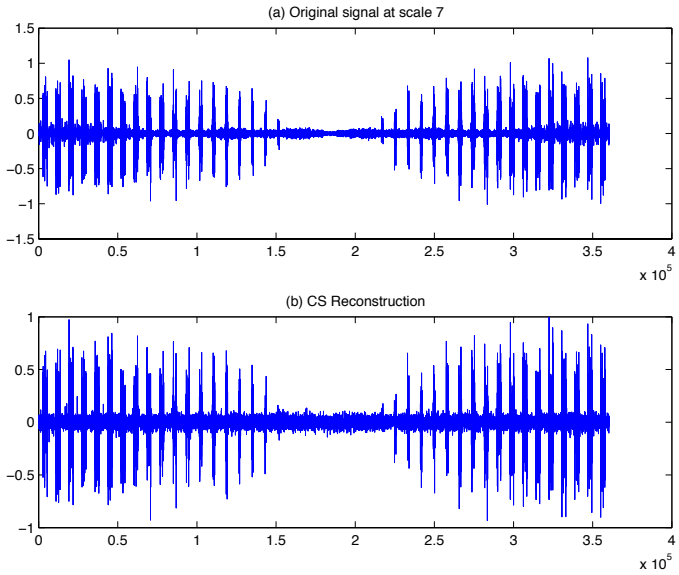

Figure 17: (a) Exact Curvelet coefficients of *Shepp-Logan* at scale 7, (b) Reconstructed coefficients in CS framework.

9.2 Extensions

Our main effort in this article has been to go beyond these initial observations by considering several extensions of the CS framework. These extensions include,

- *Postprocessing noise removal.* In order to defeat the appearance of visual noise in the CS reconstructions, we applied post-processing to the CS reconstruction by translation-invariant de-noising with a level-dependent threshold. We found the appearance of visual noise dramatically reduced (for objects *Blocks* and *Bumps*).

- *Allowing noise in the measurements.* The basic CS framework does not allow for noise in the observations. To handle the case where noise is present, we suggested the use of a noise-tolerant ℓ^1 minimization, essentially using 'Basis-Pursuit Denoising' in place of Basis Pursuit, on which CS is based. We presented examples showing that this can in our examples substantially reduce the effects of noise on the reconstruction.

- *Multiscale Deployment.* More significantly, we considered the use of CS *not* as the only way to gather information about an object, but as a tool to be used in conjunction with classical linear sampling at coarse scales. We considered a hybrid method, already discussed in [13] from a theoretical perspective, in which very coarse-scale linear sampling was combined with CS applied at all finer scales. We also introduced a fully multiscale method in which CS was applied at each scale separately of all others, and linear sampling only at the coarser scale. Both approaches can significantly outperform linear sampling, giving comparable accuracy of reconstruction with far fewer samples in our examples. This tends to support the theoretical claim in [13] that compressed sampling, deployed in a Hybrid fashion, can, for certain kinds of objects, produce an order-of-magnitude improvement over classical linear sampling.

- *Alternate CS Matrices.* Perhaps of most significance is our empirical finding that several apparently different random matrix ensembles all perform similarly when used in the CS

framework. Such a finding, if true in general, is important for algorithmic reasons. To apply interior-point methods to solve the linear program (1.1) requires many matrix vector products Φu and $\Phi^T v$ for strategically chosen vectors u and v [5]. It is *much* faster to apply matrices Φ and Φ^T defined by the Partial Fourier and Partial Hadamard ensembles, where the cost will be $O(m \log(m))$ flops. Clearly, for large m and n one would much prefer to use such matrices over the Random Uniform Ensemble or the Random Signs ensemble, which cost $O(nm)$ flops to apply. It appears that we have many choices of matrices yielding adequate performance in the CS framework, consistent with the Theorems in [13], and that among those matrices are some which can be applied rapidly.

Compressed Sensing seems a promising strategy for sampling signals characterized by large numbers of nominal samples and yet exhibiting high compressibility by known transforms like the wavelet transform or Fourier transform. The limited experiments conducted here already show gains by factors of 2, 4, or even more in moderate-size problem sizes, and these can be enhanced by deployment in a multiscale fashion and by applying de-noising to reconstructions.

Of course the theoretical implications of the bound (1.2) seem even stronger than what we observe here, in the sense that at least for n and m large, really dramatic benefits ought to be possible for certain classes of compressible objects. When this paper was submitted, several groups were already exploring such ideas, first of course being Candès, Romberg, and Tao, who are actively pursuing the implications of [3, 4, 30], which have proved so inspiring. In addition, R.R. Coifman at Yale has conduct actual physical experiments in the CS framework using spectroscopic equipment. Clearly we can expect far more experimentation and theoretical development of such ideas in the near future.

9.3 Directions for Further Research

- *Compressed Sensing as a Compression Scheme.* Our paper has viewed CS as a tool for reducing the number of measurements rather than a tool for data compression. Our viewpoint is appropriate when measurements are expensive to make, but storing and communicating the measured values is cheap. In that setting the dominant cost is the measurement process itself. For example, measurement processes may involve long delays, radiation exposure, power consumption, or large numbers of dedicated physical device components. Our viewpoint seeks to minimize such costly factors.

  A different viewpoint takes CS as part of a compression scheme in which system bit rate (rather than measurement resource) is the driving goal. In this viewpoint, the raw CS measurements need to be quantized to achieve lower bit rates, and the reconstruction would involve the use of such quantized data.

  This alternate interpretation would focus new scrutiny of the stability of the CS concept. Some of the issues have already been addressed here in passing. We have already remarked, as was already pointed out in [13], that reconstruction by (L_1) automatically has a certain degree of stability (in ℓ^1 norm), which appears to be sufficient if the data are very accurately quantized. We also remarked, in Section 6 that $(L_{1,\epsilon})$ provides stability in ℓ^2 norm. However, much more detailed studies of this effect would be needed. We believe that Candès and his team are carrying out such studies, and encourage the interested reader to follow their work.

- *Alleviating Reconstruction Noise.* As sections 2 and 3 demonstrated, when the number of sensed samples is small we often end up with a noisy reconstruction, although the data were

noise-free. In Section 5 we tackled this issue by a postprocessing noise removal scheme. Presumably, however, this problem can be approached more directly, after more carefully modelling the statistical properties of the reconstruction errors. Such models would lead to improved denoising. Imposing other constraints, such as total variation control and positivity constraints, where relevant, should also be helpful.

- *Hybrid/Multiscale Compressed Sensing.* The proof-of-concept examples presented in Sections 7 and 8 demonstrated clear advantages to using either a Hybrid or Multiscale sensing scheme, both in terms of performance, and in terms of computational complexity. Such initial results raise a number of interesting issues worthy of further study. Among these are:

  - *How to choose a coarsest scale?* Since the coefficients in the coarsest scale of the representation are measured directly, the choice of a coarsest scale has a significant impact on the performance of the sensing scheme, on one hand, and on the total number of measurements needed, on the other hand. Our approach is motivated by the fact that wavelet transforms typically have a series of 'full' levels at the coarse scales, followed by increasingly sparse levels at increasingly fine scales. So our simulations used what amounts to the 'finest full level' as the coarse scale. This issue bears further examination.
  - *How to divide a 'budget' between scales?* Once a coarsest scale has been chosen, for multiscale CS we also need to divide our 'budget' of allowable measurements between the 'detail' scales. The approach employed in this paper divided the samples equally among all scales. This was based on the pattern visible in the wavelet transform of *Blocks*: that singularities contribute the same number of nonzero coefficients at each scale. A referee asked whether this was really the best allocation of samples, and we agree that the issue bears further study.

9.4 More Recent Theoretical Work

In the months since this paper was submitted, a number of interesting theoretical papers have been circulated elaborating the basic inequality (1.2). Candès and Tao [4] and Rudelson and Vershynin [36] give alternate approaches to(1.2), which also cover, for example, the random signs ensemble. Such large-n results support the empirical evidence in Section 4 above, that such alternate ensembles perform about equally well as the random spherical ensemble. Haupt and Nowak [34] also studied the random signs ensemble and came to similar conclusions.

Work by Tropp and Gilbert [35] shows that in the setting of Section 2 – where one posits only a few nonzeros – one can typically use greedy stepwise approximation (orthogonal matching pursuit) – which is claimed to be more computationally efficient than ℓ^1 optimization. A cursory examination of their results suggests a threshold effect at a lower threshold than seen here, i.e. that recovery by greedy methods demands higher degrees of sparsity than recovery by ℓ^1 methods. Tropp and Gilbert also reach the conclusion that the random signs ensemble behaves similarly to the Gaussian ensemble.

The phase transitions observed in Section 2 have been carefully studied both in the companion paper [29] and, more recently, in [32, 33]. There it was shown that the threshold for perfect recovery of ℓ^0-sparse objects is proportional to n, for n/m bounded away from zero, and precise numerical estimates for the large-n setting have been obtained.

Candès and Tao [31] have improved our understanding of the noise-cognizant optimization procedure ($L_{1,\epsilon}$) with a streamlined proof of stability that extends naturally to other matrix

ensembles.

9.5 More Recent Computational Work

Apparently, Candès and Romberg at CalTech were in Fall 2004 conducting experiments paralleling those reported in this paper; these have now been circulated in [30] which is closely related to our work and should be of interest to any reader who has gotten this far. Both Tropp and Gilbert [35] and Alex Petukhov (personal communication) have made extensive simulations on the behavior of the greedy algorithm in a CS setting. Haupt and Nowak [34] have also been conducting promising numerical experiments using the random-signs ensemble.

References

[1] J. Buckheit and D. L. Donoho (1995) WaveLab and reproducible research, in A. Antoniadis, Editor, *Wavelets and Statistics*, Springer, 1995.

[2] E. J. Candès and DL Donoho (2004) New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Comm. Pure and Applied Mathematics* **LVII** 219-266.

[3] E.J. Candès, J. Romberg and T. Tao. (2004) Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. Manuscript.

[4] E. J. Candès and T. Tao (2004) Near Optimal Signal Recovery From Random Projections: Universal Encoding Strategies? Manuscript.

[5] Chen, S., Donoho, D.L., and Saunders, M.A. (1999) Atomic Decomposition by Basis Pursuit. *SIAM J. Sci Comp.*, **20**, 1, 33-61.

[6] Coifman, R.R. and Donoho, D.L. (1995) Translation-invariant de-noising. In *Wavelets and Statistics*, Antoniadis, A. and Oppenheim, G. (Eds.), Lect. Notes Statist., 103, pp. 125-150, New York: Springer-Verlag.

[7] R.R. Coifman, Y. Meyer, S. Quake, and M.V. Wickerhauser (1990) Signal Processing and Compression with Wavelet Packets. in *Wavelets and Their Applications*, J.S. Byrnes, J. L. Byrnes, K. A. Hargreaves and K. Berry, eds. 1994,

[8] Donoho, D.L. and Huo, Xiaoming (2001) Uncertainty Principles and Ideal Atomic Decomposition. *IEEE Trans. Info. Thry.* **47** (no.7), Nov. 2001, pp. 2845-62.

[9] Donoho, D.L. and Elad, Michael (2002) Optimally Sparse Representation from Overcomplete Dictionaries via ℓ^1 norm minimization. *Proc. Natl. Acad. Sci. USA* March 4, 2003 **100** 5, 2197-2002.

[10] Donoho, D., Elad, M., and Temlyakov, V. (2004) Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. Submitted. URL: http://www-stat.stanford.edu/~donoho/Reports/2004.

[11] Donoho, D.L. (2004) For most large underdetermined systems of linear equations, the minimal ℓ^1 solution is also the sparsest solution. Manuscript. Submitted. URL: http://www-stat.stanford.edu/~donoho/Reports/2004

[12] Donoho, D.L. (2004) For most underdetermined systems of linear equations, the minimal ℓ^1-norm near-solution approximates the sparsest near-solution. Manuscript. Submitted. URL: http://www-stat.stanford.edu/~donoho/Reports/2004

[13] Donoho, D.L. (2004) Compressed Sensing. Manuscript. Submitted. URL: http://www-stat.stanford.edu/~donoho/Reports/2004

[14] M. Elad and A.M. Bruckstein (2002) A generalized uncertainty principle and sparse representations in pairs of bases. *IEEE Trans. Info. Thry.* **49** 2558-2567.

[15] J.J. Fuchs (2004) On Sparse Representations in Arbitrary Redundant Bases. *IEEE Trans. Info. Thry* **50** (no.6), June 2004, pp. 1341-44.

[16] J.J. Fuchs (2004) Recovery of exact sparse representations in the presence of bounded noise. Technical report IRISA no. 1618.

[17] Garnaev, A.Y. and Gluskin, E.D. (1984) On widths of the Euclidean Ball. *Soviet Mathematics – Doklady* **30** (in English) 200-203.

[18] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan and M. Strauss, (2002) Near-optimal sparse fourier representations via sampling, in *Proc 34th ACM symposium on Theory of Computing*, pp. 152–161, ACM Press.

[19] R. Gribonval and M. Nielsen. Sparse Representations in Unions of Bases. *IEEE Trans. Info. Thry* **49** (no.12), Dec. 2003, pp. 1320-25.

[20] R. Gribonval and M. Nielsen (2003). Highly Sparse Representations from Dictionaries are Unique and Independent of the Sparseness Measure. Manuscript.

[21] Boris S. Kashin (1977) Diameters of certain finite-dimensional sets in classes of smooth functions. *Izv. Akad. Nauk SSSR, Ser. Mat.* **41** (2) 334-351.

[22] S. Mallat, Z. Zhang, (1993). Matching Pursuits with Time-Frequency Dictionaries. *IEEE Trans. Sig. Proc.*, **41** (no.12), pp. 3397-3415.

[23] B.K. Natarajan (1995) Sparse Approximate Solutions to Linear Systems. *SIAM J. Comput.* **24**: 227-234.

[24] E. Novak (1996) On the power of Adaption. *Journal of Complexity* **12**, 199-237.

[25] Pinkus, A. (1985) *n-widths in Approximation Theory.* Springer-Verlag.

[26] Pinkus, A. (1986) n-widths and Optimal Recovery in *Approximation Theory*, Proceeding of Symposia in Applied Mathematics, **36**, Carl de Boor, Editor. American Mathematical Society, Providence, RI.

[27] J.A. Tropp (2003) Greed is Good: Algorithmic Results for Sparse Approximation. *IEEE Trans Info. Thry.* **50** (no.11), Oct. 2004, pp. 2231-42.

[28] J.A. Tropp (2004) Just Relax: Convex programming methods for Subset Sleection and Sparse Approximation. Manuscript.

[29] Y. Tsaig, D.L. Donoho (2004) Breakdown of Equivalence between the minimal ℓ^1-norm Solution and the Sparsest Solution. Submitted. URL: http://www-stat.stanford.edu/~donoho/Reports/2004

References Added in Revision.

[30] E. J. Candès and J. Romberg (2004) Practical Signal Recovery from Random Projections. Manuscript.

[31] E. J. Candès and T. Tao (2005) Stable Signal Recovery from noisy and incomplete observations. Manuscript.

[32] D.L. Donoho (2004) Neighborly polytopes and sparse solution of underdetermined linear equations. *Tech. report, Dept. of Statistics, Stanford Univ.* 2005-04

[33] D.L. Donoho (2005) High-Dimensional Centrosymmetric Polytopes with Neighborliness Proportional to Dimension. *Tech. report, Dept. of Statistics, Stanford Univ.* 2005-05

[34] J. Haupt and R. Nowak. (2005) Signal Reconstruction from Noisy Random Projections. Manuscript.

[35] J.A. Tropp and A. Gilbert (2005) Signal Recovery from Partial Information by Orthogonal Matching Pursuit. Manuscript.

[36] M. Rudelson & R. Vershynin (2005) Geometric approach to error-correcting codes and reconstruction of signals *Tech. report, Department of Mathematics, Univ. of California at Davis* 2005-05.