# Compressed Sensing

David L. Donoho
Department of Statistics
Stanford University

September 14, 2004

**Abstract**

Suppose $x$ is an unknown vector in $\mathbf{R}^m$ (depending on context, a digital image or signal); we plan to acquire data and then reconstruct. Nominally this 'should' require $m$ samples. But suppose we know *a priori* that $x$ is compressible by transform coding with a known transform, and we are allowed to acquire data about $x$ by measuring $n$ general linear functionals – rather than the usual pixels. If the collection of linear functionals is well-chosen, and we allow for a degree of reconstruction error, the size of $n$ can be dramatically smaller than the size $m$ usually considered necessary. Thus, certain natural classes of images with $m$ pixels need only $n = O(m^{1/4}\log^{5/2}(m))$ nonadaptive nonpixel samples for faithful recovery, as opposed to the usual $m$ pixel samples.

Our approach is abstract and general. We suppose that the object has a sparse representation in some orthonormal basis (eg. wavelet, Fourier) or tight frame (eg curvelet, Gabor), meaning that the coefficients belong to an $\ell^p$ ball for $0 < p \leq 1$. This implies that the $N$ most important coefficients in the expansion allow a reconstruction with $\ell^2$ error $O(N^{1/2-1/p})$. It is possible to design $n = O(N\log(m))$ *nonadaptive* measurements which contain the information necessary to reconstruct any such object with accuracy comparable to that which would be possible if the $N$ most important coefficients of that object were directly observable. Moreover, a good approximation to those $N$ important coefficients may be extracted from the $n$ measurements by solving a convenient linear program, called by the name Basis Pursuit in the signal processing literature. The nonadaptive measurements have the character of 'random' linear combinations of basis/frame elements.

These results are developed in a theoretical framework based on the theory of optimal recovery, the theory of $n$-widths, and information-based complexity. Our basic results concern properties of $\ell^p$ balls in high-dimensional Euclidean space in the case $0 < p \leq 1$. We estimate the Gel'fand $n$-widths of such balls, give a criterion for near-optimal subspaces for Gel'fand $n$-widths, show that 'most' subspaces are near-optimal, and show that convex optimization can be used for processing information derived from these near-optimal subspaces.

The techniques for deriving near-optimal subspaces include the use of almost-spherical sections in Banach space theory.

**Key Words and Phrases.** Integrated Sensing and Processing. Optimal Recovery. Information-Based Complexity. Gel'fand $n$-widths. Adaptive Sampling. Sparse Solution of Linear equations. Basis Pursuit. Minimum $\ell^1$ norm decomposition. Almost-Spherical Sections of Banach Spaces.

# 1 Introduction

As our modern technology-driven civilization acquires and exploits ever-increasing amounts of data, 'everyone' now knows that most of the data we acquire 'can be thrown away' with almost no perceptual loss – witness the broad success of lossy compression formats for sounds, images and specialized technical data. The phenomenon of ubiquitous compressibility raises very natural questions: *why go to so much effort to acquire* **all** *the data when* **most** *of what we get will be thrown away? Can't we just* **directly measure** *the part that won't end up being thrown away?*

In this paper we design compressed data acquisition protocols which perform *as if* it were possible to directly acquire just the important information about the signals/images – in effect not acquiring that part of the data that would eventually just be 'thrown away' by lossy compression. Moreover, the protocols are nonadaptive and parallelizable; they do not require knowledge of the signal/image to be acquired in advance - other than knowledge that the data will be compressible - and do not attempt any 'understanding' of the underlying object to guide an active or adaptive sensing strategy. The measurements made in the compressed sensing protocol are *holographic* – thus, not simple pixel samples – and must be processed *nonlinearly.*

In specific applications this principle might enable dramatically reduced measurement time, dramatically reduced sampling rates, or reduced use of Analog-to-Digital converter resources.

## 1.1 Transform Compression Background

Our treatment is abstract and general, but depends on one specific assumption which is known to hold in many settings of signal and image processing: the principle of *transform sparsity.* We suppose that the object of interest is a vector $x \in \mathbf{R}^m$, which can be a signal or image with $m$ samples or pixels, and that there is an orthonormal basis $(\psi_i : i = 1, \ldots, m)$ for $\mathbf{R}^m$ which can be for example an orthonormal wavelet basis, a Fourier basis, or a local Fourier basis, depending on the application. (As explained later, the extension to tight frames such as curvelet or Gabor frames comes for free). The object has transform coefficients $\theta_i = \langle x, \psi_i \rangle$, and these are assumed sparse in the sense that, for some $0 < p < 2$ and for some $R > 0$:

$$\|\theta\|_p \equiv (\sum_i |\theta_i|^p)^{1/p} \leq R. \tag{1.1}$$

Such constraints are actually obeyed on natural classes of signals and images; this is the primary reason for the success of standard compression tools based on transform coding [10]. To fix ideas, we mention two simple examples of $\ell^p$ constraint.

- *Bounded Variation model for images.* Here image brightness is viewed as an underlying function $f(x, y)$ on the unit square $0 \leq x, y \leq 1$ which obeys (essentially)

$$\int_0^1 \int_0^1 |\nabla f| dx dy \leq R.$$

  The digital data of interest consists of $m = n^2$ pixel samples of $f$ produced by averaging over $1/n \times 1/n$ pixels. We take a wavelet point of view; the data are seen as a superposition of contributions from various scales. Let $x^{(j)}$ denote the component of the data at scale $j$, and let $(\psi_i^j)$ denote the orthonormal basis of wavelets at scale $j$, containing $3 \cdot 4^j$ elements. The corresponding coefficients obey $\|\theta^{(j)}\|_1 \leq 4R$.

- *Bump Algebra model for spectra.* Here a spectrum (eg mass spectrum or NMR spectrum) is modelled as digital samples $(f(i/n))$ of an underlying function $f$ on the real line which is

a superposition of so-called spectral lines of varying positions, amplitudes, and linewidths. Formally,

$$f(t) = \sum_{i=1}^{\infty} a_i g((t - t_i)/s_i).$$

Here the parameters $t_i$ are line locations, $a_i$ are amplitudes/polarities and $s_i$ are linewidths, and $g$ represents a lineshape, for example the Gaussian, although other profiles could be considered. We assume the constraint where $\sum_i |a_i| \le R$, which in applications represents an energy or total mass constraint. Again we take a wavelet viewpoint, this time specifically using smooth wavelets. The data can be represented as a superposition of contributions from various scales. Let $x^{(j)}$ denote the component of the image at scale $j$, and let $(\psi_i^j)$ denote the orthonormal basis of wavelets at scale $j$, containing $2^j$ elements. The corresponding coefficients again obey $\|\theta^{(j)}\|_1 \le c \cdot R \cdot 2^{-j/2}$, [33].

While in these two examples, the $\ell^1$ constraint appeared, other $\ell^p$ constraints can appear naturally as well; see below. For some readers the use of $\ell^p$ norms with $p < 1$ may seem initially strange; it is now well-understood that the $\ell^p$ norms with such small $p$ are natural mathematical measures of sparsity [11, 13]. As $p$ decreases below 1, more and more sparsity is being required. Also, from this viewpoint, an $\ell^p$ constraint based on $p = 2$ requires no sparsity at all.

Note that in each of these examples, we also allowed for separating the object of interest into subbands, each one of which obeys an $\ell^p$ constraint. In practice below, we stick with the view that the object $\theta$ of interest is a coefficient vector obeying the constraint (1.1), which may mean, from an application viewpoint, that our methods correspond to treating various subbands separately, as in these examples.

The key implication of the $\ell^p$ constraint is sparsity of the transform coefficients. Indeed, we have trivially that, if $\theta_N$ denotes the vector $\theta$ with everything except the $N$ largest coefficients set to 0,

$$\|\theta - \theta_N\|_2 \le \zeta_{2,p} \cdot \|\theta\|_p \cdot (N+1)^{1/2-1/p}, \qquad N = 0, 1, 2, \ldots, \tag{1.2}$$

with a constant $\zeta_{2,p}$ depending only on $p \in (0, 2)$. Thus, for example, to approximate $\theta$ with error $\epsilon$, we need to keep only the $N \asymp \epsilon^{(p-2)/2p}$ biggest terms in $\theta$.

## 1.2 Optimal Recovery/Information-Based Complexity Background

Our question now becomes: if $x$ is an unknown signal whose transform coefficient vector $\theta$ obeys (1.1), can we make a reduced number $n \ll m$ of measurements which will allow faithful reconstruction of $x$. Such questions have been discussed (for other types of assumptions about $x$) under the names of *Optimal Recovery* [39] and *Information-Based Complexity* [46]; we now adopt their viewpoint, and partially adopt their notation, without making a special effort to be really orthodox. We use 'OR/IBC' as a generic label for work taking place in those fields, admittedly being less than encyclopedic about various scholarly contributions.

We have a class $X$ of possible objects of interest, and are interested in designing an information operator $I_n : X \mapsto \mathbf{R}^n$ that samples $n$ pieces of information about $x$, and an algorithm $A : \mathbf{R}^n \mapsto \mathbf{R}^m$ that offers an approximate reconstruction of $x$. Here the *information operator* takes the form

$$I_n(x) = (\langle \xi_1, x \rangle, \ldots, \langle \xi_n, x \rangle)$$

where the $\xi_i$ are sampling kernels, not necessarily sampling pixels or other simple features of the signal, however, they are nonadaptive, i.e. fixed independently of $x$. The algorithm $A_n$ is an unspecified, possibly nonlinear reconstruction operator.

We are interested in the $\ell^2$ error of reconstruction and in the behavior of optimal information and optimal algorithms. Hence we consider the minimax $\ell^2$ error as a standard of comparison:

$$E_n(X) = \inf_{A_n, I_n} \sup_{x \in X} \|x - A_n(I_n(x))\|_2,$$

So here, all possible methods of nonadaptive sampling are allowed, and all possible methods of reconstruction are allowed.

In our application, the class $X$ of objects of interest is the set of objects $x = \sum_i \theta_i \psi_i$ where $\theta = \theta(x)$ obeys (1.1) for a given $p$ and $R$. Denote then

$$X_{p,m}(R) = \{x : \|\theta(x)\|_p \le R\}.$$

Our goal is to evaluate $E_n(X_{p,m}(R))$ and to have practical schemes which come close to attaining it.

### 1.3 Four Surprises

Here is the main quantitative phenomenon of interest for this article.

**Theorem 1** *Let $(n, m_n)$ be a sequence of problem sizes with $n < m_n$, $n \to \infty$, and $m_n \sim An^\gamma$, $\gamma > 1$, $A > 0$. Then for $0 < p \le 1$ there is $C_p = C_p(A, \gamma) > 0$ so that*

$$E_n(X_{p,m}(R)) \le C_p \cdot R \cdot (n/\log(m_n))^{1/2-1/p}, \quad n \to \infty. \tag{1.3}$$

We find this surprising in four ways. First, compare (1.3) with (1.2). We see that the forms are similar, under the calibration $n = N \log(m_n)$. In words: the quality of approximation to $x$ which could be gotten by using the $N$ biggest transform coefficients can be gotten by using the $n \approx N \log(m)$ pieces of *nonadaptive* information provided by $I_n$. The surprise is that we would not know in advance which transform coefficients are likely to be the important ones in this approximation, yet the optimal information operator $I_n$ is nonadaptive, depending at most on the class $X_{p,m}(R)$ and not on the specific object. In some sense this nonadaptive information is just as powerful as knowing the $N$ best transform coefficients.

This seems even more surprising when we note that for objects $x \in X_{p,m}(R)$, the transform representation is the optimal one: no other representation can do as well at characterising $x$ by a few coefficients [11, 12]. Surely then, one imagines, the sampling kernels $\xi_i$ underlying the optimal information operator must be simply measuring individual transform coefficients? Actually, no: the information operator is measuring very complex holographic functionals which in some sense mix together all the coefficients in a big soup. Compare (6.1) below.

Another surprise is that, if we enlarged our class of information operators to allow adaptive ones, e.g. operators in which certain measurements are made in response to earlier measurements, we could scarcely do better. Define the *minimax error under adaptive information* $E_n^{Adapt}$ allowing adaptive operators

$$I_n^A = (\langle \xi_1, x \rangle, \langle \xi_{2,x}, x \rangle, \ldots \langle \xi_{n,x}, x \rangle),$$

where, for $i \ge 2$, each kernel $\xi_{i,x}$ is allowed to depend on the information $\langle \xi_{j,x}, x \rangle$ gathered at previous stages $1 \le j < i$. Formally setting

$$E_n^{Adapt}(X) = \inf_{A_n, I_n^A} \sup_{x \in X} \|x - A_n(I_n^A(x))\|_2,$$

we have

**Theorem 2** *For $0 < p \leq 1$, and $C_p > 0$*

$$E_n(X_{p,m}(R)) \leq 2^{1/p} \cdot E_n^{Adapt}(X_{p,m}(R)), \quad R > 0$$

So adaptive information is of minimal help – despite the quite natural expectation that an adaptive method ought to be able iteratively somehow 'localize' and then 'close in' on the 'big coefficients'.

An additional surprise is that, in the already-interesting case $p = 1$, Theorems 1 and 2 are easily derivable from known results in OR/IBC and approximation theory! However, the derivations are indirect; so although they have what seem to the author as fairly important implications, very little seems known at present about good nonadaptive information operators or about concrete algorithms matched to them.

Our goal in this paper is to give direct arguments which cover the case $0 < p \leq 1$ of highly compressible objects, to give direct information about near-optimal information operators and about concrete, computationally tractable algorithms for using this near-optimal information.

### 1.4 Geometry and Widths

From our viewpoint, the phenomenon described in Theorem 1 concerns the geometry of high-dimensional convex and nonconvex 'balls'. To see the connection, note that the class $X_{p,m}(R)$ is the image, under orthogonal transformation, of an $\ell^p$ ball. If $p = 1$ this is convex and symmetric about the origin, as well as being orthosymmetric with respect to the axes provided by the wavelet basis; if $p < 1$, this is again symmetric about the origin and orthosymmetric, while not being convex, but still starshaped.

To develop this geometric viewpoint further, we consider two notions of $n$-width; see [39].

**Definition 1.1** *The Gel'fand $n$-width of $X$ with respect to the $\ell_m^2$ norm is defined as*

$$d^n(X; \ell_m^2) = \inf_{V_n} \sup\{\|x\|_2 : x \in V_n^\perp \cap X\},$$

*where the infimum is over $n$-dimensional linear subspaces of $\mathbf{R}^m$, and $V_n^\perp$ denotes the orthocomplement of $V_n$ with respect to the standard Euclidean inner product.*

In words, we look for a subspace such that 'trapping' $x \in X$ in that subspace causes $x$ to be small. Our interest in Gel'fand $n$-widths derives from an equivalence between optimal recovery for nonadaptive information and such $n$-widths, well-known in the $p = 1$ case [39], and in the present setting extending as follows:

**Theorem 3** *For $0 < p \leq 1$,*

$$d^n(X_{p,m}(R)) \leq E_n(X_{p,m}(R)) \leq 2^{1/p-1} \cdot d^n(X_{p,m}(R)), \qquad R > 0. \tag{1.4}$$

Thus the Gel'fand $n$-widths either exactly or nearly equal the value of optimal information. Ultimately, the bracketing with constant $2^{1/p-1}$ will be for us just as good as equality, owing to the unspecified constant factors in (1.3). We will typically only be interested in *near-optimal* performance, i.e. in obtaining $E_n$ to within constant factors.

It is relatively rare to see the Gel'fand $n$-widths studied directly [40]; more commonly one sees results about the Kolmogorov $n$-widths:

**Definition 1.2** *Let $X \subset \mathbf{R}^m$ be a bounded set. The Kolmogorov $n$-width of $X$ with respect the $\ell_m^2$ norm is defined as*

$$d_n(X; \ell_m^2) = \inf_{V_n} \sup_{x \in X} \inf_{y \in V_n} \|x - y\|_2.$$

*where the infimum is over $n$-dimensional linear subspaces of $\mathbf{R}^m$.*

In words, $d_n$ measures the quality of approximation of $X$ possible by $n$-dimensional subspaces $V_n$.

In the case $p = 1$, there is an important duality relationship between Kolmogorov widths and Gel'fand widths which allows us to infer properties of $d^n$ from published results on $d_n$. To state it, let $d^n(X, \ell^q)$ be defined in the obvious way, based on approximation in $\ell^q$ rather than $\ell^2$ norm. Also, for given $p \geq 1$ and $q \geq 1$, let $p'$ and $q'$ be the standard dual indices $= 1 - 1/p$, $1 - 1/q$. Also, let $b_{p,m}$ denote the standard unit ball of $\ell_m^p$. Then [40]

$$d_n(b_{p,m}; \ell_m^q) = d^n(b_{q',m}, \ell_m^{p'}). \tag{1.5}$$

In particular

$$d_n(b_{2,m}; \ell_m^\infty) = d^n(b_{1,m}, \ell_m^2).$$

The asymptotic properties of the left-hand side have been determined by Garnaev and Gluskin [24]. This follows major work by Kashin [28], who developed a slightly weaker version of this result in the course of determining the Kolmogorov $n$-widths of Sobolev spaces. See the original papers, or Pinkus's book [40] for more details.

**Theorem 4 (Kashin; Garnaev and Gluskin)** *For all $n$ and $m > n$*

$$d_n(b_{2,m}, \ell_m^\infty) \asymp (n/(1 + \log(m/n)))^{-1/2}.$$

Theorem 1 now follows in the case $p = 1$ by applying KGG with the duality formula (1.5) and the equivalence formula (1.4). The case $0 < p < 1$ of Theorem 1 does not allow use of duality and the whole range $0 < p \leq 1$ is approached differently in this paper.

### 1.5 Mysteries ...

Because of the indirect manner by which the KGG result implies Theorem 1, we really don't learn much about the phenomenon of interest in this way. The arguments of Kashin, Garnaev and Gluskin show that there exist near-optimal $n$-dimensional subspaces for the Kolmogorov widths; they arise as the nullspaces of certain matrices with entries $\pm 1$ which are known to exist by counting the number of matrices lacking certain properties, the total number of matrices with $\pm 1$ entries, and comparing. The interpretability of this approach is limited.

The implicitness of the information operator is matched by the abstractness of the reconstruction algorithm. Based on OR/IBC theory we know that the so-called *central algorithm* is optimal. This "algorithm" asks us to consider, for given information $y_n = I_n(x)$ the collection of all objects $x$ which could have given rise to the data $y_n$:

$$I_n^{-1}(y_n) = \{x : I_n(x) = y_n\}.$$

Defining now the *center* of a set $S$

$$center(S) = \inf_c \sup_{x \in S} \|x - c\|_2,$$

the central algorithm is

$$\hat{x}_n^* = center(I_n^{-1}(y_n) \cap X_{p,m}(R)),$$

and it obeys, when the information $I_n$ is optimal,

$$\sup_{X_{p,m}(R)} \|x - \hat{x}_n^*\|_2 = E_n(X_{p,m}(R));$$

see Section 3 below.

This abstract viewpoint unfortunately does not translate into a practical approach (at least in the case of the $X_{p,m}(R)$, $0 < p \leq 1$). The set $I_n^{-1}(y_n) \cap X_{p,m}(R)$ is a section of the ball $X_{p,m}(R)$, and finding the center of this section does not correspond to a standard tractable computational problem. Moreover, this assumes we know $p$ and $R$ which would typically not be the case.

### 1.6 Results

Our paper develops two main types of results.

- *Near-Optimal Information.* We directly consider the problem of near-optimal subspaces for Gel'fand $n$-widths of $X_{p,m}(R)$, and introduce 3 structural conditions (CS1-CS3) on an $n$-by-$m$ matrix which imply that its nullspace is near-optimal. We show that the vast majority of $n$-subspaces of $\mathbf{R}^m$ are near-optimal, and random sampling yields near-optimal information operators with overwhelmingly high probability.

- *Near-Optimal Algorithm.* We study a simple nonlinear reconstruction algorithm: simply minimize the $\ell^1$ norm of the coefficients $\theta$ subject to satisfying the measurements. This has been studied in the signal processing literature under the name Basis Pursuit; it can be computed by linear programming. We show that this method gives near-optimal results for all $0 < p \leq 1$.

In short, we provide a large supply of near-optimal information operators and a near-optimal reconstruction procedure based on linear programming, which, perhaps unexpectedly, works even for the non-convex case $0 < p < 1$.

For a taste of the type of result we obtain, consider a specific information/algorithm combination.

- *CS Information.* Let $\Phi$ be an $n \times m$ matrix generated by randomly sampling the columns, with different columns iid random uniform on $\mathbf{S}^{n-1}$. With overwhelming probability for large $n$, $\Phi$ has properties CS1-CS3 detailed below; assume we have achieved such a favorable draw. Let $\Psi$ be the $m \times m$ basis matrix with basis vector $\psi_i$ as the $i$-th column. The *CS Information* operator $I_n^{CS}$ is the $n \times m$ matrix $\Phi\Psi^T$.

- $\ell^1$*-minimization.* To reconstruct from CS Information, we solve the convex optimization problem

$$(L_1) \qquad \min \|\Psi^T x\|_1 \text{ subject to } y_n = I_n^{CS}(x).$$

  In words, we look for the object $\hat{x}_1$ having coefficients with smallest $\ell^1$ norm that is consistent with the information $y_n$.

To evaluate the quality of an information operator $I_n$, set

$$E_n(I_n, X) \equiv \inf_{A_n} \sup_{x \in X} \|x - A_n(I_n(x))\|_2.$$

To evaluate the quality of a combined algorithm/information pair $(A_n, I_n)$, set

$$E_n(A_n, I_n, X) \equiv \sup_{x \in X} \|x - A_n(I_n(x))\|_2.$$

**Theorem 5** *Let $n, m_n$ be a sequence of problem sizes obeying $n < m_n \sim An^\gamma$, $A > 0$, $\gamma \geq 1$; and let $I_n^{CS}$ be a corresponding sequence of operators deriving from CS-matrices with underlying parameters $\eta_i$ and $\rho$ (see Section 2 below). Let $0 < p \leq 1$. There exists $C = C(p, (\eta_i), \rho, A, \gamma) > 0$ so that $I_n^{CS}$ is near-optimal*

$$E_n(I_n^{CS}, X_{p,m}(R)) \leq C \cdot E_n(X_{p,m}(R)), \quad R > 0, \quad n > n_0.$$

*Moreover the algorithm $A_{1,n}$ delivering the solution to $(L_1)$ is near-optimal:*

$$E_n(A_{1,n}, I_n^{CS}, X_{p,m}(R)) \leq C \cdot E_n(X_{p,m}(R)), \quad R > 0 \quad n > n_0.$$

Thus for large $n$ we have a simple description of near-optimal information and a tractable near-optimal reconstruction algorithm.

### 1.7 Potential Applications

To see the potential implications, recall first the Bump Algebra model for spectra. In this context, our result says that, for a spectrometer based on the information operator $I_n$ in Theorem 5, it is really only necessary to take $n = O(\sqrt{m}\log(m))$ measurements to get an accurate reconstruction of such spectra, rather than the nominal $m$ measurements. However, they must then be processed nonlinearly.

Consider the Bounded Variation model for Images. In that context, a result paralleling Theorem 5 says that for a specialized imaging device based on a near-optimal information operator it is really only necessary to take $n = O(\sqrt{m}\log(m))$ measurements to get an accurate reconstruction of images with $m$ pixels, rather than the nominal $m$ measurements.

The calculations underlying these results will be given below, along with a result showing that for cartoon-like images (which may model certain kinds of simple natural imagery, like brains) the number of measurements for an $m$-pixel image is only $O(m^{1/4}\log(m))$.

### 1.8 Contents

Section 2 introduces a set of conditions CS1-CS3 for near-optimality of an information operator. Section 3 considers abstract near-optimal algorithms, and proves Theorems 1-3. Section 4 shows that solving the convex optimization problem $(L_1)$ gives a near-optimal algorithm whenever $0 < p \leq 1$. Section 5 points out immediate extensions to weak-$\ell^p$ conditions and to tight frames. Section 6 sketches potential implications in image, signal, and array processing. Section 7 shows that conditions CS1-CS3 are satisfied for "most" information operators.

Finally, in Section 8 below we note the ongoing work by two groups (Gilbert et al. [25]) and (Candès et al [4, 5]), which although not written in the $n$-widths/OR/IBC tradition, imply (as we explain), closely related results.

## 2 Information

Consider information operators constructed as follows. With $\Psi$ the orthogonal matrix whose columns are the basis elements $\psi_i$, and with certain $n$-by-$m$ matrices $\Phi$ obeying conditions specified below, we construct corresponding information operators $I_n = \Phi\Psi^T$. Everything will be completely transparent to the orthogonal matrix $\Psi$ and hence we will assume that $\Psi$ is the identity throughout this section.

In view of the relation between Gel'fand $n$-widths and minimax errors, we may work with $n$-widths. We define the *width* of a set $X$ relative to an operator $\Phi$:

$$w(\Phi, X) \equiv \sup \|x\|_2 \text{ subject to } x \in X \cap nullspace(\Phi) \tag{2.1}$$

In words, this is the radius of the section of $X$ cut out by $nullspace(\Phi)$. In general, the Gel'fand $n$-width is the smallest value of $w$ obtainable by choice of $\Phi$:

$$d^n(X) = \inf\{w(\Phi, X) : \Phi \text{ is an } n \times m \text{ matrix}\}$$

We will show for all large $n$ and $m$ the existence of $n$ by $m$ matrices $\Phi$ where

$$w(\Phi, b_{p,m}) \le C \cdot d^n(b_{p,m}),$$

with $C$ dependent at most on $p$ and the ratio $\log(m)/\log(n)$.

### 2.1 Conditions CS1-CS3

In the following, with $J \subset \{1, \ldots, m\}$ let $\Phi_J$ denote a submatrix of $\Phi$ obtained by selecting just the indicated columns of $\Phi$. We let $V_J$ denote the range of $\Phi_J$ in $\mathbf{R}^n$. Finally, we consider a family of *quotient norms* on $\mathbf{R}^n$; with $\ell^1(J^c)$ denoting the $\ell^1$ norm on vectors indexed by $\{1, \ldots, m\} \backslash J$,

$$Q_{J^c}(v) = \min \|\theta\|_{\ell^1(J^c)} \text{ subject to } \Phi_{J^c}\theta = v.$$

These describe the minimal $\ell^1$-norm representation of $v$ achievable using only specified subsets of columns of $\Phi$.

We describe a set of three conditions to impose on an $n \times m$ matrix $\Phi$, indexed by strictly positive parameters $\eta_i$, $1 \le i \le 3$, and $\rho$.

CS1 The minimal singular value of $\Phi_J$ exceeds $\eta_1 > 0$ uniformly in $|J| < \rho n/\log(m)$.

CS2 On each subspace $V_J$ we have the inequality

$$\|v\|_1 \ge \eta_2 \cdot \sqrt{n} \cdot \|v\|_2, \quad \forall v \in V_J,$$

uniformly in $|J| < \rho n/\log(m)$.

CS3 On each subspace $V_J$

$$Q_{J^c}(v) \ge \eta_3/\sqrt{\log(m/n)} \cdot \|v\|_1, \quad v \in V_J,$$

uniformly in $|J| < \rho n/\log(m)$.

CS1 demands a certain quantitative degree of linear independence among all small groups of columns. CS2 says that linear combinations of small groups of columns give vectors that look much like random noise, at least as far as the comparison of $\ell^1$ and $\ell^2$ norms is concerned. It will be implied by a geometric fact: every $V_J$ slices through the $\ell^1_m$ ball in such a way that the resulting convex section is actually close to spherical. CS3 says that for every vector in some $V_J$, the associated quotient norm $Q_{J^c}$ is never dramatically better than the simple $\ell^1$ norm on $\mathbf{R}^n$.

It turns out that matrices satisfying these conditions are ubiquitous for large $n$ and $m$ when we choose the $\eta_i$ and $\rho$ properly. Of course, for any finite $n$ and $m$, all norms are equivalent and almost any arbitrary matrix can trivially satisfy these conditions simply by taking $\eta_1$ very small and $\eta_2$, $\eta_3$ very large. However, the definition of 'very small' and 'very large' would have to depend on $n$ for this trivial argument to work. We claim something deeper is true: it is possible to choose $\eta_i$ and $\rho$ independent of $n$ and of $m \le An^\gamma$.

Consider the set

$$\mathbf{S}^{n-1} \times \cdots \times \mathbf{S}^{n-1} \qquad \overset{\leftarrow m \text{ terms} \rightarrow}{}$$

of *all* $n \times m$ matrices having unit-normalized columns. On this set, measure frequency of occurrence with the natural uniform measure (the product measure, uniform on each factor $S^{n-1}$).

**Theorem 6** *Let $(n, m_n)$ be a sequence of problem sizes with $n \to \infty$, $n < m_n$, and $m \sim An^\gamma$, $A > 0$ and $\gamma \ge 1$. There exist $\eta_i > 0$ and $\rho > 0$ depending only on $A$ and $\gamma$ so that, for each $\delta > 0$ the proportion of all $n \times m$ matrices $\Phi$ satisfying CS1-CS3 with parameters $(\eta_i)$ and $\rho$ eventually exceeds $1 - \delta$.*

We will discuss and prove this in Section 7 below.

For later use, we will leave the constants $\eta_i$ and $\rho$ implicit and speak simply of CS matrices, meaning matrices that satisfy the given conditions with values of parameters of the type described by this theorem, namely, with $\eta_i$ and $\rho$ not depending on $n$ and permitting the above ubiquity.

## 2.2 Near-Optimality of CS-Matrices

We now show that the CS conditions imply near-optimality of widths induced by CS matrices.

**Theorem 7** *Let $(n, m_n)$ be a sequence of problem sizes with $n \to \infty$ and $m_n \sim A \cdot n^\gamma$. Consider a sequence of $n$ by $m_n$ matrices $\Phi_{n,m_n}$ obeying the conditions CS1-CS3 with $\eta_i$ and $\rho$ positive and independent of $n$. Then for each $p \in (0, 1]$, there is $C = C(p, \eta_1, \eta_2, \eta_3, \rho, A, \gamma)$ so that*

$$w(\Phi_{n,m_n}, b_{p,m_n}) \le C \cdot (n/\log(m/n))^{1/2-1/p}, \qquad n > n_0.$$

**Proof.** Consider the optimization problem

$$(Q_p) \qquad \sup \|\theta\|_2 \text{ subject to } \Phi\theta = 0, \quad \|\theta\|_p \le 1.$$

Our goal is to bound the value of $(Q_p)$:

$$val(Q_p) \le C \cdot (n/\log(m/n))^{1/2-1/p}.$$

Choose $\theta$ so that $0 = \Phi\theta$. Let $J$ denote the indices of the $k = \lfloor \rho n / \log(m) \rfloor$ largest values in $\theta$. Without loss of generality suppose coordinates are ordered so that $J$ comes first among the $m$ entries, and partition $\theta = [\theta_J, \theta_{J^c}]$. Clearly

$$\|\theta_{J^c}\|_p \le \|\theta\|_p \le 1, \tag{2.2}$$

while, because each entry in $\theta_J$ is at least as big as any entry in $\theta_{J^c}$, (1.2) gives

$$\|\theta_{J^c}\|_2 \le \zeta_{2,p} \cdot (k+1)^{1/2-1/p} \tag{2.3}$$

A similar argument for $\ell^1$ approximation gives, in case $p < 1$,

$$\|\theta_{J^c}\|_1 \le \zeta_{1,p} \cdot (k+1)^{1-1/p}. \tag{2.4}$$

Now $0 = \Phi_J\theta_J + \Phi_{J^c}\theta_1$. Hence, with $v = \Phi_J\theta_J$, we have $-v = \Phi_{J^c}\theta_{J^c}$. As $v \in V_J$ and $|J| = k < \rho n/\log(m)$, we can invoke CS3, getting

$$\|\theta_{J^c}\|_1 \ge Q_{J^c}(-v) \ge \eta_3/\sqrt{\log(m/n)} \cdot \|v\|_1.$$

On the other hand, again using $v \in V_J$ and $|J| = k < \rho n/\log(m)$ invoke CS2, getting

$$\|v\|_1 \ge \eta_2 \cdot \sqrt{n} \cdot \|v\|_2.$$

Combining these with the above,

$$\|v\|_2 \le (\eta_2\eta_3)^{-1} \cdot (\sqrt{\log(m/n)}/\sqrt{n}) \cdot \|\theta_{J^c}\|_1 \le c_1 \cdot (n/\log(m))^{1/2-1/p},$$

with $c_1 = \zeta_{1,p}\rho^{1-1/p}/\eta_2\eta_3$. Recalling $|J| = k < \rho n/\log(m)$, and invoking CS1 we have

$$\|\theta_J\|_2 \le \|\Phi_J\theta_J\|_2/\eta_1 = \|v\|_2/\eta_1.$$

In short, with $c_2 = c_1/\eta_1$,

$$\begin{aligned} \|\theta\|_2 &\le \|\theta_J\|_2 + \|\theta_{J^c}\|_2 \\ &\le c_2 \cdot (n/\log(m))^{1/2-1/p} + \zeta_{2,p} \cdot (\rho n/\log(m))^{1/2-1/p}. \end{aligned}$$

The Theorem follows with $C = (\zeta_{1,p}\eta_2\eta_3/\eta_1 + \zeta_{2,p}\rho^{1/2-1/p})$. QED

# 3 Algorithms

Given an information operator $I_n$, we must design a reconstruction algorithm $A_n$ which delivers reconstructions compatible in quality with the estimates for the Gel'fand $n$-widths. As discussed in the introduction, the optimal method in the OR/IBC framework is the so-called central algorithm, which unfortunately, is typically not efficiently computable in our setting. We now describe an alternate abstract approach, allowing us to prove Theorem 1.

### 3.1 Feasible-Point Methods

Another general abstract algorithm from the OR/IBC literature is the so-called *feasible-point method*, which aims simply to find *any* reconstruction compatible with the observed information and constraints.

As in the case of the central algorithm, we consider, for given information $y_n = I_n(x)$ the collection of all objects $x \in X_{p,m}(R)$ which could have given rise to the information $y_n$:

$$\hat{X}_{p,R}(y_n) = \{x : y_n = I_n(x), \quad x \in X_{p,m}(R)\}$$

In the feasible-point method, we simply select any member of $\hat{X}_{p,R}(y_n)$, by whatever means. A popular choice is to take an element of *least norm*, i.e. a solution of the problem

$$(P_p) \qquad \min_x \|\theta(x)\|_p \text{ subject to } y_n = I_n(x),$$

where here $\theta(x) = \Psi^T x$ is the vector of transform coefficients, $\theta \in \ell_m^p$. A nice feature of this approach is that it is not necessary to know the radius $R$ of the ball $X_{p,m}(R)$; the element of least norm will always lie inside it.

Calling the solution $\hat{x}_{p,n}$, one can show, adapting standard OR/IBC arguments in [36, 46, 40]

**Lemma 3.1**

$$\sup_{X_{p,m}(R)} \|x - \hat{x}_{p,n}\|_2 \le 2 \cdot E_n(X_{p,m}(R)), \qquad 0 < p \le 1. \tag{3.1}$$

In short, the least-norm method is within a factor two of optimal.

**Proof.** We first justify our claims for optimality of the central algorithm, and then show that the minimum norm algorithm is near to the central algorithm. Let again $\hat{x}_n^*$ denote the result of the central algorithm. Now

$$\begin{aligned} radius(\hat{X}_{p,R}(y_n)) &\equiv \inf_c \sup\{\|x - c\|_2 : x \in \hat{X}_{p,R}(y_n)\} \\ &= \sup\{\|x - \hat{x}_n^*\|_2 : x \in \hat{X}_{p,R}(y_n)\} \end{aligned}$$

Now clearly, in the special case when $x$ is only known to lie in $X_{p,m}(R)$ and $y_n$ is measured, the minimax error is exactly $radius(\hat{X}_{p,R}(y_n))$. Since this error is achieved by the central algorithm for each $y_n$, the minimax error over all $x$ is achieved by the central algorithm. This minimax error is

$$\sup\{radius(\hat{X}_{p,R}(y_n)) : y_n \in I_n(X_{p,m}(R))\} = E_n(X_{p,m}(R)).$$

Now the least-norm algorithm obeys $\hat{x}_{p,n}(y_n) \in X_{p,m}(R)$; hence

$$\|\hat{x}_{p,n}(y_n) - \hat{x}_n^*(y_n)\|_2 \le radius(\hat{X}_{p,R}(y_n)).$$

But the triangle inequality gives

$$\|x - \hat{x}_{p,n}\|_2 \le \|x - \hat{x}_n^*(y_n)\|_2 + \|\hat{x}_n^*(y_n) - \hat{x}_{p,n}\|_2;$$

hence, if $x \in \hat{X}_{p,R}(y_n)$,

$$\begin{aligned} \|x - \hat{x}_{p,n}\|_2 &\le 2 \cdot radius(\hat{X}_{p,R}(y_n)) \\ &\le 2 \cdot \sup\{radius(\hat{X}_{p,R}(y_n)) : y_n \in I_n(X_{p,m}(R))\} \\ &= 2 \cdot E_n(X_{p,m}(R)). \end{aligned}$$

QED.

More generally, if the information operator $I_n$ is only near-optimal, then the same argument gives

$$\sup_{X_{p,m}(R)} \|x - \hat{x}_{p,n}\|_2 \le 2 \cdot E_n(I_n, X_{p,m}(R)), \qquad 0 < p \le 1. \tag{3.2}$$

### 3.2 Proof of Theorem 3

Before proceeding, it is convenient to prove Theorem 3. Note that the case $p \geq 1$ is well-known in OR/IBC so we only need to give an argument for $p < 1$ (though it happens that our argument works for $p = 1$ as well). The key point will be to apply the $p$-triangle inequality

$$\|\theta + \theta'\|_p^p \leq \|\theta\|_p^p + \|\theta'\|_p^p,$$

valid for $0 < p < 1$; this inequality is well-known in interpolation theory [1] through Peetre and Sparr's work, and is easy to verify directly.

Suppose without loss of generality that there is an optimal subspace $V_n$, which is fixed and given in this proof. As we just saw,

$$E_n(X_{p,m}(R)) = \sup\{radius(\hat{X}_{p,R}(y_n)) : y_n \in I_n(X_{p,m}(R))\}.$$

Now

$$d^n(X_{p,m}(R)) = radius(\hat{X}_{p,R}(0))$$

so clearly $E_n \geq d^n$. Now suppose without loss of generality that $x_1$ and $x_{-1}$ attain the radius bound, i.e. they satisfy $I_n(x_{\pm 1}) = y_n$ and, for $c = center(\hat{X}_{p,R}(y_n))$ they satisfy

$$E_n(X_{p,m}(R)) = \|x_1 - c\|_2 = \|x_{-1} - c\|_2.$$

Then define $\delta = (x_1 - x_{-1})/2^{1/p}$. Set $\theta_{\pm 1} = \Psi^T x_{\pm 1}$ and $\xi = \Psi^T \delta$. By the $p$-triangle inequality

$$\|\theta_1 - \theta_{-1}\|_p^p \leq \|\theta_1\|_p^p + \|\theta_{-1}\|_p^p,$$

and so

$$\|\xi\|_p = \|(\theta_1 - \theta_{-1})/2^{1/p}\|_p \leq R.$$

Hence $\delta \in X_{p,m}(R)$. However, $I_n(\delta) = I_n((x_{+1} - x_{-1})/2^{1/p}) = 0$, so $\delta$ belongs to $X_{p,m}(R) \cap V_n$. Hence $\|\delta\|_2 \leq d^n(X_{p,m}(R))$ and

$$E_n(X_{p,m}(R)) = \|x_1 - x_{-1}\|_2/2 = 2^{1/p-1}\|\delta\|_2 \leq 2^{1/p-1} \cdot d^n(X_{p,m}(R)).$$

QED

### 3.3 Proof of Theorem 1

We are now in a position to prove Theorem 1 of the introduction.

First, in the case $p = 1$, we have already explained in the introduction that the theorem of Garnaev and Gluskin implies the result by duality. In the case $0 < p < 1$, we need only to show a lower bound and an upper bound of the same order.

For the lower bound, we consider the entropy numbers, defined as follows. Let $X$ be a set and let $e_n(X, \ell^2)$ be the smallest number $\epsilon$ such that an $\epsilon$-net for $X$ can be built using a net of cardinality at most $2^n$. From Carl's Theorem - see the exposition in Pisier's book - there is a constant $c > 0$ so that the Gel'fand $n$-widths dominate the entropy numbers.

$$d^n(b_{p,m}) \geq ce_n(b_{p,m}).$$

Secondly, the entropy numbers obey [43, 30]

$$e_n(b_{p,m}) \asymp (n/\log(m/n))^{1/2-1/p}.$$

At the same time the combination of Theorem 7 and Theorem 6 shows that

$$d^n(b_{p,m}) \le c(n/\log(m))^{1/2-1/p}$$

Applying now the Feasible-Point method, we have

$$E_n(X_{m,p}(1)) \le 2d^n(b_{p,m}),$$

with immediate extensions to $E_n(X_{m,p}(R))$ for all $R \neq 1 > 0$. We conclude that

$$E_n(b_{p,m}) \asymp (n/\log(m/n))^{1/2-1/p}.$$

as was to be proven.

### 3.4 Proof of Theorem 2

Now is an opportune time to prove Theorem 2. We note that in the case of $1 \le p$, this is known [38]. The argument is the same for $0 < p < 1$, and we simply repeat it. Suppose that $x = 0$, and consider the adaptively-constructed subspace according to whatever algorithm is in force. When the algorithm terminates we have an $n$-dimensional information vector 0 and a subspace $V_n^0$ consisting of objects $x$ which would all give that information vector. For all objects in $V_n^0$ the adaptive information therefore turns out the same. Now the minimax error associated with that information is exactly $radius(V_n^0 \cap X_{p,m}(R))$; but this cannot be smaller than

$$\inf_{V_n} radius(V_n \cap X_{p,m}(R)) = d^n(X_{p,m}(R)).$$

The result follows by comparing $E_n(X_{p,m}(R))$ with $d^n$.

## 4 Basis Pursuit

The least-norm method of the previous section has two drawbacks. First it requires that one know $p$; we prefer an algorithm which works for $0 < p \le 1$. Second, if $p < 1$ the least-norm problem invokes a nonconvex optimization procedure, and would be considered intractable. In this section we correct both drawbacks.

### 4.1 The Case $p = 1$

In the case $p = 1$, $(P_1)$ is a convex optimization problem. Write it out in an equivalent form, with $\theta$ being the optimization variable:

$$(P_1) \quad \min_{\theta} \|\theta\|_1 \text{ subject to } \Phi\theta = y_n.$$

This can be formulated as a linear programming problem: let $A$ be the $n$ by $2m$ matrix $[\Phi \ -\Phi]$. The linear program

$$(LP) \quad \min_{z} 1^T z \text{ subject to } Az = y_n, x \ge 0.$$

has a solution $z^*$, say, a vector in $R^{2m}$ which can be partitioned as $z^* = [u^* v^*]$; then $\theta^* = u^* - v^*$ solves $(P_1)$. The reconstruction $\hat{x}_{1,n} = \Psi\theta^*$. This linear program is typically considered computationally tractable. In fact, this problem has been studied in the signal analysis literature under the name Basis Pursuit [7]; in that work, very large-scale underdetermined problems -

e.g. with $n = 8192$ and $m = 262144$ - were solved successfully using interior-point optimization methods.

As far as performance goes, we already know that this procedure is near-optimal in case $p = 1$; from from (3.2):

**Corollary 4.1** *Suppose that $I_n$ is an information operator achieving, for some $C > 0$,*

$$E_n(I_n, X_{1,m}(1)) \le C \cdot E_n(X_{1,m}(1));$$

*then the Basis Pursuit algorithm $A_{1,n}(y_n) = \hat{x}_{1,n}$ achieves*

$$E_n(I_n, A_{1,n}, X_{1,m}(R)) \le 2C \cdot E_n(X_{1,m}(R)), \qquad R > 0.$$

In particular, we have a universal algorithm for dealing with *any* class $X_{1,m}(R)$ – i.e. any $\Psi$, any $m$, any $R$. First, apply a near-optimal information operator; second, reconstruct by Basis Pursuit. The result obeys

$$\|x - \hat{x}_{1,n}\|_2 \le C \cdot \|\theta\|_1 \cdot (n/\log m)^{-1/2},$$

for $C$ a constant depending at most on $\log(m)/\log(n)$. The inequality can be interpreted as follows. Fix $\epsilon > 0$. Suppose the unknown object $x$ is known to be highly compressible, say obeying the *a priori* bound $\|\theta\|_1 \le cm^{\alpha}$, $\alpha < 1/2$. For any such object, rather than making $m$ measurements, we only need to make $n \sim C_\epsilon \cdot m^{2\alpha} \log(m)$ measurements, and our reconstruction obeys:

$$\|x - \hat{x}_{1,\theta}\|_2 \ll \epsilon \cdot \|x\|_2.$$

While the case $p = 1$ is already significant and interesting, the case $0 < p < 1$ is of interest because it corresponds to data which are *more highly compressible* and for which, our interest in achieving the performance indicated by Theorem 1 is even greater. Later in this section we extend the same interpretation of $\hat{x}_{1,n}$ to performance over $X_{p,m}(R)$ throughout $p < 1$.

## 4.2 Relation between $\ell^1$ and $\ell^0$ minimization

The general OR/IBC theory would suggest that to handle cases where $0 < p < 1$, we would need to solve the nonconvex optimization problem $(P_p)$, which seems intractable. However, in the current situation at least, a small miracle happens: solving $(P_1)$ is again near-optimal. To understand this, we first take a small detour, examining the relation between $\ell^1$ and the extreme case $p \to 0$ of the $\ell^p$ spaces. Let's define

$$(P_0) \qquad \min \|\theta\|_0 \text{ subject to } \Phi\theta = x,$$

where of course $\|\theta\|_0$ is just the number of nonzeros in $\theta$. Again, since the work of Peetre and Sparr the importance of $\ell^0$ and the relation with $\ell^p$ for $0 < p < 1$ is well-understood; see [1] for more detail.

Ordinarily, solving such a problem involving the $\ell^0$ norm requires combinatorial optimization; one enumerates all sparse subsets of $\{1, \ldots, m\}$ searching for one which allows a solution $\Phi\theta = x$. However, when $(P_0)$ has a sparse solution, $(P_1)$ will find it.

**Theorem 8** *Suppose that $\Phi$ satisfies CS1-CS3 with given positive constants $\rho$, $(\eta_i)$. There is a constant $\rho_0 > 0$ depending only on $\rho$ and $(\eta_i)$ and not on $n$ or $m$ so that, if $\theta$ has at most $\rho_0 n/\log(m)$ nonzeros, then $(P_0)$ and $(P_1)$ both have the same unique solution.*

In words, although the system of equations is massively undetermined, $\ell^1$ minimization and sparse solution coincide - when the result is sufficiently sparse.

There is by now an extensive literature exhibiting results on equivalence of $\ell^1$ and $\ell^0$ minimization [14, 20, 47, 48, 21]. In the first literature on this subject, equivalence was found under conditions involving sparsity constraints allowing $O(n^{1/2})$ nonzeros. While it may seem surprising that any results of this kind are possible, the sparsity constraint $\|\theta\|_0 = O(n^{1/2})$ is, ultimately, disappointingly small. A major breakthrough was the contribution of Candès, Romberg, and Tao (2004) which studied the matrices built by taking $n$ rows at random from an $m$ by $m$ Fourier matrix and gave an order $O(n/\log(n))$ bound, showing that dramatically weaker sparsity conditions were needed than the $O(n^{1/2})$ results known previously. In [17], it was shown that for 'nearly all' $n$ by $m$ matrices with $n < m < An$, equivalence held for $\leq \rho n$ nonzeros, $\rho = \rho(A)$. The above result says effectively that for 'nearly all' $n$ by $m$ matrices with $m \leq An^\gamma$, equivalence held up to $O(\rho n/\log(n))$ nonzeros, where $\rho = \rho(A, \gamma)$.

Our argument, in parallel with [17], shows that the nullspace $\Phi\beta = 0$ has a very special structure for $\Phi$ obeying the conditions in question. When $\theta$ is sparse, the *only* element in a given affine subspace $\theta + nullspace(\Phi)$ which can have small $\ell^1$ norm is $\theta$ itself.

To prove Theorem 8, we first need a lemma about the non-sparsity of elements in the nullspace of $\Phi$. Let $J \subset \{1, \ldots, m\}$ and, for a given vector $\beta \in R^m$, let $1_J\beta$ denote the mutilated vector with entries $\beta_i 1_J(i)$. Define the concentration

$$\nu(\Phi, J) = \sup\{\frac{\|1_J\beta\|_1}{\|\beta\|_1} : \quad \Phi\beta = 0\}$$

This measures the fraction of $\ell^1$ norm which can be concentrated on a certain subset for a vector in the nullspace of $\Phi$. This concentration cannot be large if $|J|$ is small.

**Lemma 4.1** *Suppose that $\Phi$ satisfies CS1-CS3, with constants $\eta_i$ and $\rho$. There is a constant $\eta_0$ depending on the $\eta_i$ so that if $J \subset \{1, \ldots, m\}$ satisfies*

$$|J| \leq \rho_1 n/\log(m), \qquad \rho_1 \leq \rho,$$

*then*

$$\nu(\Phi, J) \leq \eta_0 \cdot \rho_1^{1/2}.$$

**Proof.** This is a variation on the argument for Theorem 7. Let $\beta \in nullspace(\Phi)$. Assume without loss of generality that $J$ is the most concentrated subset of cardinality $|J| < \rho n/\log(m)$, and that the columns of $\Phi$ are numbered so that $J = \{1, \ldots, |J|\}$; partition $\beta = [\beta_J, \beta_{J^c}]$. We again consider $v = \Phi_J\beta_J$, and have $-v = \Phi_{J^c}\beta_{J^c}$. We again invoke CS2-CS3, getting

$$\|v\|_2 \leq \eta_2\eta_3 \cdot (\sqrt{n}/\sqrt{\log(m/n)}) \cdot \|\theta_{J^c}\|_1 \leq c_1 \cdot (n/\log(m))^{1/2-1/p},$$

We invoke CS1, getting

$$\|\beta_J\|_2 \leq \|\Phi_J\beta_J\|_2/\eta_1.$$

Now of course $\|\beta_J\|_1 \leq \sqrt{|J|} \cdot \|\beta_J\|_2$. Combining all these

$$\|\beta_J\|_1 \leq \sqrt{|J|} \cdot \eta_1^{-1}\eta_2\eta_3 \cdot \sqrt{\frac{\log(m)}{n}} \cdot \|\beta\|_1.$$

The lemma follows, setting $\eta_0 = \eta_2\eta_3/\eta_1$. QED

**Proof of Theorem 8.** Suppose that $x = \Phi\theta$ and $\theta$ is supported on a subset $J \subset \{1, \ldots, m\}$.

We first show that if $\nu(\Phi, J) < 1/2$, $\theta$ is the only minimizer of $(P_1)$. Suppose that $\theta'$ is a solution to $(P_1)$, obeying

$$\|\theta'\|_1 \le \|\theta\|_1.$$

Then $\theta' = \theta + \beta$ where $\beta \in nullspace(\Phi)$. We have

$$0 \le \|\theta\|_1 - \|\theta'\|_1 \le \|\beta_J\|_1 - \|\beta_{J^c}\|_1.$$

Invoking the definition of $\nu(\Phi, J)$ twice,

$$\|\beta_J\|_1 - \|\beta_{J^c}\|_1 \le (\nu(\Phi, J) - (1 - \nu(\Phi, J))\|\beta\|_1.$$

Suppose now that $\nu < 1/2$. Then $2\nu(\Phi, J) - 1 < 0$ and we have

$$\|\beta\|_1 \le 0,$$

i.e. $\theta = \theta'$.

Now recall the constant $\eta_0 > 0$ of Lemma (4.1). Define $\rho_0$ so that $\eta_0\sqrt{\rho_0} < 1/2$ and $\rho_0 \le \rho$. Lemma 4.1 shows that $|J| \le \rho_0 n/\log(m)$ implies $\nu(\Phi, J) \le \eta_0\rho_0^{1/2} < 1/2$. QED.

### 4.3 Near-Optimality of BP for $0 < p < 1$

We now return to the claimed near-optimality of BP throughout the range $0 < p < 1$.

**Theorem 9** *Suppose that $\Phi$ satisifies CS1-CS3 with constants $\eta_i$ and $\rho$. There is $C = C(p, (\eta_i), \rho, A, \gamma)$ so that the solution $\hat{\theta}_{1,n}$ to $(P_1)$ obeys*

$$\|\theta - \hat{\theta}_{1,n}\|_2 \le C_p \cdot \|\theta\|_p \cdot (n/\log m)^{1/2-1/p},$$

The proof requires an $\ell^1$ stability lemma, showing the stability of $\ell^1$ minimization under small perturbations as measured in $\ell^1$ norm. For $\ell^2$ and $\ell^\infty$ stability lemmas, see [16, 48, 18]; however, note that those lemmas do not suffice for our needs in this proof.

**Lemma 4.2** *Let $\theta$ be a vector in $R^m$ and $1_J\theta$ be the corresponding mutilated vector with entries $\theta_i 1_J(i)$. Suppose that*

$$\|\theta - 1_J\theta\|_1 \le \epsilon,$$

*where $\nu(\Phi, J) \le \nu_0 < 1/2$. Consider the instance of $(P_1)$ defined by $x = \Phi\theta$; the solution $\hat{\theta}_1$ of this instance of $(P_1)$ obeys*

$$\|\theta - \hat{\theta}_1\|_1 \le \frac{2\epsilon}{1 - 2\nu_0}. \tag{4.1}$$

**Proof of Lemma.** Put for short $\hat{\theta} \equiv \hat{\theta}_1$, and set $\beta = \theta - \hat{\theta} \in nullspace(\Phi)$. By definition of $\nu$,

$$\|\theta - \hat{\theta}\|_1 = \|\beta\|_1 \le \|\beta_{J^c}\|_1/(1 - \nu_0),$$

while

$$\|\beta_{J^c}\|_1 \le \|\theta_{J^c}\|_1 + \|\hat{\theta}_{J^c}\|_1.$$

As $\hat{\theta}$ solves $(P_1)$,

$$\|\hat{\theta}_J\|_1 + \|\hat{\theta}_{J^c}\|_1 \le \|\theta\|_1,$$

and of course

$$\|\theta_J\|_1 - \|\hat{\theta}_J\|_1 \le \|1_J(\theta - \hat{\theta})\|_1.$$

Hence

$$\|\hat{\theta}_{J^c}\|_1 \le \|1_J(\theta - \hat{\theta})\|_1 + \|\theta_{J^c}\|_1.$$

Finally,

$$\|1_J(\theta - \hat{\theta})\|_1 = \|\beta_J\|_1 \le \nu_0 \cdot \|\beta\|_1 = \nu_0 \cdot \|\theta - \hat{\theta}\|_1.$$

Combining the above, setting $\delta \equiv \|\theta - \hat{\theta}\|_1$ and $\epsilon \equiv \|\theta_{J^c}\|_1$, we get

$$\delta \le (\nu_0 \delta + 2\epsilon)/(1 - \nu_0),$$

and (4.1) follows. QED

**Proof of Theorem 9** We use the same general framework as Theorem 7. Let $x = \Phi\theta$ where $\|\theta\|_p \le R$. Let $\hat{\theta}$ be the solution to $(P_1)$, and set $\beta = \hat{\theta} - \theta \in nullspace(\Phi)$.

Let $\eta_0$ as in Lemma 4.1 and set $\rho_0 = \min(\rho, (4\eta_0)^{-2})$. Let $J$ index the $\rho_0 n/\log(m)$ largest-amplitude entries in $\theta$. From $\|\theta\|_p \le R$ and (2.4) we have

$$\|\theta - 1_J\theta\|_1 \le \xi_{1,p} \cdot R \cdot |J|^{1-1/p},$$

and Lemma 4.1 provides

$$\nu(\Phi, J) \le \eta_0 \rho_0^{1/2} \le 1/4.$$

Applying Lemma 4.2,

$$\|\beta\|_1 \le c \cdot R \cdot |J|^{1-1/p}. \tag{4.2}$$

The vector $\delta = \beta/\|\beta\|_1$ lies in $nullspace(\Phi)$ and has $\|\delta\|_1 = 1$. Hence

$$\|\delta\|_2 \le c \cdot (n/\log(m))^{-1/2}.$$

We conclude by homogeneity that

$$\|\beta\|_2 \le c \cdot \|\beta\|_1 \cdot (n/\log(m))^{-1/2}.$$

Combining this with (4.2),

$$\|\beta\|_2 \le c \cdot R \cdot (n/\log(m))^{1/2-1/p}.$$

QED

# 5 Immediate Extensions

Before continuing, we mention two immediate extensions to these results of interest below and elsewhere.

## 5.1 Tight Frames

Our main results so far have been stated in the context of $(\phi_i)$ making an orthonormal basis. In fact the results hold for *tight frames*. These are collections of vectors which, when joined together as columns in an $m \times m'$ matrix $\Psi$ $(m < m')$ have $\Psi\Psi^T = I_m$. It follows that, if $\theta(x) = \Psi^T x$, then we have the Parseval relation

$$\|x\|_2 = \|\theta(x)\|_2,$$

and the reconstruction formula $x = \Psi\theta(x)$. In fact, Theorems 7 and 9 only need the Parseval relation in the proof. Hence the same results hold without change when the relation between $x$

and $\theta$ involves a tight frame. In particular, if $\Phi$ is an $n \times m$ matrix satisfying CS1-CS3, then $I_n = \Phi\Psi^T$ defines a near-optimal information operator on $R^{m'}$, and solution of the optimization problem

$$(L_1) \quad \min_x \|\psi^T x\|_1 \text{ subject to } I_n(x) = y_n,$$

defines a near-optimal reconstruction algorithm $\hat{x}_1$.

### 5.2 Weak $\ell^p$ Balls

Our main results so far have been stated for $\ell^p$ spaces, but the proofs hold for *weak* $\ell^p$ balls as well ($p < 1$). The weak $\ell^p$ ball of radius $R$ consists of vectors $\theta$ whose decreasing rearrangements $|\theta|_{(1)} \geq |\theta|_{(2)} \geq |\theta|_{(3)} \geq ...$ obey

$$|\theta|_{(N)} \leq R \cdot N^{-1/p}, \quad N = 1, 2, 3, \ldots.$$

Conversely, for a given $\theta$, the smallest $R$ for which these inequalities all hold is defined to be the norm: $\|\theta\|_{w\ell^p} \equiv R$. The "weak" moniker derives from $\|\theta\|_{w\ell^p} \leq \|\theta\|_p$. Weak $\ell^p$ constraints have the following key property: if $\theta_N$ denotes the vector $N$ with all except the $N$ largest items set to zero, then the inequality

$$\|\theta - \theta_N\|_q \leq \zeta_{q,p} \cdot R \cdot (N+1)^{1/q-1/p} \tag{5.1}$$

is valid for $p < 1$ and $q = 1, 2$, with $R = \|\theta\|_{w\ell^p}$. In fact, Theorems 7 and 9 used (5.1) in the proof, together with (implicitly) $\|\theta\|_p \geq \|\theta\|_{w\ell^p}$. Hence we can state results for spaces $Y_{p,m}(R)$ defined using only weak-$\ell^p$ norms, and the proofs apply without change.

## 6 Stylized Applications

We sketch 3 potential applications of the above abstract theory.

### 6.1 Bump Algebra

Consider the class $\mathcal{F}(B)$ of functions $f(t)$, $t \in [0, 1]$ which are restrictions to the unit interval of functions belonging to the Bump Algebra $\mathcal{B}$ [33], with bump norm $\|f\|_{\mathcal{B}} \leq B$. This was mentioned in the introduction, where it was mentioned that the wavelet coefficients at level $j$ obey $\|\theta^{(j)}\|_1 \leq C \cdot B \cdot 2^{-j}$ where $C$ depends only on the wavelet used. Here and below we use standard wavelet analysis notations as in [8, 32, 33].

We consider two ways of approximating functions in $f$. In the *classic linear scheme*, we fix a 'finest scale' $j_1$ and measure the resumé coefficients $\beta_{j_1,k} = \langle f, \varphi_{j_1,k} \rangle$ where $\varphi_{j,k} = 2^{j/2}\varphi(2^j t - k)$, with $\varphi$ a smooth function integrating to 1. Think of these as point samples at scale $2^{-j_1}$ after applying an anti-aliasing filter. We reconstruct by $P_{j_1} f = \sum_k \beta_{j_1,k}\varphi_{j_1,k}$ giving an approximation error

$$\|f - P_{j_1} f\|_2 \leq C \cdot \|f\|_{\mathcal{B}} \cdot 2^{-j_1/2},$$

with $C$ depending only on the chosen wavelet. There are $N = 2^{j_1}$ coefficients $(\beta_{j_1,k})_k$ associated with the unit interval, and so the approximation error obeys:

$$\|f - P_{j_1} f\|_2 \leq C \cdot \|f\|_{\mathcal{B}} \cdot N^{-1/2}.$$

In the *compressed sensing scheme*, we need also wavelets $\psi_{j,k} = 2^{j/2}\psi(2^j x - k)$ where $\psi$ is an oscillating function with mean zero. We pick a coarsest scale $j_0 = j_1/2$. We measure the resumé

coefficients $\beta_{j_0,k}$, – there are $2^{j_0}$ of these – and then let $\theta \equiv (\theta_\ell)_{\ell=1}^m$ denote an enumeration of the detail wavelet coefficients $((\alpha_{j,k} : 0 \le k < 2^j) : j_0 \le j < j_1)$. The dimension $m$ of $\theta$ is $m = 2^{j_1} - 2^{j_0}$. The norm

$$\|\theta\|_1 \le \sum_{j_0}^{j_1} \|(\alpha_{j,k})\|_1 \le \sum_{j_0}^{j_1} C \cdot \|f\|_{\mathcal{B}} \cdot 2^{-j} \le cB2^{-j_0}.$$

We take $n = c \cdot 2^{j_0} \log(2^{j_1})$ and apply a near-optimal information operator for this $n$ and $m$ (described in more detail below). We apply the near-optimal algorithm of $\ell^1$ minimization, getting the error estimate

$$\|\hat{\theta} - \theta\|_2 \le c\|\theta\|_1 \cdot (n/\log(m))^{-1/2} \le c2^{-2j_0} \le c2^{-j_1},$$

with $c$ independent of $f \in \mathcal{F}(B)$. The overall reconstruction

$$\hat{f} = \sum_k \beta_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{j_1-1} \hat{\alpha}_{j,k}\psi_{j,k}$$

has error

$$\|f - \hat{f}\|_2 \le \|f - P_{j_1}f\|_2 + \|P_{j_1}f - \hat{f}\|_2 = \|f - P_{j_1}f\|_2 + \|\hat{\theta} - \theta\|_2 \le c2^{-j_1},$$

again with $c$ independent of $f \in \mathcal{F}(B)$. This is of the same order of magnitude as the error of linear sampling.

The compressed sensing scheme takes a total of $2^{j_0}$ samples of resumé coefficients and $n \le c2^{j_0}\log(2^{j_1})$ samples associated with detail coefficients, for a total $\le c \cdot j_1 \cdot 2^{j_1/2}$ pieces of information. It achieves error comparable to classical sampling based on $2^{j_1}$ samples. Thus it needs dramatically fewer samples for comparable accuracy: roughly speaking, only the square root of the number of samples of linear sampling.

To achieve this dramatic reduction in sampling, we need an information operator based on some $\Phi$ satisfying CS1-CS3. The underlying measurement kernels will be of the form

$$\xi_i = \sum_{j=1}^m \Phi_{i,\ell}\phi_\ell, \qquad i = 1, \ldots, n, \tag{6.1}$$

where the collection $(\phi_\ell)_{\ell=1}^m$ is simply an enumeration of the wavelets $\psi_{j,k}$, with $j_0 \le j < j_1$ and $0 \le k < 2^j$.

## 6.2 Images of Bounded Variation

We consider now the model with images of Bounded Variation from the introduction. Let $\mathcal{F}(B)$ denote the class of functions $f(x)$ with domain$(x) \in [0,1]^2$, having total variation at most $B$ [9], and bounded in absolute value by $\|f\|_\infty \le B$ as well. In the introduction, it was mentioned that the wavelet coefficients at level $j$ obey $\|\theta^{(j)}\|_1 \le C \cdot B$ where $C$ depends only on the wavelet used. It is also true that $\|\theta^{(j)}\|_\infty \le C \cdot B \cdot 2^{-j}$.

We again consider two ways of approximating functions in $f$. The *classic linear scheme* uses a two-dimensional version of the scheme we have already discussed. We again fix a 'finest scale' $j_1$ and measure the resumé coefficients $\beta_{j_1,k} = \langle f, \varphi_{j_1,k}\rangle$ where now $k = (k_1, k_2)$ is a pair of integers

$0 \le k_1, k_2 < 2^{j_1}$. indexing position. We use the Haar scaling function $\varphi_{j_1,k} = 2^{j_1} \cdot 1_{\{2^{j_1}x-k \in [0,1]^2\}}$. We reconstruct by $P_{j_1} f = \sum_k \beta_{j_1,k} \varphi_{j_1,k}$ giving an approximation error

$$\|f - P_{j_1} f\|_2 \le 4 \cdot B \cdot 2^{-j_1/2}.$$

There are $N = 4^{j_1}$ coefficients $\beta_{j_1,k}$ associated with the unit interval, and so the approximation error obeys:

$$\|f - P_{j_1} f\|_2 \le c \cdot B \cdot N^{-1/4}.$$

In the *compressed sensing scheme*, we need also Haar wavelets $\psi^\sigma_{j_1,k} = 2^{j_1}\psi^\sigma(2^{j_1}x - k)$ where $\psi^\sigma$ is an oscillating function with mean zero which is either horizontally-oriented ($\sigma = v$), vertically oriented ($\sigma = h$) or diagonally-oriented ($\sigma = d$). We pick a 'coarsest scale' $j_0 = j_1/2$, and measure the resumé coefficients $\beta_{j_0,k}$, – there are $4^{j_0}$ of these. Then let $\theta \equiv (\theta_\ell)_{\ell=1}^m$ be the concatenation of the detail wavelet coefficients $((\alpha^\sigma_{j,k} : 0 \le k_1, k_2 < 2^j, \sigma \in \{h, v, d\}) : j_0 \le j < j_1)$. The dimension $m$ of $\theta$ is $m = 4^{j_1} - 4^{j_0}$. The norm

$$\|\theta\|_1 \le \sum_{j=j_0}^{j_1} \|(\theta^{(j)})\|_1 \le c(j_1 - j_0)\|f\|_{\mathcal{BV}}.$$

We take $n = c \cdot 4^{j_0} \log^2(4^{j_1})$ and apply a near-optimal information operator for this $n$ and $m$. We apply the near-optimal algorithm of $\ell^1$ minimization to the resulting information, getting the error estimate

$$\|\hat\theta - \theta\|_2 \le c\|\theta\|_1 \cdot (n/\log(m))^{-1/2} \le cB \cdot 2^{-j_1},$$

with $c$ independent of $f \in \mathcal{F}(B)$. The overall reconstruction

$$\hat f = \sum_k \beta_{j_0,k}\varphi_{j_0,k} + \sum_{j=j_0}^{j_1-1} \hat\alpha_{j,k}\psi_{j,k}$$

has error

$$\|f - \hat f\|_2 \le \|f - P_{j_1} f\|_2 + \|P_{j_1} f - \hat f\|_2 = \|f - P_{j_1} f\|_2 + \|\hat\theta - \theta\|_2 \le c2^{-j_1},$$

again with $c$ independent of $f \in \mathcal{F}(B)$. This is of the same order of magnitude as the error of linear sampling.

The compressed sensing scheme takes a total of $4^{j_0}$ samples of resumé coefficients and $n \le c4^{j_0}\log^2(4^{j_1})$ samples associated with detail coefficients, for a total $\le c \cdot j_1^2 \cdot 4^{j_1/2}$ pieces of measured information. It achieves error comparable to classical sampling with $4^{j_1}$ samples. Thus just as we have seen in the Bump Algebra case, we need dramatically fewer samples for comparable accuracy: roughly speaking, only the square root of the number of samples of linear sampling.

### 6.3 Piecewise $C^2$ Images with $C^2$ Edges

We now consider an example where $p < 1$, and we can apply the extensions to tight frames and to weak-$\ell^p$ mentioned earlier. Again in the image processing setting, we use the $C^2$-$C^2$ model discussed in Candès and Donoho [2, 3]. Consider the collection $\mathcal{C}^{2,2}(B, L)$ of piecewise smooth $f(x)$, $x \in [0,1]^2$, with values, first and second partial derivatives bounded by $B$, away from an exceptional set $\Gamma$ which is a union of $C^2$ curves having first and second derivatives in an appropriate parametrization $\le B$; the curves have total length $\le L$. More colorfully, such

images are *cartoons* – well-behaved except for discontinuities inside regions with nice curvilinear boundaries. They might be reasonable models for certain kinds of technical imagery – eg in radiology.

The curvelets tight frame [3] is a collection of smooth frame elements offering a Parseval relation

$$\|f\|_2^2 = \sum_\mu |\langle f, \gamma_\mu \rangle|^2$$

and reconstruction formula

$$f = \sum_\mu \langle f, \gamma_\mu \rangle \gamma_\mu.$$

The frame elements have a multiscale organization, and frame coefficients $\theta^{(j)}$ grouped by scale obey the weak $\ell^p$ constraint

$$\|\theta^{(j)}\|_{w\ell^{2/3}} \le c(B, L), \qquad f \in \mathcal{C}^{2,2}(B, L);$$

compare [3]. For such objects, classical linear sampling at scale $j_1$ by smooth 2-D scaling functions gives

$$\|f - P_{j_1} f\|_2 \le c \cdot B \cdot 2^{-j_1/2}, \qquad f \in \mathcal{C}^{2,2}(B, L).$$

This is no better than the performance of linear sampling for the BV case, despite the piecewise $C^2$ character of $f$; the possible discontinuities in $f$ are responsible for the inability of linear sampling to improve its performance over $C^{2,2}(B, L)$ compared to BV.

In the *compressed sensing scheme*, we pick a coarsest scale $j_0 = j_1/4$. We measure the resumé coefficients $\beta_{j_0,k}$ in a smooth wavelet expansion – there are $4^{j_0}$ of these – and then let $\theta \equiv (\theta_\ell)_{\ell=1}^m$ denote a concatentation of the finer-scale curvelet coefficients $(\theta^{(j)} : j_0 \le j < j_1)$. The dimension $m$ of $\theta$ is $m = c(4^{j_1} - 4^{j_0})$, with $c > 1$ due to overcompleteness of curvelets. The norm

$$\|\theta\|_{w\ell^{2/3}} \le (\sum_{j_0}^{j_1} \|(\theta^{(j)})\|_{w\ell^{2/3}}^{2/3})^{3/2} \le c(j_1 - j_0)^{3/2},$$

with $c$ depending on $B$ and $L$. We take $n = c \cdot 4^{j_0} \log^{5/2}(4^{j_1})$ and apply a near-optimal information operator for this $n$ and $m$. We apply the near-optimal algorithm of $\ell^1$ minimization to the resulting information, getting the error estimate

$$\|\hat{\theta} - \theta\|_2 \le c\|\theta\|_{2/3} \cdot (n/\log(m))^{-1} \le c' \cdot 2^{-j_1/2},$$

with $c'$ absolute. The overall reconstruction

$$\hat{f} = \sum_k \beta_{j_0,k} \varphi_{j_0,k} + \sum_{j=j_0}^{j_1-1} \sum_{\mu \in M_j} \hat{\theta}_\mu^{(j)} \gamma_\mu$$

has error

$$\|f - \hat{f}\|_2 \le \|f - P_{j_1} f\|_2 + \|P_{j_1} f - \hat{f}\|_2 = \|f - P_{j_1} f\|_2 + \|\hat{\theta} - \theta\|_2 \le c 2^{-j_1/2},$$

again with $c$ independent of $f \in \mathcal{C}^{2,2}(B, L)$. This is of the same order of magnitude as the error of linear sampling.

The compressed sensing scheme takes a total of $4^{j_0}$ samples of resumé coefficients and $n \le c4^{j_0} \log^{5/2}(4^{j_1})$ samples associated with detail coefficients, for a total $\le c \cdot j_1^{5/2} \cdot 4^{j_1/4}$ pieces of information. It achieves error comparable classical sampling based on $4^{j_1}$ samples. Thus, even more so than in the Bump Algebra case, we need dramatically fewer samples for comparable accuracy: roughly speaking, only the fourth root of the number of samples of linear sampling.

# 7 Nearly All Matrices are CS-Matrices

We may reformulate Theorem 6 as follows.

**Theorem 10** *Let $n, m_n$ be a sequence of problem sizes with $n \to \infty$, $n < m \sim An^\gamma$, for $A > 0$ and $\gamma > 1$. Let $\Phi = \Phi_{n,m}$ be a matrix with $m$ columns drawn independently and uniformly at random from $S^{n-1}$. Then for some $\eta_i > 0$ and $\rho > 0$, conditions CS1-CS3 hold for $\Phi$ with overwhelming probability for all large $n$.*

Indeed, note that the probability measure on $\mathbf{\Phi}_{n,m}$ induced by sampling columns iid uniform on $\mathbf{S}^{n-1}$ is exactly the natural uniform measure on $\mathbf{\Phi}_{n,m}$. Hence Theorem 6 follows immediately from Theorem 10.

In effect matrices $\Phi$ satisfying the CS conditions are so ubiquitous that it is reasonable to generate them by sampling *at random* from a *uniform* probability distribution.

The proof of Theorem 10 is conducted over the next three subsections; it proceeds by studying events $\Omega_n^i$, $i = 1, 2, 3$, where $\Omega_n^1 \equiv \{$ CS1 Holds $\}$, etc. It will be shown that for parameters $\eta_i > 0$ and $\rho_i > 0$

$$P(\Omega_n^i) \to 1, \qquad n \to \infty;$$

then defining $\rho = \min_i \rho_i$ and $\Omega_n = \cap_i \Omega_n^i$, we have

$$P(\Omega_n) \to 1, \qquad n \to \infty.$$

Since, when $\Omega_n$ occurs, our random draw has produced a matrix obeying CS1-CS3 with parameters $\eta_i$ and $\rho$, this proves Theorem 10.

## 7.1 Control of Minimal Eigenvalue

The following lemma allows us to choose positive constants $\rho_1$ and $\eta_1$ so that condition CS1 holds with overwhelming probability.

**Lemma 7.1** *Consider sequences of $(n, m_n)$ with $n \le m_n \sim An^\gamma$. Define the event*

$$\Omega_{n,m,\rho,\lambda} = \{\lambda_{min}(\Phi_J^T \Phi_J) \ge \lambda, \qquad \forall |J| < \rho \cdot n / \log(m)\}.$$

*Then, for each $\lambda < 1$, for sufficiently small $\rho > 0$*

$$P(\Omega_{n,m,\rho,\lambda}) \to 1, n \to \infty.$$

The proof involves three ideas. First, that for a specific subset we get large deviations bounds on the minimum eigenvalue.

**Lemma 7.2** *For $J \subset \{1, \dots m\}$, let $\Omega_{n,J}$ denote the event that the minimum eigenvalue $\lambda_{min}(\Phi_J^T \Phi_J)$ exceeds $\lambda < 1$. Then for sufficiently small $\rho_1 > 0$ there is $\beta_1 > 0$ so that for all $n > n_0$,*

$$P(\Omega_{n,J}^c) \le \exp(-n\beta_1),$$

*uniformly in $|J| \le \rho_1 n$.*

This was derived in [17] and in [18], using the concentration of measure property of singular values of random matrices, eg. see Szarek's work [44, 45].

Second, we note that the event of main interest is representable as:

$$\Omega_{n,m,\rho,\eta} = \cap_{|J| \le \rho n/\log(m)} \Omega_{n,J}.$$

Thus we need to estimate the probability of occurrence of every $\Omega_{n,J}$ simultaneously.

Third, we can combine the individual bounds to control simultaneous behavior:

**Lemma 7.3** *Suppose we have events $\Omega_{n,J}$ all obeying, for some fixed $\beta > 0$ and $\rho > 0$,*

$$P(\Omega_{n,J}^c) \le \exp(-n\beta),$$

*for each $J \subset \{1, \ldots, m\}$ with $|J| \le \rho n$. Pick $\rho_1 > 0$ with $\rho_1 < \min(\beta, \rho)$ and $\beta_1 > 0$ with $\beta_1 < \beta - \rho_1$. Then for all sufficiently large $n$,*

$$P\{\Omega_{n,J}^c \text{ for some } J \subset \{1, \ldots, m\} \text{ with } |J| \le \rho_1 n/\log(m)\} \le \exp(-\beta_1 n).$$

Lemma 7.1 follows from this immediately.

To prove Lemma 7.3, let $\mathcal{J} = \{J \subset \{1, \ldots, m\} \text{ with } |J| \le \rho_1 n/\log(m)\}$. We note that by Boole's inequality,

$$P(\cup_{\mathcal{J}} \Omega_{n,J}^c) \le \sum_{J \in \mathcal{J}} P(\Omega_{n,J}^c) \le \#\mathcal{J} \cdot \exp(-\beta n),$$

the last inequality following because each member $J \in \mathcal{J}$ is of cardinality $\le \rho n$, since $\rho_1 n/\log(m) < \rho n$, as soon as $n \ge 3 > e$. Also, of course,

$$\log \binom{m}{k} \le k \log(m),$$

so we get $\log(\#\mathcal{J}) \le \rho_1 n$. Taking $\beta_1$ as given, we get the desired conclusion. QED

## 7.2 Spherical Sections Property

We now show that condition CS2 can be made overwhelmingly likely by choice of $\eta_2$ and $\rho_2$ sufficiently small but still positive. Our approach derives from [17], which applied an important result from Banach space theory, the almost *spherical sections* phenomenon. This says that slicing the unit ball in a Banach space by intersection with an appropriate finite-dimensional linear subspace will result in a slice that is effectively spherical. We develop a quantitative refinement of this principle for the $\ell^1$ norm in $\mathbf{R}^n$, showing that, with overwhelming probability, every operator $\Phi_J$ for $|J| < \rho n/\log(m)$ affords a spherical section of the $\ell_n^1$ ball. The basic argument we use originates from work of Milman, Kashin and others [22, 28, 37]; we refine an argument in Pisier [41] and, as in [17] draw inferences that may be novel. We conclude that not only do almost-spherical sections exist, but they are so ubiquitous that every $\Phi_J$ with $|J| < \rho n/\log(m)$ will generate them.

**Definition 7.1** *Let $|J| = k$. We say that $\Phi_J$ offers an $\epsilon$-isometry between $\ell^2(J)$ and $\ell_n^1$ if*

$$(1-\epsilon) \cdot \|\alpha\|_2 \le \sqrt{\frac{\pi}{2n}} \cdot \|\Phi_J \alpha\|_1 \le (1+\epsilon) \cdot \|\alpha\|_2, \quad \forall \alpha \in \mathbf{R}^k. \tag{7.1}$$

The following Lemma shows that condition CS2 is a generic property of matrices.

**Lemma 7.4** *Consider the event $\Omega_n^2 (\equiv \Omega_n^2(\epsilon, \rho))$ that every $\Phi_J$ with $|J| \le \rho \cdot n/\log(m)$ offers an $\epsilon$-isometry between $\ell^2(J)$ and $\ell_n^1$. For each $\epsilon > 0$, there is $\rho(\epsilon) > 0$ so that*

$$P(\Omega_n^2) \to 1, \qquad n \to \infty.$$

To prove this, we first need a lemma about individual subsets $J$ proven in [17].

**Lemma 7.5** *Fix $\epsilon > 0$. Choose $\delta$ so that*

$$(1-3\delta)(1-\delta)^{-1} \ge (1-\epsilon)^{1/2} \text{ and } (1+\delta)(1-\delta)^{-1} \le (1+\epsilon)^{1/2}. \tag{7.2}$$

*Choose $\rho_1 = \rho_1(\epsilon)$ so that*

$$\rho_1 \cdot (1 + 2/\delta) < \delta^2 \frac{2}{\pi},$$

*and let $\beta(\epsilon)$ denote the difference between the two sides. For a subset $J$ in $\{1, \ldots, m\}$ let $\Omega_{n,J}$ denote the event that $\Phi_J$ furnishes an $\epsilon$-isometry to $\ell_n^1$. Then as $n \to \infty$,*

$$\max_{|J| \le \rho_1 n} P(\Omega_{n,J}^c) \le 2\exp(-\beta(\epsilon)n(1+o(1))).$$

Now note that the event of interest for Lemma 7.4 is

$$\Omega_n^2 = \cap_{|J| \le \rho n/\log(m)} \Omega_{n,J};$$

to finish apply the individual Lemma 7.5 together with the combining principle in Lemma 7.3.

### 7.3 Quotient Norm Inequalities

We now show that, for $\eta_3 = 3/4$, for sufficiently small $\rho_3 > 0$, nearly all matrices have property CS3.

Let $J$ be any collection of indices in $\{1, \ldots, m\}$; $Range(\Phi_J)$ is a linear subspace of $\mathbf{R}^n$, and on this subspace a subset $\Sigma_J$ of possible *sign patterns* can be realized, i.e. sequences of $\pm 1$'s generated by

$$\sigma(k) = \mathrm{sgn}\Big\{ \sum_I \alpha_i \phi_i(k) \Big\}, \qquad 1 \le k \le n.$$

CS3 will follow if we can show that for every $v \in Range(\Phi_I)$, some approximation $y$ to $\mathrm{sgn}(v)$ satisfies $|\langle y, \phi_i \rangle| \le 1$ for $i \in J^c$.

**Lemma 7.6 Simultaneous Sign-Pattern Embedding.** *Fix $\delta > 0$. Then for $\tau < \delta^2/32$, set*

$$\epsilon_n = (\log(m_n/(\tau n)))^{-1/2}/4.$$

*For sufficiently small $\rho_3 > 0$, there is an event $\Omega_n^3 \equiv \Omega_n^3(\rho_3, \delta)$ with $P(\Omega_n^3) \to 1$, as $n \to \infty$. On this event, for every subset $J$ with $|J| < \rho_3 n/\log(m)$, for every sign pattern in $\sigma \in \Sigma_J$, there is a vector $y (\equiv y_\sigma)$ with*

$$\|y - \epsilon_n \sigma\|_2 \le \epsilon_n \cdot \delta \cdot \|\sigma\|_2, \tag{7.3}$$

*and*

$$|\langle \phi_i, y \rangle| \le 1, \qquad i \in J^c. \tag{7.4}$$

In words, a small multiple $\epsilon_n\sigma$ of any sign pattern $\sigma$ almost lives in the dual ball $\{x : |\langle\phi_i, x\rangle| \le 1, i \in J^c\}$.

Before proving this result, we indicate how it gives the property CS3; namely, that $|J| < \rho_3 n/\log(m)$, and $v = -\Phi_{J^c}\beta_{J^c}$ imply

$$\|\beta_{J^c}\|_1 \ge \eta_3/\sqrt{\log(m/n)} \cdot \|v\|_1.$$

By the duality theorem for linear programming the value of the primal program

$$\min \|\beta_{J^c}\|_1 \text{ subject to } \Phi_{J^c}\beta_{J^c} = -v \tag{7.5}$$

is at least the value of the dual

$$\max\langle v, y\rangle \text{ subject to } |\langle\phi_i, y\rangle| \le 1, i \in J^c.$$

Lemma 7.6 gives us a supply of dual-feasible vectors and hence a lower bound on the dual program. Take $\sigma = \mathrm{sgn}(v)$; we can find $y$ which is dual-feasible and obeys

$$\langle v, y\rangle \ge \langle v, \epsilon_n\sigma\rangle - \|y - \epsilon_n\sigma\|_2\|v\|_2 \ge \epsilon_n\|v\|_1 - \epsilon_n\delta\|\sigma\|_2\|v\|_2;$$

picking $\delta$ appropriately and taking into account the spherical sections theorem, for sufficiently large $n$ we have $\delta\|\sigma\|_2\|v\|_2 \le \frac{1}{4}\|v\|_1$; (7.5) follows with $\eta_3 = 3/4$.

### 7.3.1 Proof of Simultaneous Sign-Pattern Embedding

The proof of Lemma 7.6 follows closely a similar result in [17] that considered the case $n < m < Am$. Our idea here is to adapt the argument for the $n < m < Am$ case to the $n < m \sim Am^\gamma$ case, with changes reflecting the different choice of $\epsilon$, $\delta$, and the sparsity bound $\rho n/\log(m)$. We leave out large parts of the argument, as they are identical to the corresponding parts in [17]. The bulk of our effort goes to produce the following lemma, which demonstrates approximate embedding of a *single* sign pattern in the dual ball.

**Lemma 7.7 Individual Sign-Pattern Embedding**. *Let $\sigma \in \{-1, 1\}^n$, let $y_0 = \epsilon_n\sigma$, with $\epsilon_n$, $m_n$, $\tau$, $\delta$ as in the statement of Lemma 7.6. Let $k \ge 0$. Given a collection $(\phi_i : 1 \le i \le m - k)$, there is an iterative algorithm, described below, producing a vector $y$ as output which obeys*

$$|\langle\phi_i, y\rangle| \le 1, \quad i = 1, \ldots, m - k. \tag{7.6}$$

*Let $(\phi_i)_{i=1}^{m-k}$ be iid uniform on $\mathbf{S}^{n-1}$; there is an event $\Omega_\sigma$ described below, having probability controlled by*

$$Prob(\Omega_\sigma^c) \le 2n\exp\{-n\beta\}, \tag{7.7}$$

*for $\beta > 0$ which can be explicitly given in terms of $\tau$ and $\delta$. On this event,*

$$\|y - y_0\|_2 \le \delta \cdot \|y_0\|_2. \tag{7.8}$$

Lemma 7.7 will be proven in a section of its own. We now show that it implies Lemma 7.6.

We recall a standard implication of so-called Vapnik-Cervonenkis theory [42]:

$$\#\Sigma_J \le \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{|J|}.$$

Notice that if $|J| < \rho n/\log(m)$, then

$$\log(\#\Sigma_J) \le \rho \cdot n + \log(n),$$

while also

$$\log \#\{J : |J| < \rho n/\log(m), J \subset \{1,\dots,m\}\} \le \rho n.$$

Hence, the total number of sign patterns generated by operators $\Phi_J$ obeys

$$\log \#\{\sigma : \sigma \in \Sigma_J, |J| < \rho n/\log(m)\} \le n \cdot 2\rho + \log(n).$$

Now $\beta$ furnished by Lemma 7.7 is positive, so pick $\rho_3 < \beta/2$ with $\rho_3 > 0$. Define

$$\Omega_n^3 = \cap_{|J|<\rho_3 n/\log(m)} \cap_{\sigma\in\Sigma_J} \Omega_{\sigma,J},$$

where $\Omega_{\sigma,J}$ denotes the instance of the event (called $\Omega_\sigma$ in the statement of Lemma 7.7) generated by a specific $\sigma, J$ combination. On the event $\Omega_n^3$, *every* sign pattern associated with *any* $\Phi_J$ obeying $|J| < \rho_3 n/\log(m)$ is almost dual feasible. Now

$$\begin{aligned} P((\Omega_n^3)^c) &\le \#\{\sigma : \sigma \in \Sigma_J, |J| < \rho_3 n/\log(m)\} \cdot \exp(-n\beta), \\ &\le \exp\{-n(\beta - 2\rho_3) + \log(n)\} \to 0, \qquad n \to \infty. \end{aligned}$$

QED.

## 7.4 Proof of Individual Sign-Pattern Embedding

### 7.4.1 An Embedding Algorithm

The companion paper [17] introduced an algorithm to create a dual feasible point $y$ starting from a nearby almost-feasible point $y_0$. It worked as follows.

Let $I_0$ be the collection of indices $1 \le i \le m$ with

$$|\langle \phi_i, y_0\rangle| > 1/2,$$

and then set

$$y_1 = y_0 - P_{I_0} y_0,$$

where $P_{I_0}$ denotes the least-squares projector $\Phi_{I_0}(\Phi_{I_0}^T\Phi_{I_0})^{-1}\Phi_{I_0}^T$. In effect, identify the indices where $y_0$ exceeds half the forbidden level $|\langle \phi_i, y_0\rangle| > 1$, and "kill" those indices.

Continue this process, producing $y_2$, $y_3$, etc., with stage-dependent thresholds $t_\ell \equiv 1 - 2^{-\ell-1}$ successively closer to 1. Set

$$I_\ell = \{i : |\langle \phi_i, y_\ell\rangle| > t_\ell\},$$

and, putting $J_\ell \equiv I_0 \cup \cdots \cup I_\ell$,

$$y_{\ell+1} = y_0 - P_{J_\ell} y_0.$$

If $I_\ell$ is empty, then the process terminates, and set $y = y_\ell$. Termination must occur at stage $\ell^* \le n$. At termination,

$$|\langle \phi_i, y\rangle| \le 1 - 2^{-\ell^*-1}, \quad i = 1,\dots,m.$$

Hence $y$ is definitely dual feasible. The only question is how close to $y_0$ it is.

### 7.4.2 Analysis Framework

Also in [17] bounds were developed for two key descriptors of the algorithm trajectory:

$$\alpha_\ell = \|y_\ell - y_{\ell-1}\|_2,$$

and

$$|I_\ell| = \#\{i : |\langle \phi_i, y_\ell \rangle| > 1 - 2^{-\ell-1}\}.$$

We adapt the arguments deployed there. We define bounds $\delta_\ell \equiv \delta_{\ell;n}$ and $\nu_\ell \equiv \nu_{\ell;n}$ for $\alpha_\ell$ and $|I_\ell|$, of the form

$$\delta_{\ell;n} = \|y_0\|_2 \cdot \omega^\ell, \qquad \ell = 1, 2, \ldots,$$

$$\nu_{\ell;n} = n \cdot \lambda_0 \cdot \epsilon_n^2 \cdot \omega^{2\ell+2}/4, \qquad \ell = 0, 1, 2, \ldots;$$

here $\lambda_0 = 1/2$ and $\omega = \min(1/2, \delta/2, \omega_0)$, where $\omega_0 > 0$ will be defined later below. We define sub-events

$$E_\ell = \{\alpha_j \le \delta_j, \quad j = 1, \ldots, \ell, \quad |I_j| \le \nu_j, j = 0, \ldots, \ell - 1\};$$

Now define

$$\Omega_\sigma = \cap_{\ell=1}^n E_\ell;$$

this event implies, because $\omega \le 1/2$,

$$\|y - y_0\|_2 \le (\sum \alpha_\ell^2)^{1/2} \le \|y_0\|_2 \cdot \omega/(1-\omega^2)^{1/2} \le \|y_0\|_2 \cdot \delta.$$

We will show that, for $\beta > 0$ chosen in conjunction with $\tau > 0$,

$$P(E_{\ell+1}^c | E_\ell) \le 2 \exp\{-\beta n\}. \tag{7.9}$$

This implies

$$P(\Omega_\sigma^c) \le 2n \exp\{-\beta n\},$$

and the Lemma follows. QED

### 7.4.3 Large Deviations

Define the events

$$F_\ell = \{\alpha_\ell \le \delta_{\ell;n}\}, \qquad G_\ell = \{|I_\ell| \le \nu_{\ell;n}\},$$

so that

$$E_{\ell+1} = F_{\ell+1} \cap G_\ell \cap E_\ell.$$

Put

$$\rho_0(\tau, \omega; n) = (128)^{-1} \frac{\log(An^\gamma)}{\log(An^\gamma/\tau n)} \frac{\omega^2}{1-\omega^2};$$

and note that this depends quite weakly on $n$. Recall that the event $E_\ell$ is defined in terms of $\omega$ and $\tau$. On the event $E_\ell$, $|J_\ell| \le \rho_0 n/\log(m)$. Lemma 7.1 implicitly defined a quantity $\lambda_1(\rho, A, \gamma)$ lowerbounding the minimum eigenvalue of every $\Phi_J^T \Phi_J$ where $|J| \le \rho n/\log(m)$. Pick $\rho_{1/2} > 0$ so that $\lambda_1(\rho_{1/2}, A, \gamma) > 1/2$. Pick $\omega_0$ so that

$$\rho_0(\tau, \omega_0; n) < \rho_{1/2}, \qquad n > n_0.$$

With this choice of $\omega_0$, when the event $E_\ell$ occurs, $\lambda_{min}(\Phi_{I_\ell}^T \Phi_{I_\ell}) > \lambda_0$. Also on $E_\ell$, $u_j = 2^{-j-1}/\alpha_j > 2^{-j-1}/\delta_j = v_j$ (say) for $j \le \ell$.

In [17] an analysis framework was developed in which a family $(Z_i^\ell : 1 \le i \le m, 0 \le \ell < n)$ of random variables iid $N(0, \frac{1}{n})$ was introduced, and it was shown that

$$P\{G_\ell^c | E_\ell\} \le P\{\sum_i 1_{\{|Z_i^\ell| > v_\ell\}} > \nu_\ell\},$$

and

$$P\{F_{\ell+1}^c | G_\ell, E_\ell\} \le P\{2 \cdot \lambda_0^{-1}\big[\nu_\ell + \delta_\ell^2\big(\sum_i \Big(Z_i^\ell\Big)^2 1_{\{|Z_i^\ell| > v_\ell\}}\big)\big] > \delta_{\ell+1}^2\}.$$

That paper also stated two simple Large Deviations bounds:

**Lemma 7.8** *Let $Z_i$ be iid $N(0,1)$, $k \ge 0$, $t > 2$.*

$$\frac{1}{m} \log P\{\sum_{i=1}^{m-k} Z_i^2 1_{\{|Z_i| > t\}} > m\Delta\} \le e^{-t^2/4} - \Delta/4,$$

*and*

$$\frac{1}{m} \log P\{\sum_{i=1}^{m-k} 1_{\{|Z_i| > t\}} > m\Delta\} \le e^{-t^2/2} - \Delta/4.$$

Applying this, we note that the event

$$\big\{2 \cdot \lambda_0^{-1}\big[\nu_\ell + \delta_\ell^2\big(\sum_i \Big(Z_i^\ell\Big)^2 1_{\{|Z_i^\ell| > v_\ell\}}\big)\big] > \delta_{\ell+1}^2\},$$

stated in terms of $N(0, \frac{1}{n})$ variables, is equivalent to an event

$$\big\{\sum_{i=1}^{m-k} Z_i^2 1_{\{|Z_i| > \tau_\ell\}} > m\Delta_\ell\big\},$$

stated in terms of standard $N(0,1)$ random variables, where

$$\tau_\ell^2 = n \cdot v_\ell^2 = \epsilon_n^2 (2\omega)^{-2\ell}/4,$$

and

$$\Delta_\ell = \frac{n}{m}(\lambda_0 \delta_{\ell+1}^2/2 - \nu_\ell)/\delta_\ell^2$$

We therefore have the inequality

$$\frac{1}{m} \log P\{F_{\ell+1}^c | G_\ell, E_\ell\} \le e^{-\tau_\ell^2/4} - \Delta_\ell/4.$$

Now

$$e^{-\tau_\ell^2/4} = e^{-[(16 \log(m/\tau n))/16] \cdot (2\omega)^{-2\ell}} = (\tau n/m)^{(2\omega)^{-2\ell}},$$

and

$$\Delta_\ell = \frac{n}{m}(\omega^2/4 - \omega^2/8) = \frac{n}{m}\omega^2/8.$$

Since $\omega \le 1/2$, the term of most concern in $(\tau n/m)^{(2\omega)^{-2\ell}}$ is at $\ell = 0$; the other terms are always better. Also $\Delta_\ell$ in fact does not depend on $\ell$. Focusing now on $\ell = 0$, we may write

$$\log P\{F_1^c | G_0\} \le m(\tau \cdot n/m - n/m \cdot \omega^2/8) = n(\tau - \omega^2/8).$$

Recalling that $\omega \le \delta/2$ and putting

$$\beta \equiv \beta(\tau;\delta) = (\delta/2)^2/8 - \tau,$$

we get $\beta > 0$ for $\tau < \delta^2/32$, and

$$P\{F_{\ell+1}^c | G_\ell, E_\ell\} \le \exp(-n\beta).$$

A similar analysis holds for the $G_\ell$'s. QED

# 8 Conclusion

## 8.1 Summary

We have described an abstract framework for compressed sensing of objects which can be represented as vectors $x \in \mathbf{R}^m$. We assume the $x$ of interest is *a priori* compressible so that in $\|\Psi^T x\|_p \le R$ for a known basis or frame $\Psi$ and $p < 2$. Starting from an $n$ by $m$ matrix $\Phi$ with $n < m$ satisfying conditions CS1-CS3, and with $\Psi$ the matrix of an orthonormal basis or tight frame underlying $X_{p,m}(R)$, we define the information operator $I_n(x) = \Phi\Psi^T x$. Starting from the $n$-pieces of measured information $y_n = I_n(x)$ we reconstruct $x$ by solving

$$(L_1) \quad \min_x \|\Psi^T x\|_1 \text{ subject to } y_n = I_n(x)$$

The proposed reconstruction rule uses convex optimization and is computationally tractable. Also the needed matrices $\Phi$ satisfying CS1-CS3 can be constructed by random sampling from a uniform distribution on the columns of $\Phi$.

We give error bounds showing that despite the apparent undersampling ($n < m$), good accuracy reconstruction is possible for compressible objects, and we explain the near-optimality of these bounds using ideas from Optimal Recovery and Information-Based Complexity. Examples are sketched related to imaging and spectroscopy.

## 8.2 Alternative Formulation

We remark that the CS1-CS3 conditions are not the only way to obtain our results. Our proof of Theorem 9 really shows the following:

**Theorem 11** *Suppose that an $n \times m$ matrix $\Phi$ obeys the following conditions, with constants $\rho_1 > 0$,$\eta_1 < 1/2$ and $\eta_2 < \infty$:*

**A1** *The maximal concentration $\nu(\Phi, J)$ (defined in Section 4.2) obeys*

$$\nu(\Phi, J) < \eta_1, \qquad |J| < \rho_1 n/\log(m). \tag{8.1}$$

**A2** *The width $w(\Phi, b_{1,m_n})$ (defined in Section 2) obeys*

$$w(\Phi, b_{1,m_n}) \le \eta_2 \cdot (n/\log(m_n))^{-1/2}. \tag{8.2}$$

*Let $0 < p \le 1$. For some $C = C(p, (\eta_i), \rho_1)$ and all $\theta \in b_{p,m}$ the solution $\hat\theta_{1,n}$ of $(P_1)$ obeys the estimate:*

$$\|\hat\theta_{1,n} - \theta\|_2 \le C \cdot (n/\log(m_n))^{1/2-1/p}.$$

In short, a different approach might exhibit operators $\Phi$ with good widths over $\ell^1$ balls only, and low concentration on 'thin' sets. Another way to see that the conditions CS1-CS3 can no doubt be approached differently is to compare the results in [17, 18]; the second paper proves results which partially overlap those in the first, by using a different technique.

### 8.3 The Partial Fourier Ensemble

We briefly discuss two recent articles which do not fit in the $n$-widths tradition followed here, and so were not easy to cite earlier with due prominence.

First, and closest to our viewpoint, is the breakthrough paper of Candès, Romberg, and Tao [4]. This was discussed in Section 4.2 above; the result of [4] showed that $\ell^1$ minimization can be used to exactly recover sparse sequences from the Fourier transform at $n$ randomly-chosen frequencies, whenever the sequence has fewer than $\rho^* n/\log(n)$ nonzeros, for some $\rho^* > 0$. Second is the article of Gilbert et al. [25], which showed that a different nonlinear reconstruction algorithm can be used to recover approximations to a vector in $\mathbf{R}^m$ which is nearly as good as the best $N$-term approximation in $\ell^2$ norm, using about $n = O(\log(m)\log(M)N)$ random but nonuniform samples in the frequency domain; here $M$ is (it seems) an upper bound on the norm of $\theta$.

These articles both point to the *partial Fourier Ensemble*, i.e. the collection of $n \times m$ matrices made by sampling $n$ rows out of the $m \times m$ Fourier matrix, as concrete examples of $\Phi$ working within the CS framework; that is, generating near-optimal subspaces for Gel'fand $n$-widths, and allowing $\ell^1$ minimization to reconstruct from such information for all $0 < p \leq 1$.

Now [4] (in effect) proves that if $m_n \sim An^\gamma$, then in the partial Fourier ensemble with uniform measure, the maximal concentration condition A1 (8.1) holds with overwhelming probability for large $n$ (for appropriate constants $\eta_1 < 1/2$, $\rho_1 > 0$). On the other hand, the results in [25] seem to show that condition A2 (8.2) holds in the partial Fourier ensemble with overwhelming probability for large $n$, when it is sampled with a certain non-uniform probability measure. Although the two papers [4, 25] refer to different random ensembles of partial Fourier matrices, both reinforce the idea that interesting relatively concrete families of operators can be developed for compressed sensing applications. In fact, Emmanuel Candès has informed us of some recent results he obtained with Terry Tao [6] indicating that, modulo polylog factors, A2 holds for the uniformly sampled partial Fourier ensemble. This seems a very significant advance.

## Acknowledgements

In spring 2004, Emmanuel Candès told DLD about his ideas for using the partial Fourier ensemble in 'undersampled imaging'; some of these were published in [4]; see also the presentation [5]. More recently, Candès informed DLD of the results in [6] we referred to above. It is a pleasure to acknowledge the inspiring nature of these conversations. DLD would also like to thank Anna Gilbert for telling him about her work [25] finding the B-best Fourier coefficients by nonadaptive sampling, and to thank Candès for conversations clarifying Gilbert's work.

## References

[1] J. Bergh and J. Löfström (1976) *Interpolation Spaces. An Introduction.* Springer Verlag.

[2] E. J. Candès and DL Donoho (2000) Curvelets - a surprisingly effective nonadaptice representation for objects with edges. in *Curves and Surfaces* eds. C. Rabut, A. Cohen, and L.L. Schumaker. Vanderbilt University Press, Nashville TN.

[3] 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.

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

[5] Candes, EJ. (2004) Presentation at Second International Conference on Computational Harmonic Anaysis, Nashville Tenn. May 2004.

[6] Candes, EJ. and Tao, T. (2004) Estimates for Fourier Minors, with Applications. Manuscript.

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

[8] Ingrid Daubechies (1992) *Ten Lectures on Wavelets.* SIAM.

[9] Cohen, A. DeVore, R., Petrushev P., Xu, H. (1999) Nonlinear Approximation and the space $BV(R^2)$. *Amer. J. Math.* **121**, 587-628.

[10] Donoho, DL, Vetterli, M. DeVore, R.A. Daubechies, I.C. (1998) Data Compression and Harmonic Analysis. *IEEE Trans. Information Theory.* **44**, 2435-2476.

[11] Donoho, DL (1993) Unconditional Bases are optimal bases for data compression and for statistical estimation. *Appl. Computational Harmonic analysis* **1** 100-115.

[12] Donoho, DL (1996) Unconditional Bases and bit-level compression. *Appl. Computational Harmonic analysis* **3** 388-392.

[13] Donoho, DL (2001) Sparse components of images and optimal atomic decomposition. *Constructive Approximation* **17** 353-382.

[14] 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.

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

[16] 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.

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

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

[19] A. Dvoretsky (1961) Some results on convex bodies and Banach Spaces. *Proc. Symp. on Linear Spaces.* Jerusalem, 123-160.

[20] 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.

[21] J.J. Fuchs (2002) On sparse representation in arbitrary redundant bases. Manuscript.

[22] T. Figiel, J. Lindenstrauss and V.D. Milman (1977) The dimension of almost-spherical sections of convex bodies. *Acta Math.* **139** 53-94.

[23] S. Gal, C. Micchelli, Optimal sequential and non-sequential procedures for evaluating a functional. *Appl. Anal.* **10** 105-120.

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

[25] 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.

[26] R. Gribonval and M. Nielsen. Sparse Representations in Unions of Bases. To appear *IEEE Trans Info Thry.*

[27] G. Golub and C. van Loan.(1989) *Matrix Computations*. Johns Hopkins: Baltimore.

[28] 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.

[29] M.A. Kon and E. Novak The adaption problem for approximating linear operators *Bull. Amer. Math. Soc.* **23** 159-165.

[30] Thomas Kuhn (2001) A lower estimate for entropy numbers. *Journal of Approximation Theory*, **110**, 120-124.

[31] Michel Ledoux. *The Concentration of Measure Phenomenon.* Mathematical Surveys and Monographs **89**. American Mathematical Society 2001.

[32] S. Mallat(1998). A Wavelet Tour of Signal Processing. Academic Press.

[33] Y. Meyer. (1993) *Wavelets and Operators*. Cambridge University Press.

[34] Melkman, A. A., Micchelli, C. A.; Optimal estimation of linear operators from inaccurate data. *SIAM J. Numer. Anal.* **16**; 1979; 87–105;

[35] A.A. Melkman (1980) $n$-widths of octahedra. in *Quantitative Approximation* eds. R.A. DeVore and K. Scherer, 209-216, Academic Press.

[36] Micchelli, C.A. and Rivlin, T.J. (1977) A Survey of Optimal Recovery, in *Optimal Estimation in Approximation Theory*, eds. C.A. Micchelli, T.J. Rivlin, Plenum Press, NY, 1-54.

[37] V.D. Milman and G. Schechtman (1986) *Asymptotic Theory of Finite-Dimensional Normed Spaces.* Lect. Notes Math. **1200**, Springer.

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

[39] 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.

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

[41] G. Pisier (1989) *The Volume of Convex Bodies and Banach Space Geometry.* Cambridge University Press.

[42] D. Pollard (1989) *Empirical Processes: Theory and Applications.* NSF - CBMS Regional Conference Series in Probability and Statistics, Volume 2, IMS.

[43] Carsten Schutt (1984) Entropy numbers of diagonal operators between symmetric banach spaces. *Journal of Approximation Theory*, **40**, 121-128.

[44] Szarek, S.J. (1990) Spaces with large distances to $\ell_\infty^n$ and random matrices. *Amer. Jour. Math.* **112**, 819-842.

[45] Szarek, S.J.(1991) Condition Numbers of Random Matrices.

[46] Traub, J.F., Woziakowski, H. (1980) *A General Theory of Optimal Algorithms*, Academic Press, New York.

[47] J.A. Tropp (2003) Greed is Good: Algorithmic Results for Sparse Approximation To appear, *IEEE Trans Info. Thry.*

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