Compressed Sensing

David L. Donoho, *Member, IEEE*

***Abstract*—Suppose x is an unknown vector in $\boldsymbol{R}^m$ (a digital image or signal); we plan to measure n general linear functionals of x and then reconstruct. If x is known to be compressible by transform coding with a known transform, and we reconstruct via the nonlinear procedure defined here, the number of measurements n can be dramatically smaller than the size m. 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.**

More specifically, suppose x has a sparse representation in some orthonormal basis (e.g., wavelet, Fourier) or tight frame (e.g., curvelet, Gabor)—so the coefficients belong to an ℓ_p ball for 0oo goo oo o o

N o o o n olg l og *Basis Pursuit* in signal processing. The nonadaptive measurements have the character of "random" linear combinations of basis/frame elements. Our results use the notions of optimal recovery, of n-widths, and information-based complexity. We estimate the Gel'fand n-widths of ℓ_p balls in high-dimensional Euclidean space in the case $0 < p \leq 1$, and give a criterion identifying near-optimal subspaces for Gel'fand n-widths. We show that "most" subspaces are near-optimal, and show that convex optimization (Basis Pursuit) is a near-optimal way to extract information derived from these near-optimal subspaces.

***Index Terms*—Adaptive sampling, almost-spherical sections of Banach spaces, Basis Pursuit, eigenvalues of random matrices, Gel'fand n-widths, information-based complexity, integrated sensing and processing, minimum ℓ_1-norm decomposition, optimal recovery, Quotient-of-a-Subspace theorem, sparse solution of linear equations.**

I. 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 we not just **directly measure** the part that will not 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.

A. 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 $x \in \boldsymbol{R}^m$, 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 $\boldsymbol{R}^m$ 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 $R > 0$

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

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 [1]. To fix ideas, we pixel sam-

pics 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 $\left(\psi_i^j\right)$ denote the orthonormal basis of wavelets at scale j, containing $3 \cdot 4^j$ elements. The corresponding coefficients obey $||\theta^{(j)}||_1 \leq 4R$.

Manuscript received September 18, 2004; revised December 15, 2005.

The author is with the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: donoho@stanford.edu).

Communicated by A. Høst-Madsen, Associate Editor for Detection and Estimation.

Digital Object Identifier 10.1109/TIT.2006.871582

- *Bump Algebra model for spectra*. Here a spectrum (e.g., mass spectrum or magnetic resonance spectrum) is modeled 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 spectrum at scale j, and let (ψ_i^j) denote the orthonormal basis of wavelets at scale j, containing 2^j elements. The corresponding coefficients again obey $||\theta^{(j)}||_1 \le c \cdot R \cdot 2^{-j/2}$, [2].

While in these two examples, the ℓ_1 constraint appeared, other ℓ_p constraints with $0 \le p < 1$ can appear naturally as well; see below. For some readers, the use of ℓ_p norms with $p < 1$ may seem initially strange; it is now well understood that the ℓ_p norms with such small p are natural mathematical measures of sparsity [3], [4]. As p decreases below 1, more and more sparsity is being required. Also, from this viewpoint, an ℓ_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 ℓ_p constraint. In practice, in the following we stick with the view that the object θ of interest is a coefficient vector obeying the constraint (I.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 ℓ_p constraint is sparsity of the transform coefficients. Indeed, we have trivially that, if θ_N denotes the vector θ with everything except the N largest coefficients set to 0

$$||\theta - \theta_N||_2 \le \zeta_{2,p} \cdot ||\theta||_p \cdot (N+1)^{1/2-1/p} \qquad \text{(I.2)}$$

for $N = 0, 1, 2, \ldots$, with a constant $\zeta_{2,p}$ depending only on $p \in (0, 2)$. Thus, for example, to approximate θ with error ϵ, we need to keep only the $N \asymp \epsilon^{(p-2)/2p}$ biggest terms in θ.

B. Optimal Recovery/Information-Based Complexity Background

Our question now becomes: if x is an unknown signal whose transform coefficient vector θ obeys (I.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* [5] and *Information-Based Complexity* [6]; 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 $I_n : X \mapsto \boldsymbol{R}^n$ that samples n pieces of information about x, and an algorithm $A_n : \boldsymbol{R}^n \mapsto \boldsymbol{R}^m$ 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 ξ_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 ℓ_2 error of reconstruction and in the behavior of optimal information and optimal algorithms. Hence, we consider the minimax ℓ_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 linear 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 (I.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.

C. Four Surprises

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

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

$$E_n(X_{p,m}(R)) \le C_p \cdot R \cdot (n/\log(m_n))^{1/2-1/p}. \qquad \text{(I.3)}$$

We find this surprising in four ways. First, compare (I.3) with (I.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 characterizing x by a few coefficients [3], [7]. Surely then, one imagines, the sampling kernels ξ_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 (VI.1) below. (*Holography* is a process where a three–dimensional (3-D) image generates by interferometry a two–dimensional (2-D) transform. Each value in the 2-D transform domain is influenced by each part of the whole 3-D object. The 3-D object can later be reconstructed by interferometry from all or even a part of the 2-D transform domain. Leaving aside the specifics of 2-D/3-D and the process of interferometry, we perceive an analogy here, in which an object is transformed to a compressed domain, and each compressed domain component is a combination of all parts of the original object.)

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

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

we have the following.

Theorem 2: For $0 < p \leq 1$, there is $C_p > 0$ so that for $R > 0$

$$E_n(X_{p,m}(R)) \leq 2^{1/p} \cdot E_n^{\mathrm{Adapt}}(X_{p,m}(R)).$$

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.

D. 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 ℓ_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 star-shaped.

To develop this geometric viewpoint further, we consider two notions of n-width; see [5].

Definition 1.1: The **Gel'fand n-width** of X with respect to the ℓ_2^m norm is defined as

$$d^n(X;\ell_2) = \inf_{V_n} \sup \left\{ \|x\|_2 : x \in V_n^{\perp} \cap X \right\}$$

where the infimum is over n-dimensional linear subspaces of $\boldsymbol{R}^m$, 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 [5], and in the present setting extending as follows.

Theorem 3: For $0 < p \leq 1$ and $R > 0$

$$d^n(X_{p,m}(R)) \leq E_n(X_{p,m}(R)) \tag{I.4}$$

$$\leq 2^{1/p-1} \cdot d^n(X_{p,m}(R)). \tag{I.5}$$

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 (I.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 [8]; more commonly, one sees results about the Kolmogorov n-widths.

Definition 1.2: Let $X \subset \boldsymbol{R}^m$ be a bounded set. The **Kolmogorov n-width** of X with respect the ℓ_2 norm is defined as

$$d_n(X;\ell_2) = \inf_{V_n} \sup_{x\in X} \inf_{y\in V_n} \|x - y\|_2$$

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

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 ℓ_q rather than ℓ_2 norm. Also, for given $p \geq 1$ and $q \geq 1$, let p' and q' be the standard dual indices $1/p' = 1 - 1/p, 1/q' = 1 - 1/q$. Also, let $b_{p,m}$ denote the standard unit ball of ℓ_p^m. Then [8]

$$d_n(b_{p,m};\ell_q) = d^n\left(b_{q',m}, \ell_{p'}\right). \tag{I.6}$$

In particular

$$d_n(b_{2,m};\ell_\infty) = d^n(b_{1,m},\ell_2).$$

The asymptotic properties of the left-hand side have been determined by Garnaev and Gluskin [9]. This follows major work by Kashin [10], 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 [8] for more details.

Theorem 4: (**Kashin, Garnaev, and Gluskin** (KGG)) For all n and $m > n$

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

Theorem 1 now follows in the case $p = 1$ by applying KGG with the duality formula (I.6) and the equivalence formula (I.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.

E. Mysteries ...

Because of the indirect manner by which the KGG result implies Theorem 1, we really do not 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 null spaces of certain matrices with entries ± 1 which are known to exist by counting the number of matrices lacking certain properties, the total number of matrices with ± 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 $y_n = I_n(x_0)$, 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

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

and it obeys, when the information I_n is optimal,

$$\sup_{x_0 \in X_{p,m}(R)} \|x_0 - \hat{x}_n^*\|_2 = E_n(X_{p,m}(R));$$

see Section III below.

This abstract viewpoint unfortunately does not translate into a practical approach (at least in the case of the $X_{p,m}(R)$, $0 < p < 1$). 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.

F. 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 three 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 $\boldsymbol{R}^m$ 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 ℓ_1 norm of the coefficients θ 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 nonconvex case $0 < p < 1$.

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

- *CS Information.* Let Φ be an $n \times m$ matrix generated by randomly sampling the columns, with different columns independent and identically distributed (i.i.d.) random uniform on $\boldsymbol{S}^{n-1}$. With overwhelming probability for large n, Φ has properties CS1–CS3 discussed in detail in Section II-A below; assume we have achieved such a favorable draw. Let Ψ be the $m \times m$ basis matrix with basis vector ψ_i as the ith column. The CS Information operator I_n^{CS} is the $n \times m$ matrix $\Phi\Psi^T$.
- *ℓ_1-minimization.* To reconstruct from CS Information, we solve the convex optimization problem

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

In words, we look for the object $\hat{x}_1$ having coefficients with smallest ℓ_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 η_i and ρ (see Section II 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:

$$E_n\left(I_n^{\mathrm{CS}}, X_{p,m}(R)\right) \leq C \cdot E_n(X_{p,m}(R))$$

for $R > 0$, $n > n_0$. Moreover, the algorithm $A_{1,n}$ delivering the solution to (L_1) is near-optimal:

$$E_n\left(A_{1,n}, I_n^{\mathrm{CS}}, X_{p,m}(R)\right) \le C \cdot E_n(X_{p,m}(R))$$

for $R > 0$, $n > n_0$.

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

G. 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 $n = O(m^{1/3}\log(m))$ measurements to get an accurate reconstruction of such spectra, rather than the nominal m measurements. However, they must then be processed nonlinearly.

Recall 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 $n = O(\sqrt{m}\log^3(m))$ 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 brain scans), the number of measurements for an m-pixel image is only $O(m^{1/4}\log^{5/2}(m))$.

H. Contents

Section II introduces a set of conditions CS1–CS3 for near-optimality of an information operator. Section III considers abstract near-optimal algorithms, and proves Theorems 1–3. Section IV shows that solving the convex optimization problem (L_1) gives a near-optimal algorithm whenever $0 < p \le 1$. Section V points out immediate extensions to weak-ℓ_p conditions and to tight frames. Section VI sketches potential implications in image, signal, and array processing. Section VII, building on work in [11], shows that conditions CS1–CS3 are satisfied for "most" information operators.

Finally, in Section VIII, we note the ongoing work by two groups (Gilbert *et al.* [12] and Candès *et al.* [13], [14]), which although not written in the n-widths/OR/IBC tradition, imply (as we explain), closely related results.

II. Information

Consider information operators constructed as follows. With Ψ the orthogonal matrix whose columns are the basis elements ψ_i, and with certain n-by-m matrices Φ obeying conditions specified below, we construct corresponding information operators $I_n = \Phi\Psi^T$. Everything will be completely transparent to the choice of orthogonal matrix Ψ and hence we will assume that Ψ 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. Let $ker(\Phi)$ denote as usual the nullspace $\{x : \Phi x = 0\}$. We define the *width* of a set X relative to an operator Φ

$$w(\Phi, X) \equiv \sup \|x\|_2 \text{ subject to } x \in X \cap ker(\Phi). \qquad \text{(II.1)}$$

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

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

We will show for all large n and m the existence of n by m matrices Φ where

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

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

A. Conditions CS1–CS3

In the following, with $J \subset \{1, \ldots, m\}$ let Φ_J denote a submatrix of Φ obtained by selecting just the indicated columns of Φ. We let V_J denote the range of Φ_J in $\boldsymbol{R}^n$. Finally, we consider a family of *quotient norms* on $\boldsymbol{R}^n$; with $\ell_1(J^c)$ denoting the ℓ_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 ℓ_1-norm representation of v achievable using only specified subsets of columns of Φ.

We define three conditions to impose on an $n \times m$ matrix Φ, indexed by strictly positive parameters η_i, $1 \le i \le 3$, and ρ.

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

CS2: On each subspace V_J we have the inequality

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

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

CS3: On each subspace V_J

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

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 ℓ_1 and ℓ_2 norms is concerned. It will be implied by a geometric fact: every V_J slices through the ℓ_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 smaller than the simple ℓ_1 norm on $\boldsymbol{R}^n$.

It turns out that matrices satisfying these conditions are ubiquitous for large n and m when we choose the η_i and ρ 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 η_1 very small and η_2, η_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 η_i and ρ independent of n and of $m \le An^\gamma$.

Consider the set

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

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 γ so that, for each $\delta > 0$ the proportion of all $n \times m$ matrices Φ satisfying CS1–CS3 with parameters (η_i) and ρ eventually exceeds $1 - \delta$.

We will discuss and prove this in Section VII. The proof will show that the proportion of matrices *not* satisfying the condition decays exponentially fast in n.

For later use, we will leave the constants η_i and ρ 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 η_i and ρ not depending on n and permitting the above ubiquity.

B. 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 Φ_{n,m_n} obeying the conditions CS1–CS3 with η_i and ρ 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 for $n > n_0$

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

Proof: Consider the optimization problem

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

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 θ so that $0 = \Phi\theta$. Let J denote the indices of the $k = \lfloor \rho n/\log(m) \rfloor$ largest values in θ. 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

$$\|\theta_{J^c}\|_p \le \|\theta\|_p \le 1 \tag{II.2}$$

while, because each entry in θ_J is at least as big as any entry in θ_{J^c}, (I.2) gives

$$\|\theta_{J^c}\|_2 \le \zeta_{2,p} \cdot (k+1)^{1/2-1/p}. \tag{II.3}$$

A similar argument for ℓ_1 approximation gives, in case $p < 1$

$$\|\theta_{J^c}\|_1 \le \zeta_{1,p} \cdot (k+1)^{1-1/p}. \tag{II.4}$$

Now $0 = \Phi_J\theta_J + \Phi_{J^c}\theta_{J^c}$. 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

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

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 $c_2 = c_1/\eta_1$

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

The theorem follows with

$$C = \left(\zeta_{1,p}\rho^{1-1/p}/(\eta_1\eta_2\eta_3) + \zeta_{2,p}\rho^{1/2-1/p}\right). \qquad \square$$

III. 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.

A. 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 $y_n = I_n(x_0)$, the collection of all objects $x \in X_{p,m}(R)$ which could have given rise to the information y_n

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

In the feasible-point method, we simply select any member of $\hat{X}_{p,R}(y_n)$, by whatever means. One can show, adapting standard OR/IBC arguments in [15], [6], [8] the following.

Lemma 3.1: Let $y_n = I_n(x_0)$ where $x_0 \in X_{p,m}(R)$ and I_n is an optimal information operator, and let $\hat{x}$ be any element of $\hat{X}_{p,R}(y_n)$. Then for $0 < p \leq 1$

$$\|x_0 - \hat{x}\|_2 \leq 2 \cdot E_n(X_{p,m}(R)). \tag{III.1}$$

In short, any feasible point is within a factor two of optimal.

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

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

Now clearly, in the special case when x_0 is only known to lie in $X_{p,m}(R)$ and $y_n = I_n(x_0)$ 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 such 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 feasible point obeys $\hat{x} \in X_{p,m}(R)$; hence,

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

But the triangle inequality gives

$$\|x_0 - \hat{x}\|_2 \leq \|x_0 - \hat{x}_n^*(y_n)\|_2 + \|\hat{x}_n^*(y_n) - \hat{x}\|_2$$

hence, as $x_0 \in \hat{X}_{p,R}(y_n)$

$$\begin{aligned} \|x_0 - \hat{x}\|_2 &\leq 2 \cdot radius\,(\hat{X}_{p,R}(y_n)) \\ &\leq 2 \cdot \sup\{radius\,(\hat{X}_{p,R}(y_n)) \\ &\qquad : y_n \in I_n(X_{p,m}(R))\} \\ &= 2 \cdot E_n(X_{p,m}(R)). \end{aligned}$$

□

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

$$\|x_0 - \hat{x}\|_2 \leq 2 \cdot E_n(I_n, X_{p,m}(R)). \tag{III.2}$$

A popular choice of feasible point is to take an element of *least norm*, i.e., a solution of the problem

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

where here $\theta(x) = \Psi^T x$ is the vector of transform coefficients, $\theta \in \ell_p^m$. 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. For later use, call the solution $\hat{x}_{p,n}$. By the preceding lemma, this procedure is near-minimax: with $y_n = I_n(x_0)$ where $x_0 \in X_{p,m}(R)$ for given $0 < p \leq 1$ and $R > 0$

$$\|x_0 - \hat{x}_{p,n}\|_2 \leq 2 \cdot E_n(X_{p,m}(R)).$$

B. 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 [17] 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_+ and x_- attain the radius bound, i.e., they satisfy $I_n(x_\pm) = y_n$ and, for $c = center\,(\hat{X}_{p,R}(y_n))$ they satisfy

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

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

$$\|\theta_+ - \theta_-\|_p^p \leq \|\theta_+\|_p^p + \|\theta_-\|_p^p$$

and so

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

Hence $\delta \in X_{p,m}(R)$. However,

$$I_n(\delta) = I_n((x_+ - x_-)/2^{1/p}) = 0$$

so δ belongs to $X_{p,m}(R) \cap V_n^\perp$. Hence, $\|\delta\|_2 \leq d^n(X_{p,m}(R))$ and

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

□

C. 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 ϵ such that an ϵ-net for X can be built using a net of cardinality at most 2^n. From Carl's theorem [18]—see the exposition in Pisier's book [19]—there is a constant $c > 0$ so that the Gel'fand n-widths dominate the entropy numbers.

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

Secondly, the entropy numbers obey [20], [21]

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

At the same time, the combination of Theorems 7 and 6 shows that

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

Applying now the Feasible Point method, we have

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

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.

D. Proof of Theorem 2

Now is an opportune time to prove Theorem 2. We note that in the case of $1 \leq p$, this is known [22]–[25]. 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 $radius\left(V_n^0 \cap X_{p,m}(R)\right)$; 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.

IV. 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 $0 < p \leq 1$. 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.

A. The Case $p = 1$

In the case $p = 1$, (P_1) is a convex optimization problem. Written in an equivalent form, with θ being the optimization variable, gives

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

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

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

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 $\hat{x}_{1,n}$ is $\Psi\theta^*$. 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 [26]; 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 (III.2) we have the following.

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

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

then the Basis Pursuit algorithm $A_{1,n}(y_n) = \hat{x}_{1,n}$ achieves, for all $R > 0$

$$E_n(I_n, A_{1,n}, X_{1,m}(R)) \leq 2C \cdot E_n(X_{1,m}(R)).$$

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

$$\|x_0 - \hat{x}_{1,n}\|_2 \leq c_1 \cdot \|\theta\|_1 \cdot (n/\log m)^{-1/2}$$

for c_1 a constant depending at most on $\log(m)/\log(n)$. The inequality can be put to use as follows. Fix $\epsilon > 0$. Suppose the unknown object x_0 is known to be highly compressible, say obeying the *a priori* bound $\|\theta\|_1 \leq c_2 m^{\alpha}$, $\alpha < 1/2$. Let $K_\epsilon = (c_1 c_2/\epsilon)^2$. For any such object, rather than making m measurements, we only need to make $n \sim K_\epsilon \cdot m^{2\alpha} \log(m)$ measurements, and our reconstruction obeys

$$\|x_0 - \hat{x}_{1,n}\|_2 \leq \epsilon \cdot \|x_0\|_2.$$

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, offering more impressive performance in Theorem 1, because the exponent $1/2 - 1/p$ is even stronger than in the $p = 1$ case. Later in this section, we extend the same interpretation of $\hat{x}_{1,n}$ to performance over $X_{p,m}(R)$ throughout $p < 1$.

B. Relation Between ℓ_1 and ℓ_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 ℓ_1 and the extreme case $p \to 0$ of the ℓ_p spaces. Let us define

$$(P_0) \quad \min \|\theta\|_0 \text{ subject to } \Phi\theta = y$$

where $\|\theta\|_0$ is just the number of nonzeros in θ. Again, since the work of Peetre and Sparr [16], the importance of ℓ_0 and the relation with ℓ_p for $0 < p < 1$ is well understood; see [17] for more detail.

Ordinarily, solving such a problem involving the ℓ_0 norm requires combinatorial optimization; one enumerates all sparse subsets of $\{1, \ldots, m\}$ searching for one which allows a solution $\Phi\theta = y$. However, when (P_0) has a sparse solution, (P_1) will find it.

Theorem 8: Suppose that Φ satisfies CS1–CS3 with given positive constants $\rho, (\eta_i)$. There is a constant $\rho_0 > 0$ depending only on ρ and (η_i) and not on n or m so that, if some solution to $y = \Phi\theta$ 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 underdetermined, ℓ_1 minimization and sparse solution coincide—when the result is sufficiently sparse.

There is by now an extensive literature exhibiting results on equivalence of ℓ_1 and ℓ_0 minimization [27]–[34]. In the early 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 [13] 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 [11], 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 [11], shows that the nullspace $\Phi\beta = 0$ has a very special structure for Φ obeying the conditions in question. When θ is sparse, the *only* element in a given affine subspace $\theta + ker(\Phi)$ which can have small ℓ_1 norm is θ itself.

To prove Theorem 8, we first need a lemma about the nonsparsity of elements in the nullspace of Φ. 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**

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

This measures the fraction of ℓ_1 norm which can be concentrated on a certain subset for a vector in the nullspace of Φ. This concentration cannot be large if $|J|$ is small.

Lemma 4.1: Suppose that Φ satisfies CS1–CS3, with constants η_i and ρ. There is a constant η_0 depending on the η_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 $\beta \in ker(\Phi)$. Assume without loss of generality that J is the most concentrated subset of cardinality $|J| < \rho n/\log(m)$, and that the columns of Φ are numbered so that $J = \{1, \ldots, |J|\}$; partition $\beta = [$ $\}$

Lemma 4.2: Let θ_0 be a vector in R^m and $1_J\theta_0$ be the corresponding mutilated vector with entries $\theta_0(i)1_J(i)$. Suppose that

$$||\theta_0 - 1_J\theta_0||_1 \le \epsilon$$

where $\nu(\Phi, J) \le \nu_0 < 1/2$. Consider the instance of (P_1) defined by $y = \Phi\theta_0$; the solution $\hat{\theta}_1$ of this instance of (P_1) obeys

$$||\theta_0 - \hat{\theta}_1||_1 \le \frac{2\epsilon}{1-2\nu_0}. \tag{IV.2}$$

Proof of Lemma 4.2: Put for short $\hat{\theta} \equiv \hat{\theta}_1$, and set $\beta = \theta_0 - \hat{\theta} \in ker(\Phi)$. By definition of ν

$$||\theta_0 - \hat{\theta}||_1 = ||\beta||_1 \le ||\beta_{J^c}||_1/(1-\nu_0)$$

while

$$||\beta_{J^c}||_1 \le ||\theta_{0,J^c}||_1 + ||\hat{\theta}_{J^c}||_1.$$

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

$$||\hat{\theta}_J||_1 + ||\hat{\theta}_{J^c}||_1 \le ||\theta_0||_1$$

and of course

$$||\theta_{0,J}||_1 - ||\hat{\theta}_J||_1 \le ||1_J(\theta_0 - \hat{\theta})||_1.$$

Hence,

$$||\hat{\theta}_{J^c}||_1 \le ||1_J(\theta_0 - \hat{\theta})||_1 + ||\theta_{0,J^c}||_1.$$

Finally

$$||1_J(\theta_0 - \hat{\theta})||_1 = ||\beta_J||_1 \le \nu_0 \cdot ||\beta||_1 = \nu_0 \cdot ||\theta_0 - \hat{\theta}||_1.$$

Combining the above, setting $\delta \equiv ||\theta_0 - \hat{\theta}||_1$, and $\epsilon \equiv ||\theta_{0,J^c}||_1$, we get

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

and (IV.2) follows. □

Proof of Theorem 9: We use the same general framework as in Theorem 7. Let $y = \Phi\theta_0$ where $||\theta_0||_p \le R$. Let $\hat{\theta}$ be the solution to (P_1), and set $\beta = \hat{\theta} - \theta_0 \in ker(\Phi)$.

Let η_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 θ_0. From $||\theta_0||_p \le R$ and (II.4) we have

$$||\theta_0 - 1_J\theta_0||_1 \le \xi_{1,p} \cdot R \cdot |J|^{1-1/p}$$

and Lemma 4.1 provides

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

Applying Lemma 4.2

$$||\beta||_1 \le c \cdot R \cdot |J|^{1-1/p}. \tag{IV.3}$$

The vector $\delta = \beta/||\beta||_1$ lies in $ker(\Phi)$ 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 (IV.3),

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

V. Immediate Extensions

Before continuing, we mention two immediate extensions to the results so far; they are of interest below and elsewhere.

A. Tight Frames

Our main results so far have been stated in the context of (ϕ_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 Ψ $(m < m')$ 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 θ involves a tight frame. In particular, if Φ 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 $\hat{x}_1$.

A referee remarked that there is no need to restrict attention to tight frames here; if we have a general frame the same results go through, with constants involving the frame bounds. This is true and potentially very useful, although we will not use it in what follows.

B. Weak ℓ_p Balls

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

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

Conversely, for a given θ, 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 $||\theta||_{w\ell_p} \le ||\theta||_p$. Weak ℓ_p constraints have the following key property: if θ_N denotes a mutilated version of the vector θ with all except the N largest items set to zero, then the inequality

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

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

VI. Stylized Applications

We sketch three potential applications of the above abstract theory. In each case, we exhibit that a certain class of functions

has expansion coefficients in a basis or frame that obey a particular ℓ_p or weak ℓ_p embedding, and then apply the above abstract theory.

A. 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 $\mathcal{B}$ [2], with bump norm $\|f\|_{\mathcal{B}} \leq B$. This was mentioned in the Introduction, which observed 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 later we use standard wavelet analysis notations as in [36], [37], [2].

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 φ a smooth function integrating to 1. Think of these as point samples at scale 2^{-j_1} after applying an antialiasing 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 ψ is an oscillating function with mean zero. We pick a coarsest scale $j_0 = j_1/3$. 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 $\big((\alpha_{j,k} : 0 \leq k < 2^j) : j_0 \leq j < j_1\big)$. The dimension m of θ is $m = 2^{j_1} - 2^{j_0}$. The norm $\|\theta\|_1$ satisfies

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

This establishes the constraint on ℓ_1 norm needed for our theory. 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 later). We apply the near-optimal algorithm of ℓ_1 minimization, getting the error estimate

$$\begin{aligned}\|\hat{\theta} - \theta\|_2 &\leq c\|\theta\|_1 \cdot (n/\log(m))^{-1/2} \\ &\leq c2^{-\frac{3}{2}j_0} \leq c2^{-j_1/2}\end{aligned}$$

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

$$\begin{aligned}\|f - \hat{f}\|_2 &\leq \|f - P_{j_1} f\|_2 + \|P_{j_1} f - \hat{f}\|_2 \\ &= \|f - P_{j_1} f\|_2 + \|\hat{\theta} - \theta\|_2 \\ &\leq c2^{-j_1/2}\end{aligned}$$

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 $n \leq c2^{j_0}\log(2^{j_1})$ samples associated with detail coefficients, for a total $\leq c \cdot j_1 \cdot 2^{j_1/3}$ 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 cube root of the number of samples of linear sampling.

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

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

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

B. 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 [38], and bounded in absolute value by $\|f\|_\infty \leq B$ as well. In the Introduction, it was mentioned that the wavelet coefficients at level j obey $\|\theta^{(j)}\|_1 \leq C \cdot B$ where C depends only on the wavelet used. It is also true that $\|\theta^{(j)}\|_\infty \leq C \cdot B \cdot 2^{-j}$.

We again consider two ways of approximating functions in f. The classic linear scheme uses a 2-D 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 $0 \leq k_1, k_2 < 2^{j_1}$. 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

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

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

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

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 ψ^σ 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 $((\alpha^\sigma_{j,k} : 0 \leq k_1, k_2 < 2^j, \sigma \in \{h, v, d\}) : j_0 \leq j < j_1)$. The dimension m of θ is $m = 4^{j_1} - 4^{j_0}$. The norm $\|\theta\|_1$ obeys

$$\|\theta\|_1 \leq \sum_{j=j_0}^{j_1} \|(\theta^{(j)})\|_1 \leq c(j_1 - j_0 + 1)\|f\|_{\mathcal{BV}}.$$

This establishes the constraint on ℓ_1 norm needed for applying our theory. We take $n = c \cdot 4^{j_0}\log^3(4^{j_1})$ and apply a near-optimal information operator for this n and m. We apply the near-

optimal algorithm of ℓ_1 minimization to the resulting information, getting the error estimate

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

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

$$\begin{aligned}\|f - \hat{f}\|_2 &\le \|f - P_{j_1}f\|_2 + \|P_{j_1}f - \hat{f}\|_2 \\ &= \|f - P_{j_1}f\|_2 + \|\hat{\theta} - \theta\|_2 \\ &\le c2^{-j_1/2}\end{aligned}$$

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 $n \le c4^{j_0}\log^3(4^{j_1})$ samples associated with detail coefficients, for a total $\le c \cdot j_1^3 \cdot 4^{j_1/2}$ 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.

C. 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-ℓ_p mentioned earlier. Again in the image processing setting, we use the C^2-C^2 model discussed in Candès and Donoho [39], [40]. 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 Γ which is a union of C^2 curves having first and second derivatives $\le B$ in an appropriate parametrization; the curves have total length $\le L$. More colorfully, such images are *cartoons*—patches of uniform smooth behavior separated by piecewise-smooth curvilinear boundaries. They might be reasonable models for certain kinds of technical imagery—e.g., in radiology.

The curvelets tight frame [40] 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 ℓ_p constraint

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

compare [40]. 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 Bounded Variation 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 Bounded Variation.

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 $\left(\theta^{(j)} : j_0 \le j < j_1\right)$. The dimension m of θ is $m = c(4^{j_1} - 4^{j_0})$, with $c > 1$ due to overcompleteness of curvelets. The weak $\ell^{2/3}$ "norm" obeys

$$\|\theta\|_{w\ell_{2/3}} \le \left(\sum_{j_0}^{j_1} \|(\theta^{(j)})\|_{w\ell_{2/3}}^{2/3}\right)^{3/2} \le c(j_1 - j_0 + 1)^{3/2}$$

with c depending on B and L. This establishes the constraint on weak ℓ_p norm needed for our theory. 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 ℓ_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

$$\begin{aligned}\|f - \hat{f}\|_2 &\le \|f - P_{j_1}f\|_2 + \|P_{j_1}f - \hat{f}\|_2 \\ &= \|f - P_{j_1}f\|_2 + \|\hat{\theta} - \theta\|_2 \\ &\le c2^{-j_1/2}\end{aligned}$$

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 to 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.

VII. 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 $\boldsymbol{S}^{n-1}$. Then for some $\eta_i > 0$ and $\rho > 0$, conditions CS1–CS3 hold for Φ with overwhelming probability for all large n.

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

In effect matrices Φ 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 Sections VII-A–C; it proceeds by studying events Ω_n^i, $i = 1, 2, 3$, where $\Omega_n^1 \equiv \{\text{CS1 Holds}\}$, etc. It will be shown that for parameters $\eta_i > 0$ and $\rho_i > 0$

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

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

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

Since, when Ω_n occurs, our random draw has produced a matrix obeying CS1–CS3 with parameters η_i and ρ, this proves Theorem 10. The proof actually shows that for some $\beta > 0$, $P\left(\Omega_n^c\right) \le \exp(-\beta n)$, $n > n_0$; the convergence is exponentially fast.

A. Control of Minimal Eigenvalue

The following lemma allows us to choose positive constants ρ_1 and η_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

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

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

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

The proof involves three ideas. The first idea is that the event of interest for Lemma 7.1 is representable in terms of events indexed by individual subsets J

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

Our plan is to bound the probability of occurrence of every $\Omega_{n,J}$.

The second idea is that for a specific subset J, we get large deviations bounds on the minimum eigenvalue; this can be stated as follows.

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

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

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

This was derived in [11] and in [35], using the concentration of measure property of singular values of random matrices, e.g., see Szarek's work [41], [42].

The third and final idea is that bounds for individual subsets J can control simultaneous behavior over all J. This is expressed as follows.

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

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

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

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

Our main goal of this subsection, Lemma 7.1, now follows by combining these three ideas.

It remains only 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

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

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 $m \ge n \ge 3 > e$. Also, of course

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

so we get $\log(\#\mathcal{J}) \le \rho_1 n$. Taking β_1 as given, we get the desired conclusion. □

B. Spherical Sections Property

We now show that condition CS2 can be made overwhelmingly likely for large n by choice of η_2 and ρ_2 sufficiently small but still positive. Our approach derives from [11], 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 [43], [44]. We develop a quantitative refinement of this principle for the ℓ_1 norm in $\boldsymbol{R}^n$, showing that, with overwhelming probability, every operator Φ_J for $|J| < \rho n/\log(m)$ affords a spherical section of the ℓ_1^n ball. The basic argument we use originates from work of Milman, Kashin, and others [44], [10], [45]; we refine an argument in Pisier [19] and, as in [11], draw inferences that may be novel. We conclude that not only do almost-spherical sections exist, but they are so ubiquitous that every Φ_J with $|J| < \rho n/\log(m)$ will generate them.

Definition 7.1: Let $|J| = k$. We say that Φ_J offers an ϵ-isometry between $\ell_2(J)$ and ℓ_1^n if $\forall\, \alpha \in \boldsymbol{R}^k$

$$(1-\epsilon) \cdot \|\alpha\|_2 \le \sqrt{\frac{\pi}{2n}} \cdot \|\Phi_J \alpha\|_1 \le (1+\epsilon) \cdot \|\alpha\|_2. \quad \text{(VII.1)}$$

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

Lemma 7.4: Consider the event $\Omega_n^2 \left(\equiv \Omega_n^2(\epsilon,\rho)\right)$ that *every* Φ_J with $|J| \le \rho \cdot n/\log(m)$ offers an ϵ-isometry between $\ell_2(J)$ and ℓ_1^n. For each $\epsilon > 0$, there is $\rho(\epsilon) > 0$ so that

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

To prove this, we first need a lemma about individual subsets J proven in [11].

Lemma 7.5: Fix $\epsilon > 0$. Choose δ so that

$$(1-3\delta)(1-\delta)^{-1} \ge (1-\epsilon)^{1/2} \tag{VII.2}$$

and

$$(1+\delta)(1-\delta)^{-1} \le (1+\epsilon)^{1/2}. \tag{VII.3}$$

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 $\{1,\dots,m\}$, let $\Omega_{n,J}$ denote the event that Φ_J furnishes an ϵ-isometry to ℓ_1^n. 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.

C. Quotient Norm Inequalities

We now show that, for $\eta_3 = 3/4$, for sufficiently small $\rho_3 > 0$, nearly all large n by m_n matrices have property CS3. Our argument borrows heavily from [11] which the reader may find helpful. We here make no attempt to provide intuition or to compare with closely related notions in the local theory of Banach spaces (e.g., Milman's Quotient of a Subspace Theorem [19]).

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

$$\sigma(k) = \operatorname{sgn}\left\{\sum_J \alpha_i \phi_i(k)\right\}, \qquad 1 \le k \le n.$$

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

Lemma 7.6: **Uniform Sign-Pattern Embedding.** Fix $\delta > 0$. Then for $\tau < \delta^2/128$, set

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

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

$$\|y - \epsilon_n \sigma\|_2 \le \epsilon_n \cdot \delta \cdot \|\sigma\|_2 \tag{VII.5}$$

and

$$\forall\, i \in J^c, \qquad |\langle \phi_i, y\rangle| \le 1. \tag{VII.6}$$

In words, a small multiple $\epsilon_n\sigma$ of any sign pattern σ 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

$$\|\beta_{J^c}\|_1 \ge \eta_3/\sqrt{\log(m/n)} \cdot \|v\|_1. \tag{VII.7}$$

Consider the convex optimization problem

$$\min \|\beta_{J^c}\|_1 \text{ subject to } \Phi_{J^c}\beta_{J^c} = -v. \tag{VII.8}$$

This can be written as a linear program, by the same sort of construction as given for (IV.1). By the duality theorem for linear programming, the value of the primal program is at least the value of the dual

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

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

$$\begin{aligned}\langle v, y\rangle &\ge \langle v, \epsilon_n\sigma\rangle - \|y - \epsilon_n\sigma\|_2\|v\|_2 \\ &\ge \epsilon_n\|v\|_1 - \epsilon_n\delta\|\sigma\|_2\|v\|_2;\end{aligned}$$

picking δ appropriately and taking into account the spherical sections theorem, for sufficiently large n, we have $\delta\|\sigma\|_2\|v\|_2 \le \frac{1}{4}\|v\|_1$; (VII.7) follows with $\eta_3 = 3/4$.

1) Proof of Uniform Sign-Pattern Embedding: The proof of Lemma 7.6 follows closely a similar result in [11] 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 ϵ, δ, 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 [11]. 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

$$|\langle \phi_i, y\rangle| \le 1, \qquad i = 1,\dots,m-k. \tag{VII.9}$$

Let $(\phi_i)_{i=1}^{m-k}$ be i.i.d. uniform on $\boldsymbol{S}^{n-1}$; there is an event Ω_σ described below, having probability controlled by

$$\operatorname{Prob}(\Omega_\sigma^c) \le 2n\exp\{-n\beta\} \tag{VII.10}$$

for $\beta > 0$ which can be explicitly given in terms of τ and δ. On this event

$$\|y - y_0\|_2 \le \delta \cdot \|y_0\|_2. \tag{VII.11}$$

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 [46]

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

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

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

while also

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

Hence, the total number of sign patterns generated by operators Φ_J obeys

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

Now β 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 Ω_σ in the statement of Lemma 7.7) generated by a specific σ, J combination. On the event Ω_n^3, *every* sign pattern associated with *any* Φ_J obeying $|J| < \rho_3 n/\log(m)$ is almost dual feasible. Now

$$\begin{aligned} P\left(\left(\Omega_n^3\right)^c\right) &\le \#\{\sigma : \sigma \in \Sigma_J, |J| < \rho_3 n/\log(m)\} \\ &\quad \times \exp(-n\beta) \\ &\le \exp\{-n(\beta - 2\rho_3) + \log(n)\} \end{aligned}$$

which tends to zero exponentially rapidly. □

D. Proof of Individual Sign-Pattern Embedding

1) An Embedding Algorithm: The companion paper [11] 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 $\Phi_{I_0}\left(\Phi_{I_0}^T\Phi_{I_0}\right)^{-1}\Phi_{I_0}^T$. 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_ℓ is empty, then the process terminates, and set $y = y_\ell$. Termination must occur at stage $\ell^* \le n$. At termination

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

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

2) Analysis Framework: Also in [11] bounds were developed for two key descriptors of the algorithm trajectory

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

and

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

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

$$\begin{aligned} \delta_{\ell;n} &= \|y_0\|_2 \cdot \omega^\ell, \qquad \ell = 1,2,\ldots \\ \nu_{\ell;n} &= n \cdot \lambda_0 \cdot \epsilon_n^2 \cdot \omega^{2\ell+2}/4, \qquad \ell = 0,1,2,\ldots; \end{aligned}$$

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

$$\begin{aligned} E_\ell = \{&\alpha_j \le \delta_j,\ j = 1,\ldots,\ell, \\ &|I_j| \le \nu_j,\ j = 0,\ldots,\ell-1\}. \end{aligned}$$

Now define

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

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

$$\begin{aligned} \|y - y_0\|_2 &\le \left(\sum \alpha_\ell^2\right)^{1/2} \le \|y_0\|_2 \cdot \omega/(1-\omega^2)^{1/2} \\ &\le \|y_0\|_2 \cdot \delta. \end{aligned}$$

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

$$P\left(E_{\ell+1}^c | E_\ell\right) \le 2\exp\{-\beta n\}. \tag{VII.12}$$

This implies

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

and the lemma follows. □

3) Large Deviations: Define the events

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

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_ℓ is defined in terms of ω and τ. On the event E_ℓ, $|J_\ell| \le \rho_0 n/\log(m)$. Lemma 7.1 implicitly defined a quantity $\lambda_1(\rho, A, \gamma)$ lowerbounding the minimum eigenvalue of

every $\Phi_J^T \Phi_J$ where $|J| \le \rho m/\log(m)$. Pick $\rho_{1/2} > 0$ so that $\lambda_1(\rho_{1/2}, A, \gamma) > 1/2$. Pick ω_0 so that

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

With this choice of ω_0, when the event E_ℓ occurs,

$$\lambda_{min}\left(\Phi_{I_\ell}^T \Phi_{I_\ell}\right) > \lambda_0.$$

Also, on E_ℓ, $u_j = 2^{-j-1}/\alpha_j > 2^{-j-1}/\delta_j = v_j$ (say) for $j \le \ell$.

In [11], an analysis framework was developed in which a family $\left(Z_i^\ell : 1 \le i \le m, 0 \le \ell < n\right)$ of random variables i.i.d. $N\left(0, \frac{1}{n}\right)$ was introduced, and it was shown that

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

and

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

That paper also stated two simple large deviations bounds.

Lemma 7.8: Let Z_i be i.i.d. $N(0,1)$, $k \ge 0$, $t > 2$

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

and

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

Applying this, we note that the event

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

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

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

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

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

We therefore have for $n > n_0$ the inequality

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

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

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

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 Δ_ℓ in fact does not depend on ℓ. Focusing now on $\ell = 0$, we may write

$$\log P\left\{F_1^c | G_0\right\} \le m\left(\tau \cdot n/m - n/m \cdot \omega^2/32\right) = n\left(\tau - \omega^2/32\right).$$

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

$$\beta \equiv \beta(\tau; \delta) = (\delta/2)^2/32 - \tau$$

we get $\beta > 0$ for $\tau < \delta^2/128$, and

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

A similar analysis holds for the G_ℓ's. □

VIII. C

A1: The maximal concentration $\nu(\Phi, J)$ (defined in Section IV-B) obeys

$$\nu(\Phi, J) < \eta_1, \quad |J| < \rho_1 n/\log(m_n). \qquad \text{(VIII.1)}$$

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

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

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 Φ with good widths over ℓ_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 [11], [35]; the second paper proves results which partially overlap those in the first, by using a different technique.

C. 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 [13]. This was discussed in Section IV-B; the result of [13] showed that ℓ_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.* [12], which showed that a different nonlinear reconstruction algorithm can be used to recover approximations to a vector in $\boldsymbol{R}^m$ which is nearly as good as the best N-term approximation in ℓ_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 θ.

These papers 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 Φ working within the CS framework; that is, generating near-optimal subspaces for Gel'fand n-widths, and allowing ℓ_1 minimization to reconstruct from such information for all $0 < p \le 1$.

Now [13] (in effect) proves that if $m_n \sim An^\gamma$, then in the partial Fourier ensemble with uniform measure, the maximal concentration condition A1 (VIII.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 [12] seem to show that condition A2 (VIII.2) holds in the partial Fourier ensemble with overwhelming probability for large n, when it is sampled with a certain nonuniform probability measure. Although the two papers [13], [12] 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, Candès has informed us of some recent results he obtained with Tao [47] indicating that, modulo polylog factors, A2 holds for the uniformly sampled partial Fourier ensemble. This seems a very significant advance.

Note Added in Proof

In the months since the paper was written, several groups have conducted numerical experiments on synthetic and real data for the method described here and related methods. They have explored applicability to important sensor problems, and studied applications issues such as stability in the presence of noise. The reader may wish to consult the forthcoming Special Issue on Sparse Representation of the EURASIP journal *Applied Signal Processing*, or look for papers presented at a special session in ICASSP 2005, or the workshop on sparse representation held in May 2005 at the University of Maryland Center for Scientific Computing and Applied Mathematics, or the workshop in November 2005 at Spars05, Université de Rennes.

A referee has pointed out that Compressed Sensing is in some respects similar to problems arising in data stream processing, where one wants to learn basic properties [e.g., moments, histogram] of a datastream without storing the stream. In short, one wants to make relatively few measurements while inferring relatively much detail. The notions of "Iceberg queries" in large databases and "heavy hitters" in data streams may provide points of entry into that literature.

ACKNOWLEDGMENT

In spring 2004, Emmanuel Candès told the present author about his ideas for using the partial Fourier ensemble in "undersampled imaging"; some of these were published in [13]; see also the presentation [14]. More recently, Candès informed the present author of the results in [47] referred to above. It is a pleasure to acknowledge the inspiring nature of these conversations. The author would also like to thank Anna Gilbert for telling him about her work [12] finding the B-best Fourier coefficients by nonadaptive sampling, and to thank Emmanuel Candès for conversations clarifying Gilbert's work. Thanks to the referees for numerous suggestions which helped to clarify the exposition and argumentation. Anna Gilbert offered helpful pointers to the data stream processing literature.

REFERENCES

[1] D. L. Donoho, M. Vetterli, R. A. DeVore, and I. C. Daubechies, "Data compression and harmonic analysis," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2435–2476, Oct. 1998.
[2] Y. Meyer, *Wavelets and Operators*. Cambridge, U.K.: Cambridge Univ. Press, 1993.
[3] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation," *Appl. Comput. Harmonic Anal.*, vol. 1, pp. 100–115, 1993.
[4] ——, "Sparse components of images and optimal atomic decomposition," *Constructive Approx.*, vol. 17, pp. 353–382, 2001.
[5] A. Pinkus, "n-widths and optimal recovery in approximation theory," in *Proc. Symp. Applied Mathematics*, vol. 36, C. de Boor, Ed., Providence, RI, 1986, pp. 51–66.
[6] J. F. Traub and H. Woziakowski, *A General Theory of Optimal Algorithms*. New York: Academic, 1980.
[7] D. L. Donoho, "Unconditional bases and bit-level compression," *Appl. Comput. Harmonic Anal.*, vol. 3, pp. 388–92, 1996.
[8] A. Pinkus, *n-Widths in Approximation Theory*. New York: Springer-Verlag, 1985.
[9] A. Y. Garnaev and E. D. Gluskin, "On widths of the Euclidean ball" (in English), *Sov. Math.–Dokl.*, vol. 30, pp. 200–203, 1984.
[10] B. S. Kashin, "Diameters of certain finite-dimensional sets in classes of smooth functions," *Izv. Akad. Nauk SSSR, Ser. Mat.*, vol. 41, no. 2, pp. 334–351, 1977.

[11] D. L. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ_1-norm solution is also the sparsest solution," *Commun. Pure Appl. Math.*, to be published.
[12] A. C. Gilbert, S. Guha, P. Indyk, S. Muthukrishnan, and M. Strauss, "Near-optimal sparse fourier representations via sampling," in *Proc 34th ACM Symp. Theory of Computing*, Montréal, QC, Canada, May 2002, pp. 152–161.
[13] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.," *IEEE Trans. Inf. Theory*, to be published.
[14] E. J. Candès, "Robust Uncertainty Principles and Signal Recovery," presented at the 2nd Int. Conf. Computational Harmonic Anaysis, Nashville, TN, May 2004.
[15] C. A. Micchelli and T. J. Rivlin, "A survey of optimal recovery," in *Optimal Estimation in Approximation Theory*, C. A. Micchelli and T. J. Rivlin, Eds: Plenum, 1977, pp. 1–54.
[16] J. Peetre and G. Sparr, "Interpolation of normed abelian groups," *Ann. Math. Pure Appl.*, ser. 4, vol. 92, pp. 217–262, 1972.
[17] J. Bergh and J. Löfström, *Interpolation Spaces. An Introduction*. Berlin, Germany: Springer-Verlag, 1976.
[18] B. Carl, "Entropy numbers s-numbers, and eigenvalue problems," *J. Funct. Anal.*, vol. 41, pp. 290–306, 1981.
[19] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge, U.K.: Cambridge Univ. Press, 1989.
[20] C. Schütt, "Entropy numbers of diagonal operators between symmetric Banach spaces," *J. Approx. Theory*, vol. 40, pp. 121–128, 1984.
[21] T. Kuhn, "A lower estimate for entropy numbers," *J. Approx. Theory*, vol. 110, pp. 120–124, 2001.
[22] S. Gal and C. Micchelli, "Optimal sequential and nonsequential procedures for evaluating a functional," *Appl. Anal.*, vol. 10, pp. 105–120, 1980.
[23] A. A. Melkman and C. A. Micchelli, "Optimal estimation of linear operators from inaccurate data," *SIAM J. Numer. Anal.*, vol. 16, pp. 87–105, 1979.
[24] M. A. Kon and E. Novak, "The adaption problem for approximating linear operators," *Bull. Amer. Math. Soc.*, vol. 23, pp. 159–165, 1990.
[25] E. Novak, "On the power of adaption," *J. Complexity*, vol. 12, pp. 199–237, 1996.
[26] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci Comp.*, vol. 20, no. 1, pp. 33–61, 1999.
[27] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–62, Nov. 2001.
[28] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representations in pairs of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 9, pp. 2558–2567, Sep. 2002.
[29] D. L. Donoho and M. Elad, "Optimally sparse representation from overcomplete dictionaries via ℓ_1 norm minimization," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2002, Mar. 2002.
[30] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans Inf Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.
[31] J. J. Fuchs, "On sparse representation in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, Jun. 2002.
[32] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[33] ——, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
[34] D. L. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
[35] D. L. Donoho, "For most underdetermined systems of linear equations, the minimal ℓ_1 -norm near-solution approximates the sparsest near-solution," *Commun. Pure Appl. Math.*, to be published.
[36] I. C. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.
[37] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic, 1998.
[38] A. Cohen, R. DeVore, P. Petrushev, and H. Xu, "Nonlinear approximation and the space $BV(R^2)$," *Amer. J. Math.*, vol. 121, pp. 587–628, 1999.
[39] E. J. Candès and D. L. Donoho, "Curvelets—A surprisingly effective nonadaptive representation for objects with edges," in *Curves and Surfaces*, C. Rabut, A. Cohen, and L. L. Schumaker, Eds. Nashville, TN: Vanderbilt Univ. Press, 2000.
[40] ——, "New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities," *Comm. Pure Appl. Math.*, vol. LVII, pp. 219–266, 2004.
[41] S. J. Szarek, "Spaces with large distances to ℓ_∞^n and random matrices," *Amer. J. Math.*, vol. 112, pp. 819–842, 1990.
[42] ——, "Condition numbers of random matrices," *J. Complexity*, vol. 7, pp. 131–149, 1991.
[43] A. Dvoretsky, "Some results on convex bodies and banach spaces," in *Proc. Symp. Linear Spaces*, Jerusalem, Israel, 1961, pp. 123–160.
[44] T. Figiel, J. Lindenstrauss, and V. D. Milman, "The dimension of almost-spherical sections of convex bodies," *Acta Math.*, vol. 139, pp. 53–94, 1977.
[45] V. D. Milman and G. Schechtman, *Asymptotic Theory of Finite-Dimensional Normed Spaces (Leecture Notes in Mathematics)*. Berlin, Germany: Springer-Verlag, 1986, vol. 1200.
[46] D. Pollard, *Empirical Processes: Theory and Applications*. Hayward, CA: Inst. Math. Statist., vol. 2, NSF-CBMS Regional Conference Series in Probability and Statistics.
[47] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies," Applied and Computational Mathematics, Calif. Inst. Technol., Tech. Rep., 2004.