A Probabilistic and RIPless Theory of Compressed Sensing

Emmanuel J. Candès and Yaniv Plan

Dedicated to the memory of Jerrold E. Marsden

Abstract—This paper introduces a simple and very general theory of compressive sensing. In this theory, the sensing mechanism simply selects sensing vectors independently at random from a probability distribution F; it includes all standard models—e.g., Gaussian, frequency measurements—discussed in the literature, but also provides a framework for new measurement strategies as well. We prove that if the probability distribution F obeys a simple incoherence property and an isotropy property, one can faithfully recover approximately sparse signals from a minimal number of noisy measurements. The novelty is that our recovery results do not require the restricted isometry property (RIP) to hold near the sparsity level in question, nor a random model for the signal. As an example, the paper shows that a signal with s nonzero entries can be faithfully recovered from about $s \log n$ Fourier coefficients that are contaminated with noise.

Index Terms—Compressed sensing, Dantzig selector, ℓ_1 minimization, Gross' golfing scheme, LASSO, operator Bernstein inequalities, random matrices, sparse regression, (weak) restricted isometries.

I. INTRODUCTION

THIS paper develops a novel, simple, and general theory of compressive sensing [12], [15], [20], a rapidly growing field of research that has developed protocols for acquiring certain types of signals with far fewer data bits than what is classically accepted.

A. A RIPless Theory?

The early paper [12] triggered a massive amount of research by showing that it is possible to sample signals at a rate proportional to their information content rather than their bandwidth. For instance, in a discrete setting, this theory asserts that a digital signal $x \in \mathbb{R}^n$ (which can be viewed as Nyquist samples of a continuous-time signal over a time window of interest) can be recovered from a small random sample of its Fourier coefficients provided that x is sufficiently sparse. Formally, suppose

that our signal x has at most s nonzero amplitudes at completely unknown locations—such a signal is called *s-sparse*—and that we are given the value of its discrete Fourier transform (DFT) at m frequencies selected uniformly at random (we think of m as being much smaller than n). Then [12] showed that one can recover x by solving an optimization problem which simply finds, among all candidate signals, that with the minimum ℓ_1 norm; the number of samples we need must be on the order of $s \log n$. In other words, if we think of s as a measure of the information content, we can sample *nonadaptively* nearly at the information rate without information loss. By swapping time and frequency, this also says that signals occupying a very large bandwidth but with a sparse spectrum can be sampled (at random time locations) at a rate far below the Shannon-Nyquist rate.

Despite considerable progress in the field, some important questions are still open. We discuss two that have both a theoretical and practical appeal.

Is it possible to faithfully recover a nearly sparse signal $x \in \mathbb{R}^n$, one which is well approximated by its s largest entries, from about $s \log n$ of its Fourier coefficients? Is it still possible when these coefficients are further corrupted by noise?

These issues are paramount since in real-world applications, signals are never exactly sparse, and measurements are never perfect either. Now the traditional way of addressing these types of problems in the field is by means of the *restricted isometry property* (RIP) [14]. The trouble here is that it is unknown whether or not this property holds when the sample size m is on the order of $s \log n$. In fact, answering this one way or the other is generally regarded as extremely difficult, and so the restricted isometry machinery does not directly apply in this setting.

This paper proves that the two questions formulated above have positive answers. In fact, we introduce recovery results which are—up to a logarithmic factor—as good as those one would get if the restricted isometry property were known to be true. To fix ideas, suppose we observe m noisy discrete Fourier coefficients about an s-sparse signal x

$$\tilde{y}_k = \sum_{t=0}^{n-1} e^{-i2\pi\omega_k t} x[t] + \sigma z_k, \quad k = 1, \ldots, m. \qquad \text{(I.1)}$$

Here, the frequencies ω_k are chosen uniformly at random in $\{0, 1/n, 2/n, \ldots, (n-1)/n\}$ and z_k is white noise with unit

variance. Then if the number of samples m is on the order of $s \log n$, it is possible to get an estimate $\hat{x}$ obeying

$$\|\hat{x} - x\|_{\ell_2}^2 = \text{polylog}(n)\,\frac{s}{m}\sigma^2 \qquad (\text{I.2})$$

by solving a convex ℓ_1-minimization program. (Note that when the noise vanishes, the recovery is exact.) Up to the logarithmic factor, which may sometimes be on the order of $\log n$ and at most a small power of this quantity, this is optimal. Now if the RIP held, one would get a squared error bounded by $O(\log n)\frac{s}{m}\sigma^2$ [16], [5] and, therefore, the "RIPless" theory developed in this paper roughly enjoys the same performance guarantees.

B. A General Theory

The estimate we have just seen is not isolated and the real purpose of this paper is to develop a theory of compressive sensing which is both as simple and as general as possible.

At the heart of compressive sensing is the idea that randomness can be used as an effective sensing mechanism. We note that random measurements are not only crucial in the derivation of many theoretical results, but also generally seem to give better empirical results as well. Therefore, we propose a mechanism whereby sensing vectors are independently sampled from a population F. Mathematically, we observe

$$\tilde{y}_k = \langle a_k, x \rangle + \sigma z_k, \quad k = 1, \ldots, m \qquad (\text{I.3})$$

where $x \in \mathbb{R}^n$, $\{z_k\}$ is a noise sequence, and the sensing vectors $a_k \overset{iid}{\sim} F$ (note that a_k is a "column vector"). For example, if F is the family of complex sinusoids, this is the Fourier sampling model introduced earlier. All we require from F is an isotropy property and an incoherence property.

Isotropy property: We say that F obeys the isotropy property if

$$\mathbb{E}aa^* = I, \quad a \sim F. \qquad (\text{I.4})$$

If F has mean zero (we do not require this), then $\mathbb{E}aa^*$ is the covariance matrix of F. In other words, the isotropy condition states that the components of $a \sim F$ have unit variance and are uncorrelated. This assumption may be weakened a little, as we shall see later.

Incoherence property: We may take the coherence parameter $\mu(F)$ to be the smallest number such that with $a = (a[1], \ldots, a[n]) \sim F$,

$$\max_{1 \le t \le n} |a[t]|^2 \le \mu(F) \qquad (\text{I.5})$$

holds either deterministically or stochastically in the sense discussed below. The smaller $\mu(F)$, i.e., the more incoherent the sensing vectors, the fewer samples we need for accurate recovery. When a simple deterministic bound is not available, one can take the smallest scalar μ obeying

$$\begin{aligned} \mathbb{E}[n^{-1}\|a\|_{\ell_2}^2\,\mathbb{1}_{E^c}] &\le \tfrac{1}{20}n^{-3/2} \qquad \text{and} \\ \mathbb{P}(E^c) &\le (nm)^{-1}, \end{aligned} \qquad (\text{I.6})$$

where E is the event $\{\max_{1 \le t \le n} |a[t]|^2 \le \mu\}$.

Suppose for instance that the components are i.i.d. $\mathcal{N}(0,1)$. Then a simple calculation we shall not detail shows that

$$\mathbb{E}[n^{-1}\|a\|_{\ell_2}^2\,\mathbb{1}_{E^c}] \le 2n\mathbb{P}(Z > \sqrt{\mu}) + 2\sqrt{\mu}\,\phi(\sqrt{\mu})$$
$$\mathbb{P}(E^c) \le 2n\mathbb{P}(Z \ge \sqrt{\mu}) \qquad (\text{I.7})$$

where Z is standard normal and ϕ is its density function. The inequality $P(Z > t) \le \phi(t)/t$ shows that one can take $\mu(F) \le 6 \log n$ as long as $n \ge 16$ and $m \le n$. More generally, if the components of a are i.i.d. samples from a sub-Gaussian distribution, $\mu(F)$ is at most a constant times $\log n$. If they are i.i.d. from a subexponential distribution, $\mu(F)$ is at most a constant times $\log^2 n$. In what follows, however, it might be convenient for the reader to assume that the deterministic bound (I.5) holds.

It follows from the isotropy property that $\mathbb{E}|a[t]|^2 = 1$, and thus $\mu(F) \ge 1$. This lower bound is achievable by several distributions and one such example is obtained by sampling a row from the DFT matrix as before, so that

$$a[t] = e^{i2\pi kt/n}$$

where k is chosen uniformly at random in $\{0, 1, \ldots, n-1\}$. Then another simple calculation shows that $\mathbb{E}aa^* = I$ and $\mu(F) = 1$ since $|a[t]|^2 = 1$ for all t. At the other extreme, suppose the measurement process reveals one entry of x selected uniformly at random so that $a = \sqrt{n}\,e_i$ where i is uniform in $\{1, \ldots, n\}$; the normalization ensures that $\mathbb{E}aa^* = I$. This is a lousy acquisition protocol because one would need to sample on the order of $n \log n$ times to recover even a 1-sparse vector (the logarithmic term comes from the coupon collector effect). Not surprisingly, this distribution is in fact highly coherent as $\mu(F) = n$.

With the assumptions set, we now give a representative result of this paper: suppose x is an arbitrary but fixed s-sparse vector and that one collects information about this signal by means of the random sensing mechanism (I.3), where z is white noise. Then if the number of samples is on the order $\mu(F)s \log n$, one can invoke ℓ_1 minimization to get an estimator $\hat{x}$ obeying

$$\|\hat{x} - x\|_{\ell_2}^2 \le \text{polylog}(n)\,\frac{s}{m}\sigma^2.$$

This bound is sharp. It is not possible to substantially reduce the number of measurements and get a similar bound, no matter how intractable the recovery method might be. Further, with this many measurements, the upper bound is optimal up to logarithmic factors. Finally, we will see that when the signal is not exactly sparse, we just need to add an approximation error to the upper bound.

To summarize, this paper proves that one can faithfully recover approximately s-sparse signals from about $s \log n$ random incoherent measurements for which $\mu(F) = O(1)$.

C. Examples of Incoherent Measurements

We have seen through examples that sensing vectors with low coherence are global or spread out. Incoherence alone, however, is not a sufficient condition: if F were a constant distribution (sampling from F would always return the same vector), one would not learn anything new about the signal by taking more samples regardless of the level of incoherence. However,

as we will see, the incoherence and isotropy properties together guarantee that sparse vectors lie away from the nullspace of the sensing matrix whose rows are the a_k^*'s.

The role of the isotropy condition is to keep the measurement matrix from being rank deficient when sufficiently many measurements are taken (and similarly for subsets of columns of A). Specifically, one would hope to be able to recover *any* signal from an arbitrarily large number of measurements. However, if $\mathbb{E}aa^*$ were rank deficient, there would be signals $x \in \mathbb{R}^n$ that would not be recoverable from an arbitrary number of samples; just take $x \neq 0$ in the nullspace of $\mathbb{E}aa^*$. The nonnegative random variable x^*aa^*x has vanishing expectation, which implies $a^*x = 0$ almost surely. (Put differently, all of the measurements would be zero almost surely.) In contrast, the isotropy condition implies that $\frac{1}{m}\sum_{k=1}^{m} a_k a_k^* \rightarrow I$ almost surely as $m \rightarrow \infty$ and, therefore, with enough measurements, the sensing matrix is well conditioned and has a left-inverse.[1]

We now provide examples of incoherent and isotropic measurements.

- **Sensing vectors with independent components**. Suppose the components of $a \sim F$ are independently distributed with mean zero and unit variance. Then F is isotropic. In addition, if the distribution of each component is light-tailed, then the measurements are clearly incoherent.
  A special case concerns the case where $a \sim N(0, I)$, also known in the field as the *Gaussian measurement ensemble*, which is perhaps the most commonly studied. Here, one can take $\mu(F) = 6\log n$ as seen before.
  Another special case is the *binary measurement ensemble* where the entries of a are symmetric Bernoulli variables taking on the values ± 1. A shifted version of this distribution is the sensing mechanism underlying the single pixel camera [23].
- **Subsampled orthogonal transforms**: Suppose we have an orthogonal matrix obeying $U^*U = nI$. Then consider the sampling mechanism picking rows of U uniformly and independently at random. In the case where U is the DFT, this is the random frequency model introduced earlier. Clearly, this distribution is isotropic and $\mu(F) = \max_{ij} |U_{ij}|^2$. In the case where U is a Hadamard matrix, or a complex Fourier matrix, $\mu(F) = 1$.
- **Random convolutions**: Consider the circular convolution model $y = Gx$ in which

$$G = \begin{bmatrix} g[0] & g[1] & g[2] & \dots & g[n-1] \\ g[n-1] & g[0] & g[1] & \dots & \\ & & & & \\ g[1] & & \dots & g[n-1] & g[0] \end{bmatrix}.$$

Because a convolution is diagonal in the Fourier domain (we just multiply the Fourier components of x with those of g), G is an isometry if the Fourier components of $g = (g[0], \dots, g[n-1])$ have the same magnitude. In this case, sampling a convolution product at randomly selected time locations is an isotropic and incoherent

process provided g is spread out $(\mu(F) = \max_t |g(t)|^2)$. This example extends to higher dimensions; e.g., to spatial 3-D convolutions. In fact, certain random convolutions may be used universally to efficiently sample a signal that is sparsely represented in *any* orthonormal basis while still giving strong theoretical guarantees for signal recovery [43]—an appealing feature due to the fast application of a convolution matrix. For a quick comparison, the theory in [43] demonstrates that about $s\log n$ measurements are sufficient for noiseless recovery and about $s\log^5 n$ are sufficient for stable recovery. The theory in this paper may also be applied to the random convolution described in [43], demonstrating that about $s\log^2 n$ measurements are sufficient for stable recovery.

- **Subsampled tight or continuous frames**: We can generalize the example above by subsampling a tight frame or even a continuous frame. An important example might be the Fourier transform with a continuous frequency spectrum. Here

$$a(t) = e^{i2\pi\omega t}$$

where ω is chosen uniformly at random in $[0, 1]$ (instead of being on an equispaced lattice as before). This distribution is isotropic and obeys $\mu(F) = 1$. A situation where this arises is in magnetic resonance imaging (MRI) as frequency samples rarely fall on an equispaced Nyquist grid. By swapping time and frequency, this is equivalent to sampling a nearly sparse trigonometric polynomial at randomly selected time points in the unit interval [38]. More generally, as described in [41, Chapter 4], this model applies to random sampling of signals that have a sparse expansion in a bounded orthonormal system of functions.

These examples could of course be multiplied, and we hope we have made clear that our framework is general and encompasses many of the measurement models commonly discussed in compressive sensing—and perhaps many more.

Our setup simply consists in selecting each measurement vector independently from the others. For completeness, however, we describe other measurement ensembles discussed in the compressive sensing literature that do not assume this model. Whether the theory in this paper can be extended to these models is open to further research. A reversal of the random convolution example described above is convolution with a random vector at *fixed* locations (rather than convolution with a fixed vector at random locations). In fact, such a measurement mechanism has been studied in detail [40], [28], [41], [42], with strong theoretical results. Similarly to the theory in this paper, near optimal theoretical guarantees on sparse-signal recovery are non-uniform [40], [41], whereas uniform recovery through RIP machinery has been quite suboptimal. Similar results are available for random Gabor matrices [34].

D. Matrix Notation

Before continuing, we pause to introduce some useful matrix notation. Divide both sides of (I.3) by $\sqrt{m}$, and rewrite our statistical model as

$$y = Ax + \sigma_m z; \qquad (I.8)$$

---

[1]One could require "near isotropy," i.e., $\mathbb{E}aa^* \approx I$. If the approximation were tight enough, our theoretical results would still follow with minimal changes to the proof.

the kth entry of y is $\tilde{y}_k$ divided by $\sqrt{m}$, the kth row of A is a_k^* divided by $\sqrt{m}$, and σ_m is σ divided by $\sqrt{m}$. This normalization implies that the columns of A are approximately unit-normed, and is most used in the compressive sensing literature.

E. Incoherent Sampling Theorem

For pedagogical reasons, we introduce our results by first presenting a recovery result from noiseless data. The recovered signal is obtained by the standard ℓ_1-minimization program

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

(Recall that the rows of A are normalized independent samples from F.)

Theorem 1.1 (Noiseless Incoherent Sampling): Let x be a fixed but otherwise arbitrary s-sparse vector in $\mathbb{R}^n$ and pick any scalar $\beta > 0$. Then with probability at least $1 - 5/n - e^{-\beta}$, x is the unique minimizer to (I.9) with $y = Ax$ provided that

$$m \geq C_\beta \cdot \mu(F) \cdot s \cdot \log n.$$

More precisely, C_β may be chosen as $C_0(1 + \beta)$ for some positive numerical constant C_0.

Among other things, this theorem states that one can perfectly recover an arbitrary sparse signal from about $s \log n$ convolution samples, or a signal that happens to be sparse in the wavelet domain from about $s \log n$ randomly selected noiselet coefficients. It extends an earlier result [11], which assumed a subsampled orthogonal model, and strengthens it since that reference could only prove the claim for randomly signed vectors x. Here, x is arbitrary, and we do not make any distributional assumption about its support or its sign pattern.

This theorem is also about a fundamental information theoretic limit: the number of samples for perfect recovery has to be on the order of $\mu(F) \cdot s \cdot \log n$, and cannot possibly be much below this number. More precisely, suppose we are given a distribution F with coherence parameter $\mu(F)$. Then there exist s-sparse vectors that cannot be recovered with probability at least $1 - 1/n$, say, from fewer than a constant times $\mu(F) \cdot s \cdot \log n$ samples. When $\mu(F) = 1$, this has been already established since [12] proves that some s sparse signals cannot be recovered from fewer than a constant times $s \cdot \log n$ random DFT samples. Our general claim follows from a modification of the argument in [12]. Assume, without loss of generality, that $\mu(F)$ is an integer and consider the isotropic process that samples rows from an $n \times n$ block diagonal matrix, each block being a DFT of a smaller size; that is, of size n/ℓ where $\mu(F) = \ell$. Then if $m \leq c_0 \cdot \mu(F) \cdot s \cdot \log n$, one can construct s-sparse signals just as in [12] for which $Ax = 0$ with probability at least $1/n$. We omit the details.

The important aspect, here, is the role played by the coherence parameter $\mu(F)$. In general, the minimal number of samples must be on the order of the coherence times the sparsity level s times a logarithmic factor. Put differently, *the coherence completely determines the minimal sampling rate.*

F. Main Results

We assume for simplicity that we are undersampling so that $m \leq n$. Our general result deals with: 1) arbitrary signals which are not necessarily sparse (images are never exactly sparse even in a transformed domain) and 2) noise. To recover x from the data y and the model (I.8), we consider the unconstrained LASSO which solves the ℓ_1 regularized least-squares problem

$$\min_{\bar{x} \in \mathbb{R}^n} \ \frac{1}{2} \|A\bar{x} - y\|_{\ell_2}^2 + \lambda \sigma_m \|\bar{x}\|_{\ell_1}. \qquad (\text{I.10})$$

We assume that z is Gaussian $z \sim N(0, I)$. However, the theorem below remains valid as long as $\|A^* z\|_{\ell_\infty} \leq \lambda_n$ for some $\lambda_n \geq 0$, and thus many other noise models would work as well. In what follows, x_s is the best s-sparse approximation of x or, equivalently, the vector consisting of the s largest entries of x in magnitude.

Theorem 1.2: Let x be an arbitrary fixed vector in $\mathbb{R}^n$ and pick any scalar $\beta > 0$. Then with probability at least $1 - 6/n - 6e^{-\beta}$, the solution $\hat{x}$ to (I.10) with $\lambda = 10\sqrt{\log n}$ obeys

$$\|\hat{x} - x\|_{\ell_2} \leq \min_{1 \leq s \leq \bar{s}} C(1 + \alpha) \left[\frac{\|x - x_s\|_{\ell_1}}{\sqrt{s}} + \sigma\sqrt{\frac{s \log n}{m}} \right] \qquad (\text{I.11})$$

provided that $m \geq C_\beta \cdot \mu(F) \cdot \bar{s} \cdot \log n$. If one measures the error in the ℓ_1 norm, then

$$\|\hat{x} - x\|_{\ell_1} \leq \min_{1 \leq s \leq \bar{s}} C(1 + \alpha) \left[\|x - x_s\|_{\ell_1} + s\sigma\sqrt{\frac{\log n}{m}} \right]. \qquad (\text{I.12})$$

Above, C is a numerical constant, C_β can be chosen as before, and $\alpha = \sqrt{\frac{(1+\beta)s\mu \log n \log m \log^2 s}{m}}$ which is never greater than $\log^{3/2} n$ in this setup.

Above, $\bar{s}$ may be interpreted as an upper bound on allowable sparsity levels s that still lead to stable reconstruction (according to our proofs). We may take $\bar{s} = \frac{m}{C_\beta \cdot \mu(F) \cdot \log n}$.

These robust error bounds do not require either: 1) a random model on the signal or 2) the RIP nor one of a few closely related strong conditions such as the RIP-1 [25], the restricted eigenvalue assumption [5] or the compatibility condition [49]—all conditions that would imply uniform recovery. In contrast, our arguments work through a new machinery which guarantees *fixed* sparse signal recovery with high probability. It is currently unknown whether our conditions are strong enough to imply uniform recovery. Further, the error bound is within at most a $\log^{3/2} n$ factor of what has been established using the RIP since a variation on the arguments in [16] would give an error bound proportional to the quantity inside the square brackets in (I.11). As a consequence, the error bound is within a polylogarithmic factor of what is achievable with the help of an oracle that would reveal the locations of the significant coordinates of the unknown signal [16]. In other words, it cannot be substantially improved.

Because much of the compressive sensing literature works with restricted isometry conditions—we shall discuss exceptions such as [22], [4] in Section I-G—we pause here to discuss

these conditions and to compare them to our own. We say that an $m \times n$ matrix A obeys the RIP with parameters s and δ if

$$(1 - \delta)\|v\|_{\ell_2}^2 \leq \|Av\|_{\ell_2}^2 \leq (1 + \delta)\|v\|_{\ell_2}^2 \qquad \text{(I.13)}$$

for all s-sparse vectors v. In other words, all the submatrices of A with at most s columns are well conditioned. When the RIP holds with parameters $2s$ and $\delta < 0.414\ldots$ [8] or even $\delta \leq 0.465\ldots$ [24], it is known that the error bound (I.11) holds (without the factor $(1 + \alpha)$). This δ is sometimes referred to as the restricted isometry constant.

Bounds on the restricted isometry constant have been established in [15] and in [46] for partial DFT matrices, and by extension, for partial subsampled orthogonal transforms. For instance, [46] proves that if A is a properly normalized partial DFT matrix, or a unitary matrix with bounded entries, then the RIP with $\delta = 1/4$ holds with high probability if $m \geq C \cdot s \log^4 n$ (C is some positive constant). Rauhut [41, Ch. 4 and 8] extended these results to the measurement matrices consisting of sampled bounded orthogonal systems. While such measurements are a special case of those considered in this paper, the theory in [41, Ch. 8] extends to the measurement model considered here. Small adjustments need to be made in the case of stochastic incoherence, but the end result is that the RIP holds with high probability when $m \geq C \cdot \mu(F) \cdot s \mu(F) \log^2 s \log n \log m$, i.e., about $s \mu \log^4 n$ measurements are sufficient. Thus, our result bridges the gap between the region where the RIP is known to hold and the region in which one has the minimum number of measurements needed to prove perfect recovery of exactly sparse signals from noisy data, which is on the order of $\mu(F) \cdot s \log n$. The gap is bridged smoothly in the sense that our error bounds match those available with the RIP as soon as the RIP is known to hold, and grow slightly larger when there are fewer measurements.

The careful reader will no doubt remark that for very specific models such as the Gaussian measurement ensemble, it is known that on the order $s \log(n/s)$ samples are sufficient for stable recovery while our result asserts that on the order of $s \log^2 n$ are sufficient (and $s \log n$ for the binary measurement ensemble). This slight loss is a small price to pay for a very simple general theory, which accommodates a wide array of sensing strategies. Having said this, the reader will also verify that specializing our proofs below gives an optimal result for the Gaussian ensemble; i.e., establishes a near-optimal error bound from about $s \log n$ observations (the $\log(n/s)$ factor, however, may still require a different approach).

Finally, another frequently discussed algorithm for sparse regression is the Dantzig selector [16]. Here, the estimator is given by the solution to the linear program

$$\min_{\bar{x} \in \mathbb{R}^n} \|\bar{x}\|_{\ell_1} \quad \text{subject to} \quad \|A^*(A\bar{x} - y)\|_{\ell_\infty} \leq \lambda \sigma_m. \tag{I.14}$$

We show that the Dantzig selector obeys nearly the same error bound.

Theorem 1.3: The Dantzig selector, with $\lambda = 10\sqrt{\log n}$ and everything else the same as in Theorem 1.2, obeys

$$\|\hat{x} - x\|_{\ell_2} \leq \min_{s \leq \bar{s}} C(1 + \alpha^2) \left[\frac{\|x - x_s\|_{\ell_1}}{\sqrt{s}} + \sigma\sqrt{\frac{s \log n}{m}} \right] \tag{I.15}$$

$$\|\hat{x} - x\|_{\ell_1} \leq \min_{s \leq \bar{s}} C(1 + \alpha^2) \left[\|x - x_s\|_{\ell_1} + s\sigma\sqrt{\frac{\log n}{m}} \right] \tag{I.16}$$

with the same probabilities as before.

The only difference is α^2 instead of α in the right-hand sides.

G. Our Contribution

The main contribution of this paper is to provide a simple framework which applies to all the standard compressive sensing models and some new ones as well. The results in this paper also reduce the minimal number of measurements theoretically required in some standard sensing models such as Fourier measurements, or, more generally, sensing matrices obtained by sampling a few rows from an orthogonal matrix. We establish that the restricted isometry property is not necessarily needed to accurately recover fixed nearly sparse vectors from noisy compressive samples. This may be of interest because in many cases RIP-based results have needed strong requirements on relationship between the number of measurements required and the sparsity level. Thus our work is a significant departure from the majority of the literature, which establishes good noisy recovery properties via the RIP machinery. This literature is, of course, extremely large and we cannot cite all contributions but a partial list would include [15], [21], [46], [13], [16], [5], [19], [50], [51], [3], [37], [29], [7], [2].

The reason why one can get strong error bounds, which are within a polylogarithmic factor of what is available with the aid of an 'oracle,' without the RIP is that our results do not imply universality. That is, we are not claiming that if A is randomly sampled and then fixed once for all, then the error bounds from Section I-F hold for all signals x. What we are saying is that if we are given an arbitrary x, and then collect data by applying our random scheme, then the recovery of *this* x will be accurate.

If one wishes to establish universal results holding for *all* x simultaneously, then we would need the RIP or a property very close to it. As a consequence, we cannot possibly be in this setup and guarantee universality since we are not willing to assume that the RIP holds. To be sure, suppose we had available an oracle informing us about the support T of x. Then we would need the pseudo-inverse of the submatrix with columns in T to be bounded. In other words, the minimum singular value of this submatrix would have to be away from zero. For a universal result, this would need to be true for all subsets of cardinality s; that is, the minimum singular value of all submatrices with s columns would have to be away from zero. This essentially is the restricted isometry property.

To the best of our knowledge, only a few other papers have addressed non-universal stability (the literature grows so rapidly that an inadvertent omission is entirely possible). In an earlier work [9], the authors also considered weak conditions that allow stable recovery; in this case the authors assumed that the signal was sampled according to a random model, but in return the measurement matrix A could be deterministic. In the asymptotic case, stable signal recovery has been demonstrated for the Gaussian measurement ensemble in a regime in which the RIP does not necessarily hold [22], [4]. In fact, the authors of [22], [4] are able to give exact limits on the error rather than error

bounds. Aside from these papers and the work in progress [10], it seems that that the literature regarding stable recovery with conditions weak enough that they do not imply universality is extremely sparse.

Lastly, we would like to point to two important papers in the literature which together give a strong result for stable RIP-less recovery with subsampled discrete Fourier transforms [27], [52]. The authors of [27] give an useful equivalence between the ℓ_1 and ℓ_2 norm of any vector which is the sum of a subset of the columns of a discrete Fourier transform. These results, have important implications for Gelfand widths, and also may be used to characterize the null space of a subsampled DFT. This theory may be combined with null-space conditions of [52] to imply stable of recovery of sparse signals by the (constrained) Lasso; the required number of measurements by this theory is

$$m \geq O(1) \cdot s \log m \log^5[(n/m) \log m].$$

In the case when $m \approx n$, this leads to the necessity of only about $s \log m \log^5 \log m$ measurements, which nearly matches the requirement in this paper of about $s \log m$ measurements. However, the error bounds are quite different from those presented in our paper. See [35, Section 2.2.3] for more details. Finally and to be complete, we would like to mention that earlier works have considered the recovery of perfectly sparse signals from subsampled orthogonal transforms [11], and of sparse trigonometric polynomials from random time samples [38].

H. Organization of the Paper

The paper is organized as follows. In Section II, we introduce several fundamental estimates which our arguments rely upon, but which also could be useful tools for other results in the field. In Section 3, we prove the noiseless recovery result, namely, Theorem 1.1. In Section IV, we prove our main results, Theorems 1.2 and 1.3. Now all these sections assume for simplicity of exposition that the coherence bound holds deterministically (I.5). We extend the proof to distributions obeying the coherence property in the stochastic sense (I.6) in the Appendix. This Appendix also contains another important technical piece, namely, a difficult proof of an intermediate result (weak RIP property). Finally, we conclude the main text with some final comments in Section V.

I. Notation

We provide a brief summary of the notations used throughout the paper. For an $m \times n$ matrix A and a subset $T \subset \{1, \ldots, n\}$, A_T denotes the $m \times |T|$ matrix with column indices in T. Also, $A_{\{i\}}$ is the i-th column of A. Likewise, for a vector $v \in \mathbb{R}^n$, v_T is the restriction of v to indices in T. Thus, if v is supported on T, $Av = A_T v_T$. In particular, $a_{k,T}$ is the vector a_k restricted to T. The operator norm of a matrix A is denoted $\|A\|$; also, $\|A\|_{1,2}$ is the maximum column norm of A. The identity matrix, in any dimension, is denoted I. Further, e_i always refers to the i-th standard basis element, e.g., $e_1 = (1, 0, \ldots, 0)$. For a scalar t, $\mathrm{sgn}(t)$ is the sign of t if $t \neq 0$ and is zero otherwise. For a vector x, $\mathrm{sgn}(x)$ applies the sign function componentwise. We shall also use μ as a shorthand for $\mu(F)$ whenever convenient.

Throughout, C is a constant whose value may change from instance to instance.

II. FUNDAMENTAL ESTIMATES

Our proofs rely on several estimates, and we provide an interpretation of each whenever possible. The first estimates **E1**–**E4** are used to prove the noiseless recovery result; when combined with the weak RIP, they imply stability and robustness. Throughout this section, δ is a parameter left to be fixed in later sections; it is always less than or equal to one.

A. Local Isometry

Let T of cardinality s be the support of x in Theorem 1.1, or the support of the best s-sparse approximation of x in Theorem 1.2. We shall need that with high probability

$$\|A_T^* A_T - I\| \leq \delta \qquad \text{(II.1)}$$

with $\delta \leq 1/2$ in the proof of Theorem 1.1 and $\delta \leq 1/4$ in that of Theorem 1.2. Put differently, the singular values of A_T must lie away from zero. This condition essentially prevents A_T from being singular as, otherwise, there would be no hope of recovering our sparse signal x. Indeed, letting h be any vector supported on T and in the null space of A, we would have $Ax = A(x + h)$ and thus, recovery would be impossible even if one knew the support of x. The condition (II.1) is much weaker than the restricted isometry property because it does not need to hold uniformly over all sparse subsets—only on the support set.

Lemma 2.1 (E1: Local Isometry): Let T be a fixed set of cardinality s. Then for $\delta > 0$,

$$\mathbb{P}(\|A_T^* A_T - I\| \geq \delta) \leq 2s \exp\left(-\frac{m}{\mu(F)s} \cdot \frac{\delta^2}{2(1 + \delta/3)}\right). \qquad \text{(II.2)}$$

In particular, if $m \geq \frac{56}{3} \mu(F) \cdot s \cdot \log n$, then

$$\mathbb{P}(\|A_T^* A_T - I\| \geq 1/2) \leq 2/n.$$

Note that $\|A_T^* A_T - I\| \leq \delta$ implies that $\|(A_T^* A_T)^{-1}\| \leq 1/(1 - \delta)$, a fact that we will use several times.

In compressive sensing, the classical way of proving such estimates is via Rudelson's selection theorem [44]. Here, we use a more modern technique based on the matrix Bernstein inequality of Ahlswede and Winter [1], developed for this setting by Gross [26], and tightened in [48] by Tropp and in [32] by Oliveira. We present the version in [48].

Theorem 2.2 (Matrix Bernstein Inequality): Let $\{X_k\} \in \mathbb{R}^{d \times d}$ be a finite sequence of independent random self-adjoint matrices. Suppose that $\mathbb{E}X_k = 0$ and $\|X_k\| \leq B$ a.s. and put

$$\sigma^2 := \left\|\sum_k \mathbb{E}X_k^2\right\|.$$

Then for all $t \geq 0$

$$\mathbb{P}\left(\left\|\sum_k X_k\right\| \geq t\right) \leq 2d \exp\left(\frac{-t^2/2}{\sigma^2 + Bt/3}\right). \qquad \text{(II.3)}$$

Proof of Estimate **E1**: Decompose $A_T^* A_T - I$ as

$$A_T^* A_T - I = m^{-1} \sum_{k=1}^{m} (a_{k,T} a_{k,T}^* - I) = m^{-1} \sum_{k=1}^{m} X_k$$

where $X_k := a_{k,T} a_{k,T}^* - I$. The isotropy condition implies $\mathbb{E} X_k = 0$, and since $\|a_T\|_{\ell_2}^2 \leq \mu(F) \cdot s$, we have $\|X_k\| = \max(\|a_{i,T}\|_{\ell_2}^2 - 1, 1) \leq \mu(F) \cdot s$. Lastly, $0 \preceq \mathbb{E} X_k^2 = \mathbb{E}(a_{k,T} a_{k,T}^*)^2 - I \preceq \mathbb{E}(a_{k,T} a_{k,T}^*)^2 = \mathbb{E}\|a_{k,T}\|^2 a_{k,T} a_{k,T}^*$. However

$$\mathbb{E}\|a_{k,T}\|^2 a_{k,T} a_{k,T}^* \preceq \mu(F) \cdot s \cdot \mathbb{E} a_{k,T} a_{k,T}^* = \mu(F) \cdot s \cdot I$$

and, therefore, $\sum_k \mathbb{E} X_k^2 \preceq m \cdot \mu(F) \cdot s \cdot I$ so that σ^2 is bounded above by $m \cdot \mu(F) \cdot s$. Plugging $t = \delta m$ into (II.3) gives the lemma. ∎

Instead of having A act as a near isometry on all vectors supported on T, we could ask that it preserves the norm of an arbitrary fixed vector (with high probability), i.e., $\|Av\|_{\ell_2} \approx \|v\|_{\ell_2}$ for a fixed v supported on T. Not surprisingly, this can be proved with generally (slightly) weaker requirements.

Lemma 2.3 (E2: Low-Distortion): Let w be a fixed vector supported on a set T of cardinality at most s. Then for each $t \leq 1/2$,

$$\mathbb{P}(\|(A_T^* A_T - I) w_T\|_{\ell_2} \geq t\|w\|_{\ell_2})$$
$$\leq \exp\left(-\frac{1}{4}\left(t\sqrt{\frac{m}{\mu(F)s}} - 1\right)^2\right).$$

The estimate follows from the vector Bernstein inequality, proved by Gross [26, Theorem 11]. We use a slightly weaker version, which we find slightly more convenient.

Theorem 2.4 (Vector Bernstein Inequality): Let $\{v_k\} \in \mathbb{R}^d$ be a finite sequence of independent random vectors. Suppose that $\mathbb{E} v_k = 0$ and $\|v_k\|_{\ell_2} \leq B$ a.s. and put $\sigma^2 \geq \sum_k \mathbb{E}\|v_k\|_{\ell_2}^2$. Then for all $0 \leq t \leq \sigma^2/B$

$$\mathbb{P}\left(\left\|\sum_k v_k\right\|_{\ell_2} \geq t\right) \leq \exp\left(-\frac{(t/\sigma - 1)^2}{4}\right) \quad \text{(II.4)}$$
$$\leq \exp\left(-\frac{t^2}{8\sigma^2} + \frac{1}{4}\right).$$

Note that the bound does not depend on the dimension d.

Proof of Estimate **E2**: We have

$$(A_T^* A_T - I) w_T = \frac{1}{m} \sum_{i=1}^{m} a_{i,T} \langle a_{i,T}, w_T \rangle - w_T =: \frac{1}{m} \sum_{i=1}^{m} v_k$$

where $v_k := \langle a_{i,T}, w_T \rangle - w_T$. By isotropy, $\mathbb{E} v_k = 0$, and by the Cauchy-Schwarz inequality together with the triangle inequality,

$$\|v_k\|_{\ell_2} \leq (\|a_{i,T}\|_{\ell_2} + 1)\|w_T\|_{\ell_2} \leq (\sqrt{s\mu} + 1)\|w_T\|_{\ell_2} =: B.$$

The second inequality follows from the coherence condition. For the variance σ^2,

$$\mathbb{E}\|v_k\|_{\ell_2}^2 = \mathbb{E}\|a_{i,T}\|_{\ell_2}^2 \langle a_{i,T}, w_T \rangle^2 - \|w_T\|_{\ell_2}^2$$
$$\leq \mathbb{E}\|a_{i,T}\|_{\ell_2}^2 \langle a_{i,T}, w_T \rangle^2 \leq s\mu \mathbb{E}\langle a_{i,T}, w_T \rangle^2 = s\mu\|w_T\|_{\ell_2}^2.$$

The last equality follows by the isotropy condition. Thus, σ^2 is bounded by $ms\mu\|w_T\|_{\ell_2}^2$. The proof now follows by plugging in B and σ^2 into the vector Bernstein inequality. ∎

B. Off-Support Incoherence

Lemma 2.5 (E3: Off-Support Incoherence): Let v be supported on T with $|T| = s$. Then for each $t > 0$

$$\mathbb{P}(\|A_{T^c}^* Av\|_{\ell_\infty} \geq t\|v\|_{\ell_2})$$
$$\leq 2n \exp\left(-\frac{m}{2\mu(F)} \cdot \frac{t^2}{1 + \frac{1}{3}\sqrt{st}}\right).$$

This lemma says that if $v = x$, then $\max_{i \in T^c} |\langle A_{\{i\}}, Ax \rangle|$ cannot be too large so that the off-support columns do not correlate too well with Ax. The proof of **E3** is an application of Bernstein's inequality—the matrix Bernstein inequality with $d = 1$—together with the union bound.[2]

Proof: We have

$$\|A_{T^c}^* Av\|_{\ell_\infty} = \max_{i \in T^c} |\langle e_i, A^* Av \rangle|.$$

Assume without loss of generality that $\|v\|_{\ell_2} = 1$, fix $i \in T^c$ and write

$$\langle e_i, A^* Av \rangle = \frac{1}{m} \sum_k w_k, \quad w_k := \langle e_i, a_k a_k^* v \rangle.$$

Since $i \in T^c$, $\mathbb{E} w_k = 0$ by the isotropy property. Next, the Cauchy-Schwarz inequality gives $|w_k| = |\langle e_i, a_k \rangle \cdot \langle a_k, v \rangle| \leq |\langle e_i, a_k \rangle|\|a_{k,T}\|_{\ell_2}$. Since $|\langle e_i, a_k \rangle| \leq \sqrt{\mu(F)}$ and $\|a_{k,T}\|_{\ell_2} \leq \sqrt{\mu(F)s}$, we have $|w_k| \leq \mu(F)\sqrt{s}$. Lastly, for the total variance, we have

$$\mathbb{E} w_k^2 \leq \mu(F)\mathbb{E}\langle a_{k,T}, v \rangle^2 = \mu(F)$$

where the equality follows from the isotropy property. Hence, $\sigma^2 \leq m\mu(F)$, and Bernstein's inequality gives

$$\mathbb{P}(|\langle e_i, A^* Av \rangle| \geq t) \leq 2 \exp\left(-\frac{m}{2\mu(F)} \cdot \frac{t^2}{1 + \frac{1}{3}\sqrt{st}}\right).$$

Combine this with the union bound over all $i \in T^c$ to give the desired result. ∎

We also require the following related bound:

$$\max_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} \leq \delta.$$

In other words, none of the column vectors of A outside of the support of x should be well approximated by *any* vector sharing the support of x.

---

[2]To extend this Lemma to the complex case, one can split $A_{T^c}^* Av$ into real and complex parts, treating each separately with Bernstein's inequality (similarly to [17, Lemma 8.2]).

Lemma 2.6 (E4: Uniform Off-Support Incoherence): Let T be a fixed set of cardinality s. For any $0 \leq t \leq \sqrt{s}$,

$$\mathbb{P}\left(\max_{i \in T^c}\|A_T^* A_{\{i\}}\|_{\ell_2} \geq t\right) \leq n \exp\left(-\frac{mt^2}{8\mu(F)s} + \frac{1}{4}\right).$$

In particular, if $m \geq 8\mu(F) \cdot s \cdot (2\log n + 1/4)$, then

$$\mathbb{P}\left(\max_{i \in T^c}\|A_T^* A_{\{i\}}\|_{\ell_2} \geq 1\right) \leq 1/n.$$

Proof: We apply the vector Bernstein inequality (Theorem 2.4). Fix $i \in T^c$ and write

$$A_T^* A_{\{i\}} = \frac{1}{m}\sum_{k=1}^{m} a_{k,T}^* \langle a_k, e_i \rangle := \frac{1}{m}\sum_{k=1}^{m} v_k.$$

As before, $\mathbb{E}v_k = \mathbb{E}a_{k,T}^*\langle a_k, e_i \rangle = 0$ since $i \in T^c$. Also, $\|v_k\|_{\ell_2} = \|a_{k,T}\|_{\ell_2}|\langle a_k, e_i \rangle| \leq \mu(F)\sqrt{s}$. Lastly, we calculate the sum of expected squared norms

$$\sum_{k=1}^{m} \mathbb{E}\|v_k\|_{\ell_2}^2 = m\mathbb{E}\|v_1\|_{\ell_2}^2 \leq m\mathbb{E}\left[\|a_{1,T}\|_{\ell_2}^2\langle e_i, a_1\rangle^2\right]$$

$$\leq m\mu(F)s \cdot \mathbb{E}\langle e_i, a_1\rangle^2 = m\mu(F)s.$$

As before, the last equality follows from the isotropy property. Bernstein's inequality together with the union bound give the lemma. ∎

C. Weak RIP

In the nonsparse and noisy setting, we shall make use of a variation on the restricted isometry property to control the size of the reconstruction error. This variation is as follows:

Theorem 2.7 (E5: Weak RIP): Let T be a fixed set of cardinality s, fix $\delta > 0$, let r be a natural number, and suppose that

$$m \geq C_\delta \cdot \beta \cdot \mu(F) \cdot \max(s \log m, r \log n \log^2 r \log m). \tag{II.5}$$

Then with probability at least $1 - 5e^{-\beta}$

$$(1-\delta)\|v\|_{\ell_2}^2 \leq \|Av\|_{\ell_2}^2 \leq (1+\delta)\|v\|_{\ell_2}^2 \tag{II.6}$$

for all v supported on $T \cup R$, where R is any set of cardinality $|R| \leq r$. Here C_δ is a fixed numerical constant which only depends upon δ.

Note that we could equivalently replace the right-hand side of (II.5) with

$$C_\delta \cdot \beta \cdot \mu \cdot \max(s \log(\beta s \mu), r \log n \log^2 r \log(\beta r \mu))$$

provided that $\beta r \mu \geq C \log n$.

This theorem is proved in the Appendix using Talagrand's generic chaining[3], and combines the framework and results of

Rudelson and Vershynin in [46] and [44]. In the proof of Theorem 1.2, we take $\delta = 1/4$.

The condition says that the column space of A_T should not be too close to that spanned by another small disjoint set R of columns. To see why a condition of this nature is necessary for any recovery algorithm, suppose that x has fixed support T and that there is a single column $A_{\{i\}}$ which is a linear combination of columns in T, i.e., $A_{T \cup \{i\}}$ is singular. Let $h \neq 0$ be supported on $T \cup \{i\}$ and in the null space of A. Then $Ax = A(x + th)$ for any scalar t. Clearly, there are some values of t such that $x + th$ is at least as sparse as x, and thus one should not expect to be able to recover x by any method. In general, if there were a vector v as above obeying $\|Av\|_{\ell_2} \ll \|v\|_{\ell_2}$ then one would have $A_T v_T \approx -A_R v_R$. Thus, if the signal x were the restriction of v to T, it would be very difficult to distinguish it from that of $-v$ to R under the presence of noise.

The weak RIP is a combination of the RIP and the local conditioning estimate **E1**. When $r = 0$, this is **E1** whereas this is the restricted isometry property when $s = 0$. The point is that we do not need the RIP to hold for sparsity levels on the order of $m/[\mu(F)\log n]$. Instead we need the following property: consider an arbitrary submatrix formed by concatenating columns in T with r other columns from A selected in any way you like; then we would like this submatrix to be well conditioned. Because T is fixed, one can prove good conditioning when s is significantly larger than the maximum sparsity level considered in the standard RIP.

D. Implications

The careful reader may ask why we bothered to state estimates **E1**–**E4** since they are all implied by the weak RIP! Our motivation is threefold: 1) some of these estimates, e.g., **E2** hold with better constants and weaker requirements than those implied by the weak RIP machinery; 2) the weak RIP requires an in-depth proof whereas the other estimates are simple applications of well-known theorems, and we believe that these theorems and the estimates should be independently useful tools to other researchers in the field; and 3) the noiseless theorem does not require the weak RIP.

III. Noiseless and Sparse Recovery

For simplicity, we state our model and the proofs for the real case, but it is straightforward to generalize our results to complex signals, noise, and measurements, while only changing the constants involved in the proofs. The inner-products involved in several of the bounds below would be replaced by their absolute value or their real value. In particular, Lemmas 4.4 and 4.3 would have a slightly different presentation.

This section proves the noiseless recovery theorem, namely, Theorem 1.1. Our proof essentially adapts the arguments of David Gross [26] from the low-rank matrix recovery problem.

A. Dual Certificates

The standard method for establishing exact recovery is to exhibit a *dual certificate*; that is to say, a vector v obeying the two properties below.

---

[3]We use the generic chaining, which provides a way of invoking the majorizing measure theorem without creating a measure. A simpler approach would be to go through Dudley's inequality—but according to our proof the majorizing measures theorem provides a tighter result by a logarithmic factor.

Lemma 3.1 (Exact Duality): Set $T = \operatorname{supp}(x)$ with x feasible for (I.9), and assume A_T has full column rank. Suppose there exists $v \in \mathbb{R}^n$ in the row space of A obeying

$$v_T = \operatorname{sgn}(x_T) \quad \text{and} \quad \|v_{T^c}\|_{\ell_\infty} < 1. \qquad \text{(III.1)}$$

Then x is the unique ℓ_1 minimizer to (I.9).

The proof is now standard, see [15]. Roughly, the existence of a dual vector implies that there is a subgradient of the ℓ_1 norm at x that is perpendicular to the feasible set. This geometric property shows that x is solution. Following Gross, we slightly modify this definition as to make use of an "inexact dual vector."

Lemma 3.2 (Inexact Duality): Set $T = \operatorname{supp}(x)$ where x is feasible, and assume that

$$\|(A_T^* A_T)^{-1}\| \leq 2 \quad \text{and} \quad \max_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} \leq 1. \quad \text{(III.2)}$$

Suppose there exists $v \in \mathbb{R}^n$ in the row space of A obeying

$$\|v_T - \operatorname{sgn}(x_T)\|_{\ell_2} \leq 1/4 \quad \text{and} \quad \|v_{T^c}\|_{\ell_\infty} \leq 1/4. \qquad \text{(III.3)}$$

Then x is the unique ℓ_1 minimizer to (I.9).

Proof: Let $\hat{x} = x + h$ be a solution to (I.9) and note that $Ah = 0$ since both x and $\hat{x}$ are feasible. To prove the claim, it suffices to show that $h = 0$. We begin by observing that

$$\begin{aligned}\|\hat{x}\|_{\ell_1} &= \|x_T + h_T\|_{\ell_1} + \|h_{T^c}\|_{\ell_1} \\ &\geq \|x_T\|_{\ell_1} + \langle \operatorname{sgn}(x_T), h_T \rangle + \|h_{T^c}\|_{\ell_1}.\end{aligned}$$

Letting $v = A^* w$ be our (inexact) dual vector, we have

$$\begin{aligned}\langle \operatorname{sgn}(x_T), h_T \rangle &= \langle \operatorname{sgn}(x_T) - v_T, h_T \rangle + \langle v_T, h_T \rangle \\ &= \langle \operatorname{sgn}(x_T) - v_T, h_T \rangle - \langle v_{T^c}, h_{T^c} \rangle\end{aligned}$$

where we used $\langle v_T, h_T \rangle = \langle v, h \rangle - \langle v_{T^c}, h_{T^c} \rangle = -\langle v_{T^c}, h_{T^c} \rangle$ since $\langle v, h \rangle = \langle w, Ah \rangle = 0$. The Cauchy-Schwarz inequality together with the properties of v yield

$$|\langle \operatorname{sgn}(x_T), h_T \rangle| \leq \frac{1}{4}(\|h_T\|_{\ell_2} + \|h_{T^c}\|_{\ell_1})$$

and, therefore

$$\|\hat{x}\|_{\ell_1} \geq \|x\|_{\ell_1} - \frac{1}{4}\|h_T\|_{\ell_2} + \frac{3}{4}\|h_{T^c}\|_{\ell_1}.$$

We now bound $\|h_T\|_{\ell_2}$. First, it follows from

$$h_T = (A_T^* A_T)^{-1} A_T^* A_T h_T = -(A_T^* A_T)^{-1} A_T^* A_{T^c} h_{T^c}$$

that $\|h_T\|_{\ell_2} \leq 2\|A_T^* A_{T^c} h_{T^c}\|_{\ell_2}$. Second

$$\begin{aligned}\|A_T^* A_{T^c} h_{T^c}\|_{\ell_2} &\leq \sum_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} |h_i| \\ &\leq \max_{i \in T^c} \|A_T^* A_{\{i\}}\|_{\ell_2} \|h_{T^c}\|_{\ell_1} \leq \|h_{T^c}\|_{\ell_1}.\end{aligned}$$

In conclusion, $\|h_T\|_2 \leq 2\|h_{T^c}\|_1$ and thus

$$\|\hat{x}\|_{\ell_1} \geq \|x\|_{\ell_1} + \frac{1}{4}\|h_{T^c}\|_{\ell_1}.$$

This implies $h_{T^c} = 0$, which in turn implies $h_T = 0$ since we must have $A_T h_T = Ah = 0$ (and A_T has full rank). ∎

Lemma 3.3 (Existence of a Dual Certificate): Under the hypotheses of Theorem 1.1, one can find $v \in \mathbb{R}^n$ obeying the conditions of Lemma 3.2 with probability at least $1 - e^{-\beta} - 1/n$.

This lemma, which is proved next, implies Theorem 1.1. The reason is that we just need to verify conditions (III.2). However, by Lemmas 2.1 and 2.6, they jointly hold with probability at least $1 - 3/n$ provided that $m \geq \mu \cdot s \cdot (19 \log n + 2)$ (recall that μ is a shorthand for $\mu(F)$).

B. Proof of Lemma 3.3

The proof uses the clever *golfing scheme* introduced in [26]. Partition A into row blocks so that from now on, A_1 are the first m_1 rows of the matrix A, A_2 the next m_2 rows, and so on. The ℓ matrices $\{A_i\}_{i=1}^\ell$ are independently distributed, and we have $m_1 + m_2 + \ldots + m_\ell = m$. As before, $A_{i,T}$ is the restriction of A_i to the columns in T.

The golfing scheme then starts with $v_0 = 0$, inductively defines

$$v_i = \frac{m}{m_i} A_i^* A_{i,T}(\operatorname{sgn}(x_T) - v_{i-1,T}) + v_{i-1}$$

for $i = 1, \ldots, \ell$, and sets $v = v_\ell$. Clearly, v is in the row space of A. To simplify notation, let $q_i = \operatorname{sgn}(x_T) - v_{i,T}$, and observe the two identities

$$q_i = \left(I - \frac{m}{m_i} A_{i,T}^* A_{i,T}\right) q_{i-1} \qquad \text{(III.4)}$$

$$= \prod_{j=1}^i \left(I - \frac{m}{m_j} A_{j,T}^* A_{j,T}\right) \operatorname{sgn}(x_T) \qquad \text{(III.5)}$$

and

$$v = \sum_{i=1}^\ell \frac{m}{m_i} A_i^* A_{i,T} q_{i-1} \qquad \text{(III.6)}$$

which shall be used frequently. From (III.4) and the fact that $I - \frac{m}{m_i} A_{i,T}^* A_{i,T}$ should be a contraction (local isometry **E1**), we see that the norm of q_i decreases geometrically fast—the terminology comes from this fact since each iteration brings us closer to the target just as each golf shot would bring us closer to the hole—so that v_T should be close to $\operatorname{sgn}(x_T)$. Hopefully, the process keeps the size of v_{T^c} under control as well.

To control the size of v_{T^c} and that of $\operatorname{sgn}(x_T) - v_T$, we claim that the following inequalities hold for each i with high probability: first,

$$\|q_i\|_{\ell_2} \leq c_i \|q_{i-1}\|_{\ell_2} \qquad \text{(III.7)}$$

and, second,

$$\left\|\frac{m}{m_i} A_{i,T^c}^* A_{i,T} q_{i-1}\right\|_{\ell_\infty} \leq t_i \|q_{i-1}\|_{\ell_2} \qquad \text{(III.8)}$$

(the values of the parameters t_i and c_i will be specified later). Let $p_1(i)$ (resp. $p_2(i)$) be the probability that the bound (III.7) (resp. (III.8)) does not hold. Lemma 2.3 (Estimate **E2**) gives

$$p_1(i) \leq \exp\left(-\frac{1}{4}(c_i \sqrt{m_i/(s\mu)} - 1)^2\right). \qquad \text{(III.9)}$$

Thus, if

$$m_i \geq \frac{2 + 8(\beta + \log \alpha)}{c_i^2} s\mu \qquad \text{(III.10)}$$

then $p_1(i) \leq \frac{1}{\alpha} e^{-\beta}$. Next, Lemma 2.5 (Estimate **E3**) gives

$$p_2(i) \leq 2n \exp\left(-\frac{3t_i^2 m_i}{6\mu + 2\mu\sqrt{s}t_i}\right). \qquad \text{(III.11)}$$

Thus, if

$$m_i \geq \left(\frac{2}{t_i^2 s} + \frac{2}{3t_i\sqrt{s}}\right)(\beta + \log(2\alpha) + \log n)s\mu \quad \text{(III.12)}$$

then $p_2(i) \leq \frac{1}{\alpha}e^{-\beta}$.

It is now time to set the number of blocks ℓ, the block sizes m_i and the values of the parameters c_i and t_i. These are as follows:

- $\ell = \lceil (\log_2 s)/2 \rceil + 2$;
- $c_1 = c_2 = 1/[2\sqrt{\log n}]$ and $c_i = 1/2$ for $3 \leq i \leq \ell$;
- $t_1 = t_2 = 1/[8\sqrt{s}]$ and $t_i = \log n/[8\sqrt{s}]$ for $3 \leq i \leq \ell$;
- $m_1, m_2 \geq 35(1 + \log 4 + \beta)s\mu c_i^{-2}$ and $m_i \geq 35(1 + \log 6 + \beta)s\mu c_i^{-2}$ for $3 \leq i \leq \ell$.

Note that under the assumptions of the lemma, $m \geq \sum_i m_i$ (extra rows need not be used in the construction of the dual certificate). To see why v is a valid certificate, suppose first that for each i, (III.7) and (III.8) hold. Then (III.4) gives

$$\|\mathrm{sgn}(x_T) - v_T\|_{\ell_2} = \|q_\ell\|_{\ell_2} \leq \|\mathrm{sgn}(x_T)\|_{\ell_2} \prod_{i=1}^{\ell} c_i \leq \frac{\sqrt{s}}{2^\ell}$$

which is bounded by $\frac{1}{4}$ as desired. Further, (III.6) yields

$$\|v_{T^c}\|_{\ell_\infty} \leq \sum_{i=1}^{\ell} \left\| \frac{m}{m_i} A_{i,T^c}^* A_{i,T} q_{i-1} \right\|_{\ell_\infty}$$
$$\leq \sum_{i=1}^{\ell} t_i \|q_{i-1}\|_{\ell_2} \leq \sqrt{s} \sum_{i=1}^{\ell} t_i \prod_{j=1}^{i-1} c_i.$$

(For $i = 0$, we take $\prod_{j=1}^{i-1} c_i = 1$.) Now with our choice of parameters, the right-hand side is bounded above by

$$\frac{1}{8}\left(1 + \frac{1}{2\sqrt{\log n}} + \frac{\log n}{4\log n} + \cdots\right) < \frac{1}{4}$$

which is the desired conclusion.

Now we must show that the bounds (III.7), (III.8) hold with probability at least $1 - e^{-\beta} - 1/n$. It follows from (III.10) and (III.12) that $p_1(i), p_2(i) \leq \frac{1}{4}e^{-\beta}$ for $i = 1, 2$ and $p_1(i), p_2(i) \leq \frac{1}{6}e^{-\beta} \leq 1/6$ for $i \geq 3$. Thus, $p_1(1) + p_1(2) + p_2(1) + p_2(2) \leq e^{-\beta}$ and $p_1(i) + p_2(i) \leq 1/3$ for $i \geq 3$. Now the union bound would never show that (III.7) and (III.8) hold with probability at least $1 - 1/n$ for all $i \geq 3$ because of the weak bound on $p_1(i) + p_2(i)$. However, using a clever idea in [26], it is not necessary for each subset of rows A_i to 'succeed' and give the desired bounds. Instead, one can sample a 'few' extra batches of rows, and throw out those that fail our requirements. We only need $\ell - 2$ working batches, after the first 2. In particular, pick $\ell' + 2 > \ell$ batches of rows, so that we require $m \geq 2 \cdot \lceil 140(1 + \log 4 + \beta) \cdot$

$\mu \cdot s \cdot \log n \rceil + \ell \cdot \lceil 140(1 + \log 6 + \beta)s\mu \rceil$ (note that we have made no attempt to optimize constants). Now as in [26], let N be the the number of batches—after the first 2—obeying (III.7) and (III.8); this N is larger (probabilistically) than a binomial $(\ell', 2/3)$ random variable. Then by Hoeffding's inequality [31, Theorem 2.3a]

$$\mathbb{P}(N < \ell - 2) \leq \exp\left(-2\frac{\left(\frac{2}{3}\ell' - \ell + 2\right)^2}{\ell'}\right)$$

and thus if we pick $\ell' = 3\lceil \log n \rceil + 1$, we have

$$\mathbb{P}(N < \ell - 2) \leq 1/n.$$

In summary, from $p_1(1) + p_2(1) + p_1(2) + p_2(2) \leq e^{-\beta}$ and the calculation above, the dual certificate v obeys the required properties with probability at least $1 - 1/n - e^{-\beta}$, provided that $m \geq C(1 + \beta) \cdot \mu \cdot s \cdot \log n$.

IV. GENERAL SIGNAL RECOVERY FROM NOISY DATA

We prove the general recovery theorems from Section I-F under the assumption of Gaussian white noise but would like to emphasize that the same result would hold for other noise distributions. Specifically, suppose we have the noisy model

$$y = Ax + z, \quad \text{where} \quad \|A^* z\|_{\ell_\infty} \leq \lambda_n \qquad \text{(IV.1)}$$

holds with high probability. Then the conclusions of Theorem 1.3 remain valid. In details, the Dantzig selector with constraint $\|A^*(y - A\bar{x})\|_{\ell_\infty} \leq 4\lambda_n$ obeys

$$\|\hat{x} - x\|_{\ell_2} \leq C_1(1 + \alpha^2)\left[\frac{\|x - x_s\|_{\ell_1}}{\sqrt{s}} + \lambda_n\sqrt{s}\right] \qquad \text{(IV.2)}$$

with high probability. Hence, (I.15) is a special case corresponding to $\lambda_n = 2.5\sigma_m\sqrt{\log n} = 2.5\sigma\sqrt{\frac{\log n}{m}}$. Likewise, the bound on the ℓ_1 loss (I.16) with λ_n in place of $\sigma\sqrt{\frac{\log n}{m}}$ holds as well. A similar generality applies to the LASSO as well, although in this case we need a second noise correlation bound, namely,

$$\|A_{T^c}^*(I - P)z\|_{\ell_\infty} \leq \lambda_n,$$

where $P := A_T(A_T^* A_T)^{-1} A_T^*$ is the projection onto the range of A_T.

Now when $z \sim \mathcal{N}(0, I)$ and A is a fixed matrix, we have

$$\|A^* z\|_{\ell_\infty} \leq 2\|A\|_{1,2}\sqrt{\log n} \qquad \text{(IV.3)}$$

with probability at least $1 - 1/2n$; recall that $\|A\|_{1,2}$ is the maximum column norm of A. Indeed, the ith component of $A^* z$ is distributed as $\mathcal{N}(0, \|A_{\{i\}}\|_{\ell_2}^2)$ and, therefore, the union bound gives

$$\mathbb{P}(\|A^* z\|_{\ell_\infty} > 2\|A\|_{1,2}\sqrt{\log n})$$
$$\leq n\mathbb{P}(|\mathcal{N}(0,1)| > 2\sqrt{\log n}).$$

The conclusion follows for $n \geq 2$ from the well-known tail bound $\mathbb{P}(|\mathcal{N}(0,1)| > t) \leq 2\phi(t)/t$, where ϕ is the density of

the standard normal distribution. The same steps demonstrate that

$$\|A^*(I - P)z\|_{\ell_\infty} \leq 2\|(I - P)A\|_{1,2}\sqrt{\log n}$$
$$\leq 2\|A\|_{1,2}\sqrt{\log n} \qquad \text{(IV.4)}$$

with probability at least $1 - 1/2n$.

A. Proof of Theorem 1.2

We begin with a few simplifying assumptions. First, we assume in the proof that $\sigma_m = 1$ since the general result follows from a simple rescaling. Second, because we are interested in situations where m is much smaller than n, we assume for simplicity of presentation that $m \leq n$ although our results extend with only a change to the numerical constants involved if $m \leq n^{O(1)}$. In truth, they extend without any assumption on the relation between m and n, but the general presentation becomes a bit more complicated.

Fix s obeying $s \leq \bar{s}$, and let $T = \text{supp}(x_s)$. We prove the error bounds of Theorem 1.2 with s fixed, and the final result follows by considering that s which minimizes either the ℓ_2 (I.11) or ℓ_1 (I.12) error bound. This is proper since the minimizing s has a deterministic value. With T as above, we assume in the rest of the proof that:

(i) all of the requirements for noiseless recovery in Lemma 3.2 are met (and in particular, the inexact dual vector is constructed for x_T);

(ii) and that the inexact dual vector v of Section III is successfully constructed.

All of this occurs with probability at least $1 - 4/n - e^{-\beta}$. Further, we assume that:

(iii) the weak RIP holds with $\delta = 1/4$, $r = \frac{m}{C(1+\beta)\cdot\mu\cdot\log n \log m \log^2 s} \vee s$ and T is as above.

This occurs with probability at least $1 - 5e^{-\beta}$, and implies the RIP at sparsity level r and restricted isometry constant $\delta = 1/4$. Lastly, we assume:

(iv) the noise correlation bound

$$\|A^*z\|_{\ell_\infty} \leq 2.5\sqrt{\log n}. \qquad \text{(IV.5)}$$

Assuming the weak RIP above (Estimate **E5**), which implies $\|A\|_{1,2} \leq 5/4$, the conditional probability that this occurs is at least $1 - 1/2n$ because of (IV.3). Because the weak RIP implies the local isometry condition **E1** with $\delta = 1/4$, all of these conditions together hold with probability at least $1 - 4/n - 6e^{-\beta}$. All of the steps in the proof are now deterministic consequences of (i)–(iv); from now on, we will assume they hold.

With $h = \hat{x} - x$, our goal is to bound both the ℓ_2 and ℓ_1 norms of h. We will do this with a pair of lemmas. The first is frequently used (recall that λ is set to $10\sqrt{\log n}$).

Lemma 4.1 (Tube Constraint): The error h obeys

$$\|A^*Ah\|_{\ell_\infty} \leq \frac{5\lambda}{4}.$$

Proof: As shown in [9, Lemma 3.1], writing that the zero vector is a subgradient of the LASSO functional $\frac{1}{2}\|y - A\bar{x}\|_{\ell_2}^2 + \lambda\|\bar{x}\|_{\ell_1}$ at $\bar{x} = \hat{x}$ gives

$$\|A^*(y - A\hat{x})\|_{\ell_\infty} \leq \lambda.$$

Then it follows from the triangle inequality that

$$\|A^*Ah\|_{\ell_\infty} \leq \|A^*(y - A\hat{x})\|_{\ell_\infty} + \|A^*z\|_{\ell_\infty} \leq \lambda + \|A^*z\|_{\ell_\infty}$$

where z is our noise term. The claim is a consequence of (IV.5). ∎

Lemma 4.2: The error h obeys

$$\|h_{T^c}\|_{\ell_1} \leq C_0(s\lambda + \|x_{T^c}\|_{\ell_1}) \qquad \text{(IV.6)}$$

for some numerical constant C_0.

Before proving this lemma, we show that it gives Theorem 1.2. Some of the steps are taken from the proof of [16, Theorem 1.1].

Proof of Theorem 1.2: Set r as in (iii) above. We begin by partitioning T^c and let T_1 be the indices of the r largest entries of h_{T^c}, T_2 be those of the next r largest, and so on. We first bound $\|h_{T \cup T_1}\|_{\ell_2}$ and set $\bar{T}_1 = T \cup T_1$ for short. The weak RIP assumption (iii) gives

$$\frac{3}{4}\|h_{\bar{T}_1}\|_{\ell_2}^2 \leq \|A_{\bar{T}_1}h_{\bar{T}_1}\|_{\ell_2}^2$$
$$= \langle A_{\bar{T}_1}h_{\bar{T}_1}, Ah \rangle - \langle A_{\bar{T}_1}h_{\bar{T}_1}, A_{\bar{T}_1^c}h_{\bar{T}_1^c} \rangle. \text{(IV.7)}$$

From Lemma 4.1, we have

$$\langle A_{\bar{T}_1}h_{\bar{T}_1}, Ah \rangle = \langle h_{\bar{T}_1}, A_{\bar{T}_1}^*Ah \rangle$$
$$\leq \|h_{\bar{T}_1}\|_{\ell_1}\|A_{\bar{T}_1}^*Ah\|_{\ell_\infty} \leq \frac{5}{4}\lambda\|h_{\bar{T}_1}\|_{\ell_1}.$$

Since $\bar{T}_1$ has cardinality at most $2s$, the Cauchy-Schwarz inequality gives

$$\langle A_{\bar{T}_1}h_{\bar{T}_1}, Ah \rangle \leq \frac{5}{4}\lambda\sqrt{2s}\|h_{\bar{T}_1}\|_{\ell_2}. \qquad \text{(IV.8)}$$

Next, we bound $|\langle A_{\bar{T}_1}h_{\bar{T}_1}, A_{\bar{T}_1^c}h_{\bar{T}_1^c} \rangle| \leq |\langle A_Th_T, A_{\bar{T}_1^c}h_{\bar{T}_1^c} \rangle| + |\langle A_{T_1}h_{T_1}, A_{\bar{T}_1^c}h_{\bar{T}_1^c} \rangle|$. We have

$$\langle A_Th_T, A_{\bar{T}_1^c}h_{\bar{T}_1^c} \rangle \leq \sum_{j\geq 2}|\langle A_Th_T, A_{T_j}h_{T_j} \rangle|. \qquad \text{(IV.9)}$$

As shown in [14, Lemma 1.2], the parallelogram identity together with the weak RIP imply that

$$|\langle A_Th_T, A_{T_j}h_{T_j} \rangle| \leq \frac{1}{4}\|h_T\|_{\ell_2}\|h_{T_j}\|_{\ell_2}$$

and, therefore

$$\langle A_Th_T, A_{\bar{T}_1^c}h_{\bar{T}_1^c} \rangle \leq \frac{1}{4}\|h_T\|_{\ell_2}\sum_{j\geq 2}\|h_{T_j}\|_{\ell_2}. \qquad \text{(IV.10)}$$

To bound the summation, we use the now standard result [16, (3.10)]

$$\sum_{j \geq 2} \left\| h_{T_j} \right\|_{\ell_2} \leq r^{-1/2} \left\| h_{T^c} \right\|_{\ell_1} \qquad \text{(IV.11)}$$

which gives

$$\left| \langle A_T h_T, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle \right| \leq \frac{1}{4} r^{-1/2} \| h_T \|_{\ell_2} \| h_{T^c} \|_{\ell_1}.$$

The same analysis yields $\left| \langle A_{T_1} h_{T_1}, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle \right| \leq \frac{1}{4} r^{-1/2} \| h_{T_1} \|_{\ell_2} \| h_{T^c} \|_{\ell_1}$ and thus

$$\left| \langle A_{\bar{T}_1} h_{\bar{T}_1}, A_{\bar{T}_1^c} h_{\bar{T}_1^c} \rangle \right| \leq \frac{\sqrt{2}}{4} r^{-1/2} \left\| h_{\bar{T}_1} \right\|_{\ell_2} \| h_{T^c} \|_{\ell_1}.$$

Plugging these estimates into (IV.7) gives

$$\left\| h_{\bar{T}_1} \right\|_{\ell_2} \leq \frac{5}{3} \sqrt{2s} \lambda + \frac{\sqrt{2}}{3} r^{-1/2} \| h_{T^c} \|_{\ell_1}. \qquad \text{(IV.12)}$$

The conclusion is now one step away. Obviously

$$\| h \|_{\ell_2} \leq \left\| h_{\bar{T}_1} \right\|_{\ell_2} + \sum_{j \geq 2} \left\| h_{T_j} \right\|_{\ell_2} \leq \left\| h_{\bar{T}_1} \right\|_{\ell_2} + r^{-1/2} \| h_{T^c} \|_{\ell_1}$$

$$\leq \frac{5}{3} \sqrt{2s} \lambda + \frac{3 + \sqrt{2}}{3} r^{-1/2} \| h_{T^c} \|_{\ell_1}$$

where the second line follows from (IV.12). Lemma 4.2 completes the proof for the ℓ_2 error. For the ℓ_1 error, note that by the Cauchy-Schwarz inequality

$$\| h \|_{\ell_1} = \| h_T \|_{\ell_1} + \| h_{T^c} \|_{\ell_1} \leq \sqrt{s} \| h_T \|_{\ell_2} + \| h_{T^c} \|_{\ell_1}$$
$$\leq \sqrt{s} \| h_{\bar{T}_1} \|_{\ell_2} + \| h_{T^c} \|_{\ell_1}.$$

Combine this with (IV.12) and Lemma 4.2.

(Note that if we had not set $\sigma_m = 1$ at the beginning of the proof, we would have $\lambda = 10 \sigma \frac{\sqrt{\log n}}{m}$, leading to the general error bound.) ∎

B. Proof of Lemma 4.2

Since $\hat{x}$ is the minimizer to (I.10), we have

$$\frac{1}{2} \| A\hat{x} - y \|_{\ell_2}^2 + \lambda \| \hat{x} \|_{\ell_1} \leq \frac{1}{2} \| Ax - y \|_{\ell_2}^2 + \lambda \| x \|_{\ell_1},$$

which can be massaged into the more convenient form

$$\frac{1}{2} \| Ah \|_{\ell_2}^2 + \lambda \| \hat{x} \|_{\ell_1} \leq \langle Ah, z \rangle + \lambda \| x \|_{\ell_1}.$$

We will manipulate this inequality to give the desired bound on $\| h_{T^c} \|_{\ell_1}$. We begin with the following Lemma.

Lemma 4.3: $\| \hat{x} \|_{\ell_1}$ can be lower bounded as follows:

$$\| \hat{x} \|_{\ell_1} \geq \| x \|_{\ell_1} + \langle h_T, sgn(x_T) \rangle + \| h_{T^c} \|_{\ell_1} - 2 \| x_{T^c} \|_{\ell_1}.$$

Proof: We have $\| \hat{x} \|_{\ell_1} = \| x_T + h_T \|_1 + \| x_{T^c} + h_{T^c} \|_1 \geq \langle x_T + h_T, sgn(x_T) \rangle + \| x_{T^c} + h_{T^c} \|_{\ell_1}$ and the claim follows from the triangle inequality. ∎

It follows from this that

$$\frac{1}{2} \| Ah \|_{\ell_2}^2 + \lambda \| h_{T^c} \|_{\ell_1}$$
$$\leq \langle Ah, z \rangle - \lambda \langle h_T, sgn(x_T) \rangle + 2\lambda \| x_{T^c} \|_{\ell_1}. \qquad \text{(IV.13)}$$

The proof of Lemma 4.2 follows by plugging the bounds given in Lemmas 4.4 and 4.5 below back into (IV.13).

Lemma 4.4: $\langle Ah, z \rangle$ can be upper bounded as follows:

$$\langle Ah, z \rangle \leq \frac{5}{12} s \lambda^2 + \frac{\lambda}{4} \| h_{T^c} \|_{\ell_1}. \qquad \text{(IV.14)}$$

Proof: The proof is similar to an argument in [9]. Let $P = A_T (A_T^* A_T)^{-1} A_T^*$ be the orthogonal projection onto the range of A_T. Then

$$\langle Ah, z \rangle = \langle PAh, z \rangle + \langle (I - P)A_{T^c} h_{T^c}, z \rangle$$
$$= \langle A_T^* Ah, (A_T^* A_T)^{-1} A_T^* z \rangle + \langle h_{T^c}, A_{T^c}^* (I - P) z \rangle$$
$$\leq \| A_T^* Ah \|_{\ell_\infty} \left\| (A_T^* A_T)^{-1} A_T^* z \right\|_{\ell_1}$$
$$+ \| h_{T^c} \|_{\ell_1} \| A_{T^c}^* (I - P) z \|_{\ell_\infty}$$
$$\leq \frac{5}{4} \lambda \left\| (A_T^* A_T)^{-1} A_T^* z \right\|_{\ell_1} + 2.5 \sqrt{\log n} \| h_{T^c} \|_{\ell_1}. \qquad \text{(IV.15)}$$

The last line follows from Lemma 4.1, (IV.4), and the fact that $\|A\|_{1,2} \leq 5/4$ (this is an easy consequence of the RIP holding with $\delta = 1/4$). We now bound the first term, and write

$$\left\| (A_T^* A_T)^{-1} A_T^* z \right\|_{\ell_1} \leq \sqrt{s} \left\| (A_T^* A_T)^{-1} A_T^* z \right\|_{\ell_2}$$
$$\leq \frac{4}{3} \sqrt{s} \| A_T^* z \|_{\ell_2}$$
$$\leq \frac{4}{3} s \| A_T^* z \|_{\ell_\infty} \leq \frac{1}{3} s \lambda. \qquad \text{(IV.16)}$$

The first inequality follows from Cauchy-Schwarz, the second from $\| (A_T^* A_T)^{-1} \| \leq 4/3$, and the fourth from $\| A^* z \|_{\ell_\infty} \leq \lambda/4$ (see (IV.5)). Inequality (IV.15) establishes the claim. ∎

Lemma 4.5:

$$|\langle h_T, sgn(x_T) \rangle| \leq Cs\lambda + \frac{7}{12} \| h_{T^c} \|_{\ell_1} + \frac{1}{2\lambda} \| Ah \|_{\ell_2}^2. \qquad \text{(IV.17)}$$

Proof: Let v be the inexact dual vector, and decompose $\langle h_T, \text{sgn}(x_T) \rangle$ as

$$|\langle h_T, \text{sgn}(x_T) \rangle| \leq |\langle h_T, \text{sgn}(x_T) - v_T \rangle| + |\langle h_T, v_T \rangle|$$
$$\leq |\langle h_T, \text{sgn}(x_T) - v_T \rangle| + |\langle h, v \rangle| + |\langle h_{T^c}, v_{T^c} \rangle|. \qquad \text{(IV.18)}$$

First

$$|\langle h_T, \text{sgn}(x_T) - v_T \rangle| \leq \| h_T \|_{\ell_2} \| \text{sgn}(x_T) - v_T \|_{\ell_2}$$
$$\leq \frac{1}{4} \| h_T \|_{\ell_2}.$$

Now

$$\| h_T \|_{\ell_2} \leq \| (A_T^* A_T)^{-1} \| \| A_T^* A_T h_T \|_{\ell_2} \leq \frac{4}{3} \| A_T^* A_T h_T \|_{\ell_2}$$
$$\leq \frac{4}{3} \| A_T^* Ah \|_{\ell_2} + \frac{4}{3} \| A_T^* A_{T^c} h_{T^c} \|_{\ell_2}$$
$$\leq \frac{4}{3} \sqrt{s} \| A_T^* Ah \|_{\ell_\infty} + \frac{4}{3} \| h_{T^c} \|_{\ell_1} \max_{j \in T^c} \| A_T^* A_{\{j\}} \|_{\ell_2}$$
$$\leq \frac{5}{3} \sqrt{s} \lambda + \frac{4}{3} \| h_{T^c} \|_{\ell_1} \qquad \text{(IV.19)}$$

where the last line follows from Lemma 4.1 and (III.2). Second, it follows from the definition of v that

$$|\langle h_{T^c}, v_{T^c}\rangle| \le \|h_{T^c}\|_{\ell_1}\|v_{T^c}\|_{\ell_\infty} \le \frac{1}{4}\|h_{T^c}\|_{\ell_1}.$$

Hence, we established

$$|\langle h_T, \operatorname{sgn}(x_T)\rangle| \le \frac{5}{12}\sqrt{s}\lambda + \frac{7}{12}\|h_{T^c}\|_{\ell_1} + |\langle h, v\rangle|. \tag{IV.20}$$

Third, we bound $|\langle h, v\rangle|$ by Lemma 4.6 below. With the notation of this lemma

$$\begin{aligned}|\langle h, v\rangle| &= |\langle h, A^*w\rangle| = |\langle Ah, w\rangle|\\ &\le \|Ah\|_{\ell_2}\|w\|_{\ell_2} \le C_0\sqrt{s}\|Ah\|_{\ell_2}\end{aligned}$$

for some $C_0 > 0$. By (IV.23) stated below

$$\|Ah\|_{\ell_2}\sqrt{s} \le \frac{\|Ah\|_{\ell_2}^2}{2C_0\lambda} + \frac{C_0 s\lambda}{2}$$

and it follows that

$$|\langle h, v\rangle| \le \frac{C_0^2}{2}s\lambda + \frac{1}{2\lambda}\|Ah\|_{\ell_2}^2. \tag{IV.21}$$

Plugging this into (IV.20) finishes the proof. ∎

Lemma 4.6: The inexact dual certificate from Section III is of the form $v = A^*w$ where $\|w\|_{\ell_2} \le C_0\sqrt{s}$ for some positive numerical constant C_0.

Proof: For notational simplicity, assume without loss of generality that the first ℓ batches of rows were those used in constructing the inexact dual vector v (none were thrown out) so that

$$v = \sum_{i=1}^{\ell} \frac{m}{m_i} A_i^* A_{i,T}q_{i-1}.$$

Hence, $v = A^*w$ with $w^* = (w_1^*, \ldots, w_\ell^*, 0, \ldots, 0)$ and $w_i := \frac{m}{m_i}A_{i,T}q_{i-1}$ so that $\|w\|_{\ell_2}^2 = \sum_{i=1}^{\ell}\|w_i\|_{\ell_2}^2$. We have

$$\begin{aligned}&\frac{m}{m_i}\|A_{i,T}\,q_{i-1}\|_{\ell_2}^2\\ &= \left\langle \frac{m}{m_i}A_{i,T}^*A_{i,T}q_{i-1}, q_{i-1}\right\rangle\\ &= \left\langle \left(\frac{m}{m_i}A_{i,T}^*A_{i,T} - I\right)q_{i-1}, q_{i-1}\right\rangle + \|q_{i-1}\|_{\ell_2}^2\\ &\le \|q_i\|_{\ell_2}\|q_{i-1}\|_{\ell_2} + \|q_{i-1}\|_{\ell_2}^2\\ &\le 2\|q_{i-1}\|_{\ell_2}^2\\ &\le 2s\prod_{j=1}^{i-1}c_j^2.\end{aligned} \tag{IV.22}$$

The last line follows from $\|q_0\|_{\ell_2}^2 = \|\operatorname{sgn}(x_T)\|_{\ell_2}^2 = s$.
It now follows that

$$\|w\|_{\ell_2}^2 \le 2s\cdot\sum_{i=1}^{\ell}\frac{m}{m_i}\prod_{j=1}^{i-1}c_j^2.$$

Assume that $m \le C(1+\beta)\mu s\log n$ so that m is just large enough to satisfy the requirements of Theorem 1.2 (up to a constant). Then recall that $m_i \ge C(1+\beta)\mu s c_i^{-2}$. Combine the bound for m and for m_i to give $\frac{m}{m_i} \le Cc_i^2\log n$. (If m is much larger, rescale each m_i proportionally to achieve the same ratio.) This gives

$$\|w\|_{\ell_2}^2 \le Cs\log n\sum_{i=1}^{\ell}\prod_{j=1}^{i}c_j^2 \le Cs\sum_{i=1}^{\ell}\prod_{j=2}^{i}c_j^2$$

since $c_1 = (2\sqrt{\log n})^{-1}$ (we take $\prod_{j=2}^{1}c_j^2 = 1$). For $i \ge 2, \prod_{j=2}^{i}c_j^2 \le 4^{-i-1}$, and the conclusion follows. ∎

C. Proof of Theorem 1.3

Proof: Fix s and T as in Section IV-A and assume that (i)–(iv) hold. The proof parallels that for the LASSO; this is why we only sketch the important points and reuse the earlier techniques with minimal extra explanation. We shall repeatedly use the inequality

$$ab \le ca^2/2 + b^2/(2c), \tag{IV.23}$$

which holds for positive scalars a, b, c. Our first intermediate result is analogous to Lemma 4.2.
Lemma 4.7: The error $h = \hat{x} - x$ obeys

$$\|h_{T^c}\|_{\ell_1} \le C(s\lambda + \|x_{T^c}\|_{\ell_1} + \sqrt{s}\|Ah\|_{\ell_2}).$$

Proof: Since x is feasible, $\|\hat{x}\|_{\ell_1} \le \|x\|_{\ell_1}$ and it follows from Lemma 4.3 that

$$\|h_{T^c}\|_{\ell_1} \le -\langle h_T, \operatorname{sgn}(x_T)\rangle + 2\|x_{T^c}\|_{\ell_1}. \tag{IV.24}$$

We bound $|\langle h_T, \operatorname{sgn}(x_T)\rangle|$ in exactly the same way as before, but omitting the last step, and obtain

$$|\langle h_T, \operatorname{sgn}(x_T)\rangle| \le Cs\lambda + \frac{7}{12}\|h_{T^c}\|_{\ell_1} + C\sqrt{s}\|Ah\|_{\ell_2}.$$

This concludes the proof. ∎

The remainder of this section proves Theorem 1.3. Observe that $\|A^*Ah\|_{\ell_\infty} \le \frac{5}{4}\lambda$ (Lemma 4.1) since the proof is identical (we do not even need to consider subgradients). Partitioning the indices as before, one can repeat the earlier argument leading to (IV.12). Then combining (IV.12) with Lemma 4.7 gives

$$\|h_{\bar{T}_1}\|_{\ell_2} \le C\sqrt{s}\lambda + Cr^{-1/2}(s\lambda + \|x_{T^c}\|_{\ell_1} + \sqrt{s}\|Ah\|_{\ell_2}). \tag{IV.25}$$

The term proportional to $\sqrt{s/r}\|Ah\|_{\ell_2}$ in the right-hand side was not present before, and we must develop an upper bound for it. Write

$$\begin{aligned}\|Ah\|_{\ell_2}^2 &= \langle A^*Ah, h\rangle \le \|A^*Ah\|_{\ell_\infty}\|h\|_{\ell_1}\\ &\le \frac{5}{4}\lambda(\|h_T\|_{\ell_1} + \|h_{T^c}\|_{\ell_1})\end{aligned}$$

and note that (IV.24) gives

$$\|h_{T^c}\|_{\ell_1} \le \|h_T\|_{\ell_1} + 2\|x_{T^c}\|_{\ell_1}.$$

These last two inequalities yield $\|Ah\|_{\ell_2}^2 \le \frac{5}{2}\lambda(\|h_T\|_{\ell_1} + \|x_{T^c}\|_{\ell_1})$, and since $\sqrt{\lambda\|x_{T^c}\|_{\ell_1}} \le \frac{1}{2}\lambda\sqrt{s} + \frac{1}{2\sqrt{s}}\|x_{T^c}\|_{\ell_1}$ because of (IV.23), we have

$$\|Ah\|_{\ell_2} \le \sqrt{\frac{5}{2}\lambda}(\sqrt{\|h_T\|_{\ell_1}} + \sqrt{\|x_{T^c}\|_{\ell_1}})$$
$$\le \sqrt{\frac{5}{2}}\left(\sqrt{\lambda\|h_T\|_{\ell_1}} + \frac{1}{2}\lambda\sqrt{s} + \frac{1}{2\sqrt{s}}\|x_{T^c}\|_{\ell_1}\right).$$

In short

$$\left\|h_{\bar{T}_1}\right\|_{\ell_2} \le C\left(\sqrt{s}\lambda + r^{-1/2}\left(s\lambda + \|x_{T^c}\|_{\ell_1} + \sqrt{s\lambda\|h_T\|_{\ell_1}}\right)\right).$$

The extra term on the right-hand side has been transmuted into $C\sqrt{\frac{s}{r}\lambda\|h_T\|_{\ell_1}}$, which may be bounded via (IV.23) as

$$C\sqrt{\frac{s}{r}\lambda\|h_T\|_{\ell_1}} \le C^2\frac{s}{r}\sqrt{s}\lambda + \frac{1}{2\sqrt{s}}\|h_T\|_{\ell_1}$$
$$\le C^2\frac{s}{r}\sqrt{s}\lambda + \frac{1}{2}\|h_T\|_{\ell_2}.$$

Since $\|h_T\|_{\ell_2} \le \|h_{\bar{T}_1}\|_{\ell_2}$, we have

$$\left\|h_{\bar{T}_1}\right\|_{\ell_2} \le C\left(1 + \sqrt{\frac{s}{r}} + \frac{s}{r}\right)\sqrt{s}\lambda + C\frac{\|x_{T^c}\|_{\ell_1}}{\sqrt{r}}.$$

The remaining steps are the same as those in the proof for the LASSO. ∎

V. DISCUSSION

This paper developed a very simple and general theory of compressive sensing, in which sensing vectors are drawn independently at random from a probability distribution. In addition to establishing a general framework, we showed that nearly sparse signals could be accurately recovered from a small number of noisy compressive samples by means of tractable convex optimization. For example, s-sparse signals can be recovered accurately from about $s \log n$ DFT coefficients corrupted by noise. Our analysis shows that stable recovery is possible from a minimal number of samples, and improves on previously known results. This improvement comes from novel stability arguments, which do not require the restricted isometry property to hold.

We have seen that the isotropy condition is not really necessary, and it would be interesting to know the extent in which it can be relaxed. In particular, for which values of α and β obeying $\alpha I \preceq \mathbb{E} aa^* \preceq \beta I$ would our results continue to hold? Also, we have assumed that the sensing vectors are sampled independently at random, and although the main idea in compressive sensing is to use randomness as a sensing mechanism, it would be interesting to know how the results would change if one were to introduce some correlations.

APPENDIX

A. Proof of Theorem 2.7 (the Weak RIP)

Our proof uses some the results and techniques of [44] and [46]. Recall that A is a matrix with rows a_i^*; each a_i is drawn

independently from a probability distribution F obeying the isotropy and incoherence conditions, and we wish to show that for any fixed $0 \le \delta < 1$,

$$(1-\delta)\|v\|_{\ell_2}^2 \le \|Av\|_{\ell_2}^2 \le (1+\delta)\|v\|_{\ell_2}^2.$$

These inequalities should hold with high probability, uniformly over all vectors v obeying $\text{supp}(v) \subset T \cup R$ where T is fixed, R may vary, and

$$|T| \le c\frac{m}{\mu \log m}, \qquad |R| \le c\frac{m}{\mu \log n \log m \log^2(|R|)}.$$

To express this in another way, set

$$X := \sup_{v \in V}\left|\|Av\|_{\ell_2}^2 - \|v\|_{\ell_2}^2\right|$$

where

$$V = \{v : \|v\|_{\ell_2} = 1, \text{supp}(v) \subset T \cup R, |R| \le r, T \cap R = \emptyset\}. \tag{A.1}$$

In words, v is a unit-normed vector supported on $T \cup R$ where T is fixed of cardinality $s \le cm/(\mu \log m)$, and R is any set disjoint from T of cardinality at most $r \le cm/(\mu \log n \log m \log^2 r)$. We wish to show that $X \le \delta$ with high probability. We will first bound this random variable in expectation and then show that it is unlikely to be much larger than its expectation. The bound in expectation is contained in the following lemma.

Lemma A.1: Fix $\epsilon > 0$. Suppose $m \ge C\mu[s \log m \vee r \log n \log m \log^2 r]$, where C is a constant only depending on ϵ. Then

$$\mathbb{E}X \le \epsilon.$$

To begin the proof, note that for any v with $\text{supp}(v) \subset T \cup R$, we have

$$\|Av\|_{\ell_2}^2 = \|A_T v_T\|_{\ell_2}^2 + \|A_R v_R\|_{\ell_2}^2 + 2\langle v_T, A_T^* A_R v_R\rangle.$$

The first two terms are easily dealt with using prior results. To be sure, under the conditions of Lemma A.1, a slight modification of the proof of [46, Theorem 3.4] gives[4]

$$\mathbb{E}\sup_{v_R:|R|\le r}\left|\|A_R v_R\|_{\ell_2}^2 - \|v_R\|_{\ell_2}^2\right| \le \frac{\epsilon}{4}\|v_R\|_{\ell_2}^2. \tag{A.2}$$

Next, it follows from [45], or the matrix Bernstein inequality in Estimate 1, that

$$\mathbb{E}\sup_{v_T}\left|\|A_T v_T\|_{\ell_2}^2 - \|v_T\|_{\ell_2}^2\right| \le \frac{\epsilon}{4}\|v_T\|_{\ell_2}^2. \tag{A.3}$$

Thus, to prove Lemma A.1, it suffices to prove that

$$\mathbb{E}\max_R\|A_R^* A_T\| \le \epsilon/4.$$

This is the content of the following theorem.

---
[4]This modification has been worked out in [39], [41].

Theorem A.2: Under the assumptions of Lemma A.1, we have

$$
\mathbb{E} \max_R \|A_R^* A_T\|
$$
$$
\leq C \left(\sqrt{\frac{s\mu \log m}{m}} + \sqrt{\frac{r\mu \log n \log m \log^2 r}{m}} \right).
$$

Put differently, the theorem develops a bound on

$$
\mathbb{E} \max_{(x,y)\in B_T \times D} \frac{1}{m} \sum_{i=1}^m \langle a_i, x\rangle \langle a_i, y\rangle \qquad \text{(A.4)}
$$

in which

$$
B_T := \{x : \|x\|_{\ell_2} \leq 1, \mathrm{supp}(x) \subset T\},
$$
$$
D := \{y : \|y\|_{\ell_2} \leq 1, \mathrm{supp}(y) \cap T = \emptyset, |\mathrm{supp}(y)| \leq r\}.
$$

By symmetrization followed by a comparison principle—both of which follow by Jensen's inequality (see [30, Lemma 6.3] followed by [30, inequality (4.8)]—(A.4) is less or equal to a numerical constant times

$$
\mathbb{E} \max_{(x,y)\in B_T \times D} \frac{1}{m} \sum_{i=1}^m g_i \langle a_i, x\rangle \langle a_i, y\rangle
$$

where the g_i's are independent $\mathcal{N}(0,1)$ random variables. The main estimate is a bound on the conditional expectation of the right-hand side; that is, holding the vectors a_i fixed.

Lemma A.3 (Main Lemma): Fix vectors $\{a_i\}_{i=1}^m$ and let

$$
R_1 := \max_{x\in B_T} \frac{1}{m} \sum_{i=1}^m \langle a_i, x\rangle^2, \qquad R_2 := \max_{y\in D} \frac{1}{m} \sum_{i=1}^m \langle a_i, y\rangle^2.
$$

Then we have

$$
\mathbb{E} \max_{(x,y)\in B_T \times D} \frac{1}{m} \sum_{i=1}^m g_i \langle a_i, x\rangle \langle a_i, y\rangle
$$
$$
\leq C \sqrt{\frac{(1+R_2)(1+R_1)s\mu \log m}{m}}
$$
$$
+ C \sqrt{\frac{(1+R_1)r\mu \log n \log m \log^2 r}{m}}.
$$

Proof of Theorem A.2: We have $\sqrt{(1+R_2)(1+R_1)} \leq \frac{1}{2}(2 + R_1 + R_2)$. Now, under the assumptions of the theorem, it follows from the results in [46], [39], [41] that $\mathbb{E}R_2 \leq C$. Likewise, the results in [45] and give $\mathbb{E}R_1 \leq C$, and thus $\mathbb{E}\sqrt{1+R_1} \leq C$. (These inequalities were also noted, in a different form, in (A.3) and (A.2)). Hence, Lemma A.3 implies

$$
\mathbb{E} \max_R \|A_R^* A_T\|
$$
$$
\leq C \left(\sqrt{\frac{s\mu \log m}{m}} + \sqrt{\frac{r\mu \log n \log m \log^2 r}{m}} \right)
$$

where the expectation is taken over the randomly selected rows $\{a_i^*\}$. ∎

B. Proof of Lemma A.3

We need to develop a bound about the expected maximum of a Gaussian process, namely

$$
\mathbb{E} \max_{(x,y)\in B_T \times D} F(x,y)
$$

where

$$
F(x,y) := \sum_{i=1}^m g_i \langle a_i, x\rangle \langle a_i, y\rangle.
$$

We shall do this by means of a certain version of the majorizing measure theorem, which may be found in [44] and is attributed to Talagrand. It is derived as a combination of a majorizing measure construction of Talagrand (combine Theorem 4.1 with Propositions 2.3 and 4.4 in [47]) and the majorizing measure theorem of Fernique [30].

From now on, (M, d) is a metric space and $B(t, \epsilon)$ is the ball of center t and radius ϵ under the metric d.

Theorem A.4 (Majorizing Measure Theorem): Let $(X_t)_{t\in M}$ be a collection of zero-mean random variables obeying the sub-gaussian tail estimate

$$
\mathbb{P}(|X_t - X_{t'}| > u) \leq \exp\left(-c \frac{u^2}{d^2(t,t')} \right) \qquad \text{(A.5)}
$$

for all $u > 0$. Fix $\rho > 1$ and let k_0 be an integer so that the diameter of M is less than ρ^{-k_0}. Suppose there exist $\sigma > 0$ and a sequence of functions $\{\varphi_k\}_{k=k_0}^\infty$, $\varphi_k : M \to \mathbb{R}^+$, with the following two properties: 1) the sequence is uniformly bounded by a constant depending only on ρ; 2) for each k and for any $t \in M$ and any points $t_1, \cdots, t_{\tilde N} \in B(t, \rho^{-k})$ with mutual distances at least ρ^{-k-1}, we have

$$
\max_{j=1,\cdots,\tilde N} \varphi_{k+2}(t_j) \geq \varphi_k(t) + \sigma \rho^{-k} \sqrt{\log \tilde N}. \qquad \text{(A.6)}
$$

Then

$$
\mathbb{E} \sup_{t\in M} X_t \leq C(\rho) \cdot \sigma^{-1}. \qquad \text{(A.7)}
$$

To apply this theorem, we begin by bounding the variance between increments in order to ascertain the metric we need to use (the induced metric). We compute

$$
d((x,y),(x',y')) := \sqrt{\mathrm{Var}(F(x,y) - F(x',y'))}
$$
$$
= \sqrt{\sum_{i=1}^m \left(\langle a_i, x\rangle \langle a_i, y\rangle - \langle a_i, x'\rangle \langle a_i, y'\rangle\right)^2}.
$$

Before continuing, we record two useful lemmas for bounding $\tilde N$. Here and below, $N(M, d, \epsilon)$ is the covering number of M in the metric d with balls of radius ϵ. In other words, it is the minimal number of translates of such balls whose union includes M.

Lemma A.5 (Packing Number Bound): Let $t_1, t_2, \cdots, t_{\tilde{N}} \in M$ be points with mutual distances at least 2ϵ under the metric d. Then

$$\tilde{N} \leq N(M, d, \epsilon).$$

This is a standard result proved by creating an injective mapping from the points $\{t_j\}$ to those in the cover set (map each t_j to the nearest point in the cover). See, e.g., [36, p. 199].

The next lemma is a standard tool used to obtain bounds on covering numbers, see [33] and [6] for a more general statement.

Lemma A.6 (Dual Sudakov Minorization): Let B_{ℓ_2} be the unit ℓ_2 ball in $\mathbb{R}^d$, and let $\|\cdot\|$ be a norm. Let $z \in \mathbb{R}^d$ be a Gaussian vector with independent $\mathcal{N}(0,1)$ entries. Then there is a numerical constant $C > 0$ such that

$$\sqrt{\log N(B_{\ell_2}, \|\cdot\|, \epsilon)} \leq \frac{C}{\epsilon} \sqrt{\mathbb{E} \|z\|^2}.$$

We now invoke the majorizing measure theorem to prove Lemma A.3 for $B_T \times D$ under the metric d. We start by bounding the diameter of $B_T \times D$ under the metric d. By Cauchy-Schwarz, for any $y \in D$, we have $|\langle a_i, y \rangle| \leq \sqrt{r\mu}$. This may be used to derive the following bound on the diameter:

$$d((x,y),(x',y')) \leq 2\sqrt{2r\mu m R_1}.$$

Thus set k_0 to be the largest integer such that

$$\rho^{-k_0} \geq 2\sqrt{2r\mu m R_1}.$$

We also set k_1—whose meaning will become apparent in a moment—to be the largest integer such that

$$\rho^{-k_1} \geq 2\sqrt{2m R_1 \mu}.$$

We now define φ_k on coarse and fine scales. In what follows, we may take $\rho = 8$ so that $C(\rho)$ (A.7) is an absolute constant.

Coarse scales: for $k = k_0, k_0 + 1, \ldots, k_1 - 1$,

$$\varphi_k(x,y) := \min\{\|u\|_{\ell_2}^2 : (u,v) \in B((x,y), 2\rho^{-k})\} + \frac{k - k_0}{\log r}.$$

Fine scales: for $k \geq k_1$, φ_k is a constant function given by

$$\varphi_k(x,y) = \varphi_k := 3\rho\sigma \int_{\rho^{-k}}^{\rho^{-k_1}} \sqrt{\log N(B_T \times D, d, \epsilon)} d\epsilon + 2.$$

Last, set

$$\sigma^{-1} := C\sqrt{m}\sqrt{(1 + R_2)(1 + R_1)s\mu \log m} + C\sqrt{m}\sqrt{(1 + R_1)r\mu \log n \log m \log^2 r}.$$

Our definition of φ_k is closely related to—and inspired by—the functions defined in [44]. We need to show that these functions are uniformly bounded and obey (A.6) for all k. We begin by verifying these properties for fine scale elements as this is the less subtle calculation.

C. Fine Scale: $k \geq k_1$

To show that (A.6) holds, observe that

$$\begin{aligned}
\varphi_{k+2} - \varphi_k &= 3\sigma\rho \int_{\rho^{-(k+2)}}^{\rho^{-k}} \sqrt{\log N(B_T \times D, d, \epsilon)} d\epsilon \\
&\geq 3\sigma\rho \int_{\rho^{-(k+2)}}^{\frac{1}{2}\rho^{-(k+1)}} \sqrt{\log N(B_T \times D, d, \epsilon)} d\epsilon \\
&\geq 3\sigma\rho \left(\frac{1}{2}\rho^{-(k+1)} - \rho^{-(k+2)} \right) \\
&\quad \cdot \sqrt{\log N\left(B_T \times D, d, \frac{1}{2}\rho^{-(k+1)} \right)} \\
&\geq \sigma\rho^{-k}\sqrt{\log \tilde{N}}.
\end{aligned}$$

The last line follows from $\rho \geq 6$ and the packing number bound (Lemma A.5). Note that this same calculation holds when $k = k_1 - 1, k_1 - 2$ because for $k \leq k_1 - 1$, $\varphi_k \leq 3$ (see Section D).

We now show that φ_k is bounded. Since

$$\begin{aligned}
\varphi_k &\leq 3\rho\sigma \int_0^{\rho^{-k_1}} \sqrt{\log N(B_T \times D, d, \epsilon)} d\epsilon + 3 \\
&\leq 3\rho\sigma \int_0^{\rho\sqrt{8mR_1\mu}} \sqrt{\log N(B_T \times D, d, \epsilon)} d\epsilon + 3
\end{aligned}$$

$$(\text{A.8})$$

it suffices to show that the right-hand side is bounded. This follows from crude upper bounds on the covering number. Indeed, observe that

$$\begin{aligned}
&d((x,y),(x',y')) \\
&\leq \sqrt{2\sum_{i=1}^m \langle a_i, x - x' \rangle^2 \langle a_i, y \rangle^2} \\
&\quad + \sqrt{2\sum_{i=1}^m \langle a_i, x' \rangle^2 \langle a_i, y - y' \rangle^2} \\
&\leq \sqrt{2mR_2} \max_{1 \leq i \leq m} |\langle a_i, x - x' \rangle| \\
&\quad + \sqrt{2mR_1} \max_{1 \leq i \leq m} |\langle a_i, y - y' \rangle| \\
&\leq \sqrt{2mR_2 s\mu} \|x - x'\|_{\ell_2} + \sqrt{2mR_1 r\mu} \|y - y'\|_{\ell_2} \\
&\leq m^{3/2} \|x - x'\|_{\ell_2} + m^{3/2} \|y - y'\|_{\ell_2}.
\end{aligned}$$

Thus

$$\begin{aligned}
&N(B_T \times D, d, \epsilon) \\
&\leq N\left(B, \|\cdot\|_{\ell_2}, \frac{\epsilon}{m^{3/2}} \right) \cdot N\left(D, \|\cdot\|_{\ell_2}, \frac{\epsilon}{m^{3/2}} \right) \\
&\leq \left(\frac{2m^{3/2}}{\epsilon} \right)^s \cdot \binom{n}{r} \left(\frac{2m^{3/2}}{\epsilon} \right)^r.
\end{aligned}$$

The second line comes from the standard volumetric estimate $N(B, \|\cdot\|_{\ell_2}, \epsilon) \leq \left(\frac{3}{\epsilon} \right)^s$ for $\epsilon \leq 1$ (see e.g., [41, Proposition 10.1]). The factor $\binom{n}{r}$ arises from decomposing D as the union

of $\binom{n-s}{r} \leq \binom{n}{r}$ sets of the same form as B_T, but with support size bounded by r. Now, in order to bound the last inequality, we further write $\binom{n}{r} \leq n^r$. Plugging this in, we obtain

$$\sqrt{\log N(B_T \times D, d, \epsilon)} \\ \leq C(\sqrt{s \log(m/\epsilon)} + \sqrt{r \log(mn/\epsilon)}).$$

To conclude, bounding the integration gives

$$\int_0^{\sqrt{8R_1 m\mu}} \sqrt{s \log(m/\epsilon)} + \sqrt{r \log(mn/\epsilon)} d\epsilon \\ \leq C\sqrt{8R_1 m s\mu}(\sqrt{\log(m)} + 1) \\ + \sqrt{8R_1 m r\mu}(\sqrt{\log(mn)} + 1)$$

which establishes the claim since the right-hand side is dominated by σ^{-1}.

D. Coarse Scale: $k \leq 0$

This section contains the crucial estimates, which must be developed very carefully. To show that φ_k is bounded, observe that by definition, $\rho^{k_1 - 1 - k_0} \leq \sqrt{r}$, and thus $(k_1 - 1 - k_0) \leq \frac{1}{2} \log r$). It follows that $\varphi_k \leq 2$.

Next, we show that the more subtle bound (A.6) holds. Let $\{(x_i, y_i)\}$ be the points in the definition of the majorizing measure theorem with mutual distances at least ρ^{-k-1}, so that $\tilde{N} = |\{(x_i, y_i)\}|$. Let (z_x, v_x) be the point of $B((x, y), \rho^{-k})$ which minimizes the value of $\|z_x\|_{\ell_2}^2$. Let (z_j, v_j) be the similar points of $B((x_i, y_i), \rho^{-k-2})$. Finally, introduce the pivotal quantity

$$\theta := \max_{1 \leq j \leq \tilde{N}} \|z_j\|_{\ell_2}^2 - \|z_x\|_{\ell_2}^2.$$

It suffices to show that

$$\rho^{-k} \sigma \sqrt{\log \tilde{N}} \leq \max_{1 \leq j \leq \tilde{N}} \varphi_{k+2}(x_j, y_j) - \varphi_k(x, y) \\ = \theta + 2/\log r.$$

In order to bound $\tilde{N}$, we consider the points (z_j, v_j) and note that $\tilde{N} = |\{(z_j, v_j)\}|$.

We shall need two key properties of the points (z_j, v_j). First, these points are well separated. Indeed, the triangle inequality, gives for $i \neq j$

$$d((z_i, v_i), (z_j, v_j)) \\ \geq d((x_i, y_i), (x_j, y_j)) - d((x_i, y_i), (z_i, v_i)) \\ - d((x_j, y_j), (z_j, v_j)) \\ \geq \rho^{-k-1} - 4\rho^{-k-2} \\ \geq \frac{1}{2} \rho^{-k-1}$$

provided that $\rho \geq 8$. Second, each point (z_j, v_j) is close to (x, y) in the sense that

$$d((x, y), (z_j, v_j)) \\ \leq d((x, y), (x_j, y_j)) + d((x_j, y_j), (z_j, v_j)) \\ \leq \rho^{-k} + 2\rho^{-k-2} \leq 2\rho^{-k}$$

provided that $\rho \geq 2$. In other words $(z_j, v_j) \in B((x, y), 2\rho^{-k})$, and thus $\|z_j\|_{\ell_2} \geq \|z_x\|_{\ell_2}$.

Now, the benefit of the special construction of φ_k on the coarse scale is that the size of θ restricts the space that $\{z_j\}$ can inhabit. To demonstrate this, since $B((x, y), 2\rho^{-k})$ is convex, $(\frac{z_x + z_j}{2}, \frac{v_x + v_j}{2}) \in B((x, y), 2\rho^{-k})$. Now combine $\left\| \frac{z_x + z_j}{2} \right\|_{\ell_2} \geq \|z_x\|_{\ell_2}$ with $\|z_j\|_{\ell_2} \geq \|z_x\|_{\ell_2}$ to give

$$\left\| \frac{z_j - z_x}{2} \right\|_{\ell_2}^2 = \frac{1}{2}\|z_j\|_{\ell_2}^2 + \frac{1}{2}\|z_x\|_{\ell_2}^2 - \left\| \frac{z_j + z_x}{2} \right\|_{\ell_2}^2 \\ \leq \|z_j\|_{\ell_2}^2 - \|z_x\|_{\ell_2}^2 \leq \theta.$$

Hence,

$$\|z_j - z_x\|_{\ell_2} \leq 2\sqrt{\theta}.$$

Combined with Lemma A.5, we obtain

$$\tilde{N} \leq N(B_T(z_x, 2\sqrt{\theta}) \times D, d, \rho^{-k-1}/4) \qquad \text{(A.9)}$$

where $B_T(z_x, 2\sqrt{\theta}) := \{x : \text{supp}(x) \subset T, \|x\|_{\ell_2} \leq 1, \|x - z_x\|_{\ell_2} \leq 2\sqrt{\theta}\}$.

We now bound the metric, d, but we will do so more carefully than in the fine scale. We have

$$d((x, y), (x', y')) \\ \leq d((x, y), (x', y)) + d((x', y), (x', y')) \\ \leq \|x - x'\|_y + \sqrt{mR_1}\|y - y'\|_X$$

where

- $\|y\|_X := \max_{1 \leq i \leq m} |\langle a_i, x \rangle|$

- $\|x\|_y := \sqrt{\sum_{i=1}^m \langle a_i, x \rangle^2 \langle a_i, y \rangle^2}.$

Note that the norm $\|\cdot\|_y$ varies with y. Also note that they may be semi-norms, but this makes no difference to the proof. All of the utilized lemmas and theorems generalize to semi-norms.

We cover $B_T(z_x, 2\sqrt{\theta}) \times D_S$ to precision ϵ by covering D_S to precision $\epsilon/2$ under the norm $\sqrt{mR_1}\|\cdot\|_X$, and then, for each y contained in this cover, we cover $B_T(u, 2\sqrt{\theta})$ to precision $\epsilon/2$ under the norm $\|\cdot\|_y$.

Thus

$$\sqrt{\log \tilde{N}} \leq \sqrt{\log N(D_S, \sqrt{mR_1}, \|\cdot\|_X, \epsilon/2)} \\ + \max_{y' \in D_S} \sqrt{\log N(B_T(u, 2\sqrt{\theta}), \|\cdot\|_y, \epsilon/2)} \\ := K_1 + \max_{y \in D} K_2(y). \qquad \text{(A.10)}$$

Using [46, Lemma 3.7] (which follows from an argument of [18]) gives

$$K_1 \leq C\sqrt{mR_1 r\mu \log n \log m} \rho^k \qquad \text{(A.11)}$$

(We do not reproduce the proof of this lemma here, but encourage the interested reader to explore the extremely clever, short, arguments used).

Now we bound $K_2(y)$. For y fixed, using dual Sudakov minorization (Lemma A.6) and Jensen's inequality, we have

$$K_2(y) \leq C\frac{\sqrt{\theta}}{\epsilon}\sqrt{\mathbb{E}\sum_{i=1}^{m}\langle a_i, z_T\rangle^2 \langle a_i, y\rangle^2} \qquad (A.12)$$

where z is aGaussian vector with standard normal entries. Note that $\mathbb{E}\langle a_i, z_T\rangle^2 = \|a_{i,T}\|_{\ell_2}^2 \leq s\mu$, and thus

$$K_2(y) \leq C\frac{\sqrt{\theta}}{\epsilon}\sqrt{\mu s m R_2} \leq C\sqrt{\mu s m R_2}\rho^k\sqrt{\theta}. \qquad (A.13)$$

Plug in the covering number bounds (i.e., plug in (A.11) and (A.13) into (A.10)) to give

$$\sqrt{\log \tilde{N}} \leq C(\sqrt{mR_1 ru \log n \log m}\rho^k + \sqrt{\mu s m R_2}\rho^k\sqrt{\theta}).$$

Now, plug in $\sqrt{\theta} \leq \theta\sqrt{\log r} + 1/\sqrt{\log r}$, along with the definition of σ, to give

$$\rho^{-k}\sigma\sqrt{\log \tilde{N}} \leq \frac{2}{\log r} + \theta$$

as desired, thus proving Theorem A.2.

E. Concentration Around the Mean

We have now proved that $\mathbb{E}X \leq \epsilon$ for any $\epsilon > 0$ provided that $m \geq C_\epsilon \mu [s \log m \vee r \log n \log m \log^2 r]$. This already shows that for any fixed $\delta > 0$,

$$\mathbb{P}(X > \delta) \leq \frac{\epsilon}{\delta}$$

and so taking ϵ to be a small fraction of δ gives a first crude bound. However, we wish to show that if $m \geq C\mu\beta[s \log m \vee r \log n \log m \log^2 r]$ then the probability of 'failure' decreases as $e^{-\beta}$. This can be proved using a theorem of [46] which in turn is a combination of Theorem 6.17 and inequality (6.19) of [30]. We restate this theorem below.

Theorem A.7: Let $Y_1, \cdots, Y_m$ be independent symmetric random variables taking values in some Banach space. Assume that $\|Y_j\| \leq R$ for all j and for some norm $\|\cdot\|$. Then for any integer $\ell \geq q$, and any $t > 0$, the random variable

$$Z := \left\|\sum_{j=1}^{m} Y_j\right\|$$

obeys

$$\mathbb{P}(Z \geq 8q\mathbb{E}Z + 2R\ell + t) \leq \left(\frac{C}{q}\right)^\ell + 2\exp\left(-\frac{t^2}{256q(\mathbb{E}Z)^2}\right).$$

In our setup, we work with a norm on positive semidefinite matrices given by

$$\|M\| := \sup_{v \in V} v^* M v$$

where V is given by (A.1). The rest of the details of the proof of concentration around the mean follows exactly as in the steps of [46, pp. 11–12] and so we do not repeat them, but encourage the interested reader to check [46]. This is the final step in proving Theorem 2.7.

F. Stochastic Incoherence

In Sections 2–4, we have assumed that the coherence bound holds deterministically, and it is now time to prove our more general statement; that is to say, we need to extend the proof to the case where it holds stochastically. We propose a simple strategy: condition on the (likely) event that each row has 'small' entries, as to recreate the case of deterministic coherence (on this event). Outside of this event, we give no guarantees, but this is of little consequence because we will require the event to hold with probability at least $1 - 1/n$. A difficulty arises because the conditional distribution of the rows no longer obeys the isotropy condition (although the rows are still independent). Fortunately, this conditional distribution obeys a *near isotropy condition*, and all of our results can be reproved using this condition instead. In particular, all of our theorems follow (with adjustments to the absolute constants involved) from the following two conditions on the distribution of the rows:

$$\begin{aligned}\|\mathbb{E}aa^* - I\| &\leq 1/(8\sqrt{n}) && \text{(near isotropy)}\\ \max_{1\leq t\leq n}\|a[t]\|_{\ell_2}^2 &\leq \mu && \text{(deterministic coherence)}.\end{aligned}$$
$$(A.14)$$

We first illustrate how to use near isotropy to prove our results. There are several results that need to be reproved, but they are all adjusted using the same principle, so to save space we just prove that a slight variation on Lemma 2.1 still holds when requiring near isotropy, and leave the rest of the analogous calculations to the interested reader.

Set $W := \mathbb{E}aa^*$ and let $W_{T,T}$ be the restriction of W to rows and columns in T. We first show that

$$\begin{aligned}\mathbb{P}(\|A_T^* A_T &- W_{T,T}\|\\ &\geq \delta) \leq 2s\exp\left(-\frac{m}{\mu(s+1)}\frac{\delta^2}{2+2\delta/3}\right).\end{aligned} \qquad (A.15)$$

To prove this bound, we use the matrix Bernstein inequality of Section II-A, and also follow the framework of the calculations of Section II-A. Thus, we skim the steps. To begin, decompose $A_T^* A_T - W_{T,T}$ as follows:

$$m(A_T^* A_T - W_{T,T}) = \sum_{k=1}^{m}(a_{k,T}a_{k,T}^* - W_{T,T}) := \sum_{k=1}^{m}X_k.$$

We have $\mathbb{E}X_k = 0$ and $\|X_k\| \leq \|a_{k,T}a_{k,T}^* - I\| + \|I - W_{T,T}\| \leq s\mu + \frac{1}{8\sqrt{n}} \leq (s+1)\mu := B$. Also, the total variance obeys

$$\begin{aligned}\|\mathbb{E}X_k\|^2 &\leq \|\mathbb{E}(a_{k,T}a_{k,T}^*)^2\| \leq s\mu\|\mathbb{E}a_{k,T}a_{k,T}^*\|\\ &= s\mu\|W_{T,T}\| \leq s\mu\left(1 + \frac{1}{8\sqrt{n}}\right) \leq (s+1)\mu.\end{aligned}$$

Thus, $\sigma^2 \leq m(s+1)\mu$, and (A.15) follows from the matrix Bernstein inequality.

Now, it follows from $\|W_{T,T} - I\| \leq \|W - I\| \leq \frac{1}{8\sqrt{n}}$ that

$$\begin{aligned}\mathbb{P}&\left(\|A_T^* A_T - I\| \geq \frac{1}{8\sqrt{n}} + \delta\right)\\ &\leq 2s\exp\left(-\frac{m}{\mu(s+1)}\frac{\delta^2}{2+2\delta/3}\right).\end{aligned}$$

In the course of the proofs of Theorems 1.1 and 1.2, we require $\|A_T^* A_T - I\| \leq 1/2$ for noiseless results and

$\|A_T^* A_T - I\| \leq 1/4$ for noisy results. This can be achieved under the near isotropy condition by increasing the required number of measurements by a tiny bit. In fact, when proving the analogous version of Lemma 2.1, one could weaken the near isotropy condition and instead require $\|\mathbb{E} aa^* - I\| \leq 1/8$, for example. However, in extending some of the other calculations to work with the near isometry condition—such as (III.11)—the factor of $\sqrt{n}$ (or at least $\sqrt{s}$) in the denominator appears necessary; this seems to be an artifact of the method of proof, namely, the golfing scheme. It is our conjecture that all of our results could be established with the weaker requirement $\|\mathbb{E} aa^* - I\| \leq \epsilon$ for some fixed positive constant ϵ.

We now describe the details concerning the conditioning on rows having small entries. Fix the coherence bound μ and let

$$E_k = \left\{ \max_{1 \leq t \leq n} |a_k[t]|^2 \leq \mu \right\} \qquad \text{and} \qquad G = \cap_{1 \leq k \leq m} E_k.$$

Thus G is the "good" event (G is for good) on which $\max_{1 \leq t \leq n} |a_k[t]|^2 \leq \mu$ for all k. By the union bound, $\mathbb{P}(G^c) \leq m \mathbb{P}(E_1^c)$. We wish for $\mathbb{P}(G^c)$ to be bounded by $1/n$, and so we require μ to be large enough so that $\mathbb{P}(E_1^c) \leq (mn)^{-1}$.

Next, we describe how conditioning on the event G induces the near isometry condition. Because of the independence of the rows of A, we may just consider the conditional distribution of a_1 given E_1. Drop the subindex for simplicity and write

$$I = \mathbb{E}[aa^*] = \mathbb{E}[aa^* \mathbb{1}_E] + \mathbb{E}[aa^* \mathbb{1}_{E^c}]$$
$$= \mathbb{E}[aa^*|E] \mathbb{P}(E) + \mathbb{E}[aa^* \mathbb{1}_{E^c}].$$

Thus

$$\|\mathbb{E}[aa^*|E] - I\| \cdot \mathbb{P}(E) = \|(1 - \mathbb{P}(E))I - \mathbb{E}[aa^* \mathbb{1}_{E^c}]\|$$
$$\leq \mathbb{P}(E^c) + \|\mathbb{E}[aa^* \mathbb{1}_{E^c}]\|. \quad \text{(A.16)}$$

We now bound $\|\mathbb{E}[aa^* \mathbb{1}_{E^c}]\|$. By Jensen's inequality (which is a crude, but still fruitful, bound here)

$$\|\mathbb{E}[aa^* \mathbb{1}_{E^c}]\| \leq \mathbb{E}[\|aa^* \mathbb{1}_{E^c}\|] = \mathbb{E}[\|a\|_{\ell_2}^2 \mathbb{1}_{E^c}]. \quad \text{(A.17)}$$

and, therefore

$$\|\mathbb{E}[aa^*|E] - I\| \leq \frac{1}{1 - \mathbb{P}(E^c)} \left(\mathbb{P}(E^c) + \mathbb{E}[\|a\|_{\ell_2}^2 \mathbb{1}_{E^c}] \right).$$

Combine this with the requirement that $\mathbb{P}(E^c) \leq (mn)^{-1}$ to give

$$\|\mathbb{E}[aa^*|E] - I\| \leq \frac{19}{20} \left(\frac{1}{20\sqrt{n}} + \mathbb{E}[\|a\|_{\ell_2}^2 \mathbb{1}_{E^c}] \right)$$

as long as $m\sqrt{n} \geq 20$. It now follows that in order to ensure near isotropy, it is sufficient that

$$\mathbb{E}[\|a\|_{\ell_2}^2 \mathbb{1}_{E^c}] \leq \frac{1}{20\sqrt{n}}.$$

It may be helpful to note a simple way to bound the left-hand side above. If $f(t)$ is such that

$$\mathbb{P}\left(\max_{1 \leq t \leq n} |a[t]|^2 \geq t \right) \leq f(t),$$

then a straightforward calculation shows that

$$\mathbb{E}[\|a\|_{\ell_2}^2 \mathbb{1}_{E^c}] \leq n\mu f(\mu) + n \int_\mu^\infty f(t) dt.$$

ACKNOWLEDGMENT

REFERENCES

[1] R. Ahlswede and A. Winter, "Strong converse for identification via quantum channels," *IEEE Trans. Inf. Theory*, vol. 48, no. 3, pp. 569–579, Mar. 2002.

[2] B. Bah and J. Tanner, Improved Bounds on Restricted Isometry Constants for Gaussian Matrices 2010 [Online]. Available: http://arxiv.org/abs/1003.3299

[3] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Construct. Approx.*, vol. 28, no. 3, pp. 253–263, 2008.

[4] M. Bayati and A. Montanari, The LASSO Risk for Gaussian Matrices 2010 [Online]. Available: http://arxiv.org/abs/1008.2581

[5] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Ann. Statist.*, vol. 37, no. 4, pp. 1705–1732, 2009.

[6] J. Bourgain, J. Lindenstrauss, and V. Milman, "Approximation of zonoids by zonotopes," *Acta. Math.*, vol. 162, no. 1, pp. 73–141, 1989.

[7] T. T. Cai, L. Wang, and G. Xu, "New bounds for restricted isometry constants," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4388–4394, Sep. 2010.

[8] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *C. R. Math. Acad. Sci. Paris, Ser. I*, vol. 346, pp. 589–592, 2008.

[9] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Ann. Statist.*, vol. 37, pp. 2145–2177, 2009.

[10] E. J. Candès, Y. Plan, and J. A. Tropp, Personal Communication 2010.

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

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

[13] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math*, vol. 59, no. 8, p. 1207, 2006.

[14] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.

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

[16] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, Jun. 2007.

[17] E. J. Candes and J. Romberg, "Quantitative robust uncertainty principles and optimally sparse decompositions," *Found. Comp. Math.*, vol. 6, no. 2, pp. 227–254, 2006.

[18] B. Carl, "Inequalities of Bernstein-Jackson type and the degree of compactness of operators in Banach spaces," *Ann. Inst. Fourier*, vol. 35, no. 3, pp. 79–118, 1985.

[19] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *J. Amer. Math. Soc.*, vol. 22, no. 1, pp. 211–231, 2009.

[20] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[21] D. L. Donoho, "For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution," *Comm. Pure Appl. Math.*, vol. 59, no. 7, pp. 907–934, 2006.

[22] D. L. Donoho, A. Maleki, and A. Montanari, The Noise Sensitivity Phase Transition in Compressed Sensing 2010 [Online]. Available: http://arxiv.org/abs/1004.1218

[23] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 83–91, 2008.

[24] S. Foucart, "A note on guaranteed sparse recovery via l1-minimization," *Appl. Comput. Harmon. Anal.*, vol. 29, no. 1, pp. 97–103, 2010.

[25] A. Gilbert and P. Indyk, "Sparse recovery using sparse matrices," *Proc. IEEE*, vol. 98, no. 6, pp. 937–947, Jun. 2010.

[26] D. Gross, Recovering Low-Rank Matrices From Few Coefficients in Any Basis 2009 [Online]. Available: http://arxiv.org/abs/0910.1879

[27] O. Guédon, S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann, "Majorizing measures and proportional subsets of bounded orthonormal systems," *Revista matemática iberoamericana*, vol. 24, no. 3, pp. 1075–1095, 2008.

[28] J. Haupt, W. U. Bajwa, G. Raz, and R. Nowak, "Toeplitz compressed sensing matrices with applications to sparse channel estimation," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5862–5875, Nov. 2010.

[29] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank, "Efficient and robust compressed sensing using optimized expander graphs," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4299–4308, Sep. 2009.

[30] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*.   New York: Springer, 1991.

[31] C. McDiarmid, *Concentration. Probabilistic Methods for Algorithmic Discrete Mathematics*.   New York: Springer, 1998.

[32] R. I. Oliveira, "Sums of random Hermitian matrices and an inequality by Rudelson," *Electron. J. Probab.*, vol. 15, pp. 203–212, 2010.

[33] A. Pajor and N. Tomczak-Jaegermann, "Subspaces of small codimension of finite-dimensional Banach spaces," *Proc. Amer. Math. Soc.*, vol. 97, no. 4, pp. 637–642, 1986.

[34] G. E. Pfander and H. Rauhut, "Sparsity in time-frequency representations," *J. Fourier Anal. and Appl.*, vol. 16, no. 2, pp. 233–260, 2010.

[35] Y. Plan, "Compressed Sensing, Sparse Approximation, and Low-rank Matrix Estimation," Ph.D. dissertation, California Institute of Technology, Pasadena, CA, 2011.

[36] D. Pollard, *Convergence of Stochastic Processes*.   New York: Springer, 1984.

[37] M. Raginsky, S. Jafarpour, Z. Harmany, R. Marcia, R. Willett, and R. Calderbank, Performance Bounds for Expander-Based Compressed Sensing in Poisson Noise [Online]. Available: http://arxiv.org/abs/1007.2377

[38] H. Rauhut, "Random sampling of sparse trigonometric polynomials," *Appl. Comput. Harmon. Anal.*, vol. 22, no. 1, pp. 16–42, 2007.

[39] H. Rauhut, "Stability results for random sampling of sparse trigonometric polynomials," *IEEE Trans. Inf. Theory*, vol. 54, no. 12, pp. 5661–5670, Dec. 2008.

[40] H. Rauhut, "Circulant and toeplitz matrices in compressed sensing," in *Proc. SPARS'09*, Saint-Malo, France, 2009.

[41] H. Rauhut, "Compressive sensing and structured random matrices," *Theoret. Found. Numer. Meth. for Sparse Recov.*, vol. 9, pp. 1–92, 2010.

[42] H. Rauhut, J. Romberg, and J. A. Tropp, Restricted Isometries for Partial Random Circulant Matrices 2010, Preprint.

[43] J. Romberg, "Compressive sensing by random convolution," *SIAM J. Imag. Sci.*, vol. 2, no. 4, pp. 1098–1128, 2009.

[44] M. Rudelson, "Almost orthogonal submatrices of an orthogonal matrix," *Israel J. Math.*, vol. 111, no. 1, pp. 143–155, 1999.

[45] M. Rudelson, "Random vectors in the isotropic position," *J. Funct. Anal.*, vol. 164, no. 1, pp. 60–72, 1999.

[46] M. Rudelson and R. Vershynin, "On sparse reconstruction from Fourier and Gaussian measurements," *Commun. Pure Appl. Math.*, vol. 61, no. 8, pp. 1025–1045, 2008.

[47] M. Talagrand, "Majorizing measures: The generic chaining," *Ann. Prob.*, vol. 24, no. 3, pp. 1049–1103, 1996.

[48] J. A. Tropp, User-Friendly Tail Bounds for Matrix Martingales 2010 [Online]. Available: http://arxiv.org/abs/1004.4389

[49] S. A. van de Geer and P. Bühlmann, "On the conditions used to prove oracle results for the Lasso," *Electron. J. Stat.*, vol. 3, pp. 1360–1392, 2009.

[50] P. Wojtaszczyk, "Stability and instance optimality for Gaussian measurements in compressed sensing," *Found. Comput. Math.*, vol. 10, no. 1, pp. 1–13, 2009.

[51] T. Zhang, "Some sharp performance bounds for least squares regression with L1 regularization," *Ann. Statist.*, vol. 37, no. 5A, pp. 2109–2144, 2009.

[52] Y. Zhang, Theory of Compressive Sensing via ℓ_1 Minimization: A Non-Rip Analysis and Extensions Rice University, Houston, TX, Tech. Rep., 2008. [Online]. Available: http://www.caam.rice.edu/~zhang/reports/tr0811_revised.pdf

Emmanuel J. Candès received the B.Sc. degree from the Ecole Polytechnique, France, in 1993 and the Ph.D. degree in statistics from Stanford University, Stanford, CA, in 1998. He is a Professor of Mathematics and of Statistics at Stanford University, and is the Ronald and Maxine Linde Professor of Applied and Computational Mathematics at the California Institute of Technology, Pasadena (on leave). His areas of interest are in mathematical signal processing, information theory, scientific computing, and mathematical optimization.

Dr. Candès was awarded the 2005 James H. Wilkinson Prize in Numerical Analysis and Scientific Computing by SIAM. He received the U.S. National Science Foundation's highest honor in 2006: the Alan T. Waterman Award. Other awards include the 2008 Information Theory Society Paper Award, the 2010 George Pólya Prize awarded by SIAM, and the 2011 Collatz Prize awarded by the International Council for Industrial and Applied Mathematics (ICIAM).

Yaniv Plan received his B.A. degree from the University of California, Berkeley in 2004 where he double majored in applied math and physics and received honors in applied math. In June 2011, he received his Ph.D. in the applied and computational math department at Caltech working under the supervision of Dr. Candès. He was awarded the W.P. Carey & Co. Inc. Prize for his dissertation. He will begin an NSF Postdoctoral Fellowship in conjunction with a Hildebrandt Assistant Professorship at the University of Michigan in September 2011. His areas of interest are in compressive sensing, sparse reconstruction, and low-rank matrix recovery, non-asymptotic random matrix theory, and probability in Banach spaces.