# Thresholds for the Recovery of Sparse Solutions via L1 Minimization

David L. Donoho
Department of Statistics
Stanford University
390 Serra Mall, Sequoia Hall
Stanford, CA 94305-4065
Email: donoho@stanford.edu

Jared Tanner
Department of Mathematics
University of Utah
155 South 1400 East
Salt Lake City, UT 84112-0090
Email: tanner@math.utah.edu

***Abstract*— The ubiquitous least squares method for systems of linear equations returns solutions which typically have all non-zero entries. However, solutions with the least number of non-zeros allow for greater insight. An exhaustive search for the sparsest solution is intractable, NP-hard. Recently, a great deal of research showed that linear programming can find the sparsest solution for certain 'typical' systems of equations, provided the solution is sufficiently sparse. In this note we report recent progress determining conditions under which the sparsest solution to large systems is available by linear programming, [1]–[3]. Our work shows that there are sharp thresholds on sparsity below which these methods will succeed and above which they fail; it evaluates those thresholds precisely and hints at several interesting applications.**

## I. Introduction

The recently introduced Compressed Sensing [4] paradigm portends a revolution in modern sampling theory. If an $n$ dimensional signal $x$ is known to be compressible - a central tenet of approximation theory - then dramatically fewer than $n$ non-adaptive measurements may be sufficient to capture the signal's essential information. Moreover, there is a known reconstruction technique to recover an approximation to $x$ which achieves the same order as optimal adaptive sampling.

At the heart of Compressed Sensing is replacing traditional point value measurements with random projections, and utilizing the highly non-linear reconstruction technique of constrained $\ell^1$ minimization, often referred to as Basis Pursuit or convex relaxation. More precisely, let $\Psi$ be an orthonormal basis (or tight frame) in which $x$ is compressible, and let $\Phi$ be a $d \times n$ matrix whose columns are selected iid uniformly at random from the $d-1$ dimensional unit sphere $S^{d-1}$; then, the $d$ non-adaptive samples of the signal are computed by $A^{CS}x = \Phi\Psi^T x = b$, and the reconstruction is the vector $x_{CS}$ solving

$$\min \|\Psi^T x_{CS}\|_{\ell^1} \quad \text{subject to} \quad A^{CS}x_{CS} = b.$$

This note focuses on the reconstruction technique,

$$\min \|x\|_{\ell^1} \quad \text{subject to} \quad Ax = b, \tag{1}$$

in an environment analogous to that of Compressed Sensing. Consider the stronger notion of compressibility, *sparsity*, measured as the number of non-zero entries and expressed by the $\ell^0$ 'norm'. Over the last few years it has been both observed experimentally and shown theoretically that, for particular classes of the sampling matrix $A$, if $x$ is sufficiently sparse then (1) often recovers the sparsest solution, ( [1]–[3], [5]–[13] to name a few),

$$\min \|x\|_{\ell^0} \quad \text{subject to} \quad Ax = b. \tag{2}$$

This phenomenon is referred to as $\ell^1/\ell^0$ equivalence, or alternatively LP/NP equivalence as (1) can be solved efficiently by linear programming whereas the naive exhaustive search for the sparsest solution of an underdetermined system of equations is NP-hard.

Precise phase-transitions for $\ell^1/\ell^0$ equivalence have recently been uncovered for a wide class of sampling matrices, $A$, similar to those occurring in Compressed Sensing. These thresholds allow us to precisely answer the question, how sparse is necessary so that $\ell^0$, (2), can be solved efficiently as $\ell^1$, (1)?

## II. Problem Formulation

Consider $A = A_{d,n}$, a $d \times n$ orthogonal projector from $\mathbb{R}^n$ to $\mathbb{R}^d$ selected uniformly at random from the Grassman manifold of such projectors. Consider the setting of proportional growth: the aspect ratio given by $\delta \in (0,1)$ with $d = \lfloor \delta n \rfloor$, and the sparsity a fraction of $d$, $k = \|x\|_{\ell^0} = \rho d$, for $\rho \in (0,1)$. The aspect ratio, $\delta$, measures the undersampling factor, and $\rho$ measures the fractional amount of sparsity. For $n$ large, the problem space is determined by these two parameters $(\delta, \rho)$. We catalog four sparsity phase-transitions as functions of the undersampling ratio, $\rho(\delta)$, which characterize the $\ell^1/\ell^0$ equivalence in the limit as $n \to \infty$. These phase-transitions quantify the necessary undersampling ratio, $\delta$, above which the solution to (1) is with overwhelming probability (exponential in $n$) the sparsest solution, (2).

We consider four thresholds, separately the cases of $x$ having signed entries or non-negative entries, and for each of these cases "strong" and "weak" thresholds, Table I. The strong threshold ensures that $\ell^1/\ell^0$ equivalence holds for $k \leq \rho_S(\delta) \cdot d$, with overwhelming probability in the uniform selection of $A$. For the weak threshold, $\ell^1/\ell^0$ equivalence holds for *most* $x$ where $k = \|x\|_{\ell^0} \leq \rho_W(\delta) \cdot d$ with

TABLE I

$\ell^1/\ell^0$ EQUIVALENCE THRESHOLDS

| | Strong | Weak |
|---|---|---|
| Signed | $\rho_S^{\pm}(\delta)$ | $\rho_W^{\pm}(\delta)$ |
| Non-negative | $\rho_S^{+}(\delta)$ | $\rho_W^{+}(\delta)$ |

overwhelming probability in the uniform selection of $A$. That is, for the vast majority of random orthogonal projection matrices $A$, the solution to (1) is the sparsest solution, (2), for the overwhelming fraction of $x$. These statements are made precise in the following section.

## III. $\ell^1/\ell^0$ EQUIVALENCE THRESHOLDS

*Theorem 1 (Strong, Signed [1]):* Fix $\epsilon > 0$. With overwhelming probability for large $n$, $A_{d,n}$ offers the property of $\ell^1/\ell^0$ equivalence for all $x$ satisfying $\|x\|_{\ell^0} \leq (\rho_S^{\pm}(\delta) - \epsilon) \cdot d$.

*Theorem 2 (Strong, Non-negative [2]):* Fix $\epsilon > 0$. With overwhelming probability for large $n$, $A_{d,n}$ offers the property of $\ell^1/\ell^0$ equivalence for all $x \geq 0$ satisfying $\|x\|_{\ell^0} \leq (\rho_S^{+}(\delta) - \epsilon) \cdot d$.

Formulae for calculating thresholds $\rho_S^{\pm}(\delta)$ and $\rho_S^{+}(\delta)$ were derived in [14] and [15], [16] respectively. They yield the depictions seen in Figure 1 as the dashed-lines, $\rho_S^{\pm}$ (top) and $\rho_S^{+}$ (bottom). The strong thresholds characterize the sparsity level required to ensure that *all* sparse signals can be recovered efficiently. Although the strong thresholds already establish $\ell^1/\ell^0$ equivalence for sparsity up to a fixed fraction of $d$, for most signals $\ell^1/\ell^0$ equivalence is satisfied for a substantially larger fraction of $d$. For such weak thresholds, consider the problem ensemble $(A, b)$ composed of a given $A_{d,n}$ and all $b$ which are generated by an $x$ where $k = \|x\|_{\ell^0} \leq \rho_W(\delta) \cdot d$.

*Theorem 3 (Weak, Signed [1]):* Fix $\epsilon > 0$. With overwhelming probability for large $n$, the problem $(A, b)$ offers the property of $\ell^1/\ell^0$ equivalence for the overwhelming fraction of $x$ satisfying $\|x\|_{\ell^0} \leq (\rho_W^{\pm}(\delta) - \epsilon) \cdot d$.

*Theorem 4 (Weak, Non-negative [2]):* Fix $\epsilon > 0$. With overwhelming probability for large $n$, the problem $(A, b)$ offers the property of $\ell^1/\ell^0$ equivalence for the overwhelming fraction of $x \geq 0$ satisfying $\|x\|_{\ell^0} \leq (\rho_W^{+}(\delta) - \epsilon) \cdot d$.

Formulae for calculating thresholds $\rho_W^{\pm}(\delta)$ and $\rho_W^{+}(\delta)$ were similarly derived in [14] and [16] respectively; and are reproduced in Figure 1 as the solid lines. Table II states quantitative values of the four thresholds for various values of $\delta$.

TABLE II

SELECTED VALUES OF THRESHOLDS

| | $\delta = .1$ | $\delta = .25$ | $\delta = .5$ | $\delta = .75$ | $\delta = .9$ |
|---|---|---|---|---|---|
| $\rho_S^{\pm}(\delta)$ | .048802 | .065440 | .089416 | .117096 | .140416 |
| $\rho_S^{+}(\delta)$ | .060131 | .087206 | .133457 | .198965 | .266558 |
| $\rho_W^{\pm}(\delta)$ | .188327 | .266437 | .384803 | .532781 | .677258 |
| $\rho_W^{+}(\delta)$ | .240841 | .364970 | .558121 | .765796 | .902596 |

A simple example of the type of statements which can be

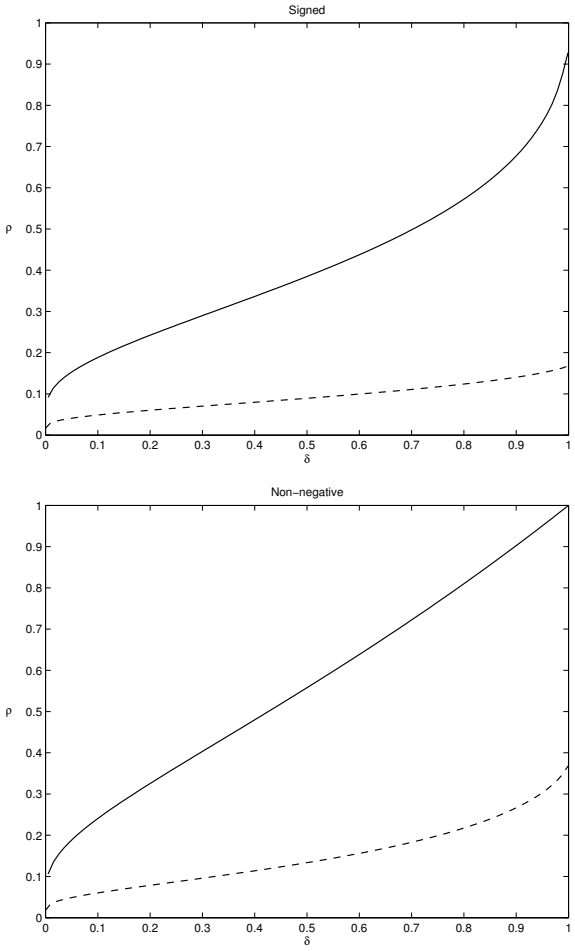


Fig. 1. Thresholds for $x$ signed (above) and non-negative (below). Weak thresholds in solid and strong thresholds in dashed.

gleaned from Table II is as follows. For a non-negative signal of length $n$ measure $n/10$ random inner-products; with high probability for $n$ large, the vast majority of signals with no more than $n/42$ non-zero entries can be recovered efficiently by linear programing, (1). Other applications include the widely discussed context of error-correcting codes, [5], [17]. In this setting the strong thresholds correspond to signal recovery from malicious errors, whereas the weak thresholds concern recoverability from random error. In either case, quantitative threshold information such as in Table II allows for precise design criteria. Application to error correcting codes will be discussed in more detail in Section VI. We conclude this section with a few remarks.

*Remark ($\ell^1/\ell^0$ Thresholds):*

*i.* Theorems 1 through 4 prove $\ell^1/\ell^0$ equivalence for sparsity proportional to $d$, further establishing the effectiveness of solving (1) to recover the sparsest solution, even for relatively non-sparse signals.

*ii.* The fact that $\rho^{+}(\delta)$ is larger than $\rho^{\pm}(\delta)$ signals value in using auxiliary information beyond sparsity; in this case non-negativity. In short, in applications we should exploit prior knowledge. Potential gains can

be achieved in determining sparse representations by utilizing prior knowledge of applications.

*iii.* The significant increase of $\rho_W(\delta)$ over $\rho_S(\delta)$ indicates the greatly improved behavior encountered *typically*, in contrast to the strict *for-all* $\ell^1/\ell^0$ equivalence.

*iv.* Results of Baryshnikov and Vitale imply that the $\ell^1/\ell^0$ equivalence thresholds stated here are equally valid for matrices $A_{d,n}$ whose columns are drawn independently from a multivariate normal distribution on $\mathbb{R}^d$ with nonsingular covariance. This follows from the fact that the uniform random projection of a spherico-regular simplex is affine equivalent in distribution to a standard Gaussian point sample, [18].

## IV. $\ell^1/\ell^0$ Equivalence Thresholds Asymptotic Behavior

Section III stated results in the setting of proportional growth; that is, each of the terms $(k, d_n, n)$ increasing with $k/d_n \to \rho$ and $d_n/n \to \delta$ as $n \to \infty$. This section similarly considers $(k, d_n, n)$ growing, but now with $d_n/n \to 0$ as $n \to \infty$. More precisely, we consider sub-exponential growth of $n$ with respect to $d_n$,

$$d_n/n \to 0, \qquad \frac{\log(n)}{d_n} \to 0, \qquad n \to \infty.$$

This regime occurs naturally in applications where there are massively underdetermined systems of equations such as sparse approximations of massive data sets, highly efficient error-correcting codes, and in the context of Compressed Sensing, [4], sparse representations from highly redundant representations (such as $n \times n^2$ Gabor systems).

In this asymptotic regime, as $\delta \to 0$, the $\ell^1/\ell^0$ equivalence thresholds are given by:

*Theorem 5 (Strong Asymptotic, Nonnegative [3]):*

$$\rho_S^+(\delta) \sim |2e \log(\delta 2\sqrt{\pi})|^{-1}, \qquad \delta \to 0 \tag{3}$$

*Theorem 6 (Strong Asymptotic, Signed [3]):*

$$\rho_S^\pm(\delta) \sim |2e \log(\delta \sqrt{\pi})|^{-1}, \qquad \delta \to 0 \tag{4}$$

*Theorem 7 (Weak Asymptotics [3]):*

$$\rho_W^+(\delta) \sim |2 \log(\delta)|^{-1}, \qquad \delta \to 0 \tag{5}$$

$$\rho_W^\pm(\delta) \sim |2 \log(\delta)|^{-1}, \qquad \delta \to 0 \tag{6}$$

Figure 2 (top) illustrates the ratio of strong thresholds and their asymptotic expressions in Theorems 5 and 6. As $\delta \to 0$ these ratios approach one. Figure 2 (bottom) shows a remarkable match between the weak thresholds, $\rho_W^+$ (dotted) and $\rho_W^\pm$ (dashed), and their leading asymptotic behavior, Theorem 7 (solid).

*Remark (Asymptotic Behavior):*

*i.* The principal difference between the strong and weak asymptotic behavior is the presence of the multiplicative factor $e$. That is, in the asymptotic regime $\delta \to 0$, $\ell^1/\ell^0$ equivalence is satisfied for the *vast majority* of $x$ with sparsity constraint $e$ times weaker than is necessary to ensure $\ell^1/\ell^0$ equivalence for all $x$.

*ii.* Remarkably, to first order, the asymptotic behavior of the thresholds is the same for both the signed and non-negative case. This is surprising since at moderate values of $\delta$ the functions are quite different.

*iii.* More subtle differences occur in the arguments to the logarithms approximating the strong thresholds, with the signed and non-negative factors differing by a multiple of 2.

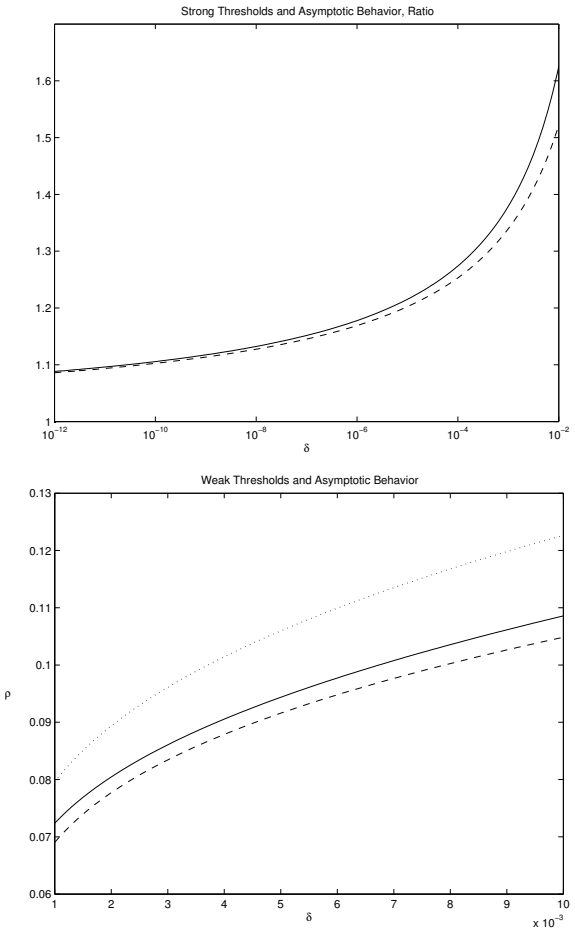


Fig. 2. Top: Ratio of strong thresholds and their asymptotic behaviors, $\rho_S^+(\delta)$ as solid line and $\rho_S^\pm(\delta)$ as dashed line. Bottom: Weak thresholds $\rho_W^+$ (dotted), $\rho_W^\pm$ (dashed), and their asymptotic behavior from Theorem 7 (solid).

## V. Empirical Evidence

The results of Theorems 1 and 2 cannot be verified computationally – it seems that this would be NP-hard. The results of Theorems 3 and 4, however, can be tested empirically by running tests of a random set of examples.

These examples used the modest value of $n = 200$. The problem domain $(\delta, \rho)$ was partitioned into a $40 \times 40$ uniform mesh of $[1/20, 99/100]$ in each direction. For each $\delta$, a random orthogonal projector $A_{d,n}$ was computed with $d = \lfloor \delta n \rfloor$, and for each value of $\rho$ a random $y$ with $\|y\|_{\ell^0} = \lfloor \rho \delta n \rfloor$ was selected. $b$ was formed from $Ay = b$ and equation (1) was

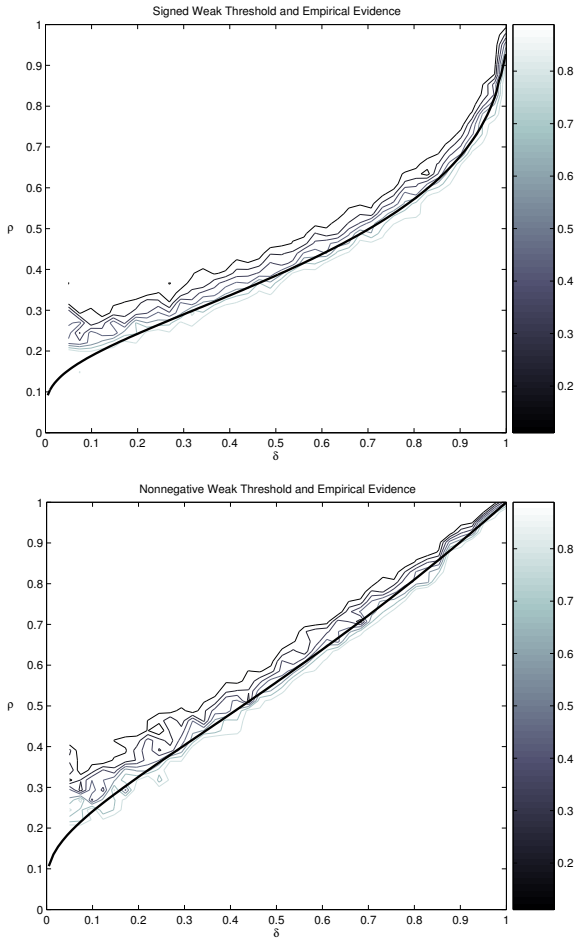


Fig. 3. Empirical verification of the weak thresholds. Fraction of $\ell^1/\ell^0$ equivalence for 50 random tests with $n = 200$ used for both signed (top) and non-negative (bottom) signals. Each plot has the associated analytic threshold, $\rho_W^{\pm}$ (top) and $\rho_W^{+}$ (bottom) overlaying the results.

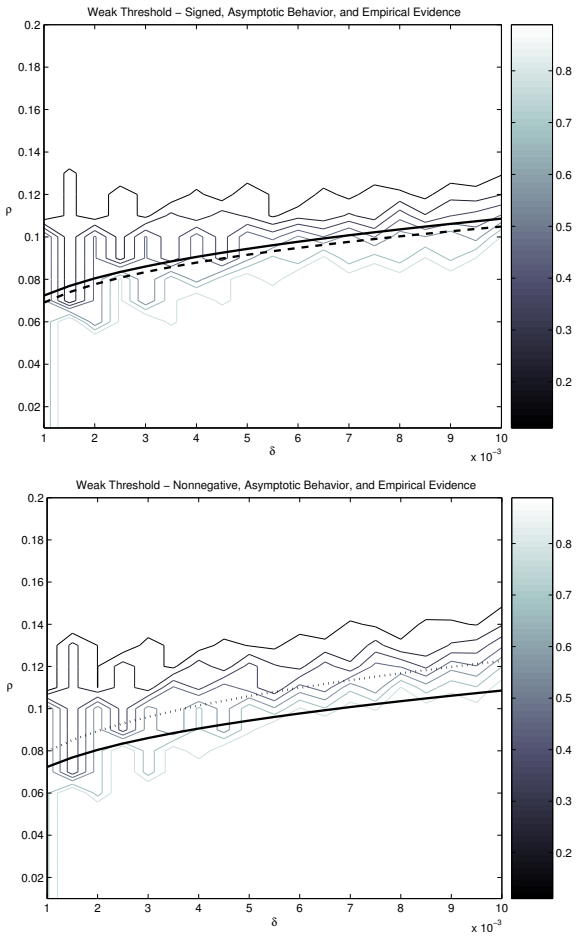


Fig. 4. Empirical verification of the weak thresholds in the asymptotic regime $\delta \in [1/1000, 1/100]$. Fraction of $\ell^1/\ell^0$ equivalence for 200 random tests with $n = 10,000$ used for both signed (top) and non-negative (bottom) signals. Each plot has the associated analytic threshold, $\rho_W^{\pm}$ (top - dashed) and $\rho_W^{+}$ (bottom - dotted) along with their asymptotic behavior from Theorem 7 (solid) overlaying the results.

then solved from $(A, b)$ to recover $y$. The fraction of successful recovery was recorded for fifty problem instances at each of the 1600 combinations of $(\delta, \rho)$. Figure 3 depicts the results of these experiments with $\rho_W^{\pm}(\delta)$ and $\rho_W^{+}(\delta)$ overlaying the corresponding signed (top) and non-negative (bottom) cases. Each of the weak thresholds track very well with the phase transition from successful to failed recovery of the sparsest solution in the empirical calculations. As these calculations were done with $n = 200$, the phase transitions are not exact step functions from success to failure, but rather steep transitions.

Similar empirical verification can be computed in the regime as $\delta \to 0$. For these experiments we use $n = 10,000$, and partition the $(\delta, \rho)$ domain as follows. Two hundred repetitions were considered for $d = 10, 15, 20, \ldots, 100$ partitioning ($\delta \in [1/1000, 1/100]$); and for each value of $d$, eleven values of the sparsity, $k$, were computed surrounding the asymptotic thresholds, $\rho_W^{\pm}(\delta) \cdot d$ and $\rho_W^{+}(\delta) \cdot d$. Figure 4 depicts the results of this experiment. The panels display the fraction of successes of the above experiments averaged over 200 random tests for each of the signed (top) and non-negative (bottom) cases. Along with the empirical results, each plot shows shows the associated analytic threshold, $\rho_W^{\pm}$ (top - dashed) and $\rho_W^{+}$ (bottom - dotted), and their asymptotic behavior, Theorem 7 - solid.

Each of the thresholds $\rho_W(\delta)$ and their asymptotic behavior as stated in Theorem 7 track very well with the phase transition from successful to failed recovery of the sparsest solution in the empirical calculations. Blocky appearance for smaller $\delta$ is due to the finite discrete values used. Fixing the region of $\delta \in [1/1000, 1/100]$ and increasing $n$ allows for both an improved achievable resolution in $k/n$, and yields a sharper transition in agreement with the analytic thresholds.

## VI. Applications to Error Correcting Codes

Consider the widely discussed application of error correcting codes, where a block of information is encoded into a longer block for redundancy against sparse errors. Unfortunately, decoding of optimal error correcting codes is NP-hard in general. Recently it has been shown that (1) allows

good, computationally feasible, decoding of certain 'random' error correcting codes; [5], [17], [19]. Let $U_0 \in \mathbb{R}^{n\times n}$ be a random unitary matrix. From $U_0$ select $n - d_n$ columns as a $n - d_n \times n$ matrix $B$ and let the remaining $d_n$ columns form $A$. These two matrices are the encoding, $B$, and recovery, $A$, matrices comprising the error correcting code. The information $\alpha \in \mathbb{R}^{n-d}$ is encoded to length $n$ by forming $s = B^T\alpha$. The redundancy allows for recovery of a corrupted signal; that is, if $s$ is transmitted, but $r = s + \eta_0$ is received, then $s$ can be recovered efficiently by solving

$$\min \|\eta\|_{\ell^1} \quad \text{subject to} \quad A\eta = b \tag{7}$$

where $b = Ar$, provided the corruption is sufficiently sparse.

For random errors, with high probability as $n \to \infty$, the vast majority of $\eta_0$ can be recovered by solving (7) provided the number of errors satisfies

$$\|\eta_0\|_{\ell^0} \le \rho^{\pm}_W(d_n/n) \cdot d_n.$$

A more interesting application is when the corruption is not random, but instead comes from a malicious source. In this setting, with high probability in the selection of $A_{d,n}$, exact recovery can by achieved efficiently by solving (7) irregardless of the magnitude or location of the corruption, provided the number of errors satisfies

$$\|\eta_0\|_{\ell^0} \le \rho^{\pm}_S(d_n/n) \cdot d_n.$$

That is, the malicious observer can know the signal being transmitted as well as the reconstruction matrix, and exact recovery is possible provided only sufficiently few entries are corrupted.

For fixed encoding redundancy $n/(n-d)$, the precise threshold $\rho^{\pm}(\delta)$ can be read from Figure 1 (top). Alternatively, for highly efficient transmission rates the redundancy approaches one, that is $d_n/n \to 0$ as $n \to \infty$. In this regime the thresholds' asymptotic behavior can be read from Theorems 6 and 7. Efficient recovery by solving (1) is possible from random errors provided the number of errors satisfies

$$\|\eta_0\|_{\ell^0} \le d_n/2\log(n/d_n) \quad n \to \infty,$$

and from malicious errors provided the number of errors satisfies

$$\|\eta_0\|_{\ell^0} \le d_n/2e\log(\sqrt{\pi}(n/d_n)) \quad n \to \infty.$$

In the limit of low redundancy, tolerance for malicious errors is $1/e$ times smaller than tolerance for random errors

## REFERENCES


[1] D. L. Donoho, "Neighborly polytopes and sparse solutions of underdetermined linear equations," *IEEE. Trans. Info. Thry.*, 2006.

[2] D. L. Donoho and J. Tanner, "Sparse nonnegative solutions of underdetermined linear equations by linear programming," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 27, pp. 9446–9451, 2005.

[3] ——, "How many random projections does one need to recover a $k-$sparse vector?" *preprint*, 2006.

[4] D. L. Donoho, "Compressed sensing," *IEEE. Trans. Info. Thry.*, 2006.

[5] E. J. Candès and T. Tao, "Near optimal signal recovery from random projections and universal encoding strategies," Applied and Computational Mathematics, California Institute of Tecnology, Tech. Rep., 2004.

[6] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, to appear.

[7] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001, reprinted from SIAM J. Sci. Comput. **20** (1998), no. 1, 33–61.

[8] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via $l^1$ minimization," *Proc. Natl. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202 (electronic), 2003.

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

[10] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inform. Theory*, vol. 48, no. 9, pp. 2558–2567, 2002.

[11] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inform. Theory*, vol. 50, no. 6, pp. 1341–1344, 2004.

[12] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inform. Theory*, vol. 49, no. 12, pp. 3320–3325, 2003.

[13] J. A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.

[14] D. L. Donoho, "High-dimensional centrally-symmetric polytopes with neighborliness proportional to dimension," *Disc. Comput. Geometry*, (online) Dec. 22, 2005.

[15] A. M. Vershik and P. V. Sporyshev, "Asymptotic behavior of the number of faces of random polyhedra and the neighborliness problem," *Selecta Math. Soviet.*, vol. 11, no. 2, pp. 181–201, 1992.

[16] D. L. Donoho and J. Tanner, "Neighborliness of randomly-projected simplices in high dimensions," *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 27, pp. 9452–9457, 2005.

[17] M. Rudelson and R. Vershynin, "Geometric approach to error-correcting codes and reconstruction of signals," Department of Mathematics, University of California, Davis, Tech. Rep., 2005. [Online]. Available: http://front.math.ucdavis.edu/math.FA/0502299

[18] Y. M. Baryshnikov and R. A. Vitale, "Regular simplices and Gaussian samples," *Discrete Comput. Geom.*, vol. 11, no. 2, pp. 141–147, 1994.

[19] E. J. Candès and T. Tao, "Decoding via linear programming," *IEEE Trans. Inform. Theory*, to appear.