

# Sharp thresholds for high-dimensional and noisy recovery of sparsity

Martin J. Wainwright  
wainwrig@eecs.berkeley.edu  
Departments of Statistics, and  
Electrical Engineering & Computer Sciences  
UC Berkeley, CA 94720

**Abstract**—The problem of consistently estimating the sparsity pattern of a vector  $\beta^* \in \mathbb{R}^p$  based on observations contaminated by noise arises in various contexts, including subset selection in regression, structure estimation in graphical models, sparse approximation, and signal denoising. Unfortunately, the natural optimization-theoretic formulation involves  $\ell_0$  constraints, which leads to NP-hard problems in general; this intractability motivates the use of relaxations based on  $\ell_1$  constraints. We analyze the behavior of  $\ell_1$ -constrained quadratic programming (QP), also referred to as the Lasso, for recovering the sparsity pattern. Our main result is to establish a sharp relation between the problem dimension  $p$ , the number  $s$  of non-zero elements in  $\beta^*$ , and the number of observations  $n$  that are required for reliable recovery. For a broad class of Gaussian ensembles satisfying mutual incoherence conditions, we establish existence and compute explicit values of thresholds  $\theta_\ell$  and  $\theta_u$  with the following properties: for any  $\nu > 0$ , if  $n > 2s(\theta_u + \nu)\log(p - s) + s + 1$ , then the Lasso succeeds in recovering the sparsity pattern with probability converging to one for large problems, whereas for  $n < 2s(\theta_\ell - \nu)\log(p - s) + s + 1$ , then the probability of successful recovery converges to zero. For the special case of the uniform Gaussian ensemble, we show that  $\theta_\ell = \theta_u = 1$ , so that the threshold is sharp and exactly determined.

**Keywords:** Quadratic programming; convex relaxation;  $\ell_0$  minimization;  $\ell_1$  relaxation; Lasso; subset selection; consistency; thresholds; sparse approximation; signal denoising; sparsity recovery; model selection.

## I. INTRODUCTION

The problem of recovering the sparsity pattern of an unknown vector  $\beta^*$ —that is, the positions of the non-zero entries of  $\beta^*$ —based on noisy observations arises in a broad variety of contexts, including subset selection in regression [1], structure estimation in graphical models [2], sparse approximation [3], and signal denoising [4]. A natural optimization-theoretic formulation of this problem is via  $\ell_0$ -minimization, where the  $\ell_0$  “norm” of a vector corresponds to the number of non-zero elements. Unfortunately, however,  $\ell_0$ -minimization problems are known to be NP-hard in general [3], so that the existence of polynomial-time algorithms is highly unlikely. This challenge motivates

the use of computationally tractable approximations or relaxations to  $\ell_0$  minimization. In particular, a great deal of research over the past decade has studied the use of the  $\ell_1$ -norm as a computationally tractable surrogate to the  $\ell_0$ -norm.

In more concrete terms, suppose that we wish to estimate an unknown but fixed vector  $\beta^* \in \mathbb{R}^p$  on the basis of a set of  $n$  observations of the form

$$Y_k = x_k^T \beta^* + W_k, \quad k = 1, \dots, n, \quad (1)$$

where  $x_k \in \mathbb{R}^p$ , and  $W_k \sim N(0, \sigma^2)$  is additive Gaussian noise. In many settings, it is natural to assume that the vector  $\beta^*$  is *sparse*, in that its *support*  $S := \{i \in \{1, \dots, p\} \mid \beta_i^* \neq 0\}$  has relatively small cardinality  $s = |S|$ . Given the observation model (1) and sparsity assumption, a reasonable approach to estimating  $\beta^*$  is by solving the  $\ell_1$ -constrained quadratic program (QP)

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{k=1}^n [Y_k - x_k^T \beta]^2 + \lambda_n \|\beta\|_1 \right\}, \quad (2)$$

where  $\lambda_n \geq 0$  is a regularization parameter. Of interest are conditions on the *ambient dimension*  $p$ , the *sparsity index*  $s$ , and the *number of observations*  $n$  for which it is possible (or impossible) to recover the support set  $S$  of  $\beta^*$ .

### A. Overview of previous work

Given the substantial literature on the use of  $\ell_1$  constraints for sparsity recovery and subset selection, we provide only a very brief (and hence necessarily incomplete) overview here. In the *noiseless version* ( $\sigma^2 = 0$ ) of the linear observation model (1), one can imagine estimating  $\beta^*$  by solving the problem

$$\min_{\beta \in \mathbb{R}^p} \|\beta\|_1 \quad \text{subject to} \quad x_k^T \beta = Y_k, \quad k = 1, \dots, n. \quad (3)$$

This problem is in fact a linear program (in disguise), and corresponds to a method in signal processing known as basis pursuit, pioneered by Chen et al. [4]. For the noiseless setting, the interesting regime is the underdetermined setting (i.e.,  $n < p$ ). With contributions from a broad range of researchers [e.g., 5], [4], [6], [7], [8], [9], there is now a fairly complete understanding of conditions on deterministic vectors  $\{x_k\}$  and sparsity index  $s$  for which the true solution  $\beta^*$  can be recovered exactly. Without going into technical details, the rough idea is that the *mutual incoherence* of the vectors  $\{x_k\}$  must be large relative to the sparsity index  $s$ , and indeed we impose similar conditions to derive our results (e.g., condition (8) in the sequel). Most closely related to the current paper—as we discuss in more detail in the sequel—are recent results by Donoho [10], as well as Candes and Tao [11] that provide high probability results

for random ensembles. More specifically, as independently established by both sets of authors using different methods, for uniform Gaussian ensembles (i.e.,  $x_k \sim N(0, I_p)$ ) with the ambient dimension  $p$  scaling linearly in terms of the number of observations (i.e.,  $p = \gamma n$ , for some  $\gamma > 1$ ), there exists a constant  $\alpha > 0$  such that all sparsity patterns with  $s \leq \alpha p$  can be recovered with high probability.

There is also a substantial body of work focusing on the noisy setting ( $\sigma^2 > 0$ ), and the use of quadratic programming techniques for sparsity recovery [e.g., 4], [12], [13], [14], [15], [2], [16]. The  $\ell_1$ -constrained quadratic program (2), also known as the Lasso [1], has been the focus of considerable research in recent years. Knight and Fu [17] analyze the asymptotic behavior of the optimal solution, not only for  $\ell_1$  regularization but for  $\ell_p$ -regularization with  $p \in (0, 2]$ . Fuchs [12] investigates optimality conditions for the constrained QP (2), and provides deterministic conditions, of the mutual incoherence form, under which a sparse solution, which is known to be within  $\epsilon$  of the observed values, can be recovered exactly. Among a variety of other results, both Tropp [13] and Donoho et al. [14] also provide sufficient conditions for the support of the optimal solution to the constrained QP (2) to be contained within the true support of  $\beta^*$ . Other authors [5], [18] have provided conditions under which estimation of a noise-contaminated signal via the Lasso is stable in the  $\ell_2$  sense; however, such  $\ell_2$ -stability does not guarantee exact recovery of the underlying sparsity pattern. Most directly related to the current paper is recent work by both Meinshausen and Bühlmann [2], focusing on Gaussian noise, and extensions by Zhao and Yu [16] to more general noise distributions, on the use of the Lasso for sparsity recovery. For the case of Gaussian noise, both papers established that under mutual incoherence conditions and appropriate choices of the regularization parameter  $\lambda_n$ , the Lasso can recover the sparsity pattern with probability converging to one for particular regimes of  $n$ ,  $p$  and  $s$ , when  $x_k$  drawn randomly from random Gaussian ensembles. We discuss connections to our results at more length in the sequel.

### B. Our contributions

Recall the linear observation model (1). For compactness in notation, let us use  $X$  to denote the  $n \times p$  matrix formed with the vectors  $x_k = (x_{k1}, x_{k2}, \dots, x_{kp}) \in \mathbb{R}^p$  as rows, and the vectors  $X_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T \in \mathbb{R}^n$  as columns. Consider the (random) set  $\mathcal{S}(X, \beta^*, W, \lambda_n)$  of optimal solutions to this constrained quadratic program (2). By convexity and boundedness of the cost function, the solution set is always non-empty. For any vector

$\beta \in \mathbb{R}^p$ , we define the sign function

$$\text{sgn}(\beta_i) := \begin{cases} +1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ 0 & \text{if } \beta_i = 0. \end{cases} \quad (4)$$

Of interest is the event that the Lasso (2) succeeds in recovering the sparsity pattern of the unknown  $\beta^*$ :

**Property  $\mathcal{R}(X, \beta^*, W, \lambda_n)$ :** There exists an optimal solution  $\hat{\beta} \in \mathcal{S}(X, \beta^*, W, \lambda_n)$  such that  $\text{sgn}(\hat{\beta}) = \text{sgn}(\beta^*)$ .

Our main result is that for a broad class of random Gaussian ensembles based on covariance matrices satisfying mutual incoherence conditions, there exist fixed constants  $0 < \theta_\ell \leq 1$  and  $1 \leq \theta_u < +\infty$  such that for all  $\nu > 0$ , property  $\mathcal{R}(X, \beta^*, W, \lambda_n)$  holds with high probability (over the choice of noise vector  $W$  and random matrix  $X$ ) whenever

$$n > 2(\theta_u + \nu) s \log(p - s) + s + 1, \quad (5)$$

and *conversely*, fails to hold with high probability whenever

$$n < 2(\theta_\ell - \nu) s \log(p - s) + s + 1. \quad (6)$$

Moreover, for the special case of the uniform Gaussian ensemble (i.e.,  $x_k \sim N(0, I_p)$ ), we show that  $\theta_\ell = \theta_u = 1$ , so that the threshold is sharp. This threshold result has a number of connections to previous work in the area that focuses on special forms of scaling. More specifically, as we discuss in more detail in Section II-B, in the special case of linear scaling (i.e.,  $n = \gamma p$  for some  $\gamma > 0$ ), this theorem provides a noisy analog of results previously established for basis pursuit in the noiseless case [10], [11]. Moreover, our result can also be adapted to an entirely different scaling regime for  $n, p$  and  $s$ , as considered by a separate body of recent work [2], [16] on the high-dimensional Lasso.

The remainder of this paper is organized as follows. Section II is devoted to the statement and proof of our main result on the asymptotic behavior of the lasso for random Gaussian ensembles. Given space constraints, we provide only an outline of the proof, with more technical lemmas only stated as results. We refer the interested reader to the complete version of the work described here that appeared earlier in technical report form [19].

## II. RECOVERY OF SPARSITY: RANDOM GAUSSIAN ENSEMBLES

We now turn to the analysis of random design matrices  $X$ , in which each row  $x_k$  is chosen as an i.i.d. Gaussian random vector with covariance matrix  $\Sigma$ . We begin by setting up and providing a precise

statement of the main result, and then discussing its connections to previous work. In the later part of this section, we provide the proof.

#### A. Statement of main result

Consider a covariance matrix  $\Sigma$  with unit diagonal, and with its minimum and maximum eigenvalues (denoted  $\Lambda_{min}$  and  $\Lambda_{max}$  respectively) bounded as

$$\Lambda_{min}(\Sigma_{SS}) \geq C_{min}, \quad \text{and} \quad \Lambda_{max}(\Sigma) \leq C_{max} \quad (7)$$

for constants  $C_{min} > 0$  and  $C_{max} < +\infty$ . Given a vector  $\beta^* \in \mathbb{R}^p$ , define its support  $S = \{i \in \{1, \dots, p\} \mid \beta_i^* \neq 0\}$ , as well as the complement  $S^c$  of its support. Suppose that  $\Sigma$  and  $S$  satisfy the conditions  $\|(\Sigma_{SS})^{-1}\|_\infty \leq D_{max}$  for some  $D_{max} < +\infty$ , and

$$\|\Sigma_{S^c S}(\Sigma_{SS})^{-1}\|_\infty \leq (1 - \epsilon) \quad (8)$$

for some  $\epsilon \in (0, 1]$ . Examples of matrix ensembles satisfying these conditions include the uniform Gaussian ensemble ( $\Sigma = I_{p \times p}$ ), Toeplitz matrix ensembles ( $\Sigma_{ij} = \mu^{|i-j|}$  for some  $\mu \in [0, 1)$ ), and certain other matrix ensembles; see Zhao and Yu [16] for further discussion. Under these conditions, we consider the linear observation model (1), where  $x_k \sim N(0, \Sigma)$  and  $W_k \sim N(0, \sigma^2)$  are independent Gaussian variables for  $k = 1, \dots, n$ . Furthermore, we define  $\rho_n := \min_{i \in S} |\beta_i^*|$ , and the sparsity index  $s = |S|$ .

**Theorem 1.** *Consider a sequence of covariance matrices  $\{\Sigma[p]\}$  and solution vectors  $\{\beta^*[p]\}$  satisfying conditions (7) and (8). Under the observation model (1), consider a sequence  $(n, p(n), s(n))$  such that  $s$ ,  $(n - s)$  and  $(p - s)$  tend to infinity. Define the thresholds*

$$\theta_\ell := \frac{(\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}})^2}{C_{max}(2 - \epsilon)^2} \leq 1, \quad \text{and}$$

$$\theta_u := \frac{\min\{2, C_{max}\}}{\epsilon^2 C_{min}} \geq 1.$$

Then for any constant  $\nu > 0$ , we have the following

- If  $n < 2(\theta_\ell - \nu)s \log(p - s) + s + 1$ , then  $\mathbb{P}[\mathcal{R}(X, \beta^*, W, \lambda_n)] \rightarrow 0$  for any non-increasing sequence  $\lambda_n > 0$ .
- Conversely, if  $n > 2(\theta_u + \nu)s \log(p - s) + s$  and  $\lambda_n \rightarrow 0$  is chosen such that  $\frac{n\lambda_n^2}{\log(p-s)} \rightarrow +\infty$  and  $\frac{1}{\rho_n} \left[ \lambda_n + \sqrt{\frac{\log s}{n}} \right] \rightarrow 0$ , then  $\mathbb{P}[\mathcal{R}(X, \beta^*, W, \lambda_n)] \rightarrow 1$ .

**Remark:** Suppose for simplicity that  $\rho_n$  remains bounded away from 0. In this case, the requirements on  $\lambda_n$  reduce to  $\lambda_n \rightarrow 0$ , and  $\lambda_n^2 n / \log(p - s) \rightarrow +\infty$ .

One suitable choice is  $\lambda_n^2 = \frac{\log(s) \log(p-s)}{n}$ , with which we have

$$\lambda_n^2 = \left( \frac{s \log(p-s)}{n} \right) \frac{\log(s)}{s} = O\left( \frac{\log s}{s} \right) \rightarrow 0,$$

and  $\frac{n\lambda_n^2}{\log(p-s)} = \log(s) \rightarrow +\infty$ . Without a bound on  $\rho_n$ , the conditions on  $\lambda_n$  constrains the rate of decrease of the minimum  $\rho_n = \min_{i \in S} |\beta_i^*|$ .

#### B. Some consequences

To develop intuition for this result, we begin by stating certain special cases as corollaries, and discussing connections to previous work.

*a) Uniform Gaussian ensemble:* First, we consider the special case of the uniform Gaussian ensemble, in which  $\Sigma = I_{p \times p}$ . Previous work by Donoho [10] as well as Candes and Tao [11] has focused on the uniform Gaussian ensemble in the noiseless ( $\sigma^2 = 0$ ) and underdetermined setting ( $n = \gamma p$  for some  $\gamma \in (0, 1)$ ). Analyzing the asymptotic behavior of the linear program (3) for recovering  $\beta^*$ , the basic result is that there exists some  $\alpha > 0$  such that all sparsity patterns with  $s \leq \alpha p$  can be recovered with high probability.

Applying Theorem 1 to the noisy version of this problem, the uniform Gaussian ensemble means that we can choose  $\epsilon = 1$ , and  $C_{min} = C_{max} = 1$ , so that the threshold constants reduce

$$\theta_\ell = \frac{(\sqrt{C_{max}} - \sqrt{C_{max} - \frac{1}{C_{max}}})^2}{C_{max}(2 - \epsilon)^2} = 1, \quad \text{and}$$

$$\theta_u = \frac{C_{max}}{\epsilon^2 C_{min}} = 1.$$

Consequently, Theorem 1 provides a sharp threshold for the behavior of the Lasso, in that failure/success is entirely determined by whether or not  $n > 2s \log(p - s) + s + 1$ . Thus, if we consider the particular linear scaling analyzed in previous work on the noiseless case [10], [11], we have:

**Corollary 1** (Linearly underdetermined setting). *Suppose that  $n = \gamma p$  for some  $\gamma \in (0, 1)$ . Then*

- If  $s = \alpha p$  for any  $\alpha \in (0, 1)$ , then  $\mathbb{P}[\mathcal{R}(X, \beta^*, W, \lambda_n)] \rightarrow 0$  for any positive sequence  $\lambda_n > 0$ .
- On the other hand, if  $s = O(\frac{p}{\log p})$ , then  $\mathbb{P}[\mathcal{R}(X, \beta^*, W, \lambda_n)] \rightarrow 1$  for any sequence  $\{\lambda_n\}$  satisfying the conditions of Theorem 1(a).

Conversely, suppose that the size  $s$  of the support of  $\beta^*$  scales linearly with the number of parameters  $p$ . The following result describes the amount of data required for the  $\ell_1$ -constrained QP to recover the sparsity pattern in the noisy setting ( $\sigma^2 > 0$ ):

**Corollary 2** (Linear fraction support). *Suppose that  $s = \alpha p$  for some  $\alpha \in (0, 1)$ . Then we require  $n > 2\alpha p \log[(1 - \alpha)p] + \alpha p$  in order to obtain exact recovery with probability converging to one for large problems.*

These two corollaries establish that there is a significant difference between recovery using basis pursuit (3) in the noiseless setting versus recovery using the Lasso (2) in the noisy setting. When the amount of data  $n$  scales only linearly with ambient dimension  $p$ , then the presence of noise means that the recoverable support size drops from a linear fraction (i.e.,  $s = \alpha p$  as in the work [10], [11]) to a sublinear fraction (i.e.,  $s = O(\frac{\log p}{p})$ , as in Corollary 1).

*b) Non-uniform Gaussian ensembles:* We now consider more general (non-uniform) Gaussian ensembles that satisfy conditions (7) and (8). As mentioned earlier, previous papers by both Meinshausen and Bühlmann [2] as well as Zhao and Yu [16] treat model selection with the high-dimensional Lasso. For suitable covariance matrices (e.g., satisfying conditions (7) and (8)), both sets of authors proved that the sparsity pattern can be recovered exactly under scaling conditions of the form

$$s = O(n^{c_1}), \quad \text{and} \quad p = O(e^{n^{c_2}}), \quad (9)$$

where  $c_1 + c_2 < 1$ . Applying Theorem 1 in this scenario, we have the following:

**Corollary 3.** *Under the scaling (9), the Lasso will recover the sparsity pattern with probability converging to one.*

In fact, under this stronger scaling (9), both papers [2], [16] proved that the probability of exact recovery converges to one at a rate exponential in some polynomial function of  $n$ . Interestingly, our results show that the Lasso can recover the sparsity pattern for a much broader range of  $(n, p, s)$  scaling.

### C. Proof Outline for Theorem 1(b)

We now turn to a proof outline for part (b) of our main result. We begin with a simple set of necessary and sufficient conditions for property  $\mathcal{R}(X, \beta^*, W, \lambda_n)$  to hold. We note that this result is not essentially new (e.g., see [12], [2], [13] for variants), and follows in a straightforward manner from optimality conditions for convex programs [20]. We define  $S := \{i \in \{1, \dots, p\} \mid \beta_i^* \neq 0\}$  to be the support of  $\beta^*$ , and let  $S^c$  be its complement. For any subset  $T \subseteq \{1, 2, \dots, p\}$ , let  $X_T$  be the  $n \times |T|$  matrix with the vectors  $\{X_i, i \in T\}$  as columns. We use the shorthand notation  $s = |S|$  and  $N = |S^c| = p - s$ .

Next define  $\vec{b} := \text{sgn}(\beta_S^*)$ , and denote by  $e_i \in \mathbb{R}^s$  the vector with 1 in the  $i^{\text{th}}$  position, and zeroes elsewhere. Consider the collections of random variables, defined for each index  $i \in S$  and  $j \in S^c$  as follows:

$$U_i := e_i^T \left( \frac{1}{n} X_S^T X_S \right)^{-1} \left[ \frac{1}{n} X_S^T W - \lambda_n \vec{b} \right], \quad (10)$$

and

$$V_j := X_j^T \left\{ X_S (X_S^T X_S)^{-1} \lambda_n \vec{b} - \left[ X_S (X_S^T X_S)^{-1} X_S^T - I_{n \times n} \right] \frac{W}{n} \right\}. \quad (11)$$

The following result provides necessary and sufficient conditions for successful recovery:

**Lemma 1.** *Assume that the matrix  $X_S^T X_S$  is invertible. Then, for any given  $\lambda_n > 0$  and noise vector  $W \in \mathbb{R}^n$ , property  $\mathcal{R}(X, \beta^*, W, \lambda_n)$  holds if and only if*

$$\max_{j \in S^c} |V_j| \leq \lambda_n, \quad \text{and} \quad (12a)$$

$$\min_{i \in S} |\beta_i^* + U_i| > 0. \quad (12b)$$

From Lemma 1, the behavior of  $\mathcal{R}(X, \beta^*, W, \lambda_n)$  can be analyzed by investigating the behavior of  $\max_{j \in S^c} |V_j|$  and  $\max_{i \in S} |U_i|$ . In particular, condition (12a) holds if and only if the event  $\mathcal{M}(V) := \{\max_{j \in S^c} |V_j| \leq \lambda_n\}$  holds. On the other hand, if we define  $\rho_n := \min_{i \in S} |\beta_i^*|$ , then the event  $\mathcal{M}(U) := \{\max_{i \in S} |U_i| \leq \rho_n\}$  is sufficient to guarantee that condition (12b) holds. Consequently, our proof is based on analyzing the asymptotic probability of these two events.

Next we note that for  $s < n$ , the random Gaussian matrix  $X_S$  will have rank  $s$  with probability one, whence the matrix  $X_S^T X_S$  is invertible with probability one. Accordingly, the necessary and sufficient conditions of Lemma 1 are applicable. The next lemma concerns the behavior of the random vector  $V = (V_1, \dots, V_N)$ , when conditioned on  $X_S$  and  $W$ . Recalling the shorthand notation  $\vec{b} := \text{sgn}(\beta^*)$ , we summarize in the following

**Lemma 2.** *Conditioned on  $X_S$  and  $W$ , the random vector  $(V \mid W, X_S)$  is Gaussian. Its mean vector is upper bounded as*

$$|\mathbb{E}[V \mid W, X_S]| \leq \lambda_n (1 - \epsilon) \mathbf{1}. \quad (13)$$

Moreover, it has conditional covariance

$$\text{cov}[V \mid W, X_S] = M_n [\Sigma_{S^c S^c} - \Sigma_{S^c S} (\Sigma_{SS})^{-1} \Sigma_{SS^c}],$$

where

$$M_n := \lambda_n^2 \vec{b}^T (X_S^T X_S)^{-1} \vec{b} + \frac{1}{n^2} W^T \left[ I_{n \times n} - X_S (X_S^T X_S)^{-1} X_S^T \right] W \quad (14)$$

is a random scaling factor.

The following lemma captures the behavior of the random scaling factor  $M_n$  defined in equation (14):

**Lemma 3.** *The random variable  $M_n$  has mean*

$$\mathbb{E}[M_n] = \frac{\lambda_n^2}{n-s-1} \vec{b}^T (\Sigma_{SS})^{-1} \vec{b} + \frac{\sigma^2 (n-s)}{n^2} \quad (15)$$

Moreover, it is sharply concentrated in that for any  $\delta > 0$ , we have  $\mathbb{P}[|M_n - \mathbb{E}[M_n]| \geq \delta \mathbb{E}[M_n]] \rightarrow 0$  as  $n \rightarrow +\infty$ .

With these preliminary results in hand, we now turn to analysis of the collections of random variables  $\{U_i, i \in S\}$  and  $\{V_j, j \in S^c\}$ .

c) *Analysis of  $\mathcal{M}(V)$ :* We begin by analyzing the behavior of  $\max_{j \in S^c} |V_j|$ . First, for a fixed but arbitrary  $\delta > 0$ , define the event

$$\mathcal{T}(\delta) := \{|M_n - \mathbb{E}[M_n]| \geq \delta \mathbb{E}[M_n]\}.$$

By conditioning on  $\mathcal{T}(\delta)$  and its complement  $[\mathcal{T}(\delta)]^c$ , we have the upper bound

$$\mathbb{P}[\max_{j \in S^c} |V_j| > \lambda_n] \leq \mathbb{P} \left[ \max_{j \in S^c} |V_j| > \lambda_n \mid [\mathcal{T}(\delta)]^c \right] + \mathbb{P}[\mathcal{T}(\delta)].$$

By the concentration statement in Lemma 3, we have  $\mathbb{P}[\mathcal{T}(\delta)] \rightarrow 0$ , so that it suffices to analyze the first term. Set  $\mu_j = \mathbb{E}[V_j | X_S]$ , and let  $Z$  be a zero-mean Gaussian vector with  $\text{cov}(Z) = \text{cov}(V | X_S, W)$ . Then we have

$$\begin{aligned} \max_{j \in S^c} |V_j| &\leq \max_{j \in S^c} [|\mu_j| + |Z_j|] \\ &\leq (1 - \epsilon) \lambda_n + \max_{j \in S^c} |Z_j|, \end{aligned}$$

where we have used the upper bound (13) on the mean. This inequality establishes the inclusion of events  $\{\max_{j \in S^c} |Z_j| \leq \epsilon \lambda_n\} \subseteq \{\max_{j \in S^c} |V_j| \leq \lambda_n\}$ , thereby showing that it suffices to prove the convergence  $\mathbb{P}[\max_{j \in S^c} |Z_j| > \epsilon \lambda_n \mid [\mathcal{T}(\delta)]^c] \rightarrow 0$ . Note that conditioned on  $[\mathcal{T}(\delta)]^c$ , the maximum value of  $M_n$  is  $v^* := (1 + \delta) \mathbb{E}[M_n]$ . Since Gaussian maxima increase with increasing variance, we have  $\mathbb{P}[\max_{j \in S^c} |Z_j| > \epsilon \lambda_n \mid [\mathcal{T}(\delta)]^c] \leq \mathbb{P}[\max_{j \in S^c} |\tilde{Z}_j| > \epsilon \lambda_n]$ , where  $\tilde{Z}$  is zero-mean Gaussian with covariance  $v^* \Sigma_{(S^c|S)}$ . It suffices to show that  $\mathbb{P}[\max_{j \in S^c} \tilde{Z}_j > \epsilon \lambda_n]$  converges to zero. Accordingly, we complete this part of the proof via the following two lemmas:

**Lemma 4.** *Under the theorem assumptions,  $\frac{v^*}{\lambda_n^2} \rightarrow 0$  and  $\lim_{n \rightarrow +\infty} \frac{1}{\lambda_n} \mathbb{E}[\max_{j \in S^c} \tilde{Z}_j] \leq \epsilon$ .*

**Lemma 5.** *For any  $\eta > 0$ , we have*

$$\mathbb{P} \left[ \max_{j \in S^c} \tilde{Z}_j > \eta + \mathbb{E}[\max_{j \in S^c} \tilde{Z}_j] \right] \leq \exp \left( -\frac{\eta^2}{2v^*} \right). \quad (16)$$

Lemma 4 implies that for all  $\delta > 0$ , we have  $\mathbb{E}[\max_{j \in S^c} \tilde{Z}_j] \leq (1 + \frac{\delta}{2}) \epsilon \lambda_n$  for all  $n$  sufficiently large. Therefore, setting  $\eta = \frac{\delta}{2} \lambda_n \epsilon$  in Lemma 5, we have for fixed  $\delta > 0$  and  $n$  sufficiently large:

$$\begin{aligned} \mathbb{P} \left[ \max_{j \in S^c} \tilde{Z}_j > (1 + \delta) \lambda_n \epsilon \right] &\leq \\ \mathbb{P} \left[ \max_{j \in S^c} \tilde{Z}_j > \frac{\delta}{2} \lambda_n \epsilon + \mathbb{E}[\max_{j \in S^c} \tilde{Z}_j] \right] &\leq \\ 2 \exp \left( -\frac{\delta^2 \lambda_n^2 \epsilon^2}{8v^*} \right). \end{aligned}$$

From Lemma 4, we have  $\lambda_n^2 / v^* \rightarrow +\infty$ , which implies that  $\mathbb{P}[\max_{j \in S^c} \tilde{Z}_j > (1 + \delta) \lambda_n \epsilon] \rightarrow 0$  for all  $\delta > 0$ . By the arbitrariness of  $\delta > 0$ , we thus have  $\mathbb{P}[\max_{j \in S^c} \tilde{Z}_j \leq \epsilon \lambda_n] \rightarrow 1$ , thereby establishing that property (12a) of Lemma 1 holds w.p. one asymptotically.

d) *Analysis of  $\mathcal{M}(U)$ :* Next we prove that  $\max_{i \in S} |U_i| < \rho_n := \min_{i \in S} |\beta_i^*|$  with probability one as  $n \rightarrow +\infty$ . Conditioned on  $X_S$ , the only random component in  $U_i$  is the noise vector  $W$ . A straightforward calculation yields that this conditioned RV is Gaussian, with mean and variance

$$\begin{aligned} Y_i &:= \mathbb{E}[U_i \mid X_S] = -\lambda_n e_i^T \left( \frac{1}{n} X_S^T X_S \right)^{-1} \vec{b}, \\ Y_i' &:= \text{var}[U_i \mid X_S] = \frac{\sigma^2}{n} e_i^T \left[ \frac{1}{n} X_S^T X_S \right]^{-1} e_i \end{aligned}$$

respectively.

Now define the event

$$\mathcal{T}(\delta) := \bigcup_{i=1}^s \left\{ |Y_i| \geq \frac{6D_{\max} n \lambda_n}{n-s-1}, \text{ or } |Y_i'| \geq 2\mathbb{E}[Y_i'] \right\}.$$

By applying the union bound and some algebra, the probability of this event is bounded as

$$\mathbb{P}[\mathcal{T}(\delta)] \leq s \frac{K}{n-s} = \frac{K}{\frac{n}{s}-1} \rightarrow 0,$$

since  $\frac{n}{s} \rightarrow +\infty$  as  $n \rightarrow +\infty$ . For convenience in notation, for any  $a \in \mathbb{R}$  and  $b \in \mathbb{R}_+$ , we use  $U_i(a, b)$  to denote a Gaussian random variable with mean  $a$  and variance  $b$ . Conditioning on the event  $\mathcal{T}(\delta)$  and

its complement, we have that  $\mathbb{P}[\max_{i \in S} U_i > \rho_n]$  is upper bounded by

$$\mathbb{P}[\max_{i \in S} U_i > \rho_n \mid \mathcal{T}(\delta)^c] + \mathbb{P}[\mathcal{T}(\delta)] \leq \mathbb{P}[\max_{i \in S} U_i(\mu_i^*, v_i^*) > \rho_n] + \frac{K}{\frac{n}{s} - 1}, \quad (17)$$

where each  $U_i(\mu_i^*, v_i^*)$  is Gaussian with mean  $\mu_i^* := 6D_{\max} \lambda_n \frac{n}{n-s-1}$  and variance  $v_i^* := 2\mathbb{E}[Y_i']$  respectively. In asserting the inequality (17), we have used the fact that the probability of the event  $\{\max_{i \in S} Y_i > \rho_n\}$  increases as the mean and variance of  $Y_i$  increase. Continuing the argument, we have

$$\begin{aligned} \mathbb{P}[\max_{i \in S} U_i(\mu_i^*, v_i^*) > \rho_n] &\leq \mathbb{P}[\max_{i \in S} |U_i(\mu_i^*, v_i^*)| > \rho_n] \\ &\leq \frac{1}{\rho_n} \mathbb{E} \left[ \max_{i \in S} |U_i(\mu_i^*, v_i^*)| \right], \end{aligned}$$

where the last step uses Markov's inequality. We now decompose  $U_i(\mu_i^*, v_i^*) \stackrel{d}{=} 2D_{\max} \lambda_n \frac{n}{n-s-1} + \tilde{U}_i(0, v_i^*)$ , and write

$$\begin{aligned} \mathbb{E} \left[ \max_{i \in S} |U_i(\mu_i^*, v_i^*)| \right] &\leq 2D_{\max} \lambda_n \frac{n}{n-s-1} \\ &\quad + \mathbb{E} \left[ \max_{i \in S} |\tilde{U}_i(0, v_i^*)| \right]. \end{aligned}$$

With this decomposition, we first bound  $v_i^* := 2\mathbb{E}[Y_i']$  and apply standard results on Gaussian maxima [21] to conclude that

$$\begin{aligned} \frac{1}{\rho_n} \mathbb{E} \left[ \max_{i \in S} |U_i(\mu_i^*, v_i^*)| \right] &\leq \\ \frac{1}{\rho_n} \left[ 2D_{\max} \lambda_n \frac{n}{n-s-1} + 3\sqrt{\frac{2\sigma^2 D_{\max} \log s}{n-s-1}} \right], \end{aligned}$$

which converges to zero by the second condition on  $\lambda_n$  in the theorem statement.

#### D. Proof Outline for Theorem 1(a)

We establish the claim by proving that under the stated conditions,  $\max_{j \in S^c} |\tilde{V}_j| > \lambda_n$  with probability one, for any positive sequence  $\lambda_n > 0$ . We begin by writing  $V_j = \mathbb{E}[V_j] + \tilde{V}_j$ , where  $\tilde{V}_j$  is zero-mean. Now

$$\begin{aligned} \max_{j \in S^c} |V_j| &\geq \max_{j \in S^c} |\tilde{V}_j| - \max_{j \in S^c} |\mathbb{E}[V_j]| \\ &\geq \max_{j \in S^c} |\tilde{V}_j| - (1-\epsilon)\lambda_n, \end{aligned}$$

where we have used Lemma 2. Consequently, the event  $\{\max_{j \in S^c} |\tilde{V}_j| > (2-\epsilon)\lambda_n\}$  implies the event  $\{\max_{j \in S^c} |V_j| > \lambda_n\}$ , so that

$$\mathbb{P}[\max_{j \in S^c} |V_j| > \lambda_n] \geq \mathbb{P}[\max_{j \in S^c} |\tilde{V}_j| > (2-\epsilon)\lambda_n].$$

From the preceding proof of Theorem 1(b), we know that conditioned on  $X_S$  and  $W$ , the random

vector  $(V_1, \dots, V_N)$  is Gaussian with covariance of the form  $M_n [\Sigma_{S^c S^c} - \Sigma_{S^c S} (\Sigma_{SS})^{-1} \Sigma_{SS^c}]$ ; thus, the zero-mean version  $(\tilde{V}_1, \dots, \tilde{V}_N)$  has the same covariance. Moreover, Lemma 3 guarantees that the random scaling term  $M_n$  is sharply concentrated. In particular, defining for any  $\delta > 0$  the event  $\mathcal{T}(\delta) := \{|M_n - \mathbb{E}[M_n]| \geq \delta \mathbb{E}[M_n]\}$ , we have  $\mathbb{P}[\mathcal{T}(\delta)] \rightarrow 0$ , and the bounds

$$\begin{aligned} \mathbb{P}[\max_{j \in S^c} |\tilde{V}_j| > (2-\epsilon)\lambda_n] &\geq \\ (1 - \mathbb{P}[\mathcal{T}(\delta)]) \mathbb{P} \left[ \max_{j \in S^c} |\tilde{V}_j| > (2-\epsilon)\lambda_n \mid \mathcal{T}(\delta)^c \right] &\geq \\ (1 - \mathbb{P}[\mathcal{T}(\delta)]) \mathbb{P} \left[ \max_{j \in S^c} |Z_j(v^*)| > (2-\epsilon)\lambda_n \right]. \end{aligned}$$

where each  $Z_j \equiv Z_j(v^*)$  is the conditioned version of  $\tilde{V}_j$  with the scaling factor  $M_n$  fixed to  $v^* := (1-\delta)\mathbb{E}[M_n]$ . (Here we have used the fact that the probability of Gaussian maxima decreases as the variance decreases, and that  $\text{var}(\tilde{V}_j) \geq v^*$  when conditioned on  $\mathcal{T}(\delta)^c$ .)

Our proof proceeds by first analyzing the expected value, and then exploiting Gaussian concentration of measure. We summarize the key results in the following:

**Lemma 6.** *Under the stated conditions, one of the following two conditions must hold:*

- (a) *either  $\frac{\lambda_n^2}{v^*} \rightarrow +\infty$ , and there exists some  $\gamma > 0$  such that  $\frac{1}{\lambda_n} \mathbb{E}[\max_{j \in S^c} Z_j] \geq (2-\epsilon)[1+\gamma]$  for all sufficiently large  $n$ , or*
- (b) *there exist constants  $\alpha, \gamma > 0$  such that  $\frac{v^*}{\lambda_n^2} \leq \alpha$  and  $\frac{1}{\lambda_n} \mathbb{E}[\max_{j \in S^c} Z_j] \geq \gamma \sqrt{\log N}$  for all sufficiently large  $n$ .*

**Lemma 7.** *For any  $\eta > 0$ , we have*

$$\mathbb{P}[\max_{j \in S^c} Z_j(v^*) < \mathbb{E}[\max_{j \in S^c} Z_j(v^*)] - \eta] \leq \exp\left(-\frac{\eta^2}{2v^*}\right). \quad (18)$$

Using these two lemmas, we complete the proof as follows. First, if condition (a) of Lemma 6 holds, then we set  $\eta = \frac{(2-\epsilon)\gamma\lambda_n}{2}$  in Lemma 7 to obtain that  $\mathbb{P}[\frac{1}{\lambda_n} \max_{j \in S^c} Z_j(v^*) \geq (2-\epsilon)(1+\frac{\gamma}{2})] \geq 1 - \exp\left(-\frac{(2-\epsilon)^2 \gamma^2 \lambda_n^2}{8v^*}\right)$ , which converges to 1 since  $\frac{\lambda_n^2}{v^*} \rightarrow +\infty$  from Lemma 6(a).

On the other hand, if condition (b) holds, then we use the bound  $\frac{1}{\lambda_n} \mathbb{E}[\max_{j \in S^c} Z_j] \geq \gamma \sqrt{\log N}$  and set  $\eta = \frac{\gamma\lambda_n \sqrt{\log N}}{2}$  in Lemma 7 to obtain that the probability  $\mathbb{P}[\frac{1}{\lambda_n} \max_{j \in S^c} Z_j(v^*) > 2(2-\epsilon)]$  is lower

bounded by

$$\mathbb{P}\left[\frac{1}{\lambda_n} \max_{j \in S^c} Z_j(v^*) \geq \frac{\gamma \sqrt{\log N}}{2}\right] \geq 1 - \exp\left(-\frac{\gamma^2 \lambda_n^2 \log N}{8v^*}\right).$$

This probability also converges to 1 since  $\frac{\lambda_n^2}{v^*} \geq 1/\alpha$  and  $\log N \rightarrow +\infty$ . Thus, in either case, we have shown that  $\lim_{n \rightarrow +\infty} \mathbb{P}\left[\frac{1}{\lambda_n} \max_{j \in S^c} Z_j(v^*) > (2 - \epsilon)\right] = 1$ , thereby completing the proof of Theorem 1(a).

### E. Simulations

We conclude with some simulations to confirm the threshold behavior predicted by Theorem 1. We consider the following three types of sparsity indices: (a) *linear sparsity*, meaning that  $s(p) = \alpha p$  for some  $\alpha \in (0, 1)$ ; (b) *sublinear sparsity*, meaning that  $s(p) = \alpha p / (\log(\alpha p))$  for some  $\alpha \in (0, 1)$ , and (c) *fractional power sparsity*, meaning that  $s(p) = \alpha p^\gamma$  for some  $\alpha, \gamma \in (0, 1)$ . For all three types of sparsity indices, we investigate the success/failure of the Lasso in recovering the sparsity pattern, where the number of observations scales as  $n = 2\theta s \log(p - s) + s + 1$ . The *control parameter*  $\theta$  is varied in the interval  $(0, 2.4)$ . For all results shown here, we fixed  $\alpha = 0.40$  for all three ensembles, and set  $\gamma = 0.75$  for the fractional power ensemble. In addition, we set  $\lambda_n = \sqrt{\frac{\log(p-s)\log(s)}{n}}$  in all cases. Here we show results for the uniform Gaussian ensemble, in which each row  $x_k$  is chosen in an i.i.d. manner from the multivariate  $N(0, I_{p \times p})$  distribution. Recall that for the uniform Gaussian ensemble, the critical value is  $\theta_u = \theta_\ell = 1$ . Figure 1 plots the control parameter  $\theta$  versus the probability of success for linear sparsity (a), sublinear sparsity pattern (b), and fractional power sparsity (c), for three different problem sizes ( $p \in \{128, 256, 512\}$ ). Each point represents the average of 200 trials. Note how the probability of success rises rapidly from 0 around the predicted threshold point  $\theta = 1$ , with the sharpness of the threshold increasing for larger problem sizes.

We now consider a non-uniform Gaussian ensemble—in particular, one in which the covariance matrices  $\Sigma$  are Toeplitz with the structure  $\Sigma_{ij} = \rho^{|i-j|}$  for some  $\rho \in (-1, 1)$ .

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{p-2} & \rho^{p-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{p-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{p-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{p-1} & \dots & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix} \quad (19)$$

for some  $\rho \in (-1, +1)$ . As shown by Zhao and Yu [16], this family of Toeplitz matrices satisfy condition (8). Moreover, the maximum and minimum eigenvalues ( $C_{min}$  and  $C_{max}$ ) can be computed using standard asymptotic results on Toeplitz matrix families [22]. Figure 2 shows representative results for this Toeplitz family with  $\rho = 0.10$ . Panel (a) corresponds to linear sparsity  $s = \alpha p$  with  $\alpha = 0.40$ , and panel (b) corresponds to sublinear sparsity ( $s = \alpha p / \log(\alpha p)$  with  $\alpha = 0.40$ ). Each panel shows three curves, corresponding to the problem sizes  $p \in \{128, 256, 512\}$ , and each point on each curve represents the average of 200 trials. The vertical lines to the left and right of  $\theta = 1$  represent the theoretical upper and lower bounds on the threshold. Once again, these simulations show good agreement with the theoretical predictions.

## III. DISCUSSION

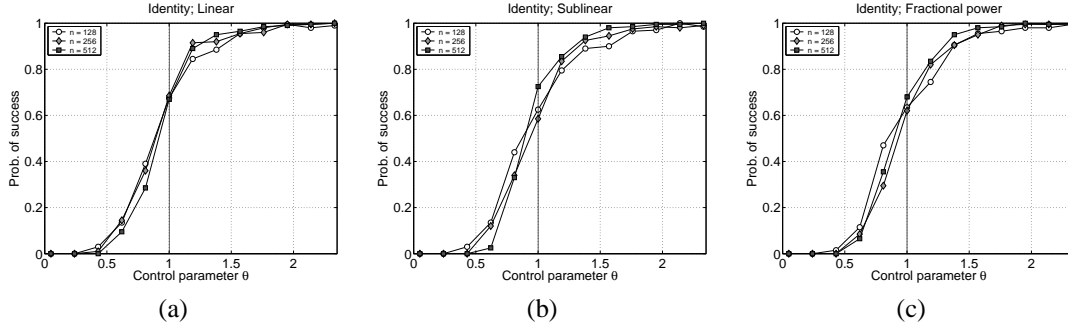
The problem of recovering the sparsity pattern of a high-dimensional vector  $\beta^*$  from noisy observations has important applications in signal denoising, graphical model selection, sparse approximation, and subset selection. This paper focuses on the behavior of  $\ell_1$ -regularized quadratic programming, also known as the Lasso, for estimating such sparsity patterns in the noisy and high-dimensional setting. The main contribution of this paper is to establish a set of general and sharp conditions on the observations  $n$ , the sparsity index  $s$  (i.e., number of non-zero entries in  $\beta^*$ ), and the ambient dimension  $p$  that characterize the success/failure behavior of the Lasso in the high-dimensional setting, in which  $n$ ,  $p$  and  $s$  all tend to infinity. For the uniform Gaussian ensemble, our threshold result is sharp, whereas for more general Gaussian ensembles, it should be possible to tighten the analysis given here.

### Acknowledgements

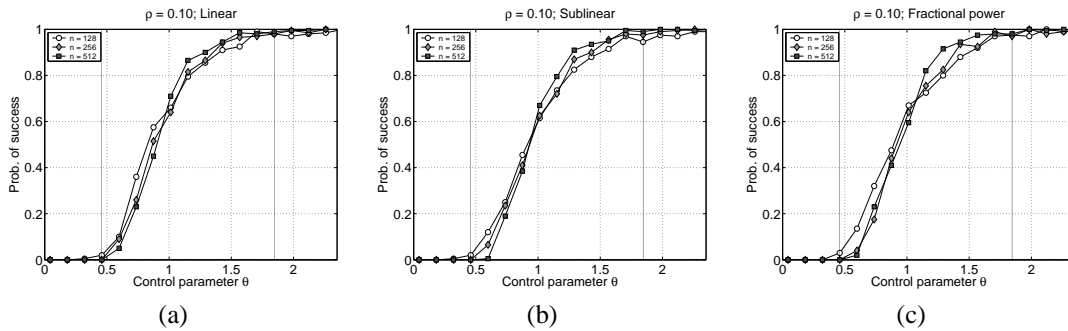
We would like to thank Noureddine El Karoui, Alyson Fletcher, Vivek Goyal and Bin Yu for helpful comments and pointers. This work was partially supported by an Alfred P. Sloan Foundation Fellowship, an Intel Corporation Equipment Grant, and NSF Grant DMS-0605165.

## REFERENCES

- [1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.
- [2] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *Annals of Statistics*, 2006, to appear.
- [3] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Computing*, vol. 24, no. 2, pp. 227–234, 1995.
- [4] S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Computing*, vol. 20, no. 1, pp. 33–61, 1998.



**Fig. 1.** Plots of the number of data samples (indexed by the control parameter  $\theta$ ) versus the probability of success in the Lasso for the uniform Gaussian ensemble. Each panel shows three curves, corresponding to the problem sizes  $p \in \{128, 256, 512\}$ , and each point on each curve represents the average of 200 trials. (a) Linear sparsity index:  $s(p) = \alpha p$ . (b) Sublinear sparsity index  $s(p) = \alpha p / \log(\alpha p)$ . (c) Fractional power sparsity index  $s(p) = \alpha p^\gamma$  with  $\gamma = 0.75$ .



**Fig. 2.** Plots of the number of data samples (indexed by the control parameter  $\theta$ ) versus the probability of success in the Lasso for the Toeplitz family with  $\rho = 0.10$ . Each panel shows three curves, corresponding to the problem sizes  $p \in \{128, 256, 512\}$ , and each point on each curve represents the average of 200 trials. (a) Linear sparsity index:  $s(p) = \alpha p$ . (b) Sublinear sparsity index  $s(p) = \alpha p / \log(\alpha p)$ . (c) Fractional power sparsity index  $s(p) = \alpha p^\gamma$  with  $\gamma = 0.75$ .

- [5] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," Caltech, Tech. Rep., 2004.
- [6] D. Donoho, "Compressed sensing," *IEEE Trans. Info Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.
- [7] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Info Theory*, vol. 48, no. 9, pp. 2558–2567, September 2002.
- [8] A. Feuer and A. Nemirovski, "On sparse representation in pairs of bases," *IEEE Trans. Info Theory*, vol. 49, no. 6, pp. 1579–1581, 2003.
- [9] J. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Info Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.
- [10] D. Donoho, "For most large undetermined system of linear equations the minimal  $l_1$ -norm near-solution is also the sparsest solution," Stanford University, Tech. Rep., 2004.
- [11] E. Candes and T. Tao, "Decoding by linear programming," *IEEE Trans. Info Theory*, vol. 51, no. 12, pp. 4203–4215, December 2005.
- [12] J. J. Fuchs, "Recovery of exact sparse representations in the presence of noise," *IEEE Trans. Info. Theory*, vol. 51, no. 10, pp. 3601–3608, October 2005.
- [13] J. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Info Theory*, vol. 52, no. 3, pp. 1030–1051, March 2006.
- [14] D. Donoho, M. Elad, and V. M. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Info Theory*, vol. 52, no. 1, pp. 6–18, January 2006.
- [15] A. K. Fletcher, S. Rangan, V. K. Goyal, and K. Ramchandran, "Denoising by sparse approximation: Error bounds based on rate-distortion theory," *Journal on Applied Signal Processing*, vol. 10, pp. 1–19, 2006.
- [16] P. Zhao and B. Yu, "Model selection with the lasso," UC Berkeley, Department of Statistics, Tech. Rep., March 2006, accepted to *Journal of Machine Learning Research*.
- [17] K. Knight and W. J. Fu, "Asymptotics for lasso-type estimators," *Annals of Statistics*, vol. 28, pp. 1356–1378, 2000.
- [18] D. Donoho, "For most large undetermined system of linear equations the minimal  $l_1$ -norm near-solution approximates the sparsest solution," Stanford University, Tech. Rep., 2004.
- [19] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy recovery of sparsity," Department of Statistics, UC Berkeley, Tech. Rep. 709, 2006.
- [20] J. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*. New York: Springer-Verlag, 1993, vol. 1.
- [21] M. Ledoux and M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes*. New York, NY: Springer-Verlag, 1991.
- [22] R. M. Gray, "Toeplitz and Circulant Matrices: A Review," Stanford University, Information Systems Laboratory, Tech. Rep., 1990.