A Note on Optimal Support Recovery in Compressed Sensing

Galen Reeves and Michael Gastpar
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Email: {greeves, gastpar}@eecs.berkeley.edu

***Abstract*—Recovery of the support set (or sparsity pattern) of a sparse vector from a small number of noisy linear projections (or samples) is a "compressed sensing" problem that arises in signal processing and statistics. Although many computationally efficient recovery algorithms have been studied, the optimality (or gap from optimality) of these algorithms is, in general, not well understood. In this note, approximate support recovery under a Gaussian prior is considered, and it is shown that optimal estimation depends on the recovery metric in general. By contrast, it is shown that in the SNR limits, there exist uniformly near-optimal estimators, namely, the ML estimate in the high SNR case, and a computationally trivial thresholding algorithm in the low SNR case.**

I. Introduction

The task of support recovery is to determine which elements of an unknown sparse vector $\mathbf{x} \in \mathbb{R}^n$ are non-zero based on a set of noisy linear observations $\mathbf{y} = A\mathbf{x}+\mathbf{w}$ where $A \in \mathbb{R}^{m \times n}$ is a known sampling matrix and $\mathbf{w} \in \mathbb{R}^m$ is an unknown error term. This problem, which is variously known as recovery of the sparsity pattern or model selection, has been studied extensively in the signal processing and statistics literature [1]–[11]. In many cases of interest, the number of observations m is far less than the signal length n and hence A is not invertible.

In the special case where the measurements are uncorrupted by noise, it is well known that $\mathbf{x}$ can be recovered exactly using using ℓ_0 minimization provided that the sampling matrix A obeys certain key properties. Although such recovery is NP hard in general, it has been shown [12]–[14] that if A obeys a few additional properties, then $\mathbf{x}$ can also be recovered exactly using a polynomial time convex relaxation (linear programming).

In the general setting, the analysis of recovery algorithms is more complex since performance depends on the assumptions about the noise, the values of the non-zero signal elements, and the recovery criteria. Although many recovery algorithms have been shown to perform well in certain settings, it is often not known if their performance can be improved significantly. Furthermore, the fundamental tradeoff between performance and computational complexity is poorly understood in general.

In this note, we consider approximate support recovery with respect to Gaussian signal priors. Our analysis reveals key insights about the computational complexity and universality of optimal recovery in the high and low SNR regimes.

II. Results

Given any vector $\mathbf{x} \in \mathbb{R}^n$, the support $\mathbf{s} \subset \{1, 2, \cdots, n\}$ is the set of integers indexing the non-zero elements of $\mathbf{x}$. We assume throughout that the sparsity $k = |\mathbf{s}|$ is known and that any estimate $\hat{\mathbf{s}}$ has size k. The distortion $d(\hat{\mathbf{s}}, \mathbf{s}) = 1 - |\hat{\mathbf{s}} \cap \mathbf{s}|/|\mathbf{s}|$ is used to measure the fraction of errors.

Proposition 1. *Suppose that the support* $\mathbf{S}$ *is distributed uniformly over all subsets of size* k*, the non-zero elements* $\{X_i\}_{i \in \mathbf{S}}$ *are i.i.d.* $\mathcal{N}(0, \frac{P}{1+P})$*, and the errors* $\{W_i\}_{i=1}^m$ *are i.i.d.* $\mathcal{N}(0, \frac{1}{1+P})$*. Given any error fraction* $\alpha \in [0, 1]$*, the probability that* $d(\hat{\mathbf{s}}(\mathbf{Y}), \mathbf{S}) > \alpha$ *is minimized by the estimate*

$$\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{y}; \alpha, P) = \arg\max_{\mathbf{s}} \sum_{\mathbf{s}' : d(\mathbf{s}', \mathbf{s}) < \alpha} e^{-\frac{1}{2}\left[\|\Sigma_{\mathbf{s}'}^{-1/2}\mathbf{y}\|^2 + \log|\Sigma_{\mathbf{s}'}|\right]}$$

where $\Sigma_{\mathbf{s}} = \frac{1}{1+P} \cdot I + \frac{P}{1+P} \cdot A_{\mathbf{s}} A_{\mathbf{s}}^T$.

In contrast to the estimate $\hat{s}_{\text{OPT}}$, which depends on the sparsity k, the error fraction α and the relative power P, the following two recovery algorithms depend only on the sparsity.

The *nearest subspace* (NS) estimate [4] corresponds to the maximum likelihood estimate of $\mathbf{x}$ and is given by the combinatorial optimization problem

$$\hat{\mathbf{s}}_{\text{NS}}(\mathbf{y}) = \arg\min_{\mathbf{s}} \|\Pi_{\mathbf{s}} \mathbf{y}\| \tag{1}$$

where $\Pi_{\mathbf{s}} = I - A_{\mathbf{s}}\big(A_{\mathbf{s}}^T A_{\mathbf{s}}\big)^{-1} A_{\mathbf{s}}^T$ if $A_{\mathbf{s}}^T A_{\mathbf{s}}$ is invertible and is equal to a matrix of zeros otherwise.

The *thresholding* (TH) estimate [10] amounts to identifying the k largest elements of the vector $A^T\mathbf{y} \in \mathbb{R}^n$ and can be expressed as

$$\hat{\mathbf{s}}_{\text{TH}}(\mathbf{y}) = \arg\max_{\mathbf{s}} \|A_{\mathbf{s}}\mathbf{y}\|^2 - \text{tr}\big(A_{\mathbf{s}}^T A_{\mathbf{s}}\big) \,. \tag{2}$$

The following theorems show that the nearest subspace and thresholding estimates correspond to the optimal estimate (for any α) at high and low SNR respectively.

Theorem 1 (High SNR)**.** *Let* A *and* $\mathbf{y}$ *be fixed. If* $\hat{\mathbf{s}}_{\text{NS}}(\mathbf{y})$ *is unique, then there exists* $P_{A,\mathbf{y}} < \infty$ *such that for all* $P > P_{A,\mathbf{y}}$ *and* $\alpha \in [0, 1)$*,*

$$\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{y}; \alpha, P) = \hat{\mathbf{s}}_{\text{NS}}(\mathbf{y}).$$

Theorem 2 (Low SNR)**.** *Let* A *and* $\mathbf{y}$ *be fixed. If,* $\hat{\mathbf{s}}_{\text{TH}}(\mathbf{y})$ *is unique, then there exists* $P_{A,\mathbf{y}} > 0$ *such that for all* $P < P_{A,\mathbf{y}}$ *and* $\alpha \in [0, 1)$*,*

$$\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{y}; \alpha, P) = \hat{\mathbf{s}}_{\text{TH}}(\mathbf{y})$$

One immediate consequence of Theorems 1 and 2 is that if the joint distribution on $\mathbf{S}$ and $\mathbf{Y}$ is given by the assumptions of Proposition 1, and if each submatrix $A_{\mathbf{s}}$ has a unique range space of dimension k, then for any $\alpha \in [0, 1)$,

$$\begin{aligned} &\Pr\left\{\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{Y};\alpha,P) = \hat{\mathbf{s}}_{\text{NS}}(\mathbf{Y})\right\} \to 1 \quad \text{as} \quad P \to \infty, \quad \text{and} \\ &\Pr\left\{\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{Y};\alpha,P) = \hat{\mathbf{s}}_{\text{TH}}(\mathbf{Y})\right\} \to 1 \quad \text{as} \quad P \to 0. \end{aligned}$$

This convergence holds for any value of α, thus showing that universal estimators, irrespective of the value of α, exist in the SNR limits. However, the *rate* at which this convergence occurs can depend strongly on the value of α, as the following simple example illustrates.

Proposition 2. *Suppose that the joint distribution on the support* $\mathbf{S}$ *and samples* $\mathbf{Y}$ *is given by the assumptions of Proposition 1, and that* $n = 4$*,* $k = 2$*,* $m = 3$ *and*

$$A = \frac{1}{2}\begin{bmatrix} -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \end{bmatrix}.$$

Then, as $P \to \infty$

$$\Pr\left\{d\big(\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{Y};\tfrac{1}{2},P),\mathbf{S}\big) > \tfrac{1}{2}\right\} = O\big(\tfrac{1}{P}\big) \tag{3}$$

$$\Pr\left\{d\big(\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{Y};0,P),\mathbf{S}\big) > \tfrac{1}{2}\right\} = \Omega\big(\tfrac{1}{\sqrt{P}}\big). \tag{4}$$

III. Discussion

To understand the significance of the results in this paper, we first consider the special case of exact recovery (i.e. $\alpha = 0$ for our recovery metric). In this setting, a great deal of previous work [1]–[4], [10] has derived necessary and sufficient conditions on scalings of the tuple (n, k, m) to ensure reliable recovery in the high dimensional setting where $n \to \infty$. In particular, it has been shown that if the per-sample SNR is a finite constant independent of n, then the *scaling conditions* of the nearest subspace and thresholding estimators are information-theoretically optimal. However, these scaling conditions do not tell the whole story since the dependence of the bounds on the SNR and the further assumptions about the values of the non-zero elements of $\mathbf{x}$ are absorbed in the (typically unknown) constants of the bounds.

In comparison to the scaling results outlined above, Theorems 1 and 2 in this paper show that, under a Gaussian prior, the nearest subspace and thresholding estimates are near-optimal in their respective SNR settings. This convergence is non-asymptotic in the dimensions (n, k, m). At low SNR, this result suggests that there is little improvement to be gained using any algorithm other than the computationally simple thresholding estimate. At high SNR, this result validates the use of computationally efficient convex relaxations to the nearest subspace algorithm such as *Basis Pursuit* [15] and *LASSO* [16].

It is interesting to note that for exact recovery, the optimal estimate under a Gaussian prior simplifies to

$$\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{y};0,P) = \arg\min_{\mathbf{s}} \|\Sigma_{\mathbf{s}'}^{-1/2}\mathbf{y}\|^2 + \log|\Sigma_{\mathbf{s}'}|,$$

and thus the convergence shown by Theorems 1 and 2 is much easier to prove than in the general case (i.e. $\alpha > 0$). However, to our knowledge, these connections between optimal estimation and the nearest subspace and thresholding estimators have not been addressed explicitly in previous work, even for exact recovery.

In many practically inspired settings, however, exact recovery is infeasible and approximate support recovery guarantees are needed. For example, if the per-sample SNR and the ratios k/n and m/n are finite constants independent of n, then any support estimator will have a constant fraction of errors as n becomes large [17]. Parallel to the setting of exact recovery, previous work [5]–[9] has focused on conditions for asymptotically reliable approximate recovery. It has been shown that if a fraction $\alpha > 0$ of errors are allowed, then the scaling conditions on (n, k, m) are fundamentally different than in the exact recovery setting and are comparable to the conditions needed for other recovery tasks such as estimation of $\mathbf{x}$ with bounded mean squared error. Moreover, it also has been shown that scaling conditions of the nearest subspace and the thresholding estimators are, again, information-theoretically optimal.

Despite the insights given by the above scaling results, there exists several important questions about optimal estimation in in the approximate recovery setting. For example, are the nearest subspace and thresholding estimates still near-optimal (with respect to constants) in the SNR limits? Can significantly better (i.e. more reliable) recovery be achieved by optimizing for a targeted error fraction?

The results in this paper provide valuable insights about the above questions. For example, the fact that Theorems 1 and 2 hold uniformly for all α shows that, under a Gaussian prior, the nearest subspace and thresholding estimators are, indeed, near-optimal in the SNR limits. However, the cautionary example given in Proposition 2 illustrates that performance (in this case, the probability that the fraction of errors exceeds α) may depend significantly on whether or not the error bound α is taken into account during estimation.

Another source of insight into approximate recovery is given by recent results [17], [18] that are complementary in nature to the results of this paper and consist of upper bounds (for the nearest subspace and thresholding estimator) and information-theoretic lower bounds on the *sampling rate* m/n needed for asymptotically reliable recovery with respect to an error fraction α. Unlike many of the previous scaling results, these bounds: 1) apply for a variety of assumptions about the non-zero signal elements; 2) are stated explicitly in terms of the SNR, the ratio k/n and various key signal properties; and 3) are shown to be relatively tight for a wide range of settings.

Figure 1 provides an illustration of the bounds discussed above for a Gaussian signal prior. Specifically, the sampling rate needed to ensure that the error fraction does not exceed $\alpha = 0.1$ is plotted as a function of the SNR. In this setting, the asymptotic bounds reinforce the main results of this paper and show the near-optimality of the nearest subspace and thresholding estimates in the SNR limits. Interestingly, the bounds also exhibit similar behavior for a variety of non-Gaussian priors which suggests the properties studied in this paper likely extend beyond the Gaussian setting.

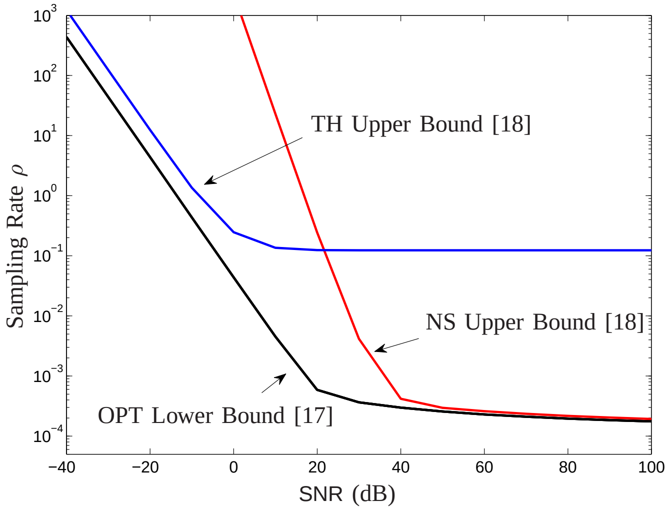

Fig. 1. Bounds on the asymptotic sampling rate $\rho = m/n$ and SNR needed to recover 90% of the support ($\alpha = 0.1$) when the non-zero signal elements are Gausisan, $k/n \rightarrow 10^{-4}$, and the sampling matrix A is constructed with i.i.d. Gaussian elements.

IV. Proofs

The following proofs require some additional notation. We define $B_\alpha(\mathbf{s}) = \{\mathbf{s}' : d(\mathbf{s}, \mathbf{s}') \leq \alpha)$ and use $A_{\mathbf{s}}$ to denote the $m \times k$ submatrix formed by the columns of A indexed by $\mathbf{s}$. For a square matrix M we use $|M|$ to denote the determinant.

A. Proof of Theorem 1

To begin, we define the *intermediate* (INT) estimate

$$\hat{\mathbf{s}}_{\text{INT}}(\mathbf{y}; \alpha, P) = \arg\max_{\mathbf{s}} \sum_{\mathbf{s}' \in B_\alpha(\mathbf{s})} e^{-\frac{1}{2}\left[(1+P)\|\Pi_{\mathbf{s}'}\mathbf{y}\|^2 + \log\left|A_{\mathbf{s}'}^T A_{\mathbf{s}'}\right|\right]}.$$

Then, we use the following two lemmas which show that the optimal estimate and nearest subspace estimate both converge to the intermediate estimate for large P.

Lemma 1. *Let A and $\mathbf{y}$ be fixed with $m > k$. Let $\alpha \in [0, 1)$ be fixed. If $\hat{\mathbf{s}}_{\text{NS}}(\mathbf{y})$ is unique then there exists $P^{(1)}_{A,\mathbf{y},\alpha} < \infty$ such that for all $P > P^{(1)}_{A,\mathbf{y},\alpha}$,*

$$\hat{\mathbf{s}}_{\text{NS}}(\mathbf{y}) = \hat{\mathbf{s}}_{\text{INT}}(\mathbf{y}; \alpha, P).$$

Proof: To begin, it is convenient to define the polynomial

$$r_{\mathbf{s}}(x) = \sum_{\mathbf{s}' \in B_\alpha(\mathbf{s})} c_{\mathbf{s}'} x^{-d_{\mathbf{s}'}}$$

where $c_{\mathbf{s}} = |A_{\mathbf{s}}^T A_{\mathbf{s}}|^{-1/2}$, and $d_{\mathbf{s}} = \frac{1}{2}\|\Pi_{\mathbf{s}}\mathbf{y}\|^2$. Observe that the intermediate estimate can be expressed as

$$\hat{\mathbf{s}}_{\text{INT}}(\mathbf{y}; \alpha, P) = \arg\max_{\mathbf{s}} r_{\mathbf{s}}\left(e^{(1+P)}\right).$$

Next, let $N = |B_\alpha(\mathbf{s})|$ and for $1 < i \leq N$ define

$$\mathbf{s}_i = \arg \min_{\mathbf{s} \in B_\alpha(\mathbf{s}_1) \setminus \cup_{j=1}^{i-1} \mathbf{s}_j} d_{\mathbf{s}}$$

where $\mathbf{s}_1 = \arg\min_{\mathbf{s}} d_{\mathbf{s}} = \hat{\mathbf{s}}_{\text{NS}}(\mathbf{y})$. Furthermore, define $B_i = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_i\}$ and observe that $B_N = B_\alpha(\mathbf{s}_1)$. Hence, to prove the desired convergence, it sufficient to show that $B_\alpha(\mathbf{s}^*) = B_{|N|}$ where

$$\mathbf{s}^* = \lim_{x\rightarrow\infty} \arg\max_{\mathbf{s}} r_{\mathbf{s}}(x)$$

We prove the above claim by induction: we first show that $B_1 \subset B_\alpha(\mathbf{s}^*)$ and then show that for each $1 < i \leq N$ the fact that $B_{i-1} \subset B_\alpha(\mathbf{s}^*)$ implies that $B_i \subseteq B_\alpha(\mathbf{s}^*)$.

To prove the first step, suppose that $\mathbf{s}_1 \notin B_\alpha(\mathbf{s}^*)$. Then,

$$\liminf_{x\rightarrow\infty} \frac{r_{\mathbf{s}_1}(x)}{r_{\mathbf{s}^*}(x)} > \liminf_{x\rightarrow\infty} \frac{c_{\mathbf{s}_1}\, x^{-d_{\mathbf{s}_1}}}{\sum_{\mathbf{s}' \in B_\alpha(\mathbf{s})} c_{\mathbf{s}'}\, x^{-d_{\mathbf{s}'}}} = \infty$$

since, by definition, $d_{\mathbf{s}_1} < d_{\mathbf{s}}$ for all $\mathbf{s} \neq \mathbf{s}_1$. Hence, we have shown by contradiction that $\mathbf{s}_1 \in B_\alpha(\mathbf{s}^*)$.

To prove the general step, assume that $B_{i-1} \subset B_\alpha(\mathbf{s}^*)$ but $\mathbf{s}_i \notin B_\alpha(\mathbf{s}^*)$. Then, $r_{\mathbf{s}_1}(x) < r_{\mathbf{s}^*}(x)$ only if

$$c_{\mathbf{s}_i}\, x^{-d_{\mathbf{s}_i}} < \sum_{\mathbf{s} \in B_\alpha(\mathbf{s}^*) \setminus B_{i-1}} c_{\mathbf{s}'}\, x^{-d_{\mathbf{s}'}}$$

However, by the definition of $\mathbf{s}_i$,

$$\liminf_{n\rightarrow\infty} \frac{c_{\mathbf{s}_i}\, x^{-d_{\mathbf{s}_i}}}{\sum_{\mathbf{s} \in B_\alpha(\mathbf{s}^*) \setminus B_{i-1}} c_{\mathbf{s}'}\, x^{-d_{\mathbf{s}'}}} = \infty.$$

and hence we have shown by contradiction that $\mathbf{s}_i \in B_\alpha(\mathbf{s})$. ■

Lemma 2. *Let A and $\mathbf{y}$ be fixed with $m > k$. Let $\alpha \in [0, 1)$ be fixed. If, $\hat{\mathbf{s}}_{\text{NS}}(\mathbf{y})$ is unique, then there exists $P^{(2)}_{A,\mathbf{y},\alpha} < \infty$ such that for all $P < P^{(2)}_{A,\mathbf{y},\alpha}$,*

$$\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{y}; \alpha, \tfrac{1}{1+P}, \tfrac{P}{1+P}) = \hat{\mathbf{s}}_{\text{INT}}(\mathbf{y}; \alpha, P).$$

Proof: To begin, define

$$f_{\mathbf{s}}(P) = \exp\left\{-\tfrac{1+P}{2}\|\Pi_{\mathbf{s}}\mathbf{y}\|^2 - \tfrac{1}{2}\log\left|A_{\mathbf{s}}^T A_{\mathbf{s}}\right|\right\},$$
$$g_{\mathbf{s}}(P) = \exp\left\{-\tfrac{1}{2}\|\Sigma_{\mathbf{s}}^{-1/2}\mathbf{y}\|^2 - \tfrac{1}{2}\log|\Sigma_{\mathbf{s}}|\right\}$$

and let $\mathbf{s}^* = \hat{\mathbf{s}}_{\text{NS}}(\mathbf{y})$. From the uniqueness of $\hat{\mathbf{s}}_{\text{NS}}(\mathbf{y})$ and Lemma 1, there exists $P^{(1)}_{A,\mathbf{y},\alpha} < \infty$ such that for any $P^{(1)}_{A,\mathbf{y},\alpha} < P < \infty$ and any $\mathbf{u} \neq \mathbf{s}^*$,

$$\frac{\sum_{\mathbf{s}' \in B_\alpha(\mathbf{s}^*)} f_{\mathbf{s}'}(P)}{\sum_{\mathbf{s}' \in B_\alpha(\mathbf{u})} f_{\mathbf{s}'}(P)} > 1$$

The above statement implies that

$$\min_{\mathbf{s} \in B_\alpha(\mathbf{s}^*) \setminus B_\alpha(\mathbf{u})} \|\Pi_{\mathbf{s}}\mathbf{y}\| < \min_{\mathbf{s} \in B_\alpha(\mathbf{u}) \setminus B_\alpha(\mathbf{s}^*)} \|\Pi_{\mathbf{s}}\mathbf{y}\|$$

and hence, for any $\mathbf{u} \neq \mathbf{s}^*$,

$$\lim_{P\rightarrow\infty} \frac{\sum_{\mathbf{s}' \in B_\alpha(\mathbf{s}^*) \setminus B_\alpha(\mathbf{u})} f_{\mathbf{s}'}(P)}{\sum_{\mathbf{s}' \in B_\alpha(\mathbf{u}) \setminus B_\alpha(\mathbf{s}^*)} f_{\mathbf{s}'}(P)} = \infty. \tag{5}$$

Now, to show the desired convergence, it is sufficient to show that a similar statement holds with $f_{\mathbf{s}}(P)$ replaced by $g_{\mathbf{s}}(P)$. In particular, note that for any $\mathbf{u} \neq \mathbf{s}$,

$$\frac{\sum_{\mathbf{s}' \in B_\alpha(\mathbf{s}^*)} g_{\mathbf{s}'}(P)}{\sum_{\mathbf{s}' \in B_\alpha(\mathbf{u})} g_{\mathbf{s}'}(P)} > 1 \tag{6}$$

if an only if

$$\frac{\sum_{\mathbf{s}'\in B_\alpha(\mathbf{s}^*)\setminus B_\alpha(\mathbf{u})} g_{\mathbf{s}'}(P)}{\sum_{\mathbf{s}'\in B_\alpha(\mathbf{u})\setminus B_\alpha(\mathbf{s}^*)} g_{\mathbf{s}'}(P)} > 1. \tag{7}$$

To show that the above inequality occurs, define

$$\Delta(P) = \max_{\mathbf{s}} \left| \log\left(f_{\mathbf{s}}(P)\right) - \log\left(g_{\mathbf{s}}(P)\right) \right|$$

and observe that

$$\begin{aligned} &\frac{\sum_{\mathbf{s}'\in B_\alpha(\mathbf{s}^*)\setminus B_\alpha(\mathbf{u})} g_{\mathbf{s}'}(P)}{\sum_{\mathbf{s}'\in B_\alpha(\mathbf{u})\setminus B_\alpha(\mathbf{s}^*)} g_{\mathbf{s}'}(P)} \\ &\geq \frac{\sum_{\mathbf{s}'\in B_\alpha(\mathbf{s}^*)\setminus B_\alpha(\mathbf{u})} f_{\mathbf{s}'}(P)}{\sum_{\mathbf{s}'\in B_\alpha(\mathbf{u})\setminus B_\alpha(\mathbf{s}^*)} f_{\mathbf{s}'}(P)} \exp\{-2\Delta(P)\}. \end{aligned} \tag{8}$$

The convergence in (5) shows that the first term on the right hand side of (8) becomes arbitrarily large as $P \to \infty$. In the following steps, we show that the second term remains bounded away from zero. By the matrix inversion lemma,

$$\Sigma_{\mathbf{s}}^{-1} = (1+P)\left[I_{m\times m} - A_{\mathbf{s}}\left(\tfrac{1}{P}I_{k\times k} + A_{\mathbf{s}}^T A_{\mathbf{s}}\right)^{-1} A_{\mathbf{s}}^T\right].$$

Applying a Taylor expansion [19] shows that,

$$\left(\tfrac{1}{P}I_{k\times k} + A_{\mathbf{s}}^T A_{\mathbf{s}}\right)^{-1} = \left(A_{\mathbf{s}}^T A_{\mathbf{s}}\right)^{-1} + O\big(\tfrac{1}{P}\big)\left(A_{\mathbf{s}}^T A_{\mathbf{s}}\right)^{-2}$$

and hence

$$\|\Sigma_{\mathbf{s}}^{-1/2}\mathbf{y}\|^2 \; = (1+P)\|\Pi_{\mathbf{s}}\mathbf{y}\|^2 + O\big(\tfrac{1+P}{P}\big). \tag{9}$$

Next, by Sylvester's determinant theorem,

$$\begin{aligned} \det\left(\Sigma_{\mathbf{s}}\right) &= \left(\tfrac{P}{1+P}\right)^m \left|\tfrac{1}{P}I_{m\times m} + A_{\mathbf{s}}A_{\mathbf{s}}^T\right| \\ &= \left(\tfrac{P}{1+P}\right)^m \left|\tfrac{1}{P}I_{k\times k} + A_{\mathbf{s}}^T A_{\mathbf{s}}\right|, \end{aligned}$$

and applying a Taylor expansion [19] shows that

$$\left|\tfrac{1}{P}I_{k\times k} + A_{\mathbf{s}}^T A_{\mathbf{s}}\right| = \left|A_{\mathbf{s}}^T A_{\mathbf{s}}\right| + O\big(\tfrac{1}{P}\big). \tag{10}$$

Combining (9) and (10) shows that $\limsup_{P\to\infty} \Delta(P)$ is finite. Thus, we have shown that for P large enough, the right hand side of (8) will be greater than one which proves the desired result.

■

To conclude the proof of Theorem 1, observe that it is sufficient to to prove convergence uniformly for all $\alpha \in [0,1)$ by applying Lemmas 1 and 2 to each α in the finite set $\mathcal{A} = \{l/k : 0 \leq l < k\}$ and letting $P_{A,\mathbf{y}} = \max_{\alpha\in\mathcal{A}} \max(P_{A,\mathbf{y},\alpha}^{(1)}, P_{A,\mathbf{y},\alpha}^{(2)})$.

B. Proof of Theorem 2

Define the exponents

$$\begin{aligned} E_{\mathbf{s}}(P) &= \tfrac{1}{2}\big[-\|\Sigma_{\mathbf{s}}^{-1/2}\mathbf{y}\|^2 - \log|\Sigma_{\mathbf{s}}| \\ &\qquad + (1+P)\|\mathbf{y}\|^2 + m\log(1+P)\big] \\ E_{\mathbf{s}} &= \tfrac{1}{2}\big[\|A_{\mathbf{s}}\mathbf{y}\|^2 - \mathrm{tr}\big(A_{\mathbf{s}}^T A_{\mathbf{s}}\big)\big]. \end{aligned}$$

and observe that the optimal estimate and the thresholding estimate can be expressed as

$$\begin{aligned} \hat{\mathbf{s}}_{\text{OPT}}(\mathbf{y};\alpha,P) &= \arg\max_{\mathbf{s}} \sum_{\mathbf{s}'\in B_\alpha(\mathbf{s})} \exp\{E_{\mathbf{s}'}(P)\} \\ \hat{\mathbf{s}}_{\text{TH}}(\mathbf{y}) &= \arg\max_{\mathbf{s}} E_{\mathbf{s}}. \end{aligned}$$

By the matrix inversion lemma,

$$\Sigma_{\mathbf{s}}^{-1} = (1+P)\left[I_{m\times m} - PA_{\mathbf{s}}\left(I_{k\times k} + PA_{\mathbf{s}}^T A_{\mathbf{s}}\right)^{-1} A_{\mathbf{s}}^T\right].$$

Applying a Taylor expansion [19] shows that,

$$\left(I_{k\times k} + PA_{\mathbf{s}}^T A_{\mathbf{s}}\right)^{-1} = I_{k\times k} + O\big(P\big)I_{k\times k}$$

and hence

$$\|\Sigma_{\mathbf{s}}^{-1/2}\mathbf{y}\|^2 = (1+P)\|\mathbf{y}\|^2 - P\|A_{\mathbf{s}}^T\mathbf{y}\|^2 + O(P^2). \tag{11}$$

Next, by Sylvester's determinant theorem,

$$\begin{aligned} \det\left(\Sigma_{\mathbf{s}}\right) &= \left(\tfrac{1}{1+P}\right)^m \left|I_{m\times m} + PA_{\mathbf{s}}A_{\mathbf{s}}^T\right| \\ &= \left(\tfrac{1}{1+P}\right)^m \left|I_{k\times k} + PA_{\mathbf{s}}^T A_{\mathbf{s}}\right|. \end{aligned}$$

Applying a Taylor expansion [19] shows that

$$\left|I_{k\times k} + PA_{\mathbf{s}}^T A_{\mathbf{s}}\right| = 1 + P\cdot\mathrm{tr}\big(A_{\mathbf{s}}^T A_{\mathbf{s}}\big) + O\big(P^2\big) \tag{12}$$

Combining (11) and (12) shows that for any $\mathbf{s}$,

$$E_{\mathbf{s}}(P) = PE_{\mathbf{s}} + O(P^2) \quad \text{as} \quad P \to 0.$$

Thus, by the approximation $\exp\{x\} = 1+x+O(x^2)$ as $x \to 0$,

$$\exp\{E_{\mathbf{s}}(P)\} = 1 + PE_{\mathbf{s}} + O(P^2) \quad \text{as} \quad P \to 0. \tag{13}$$

Using (13), the optimal estimate can be expressed as

$$\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{y};\alpha,P) = \arg\max_{\mathbf{s}} \sum_{\mathbf{s}'\in B_\alpha(\mathbf{s})} E_{\mathbf{s}'} + \delta_s(P)$$

where $\max_{\mathbf{s}} |\delta_s(P)| = O(1/P)$ as $P \to 0$. Hence, if the estimate

$$\hat{\mathbf{s}}^* = \arg\max_{\mathbf{s}} \sum_{\mathbf{s}'\in B_\alpha(\mathbf{s})} E_{\mathbf{s}'}$$

is unique, we may conclude that there exists $P_{A,\mathbf{y}} > 0$ such that for all $P < P_{A,\mathbf{y}}$,

$$\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{y};\alpha,P) = \mathbf{s}^*.$$

To conclude the proof, we will show that $\mathbf{s}^* = \mathbf{s}_{\text{TH}}(\mathbf{y})$ which proves both the uniqueness of $\mathbf{s}^*$ and the desired convergence. Observe that $E_{\mathbf{s}}$ can be decomposed as

$$\begin{aligned} \sum_{\mathbf{s}'\in B_\alpha(\mathbf{s})} 2E_{\mathbf{s}} &= \sum_{\mathbf{s}'\in B_\alpha(\mathbf{s})} \sum_{i\in\mathbf{s}'} \left[(a_i^T\mathbf{y})^2 - \|a_i\|^2\right] \\ &= \sum_{i=1}^{n} N_i(\mathbf{s}) \left[\|a_i\mathbf{y}\|^2 - \|a_i\|^2\right] \end{aligned} \tag{14}$$

where $N_i(\mathbf{s}) = |\{\mathbf{s}' \ni i : d(\mathbf{s},\mathbf{s}') \leq \alpha\}|$ obeys $N_i(\mathbf{s}) = N_j(\mathbf{s}) > N_l(\mathbf{s})$ for all $i,j \in \mathbf{s}$ and $l \notin \mathbf{s}$. Hence, the right hand side of (14) is maximized when $\mathbf{s}$ corresponds to the indices of the k largest values of $(a_i^T\mathbf{y})^2 - \|a_i\|^2$. This proves that $\mathbf{s}^* = \mathbf{s}_{\text{TH}}(\mathbf{y})$ and thus concludes the proof.

C. *Proof of Proposition 2*

To begin, it is convenient to enumerate the possible supports as $\mathbf{s}_1 = \{1,2\}$, $\mathbf{s}_2 = \{1,3\}$, $\mathbf{s}_3 = \{1,4\}$, $\mathbf{s}_4 = \{2,3\}$, $\mathbf{s}_5 = \{2,4\}$, and $\mathbf{s}_6 = \{3,4\}$. By the symmetry of A,

$$\begin{aligned} &\Pr\left\{d\big(\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{Y};\alpha,P),\mathbf{S}\big) > \tfrac{1}{2}\right\} \\ &= \Pr\left\{d\big(\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{Y};\alpha,P),\mathbf{S}\big) > \tfrac{1}{2}\big|\mathbf{S} = \mathbf{s}_i\right\} \end{aligned}$$

for any $\mathbf{s}_i$. Thus, without loss of generality, we may condition on the realization $\mathbf{S} = \mathbf{s}_1$.

Next, for each $1 \le i \le 6$, consider the singular decomposition $\Sigma_{\mathbf{s}_i} = U_i^T D_i U_i$. It may be verified $D_i = D$ where

$$\text{diag}(D) = \begin{bmatrix} P+1, & P/2+1, & 1 \end{bmatrix}.$$

Also, define $B_i = D^{-1/2} U_i^T U_1 D^{-1/2}$ and $\mathbf{V} = D^{-1/2} U_1^T \mathbf{Y}$ and observe that $\|\Sigma_{\mathbf{s}_i}\mathbf{Y}\| = \|B_i\mathbf{V}\|$ where the elements of $\mathbf{V}$ are i.i.d. $\mathcal{N}(0,1)$.

We now prove the scaling (3). Using the above properties and the definition of the $\hat{\mathbf{s}}_{\text{OPT}}$ given in Proposition 1, the probability of erroneous recovery can be bounded as

$$\begin{aligned} &\Pr\Big\{d\big(\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{Y};\tfrac{1}{2},P),\mathbf{S}\big) > \tfrac{1}{2}\Big\} \\ &= \Pr\Big\{\arg\max_i \|B_i\mathbf{V}\| = 1\Big\} \\ &\le \Pr\left\{\|B_1\mathbf{V}\| > \max_{2\le i\le 5}\|B_i\mathbf{V}\|\right\}. \end{aligned}$$

Clearly, B_1 is equal to the identity matrix. Letting $\mathbf{b}_i^T$ denote the third row of B_i, it may be verified that

$$\begin{aligned} \mathbf{b}_2^T &= \tfrac{1}{2}\begin{bmatrix} -\sqrt{P+1} & \sqrt{P+2} & 1 \end{bmatrix} \\ \mathbf{b}_3^T &= \tfrac{1}{2}\begin{bmatrix} -\sqrt{P+1} & \sqrt{P+2} & -1 \end{bmatrix} \\ \mathbf{b}_4^T &= \tfrac{1}{2}\begin{bmatrix} \sqrt{P+1} & \sqrt{P+2} & 1 \end{bmatrix} \\ \mathbf{b}_5^T &= \tfrac{1}{2}\begin{bmatrix} -\sqrt{P+1} & -\sqrt{P+2} & 1 \end{bmatrix}. \end{aligned}$$

Hence,

$$\begin{aligned} \max_{2\le i\le 5}\|B_i\mathbf{V}\|^2 &\ge \max_{2\le i\le 5}(\mathbf{b}_i^T\mathbf{V})^2 \\ &= \tfrac{1}{4}\big(\sqrt{P+1}|V_1| + \sqrt{P+2}|V_2| + |V_3|\big)^2 \end{aligned}$$

and thus

$$\Pr\Big\{\|B_1\mathbf{V}\| > \max_{2\le i\le 5}\|B_6\mathbf{V}\|\Big\} < \Pr\Big\{V_3^2 > \tfrac{P-4}{4}\big(V_1^2 + V_2^2\big)\Big\}.$$

Applying Lemma 3 completes the proof of (3).

Next, we bound the scaling (4). This time, the probability of erroneous recovery is given by

$$\Pr\Big\{d\big(\hat{\mathbf{s}}_{\text{OPT}}(\mathbf{Y};0,P),\mathbf{S}\big) > \tfrac{1}{2}\Big\} = \Pr\Big\{\arg\min_i \|B_i\mathbf{V}\| = 6\Big\}.$$

To lower bound the above probability, observe that

$$\min_{2\le i\le 5}\|B_i\mathbf{V}\|^2 \ge \min_{2\le i\le 5}(\mathbf{b}_i^T\mathbf{V})^2$$

where $\mathbf{b}_i^T$ are defined as above. Also, it may be verified that $\|B_6\mathbf{V}\|^2 = (P+1)V_1^2 + V_2^2 + \frac{1}{P+1}V_3^2$. Hence, if we we define

$$\mathcal{E} = \Big\{\tfrac{1}{64}PV_2^2 > V_3^2 > 4PV_1^2\Big\}.$$

then a bit of algebra shows that for $P > 20$,

$$\mathcal{E} \Rightarrow \big\{\|B_6\mathbf{V}\|^2 < V_2^2 + V_3^2\big\} \cap \big\{\min_{1\le i\le 5}\|B_i\mathbf{V}\|^2 \ge V_2^2 + V_3^2\big\},$$

and hence $\Pr\big\{\arg\min_i \|B_i\mathbf{V}\|^2 = 6\big\} \ge \Pr\{\mathcal{E}\}$. Using

$$\Pr\{\mathcal{E}\} > \Pr\left\{V_3^2 > 4PV_1^2\right\} - \Pr\left\{V_3^2 > \tfrac{1}{128}P(V_2^2 + V_3^2)\right\}.$$

and applying Lemma 3 completes the proof of (3).

Lemma 3. *If Z_1, Z_2, Z_3 are i.i.d. $\mathcal{N}(0,1)$, then, as $P \to \infty$,*

$$\begin{aligned} \Pr\{Z_1^2 > PZ_2^2\} &= \Theta(P^{-1/2}) \\ \Pr\{Z_1^2 > P(Z_2^2 + Z_3^2)\} &= \Theta(P^{-1}) \end{aligned}$$

Proof: The proof follows from Gaussian tail bounds. ■

ACKNOWLEDGMENT

This work was supported in part by ARO MURI No. W911NF-06-1-0076.

REFERENCES

[1] M. J. Wainwright, “Sharp thresholds for high-dimensional and noisy recovery of sparsity,” in *Proc. Allerton Conf. on Comm., Control, and Computing*, Monticello, IL, Sep. 2006.

[2] N. Meinshausen and P. Bühlmann, “High-dimensional graphs and variable selection with the lasso,” *Annals of Stat.*, vol. 34, pp. 1436–1462, 2006.

[3] P. Zhao and B. Yu, “On model selection consistency of lasso,” *J. of Machine Learning Research*, vol. 51, no. 10, pp. 2541–2563, Nov. 2006.

[4] M. J. Wainwright, “Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting,” in *Proc. IEEE Int. Symp. on Inform. Theory*, Nice, France, Jun. 2007.

[5] G. Reeves, “Sparse signal sampling using noisy linear projections,” Univ. of California, Berkeley, Dept. of Elec. Eng. and Cpmp. Sci., Tech. Rep. UCB/EECS-2008-3, Jan. 2008.

[6] S. Aeron, M. Zhao, and V. Saligrama, “On sensing capacity of sensor networks for the class of linear observation, fixed snr models,” Jun 2007, arXiv:0704.3434v3 [cs.IT].

[7] ——, “Fundamental limits on sensing capacity for sensor networks and compressed sensing,” Apr. 2008, arXiv:0804.3439v1 [cs.IT].

[8] M. Akcakaya and V. Tarokh, “Shannon theoretic limits on noisy compressive sampling,” Nov. 2007, arXiv:0711.0366v1 [cs.IT].

[9] G. Reeves and M. Gastpar, “Sampling bounds for sparse support recovery in the presence of noise,” in *Proc. IEEE Int. Symp. on Inform. Theory*, Toronto, Canada, Jul. 2008.

[10] A. K. Fletcher, S. Rangan, and V. K. Goyal, “Necessary and sufficient conditions on sparsity pattern recovery,” May 2008, arXiv:0804.1839v1 [cs.IT].

[11] W. Wang, M. J. Wainwright, and K. Ramchandran, “Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices,” in *Proc. IEEE Int. Symp. on Inform. Theory*, Toronto, Canada, Jul. 2008.

[12] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

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

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

[15] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM J. of Sci. Comp.*, vol. 20, no. 1, pp. 33–61, 1999.

[16] R. Tibshirani, “Regression shrinkage and selection via the lasso,,” *J. Royal Stat. Soc., Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[17] G. Reeves and M. Gastpar, “Approximate sparsity pattern recovery: Information-theoretic lower bounds,” Preprint.

[18] ——, “Approximate sparsity pattern recovery: Information-theoretic upper bounds,” Preprint.

[19] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.