Support Recovery for Orthogonal Matching Pursuit: Upper and Lower bounds

Raghav Somani[*]
Microsoft Research, India
t-rasom@microsoft.com

Chirag Gupta[* †]
Machine Learning Department,
Carnegie Mellon University
chiragg@andrew.cmu.edu

Prateek Jain
Microsoft Research, India
prajain@microsoft.com

Praneeth Netrapalli
Microsoft Research, India
praneeth@microsoft.com

Abstract

We study the problem of sparse regression where the goal is to learn a sparse vector that best optimizes a given objective function. Under the assumption that the objective function satisfies restricted strong convexity (RSC), we analyze orthogonal matching pursuit (OMP), a greedy algorithm that is used heavily in applications, and obtain a support recovery result as well as a tight generalization error bound for the OMP estimator. Further, we show a lower bound for OMP, demonstrating that both our results on support recovery and generalization error are tight up to logarithmic factors. To the best of our knowledge, these are the first such tight upper and lower bounds for *any* sparse regression algorithm under the RSC assumption.

1 Introduction

The goal in sparse regression is to find the optimal sparse vector that minimizes a given objective function. Sparse regression is an important problem in statistical machine learning since sparse models lead to better generalization guarantees when the feature dimension is high or data is less, eg, high-dimensional statistics [19], bioinformatics [18], etc. Sparse models also have a smaller memory footprint and are thus useful for resource-constrained machine learning [9]. For simplicity of exposition, we focus on the problem of sparse *linear* regression (SLR), which is a representative problem in this domain. Results for this problem typically extend easily to the general case. Given $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\mathbf{y}$, the goal of SLR is to recover a sparse vector $\bar{\mathbf{x}}$ that minimizes $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$.

The unconditional version of sparse regression can be shown to be NP-hard via a reduction to 3-set cover [14]. However, the problem has been studied heavily under a variety of assumptions such as incoherence [7], null-space property [8], restricted isometry property (RIP) or restricted strong convexity (RSC) [4, 15]. RSC, in particular, is one of the weakest and most popular assumptions for sparse regression problems and has been studied in the context of various algorithms [27, 11, 1, 13]. In this paper, we study the SLR problem under RSC condition.

Typically SLR is studied with one of two goals: a) support recovery, i.e., recovering support (or features) of $\bar{\mathbf{x}}$ and b) bounding generalization error $\left(\|\mathbf{A}(\mathbf{x}-\bar{\mathbf{x}})\|_2^2/n\right)$ which bounds excess error on unseen test points if each row of $\mathbf{A}$ is sampled from a fixed distribution. In general, support recovery

---

[*]Equal contribution
[†]Work done in part while Chirag Gupta was a Research Fellow at Microsoft Research, India

is a more fundamental and challenging problem as a strong support recovery result usually tends to provide strong generalization error bound.

Existing sparse regression algorithms can be broadly categorized into three categories: a) ℓ_1 minimization or LASSO based algorithms [6, 5, 1], b) non-convex penalty based methods [2, 11, 13], c) greedy methods [22, 17, 27, 12]. In this work, we focus on OMP which is a greedy method that incrementally adds elements to support based on the amount of reduction in training error. Owing to its simplicity, flexibility, and strong practical performance, OMP is one of the most celebrated and practically used algorithms for sparse regression.

OMP has been shown to provide support recovery in noiseless settings, i.e., when $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}}$, under various conditions like incoherence [8], null-space property, RIP/RSC [28] etc. In the noisy setting, while the generalization error of OMP has been studied [28] under RSC, these bounds do not match known lower bounds [29] in terms of the restricted strong convexity constant. In fact, the tightest known generalization error upper bound for polynomial time algorithms is a factor of restricted strong convexity constant worse than the known lower bound [29, 28, 11, 30]. Furthermore, strong support recovery results under RSC are also known only for a non-convex SCAD/MCP penalty based method [13]. For greedy methods, there have been several recent works [20, 21, 25] that consider the problem of support recovery. However, none of these works give strong results for this problem.

In this work, we significantly improve upon these support recovery results for OMP. We show that if the smallest element of $\bar{\mathbf{x}}$ is larger than an appropriate noise level, then OMP recovers the *full* support of $\bar{\mathbf{x}}$ (see Theorem 3.1). As noted in remarks 3 and 4 below the theorem, our result has a better dependence on the restricted condition number than the ones in [20, 21, 25]. The proof of Theorem 3.1 exploits the fact that if a certain element of $\bar{\mathbf{x}}$ is not included in the current support set, then a single step of OMP should lead to a large *additive* decrease in the error. In addition, we present a generalization error analysis for OMP.

Finally, we provide matching lower bounds for our support recovery and generalization error results. To this end, we construct a design matrix that ensures that OMP picks incorrect indices until a large number of elements are added to the support set (see Theorems 4.2, 4.3). As the support set size has to increase arbitrarily for recovery, this also implies poor generalization error (see Theorem 4.3).

We note that our lower bound results are unconditional and are directly applicable to OMP. In contrast, existing lower bounds such as [29] obtain a lower bound for generalization error of *any* polynomial time algorithm assuming $\mathbf{NP} \not\subset \mathbf{P/poly}$. Moreover, these lower bound results are restricted to algorithms which recover *exactly* s^*-sparse vectors, where $s^* = |\mathrm{supp}(\bar{\mathbf{x}})|$ and hence do not apply to OMP if it adds more than s^* elements to the support set, which is the more meaningful scenario to consider. Moreover, if each element of $\bar{\mathbf{x}}$ is *large*, then the claim of [29] is almost vacuous as one can recover the support exactly which is the main problem in SLR. In that case, while the generalization error lower bound of [29] holds, it does not preclude the OMP algorithm from recovering the correct support (see Section 4).

Notation: Matrices are typically written in bold capital letters (such as $\mathbf{A}$ and $\boldsymbol{\Sigma}$), vectors are typically written in bold small case letters (such as $\mathbf{x}$ and $\boldsymbol{\eta}$) and universal constants independent of problem parameters are written as C_1, C_2, etc. For a matrix $\mathbf{A}$, $\mathbf{A}_i$ represents its ith column and $\mathbf{A_S}$ represent the sub-matrix of $\mathbf{A}$ with columns in the index set $\mathbf{S}$. $\rho_s^+(\mathbf{A}^T\mathbf{A})$, $\rho_s^-(\mathbf{A}^T\mathbf{A})$ are restricted smoothness and restricted strong convexity constants of the matrix $\mathbf{A}$ (defined below). $\widetilde{\kappa}_s(\mathbf{A}^T\mathbf{A}) := \rho_1^+(\mathbf{A}^T\mathbf{A})/\rho_s^-(\mathbf{A}^T\mathbf{A})$ for all $s > 0$. ρ_s^+, ρ_s^- and $\widetilde{\kappa}_s$ when used without parameter, represent $\rho_s^+(\mathbf{A}^T\mathbf{A})$, $\rho_s^-(\mathbf{A}^T\mathbf{A})$ and $\widetilde{\kappa}_s(\mathbf{A}^T\mathbf{A})$ respectively. The non-zero element of $\bar{\mathbf{x}}$ with the least absolute value is denoted as $\bar{x}_{\min}$.

2   Preliminaries and Setting

In this section, we will present some preliminaries and the problem setting considered in this paper. Broadly, we are interested in sparse estimation problems where we are given a function $Q(\cdot)$ and we wish to solve $\min_{\mathbf{x}:\|\mathbf{x}\|_0 \le s^*} Q(\mathbf{x})$. This problem is in general NP-hard even when $Q(\cdot)$ is a quadratic function. So, we consider this problem under *restricted strong convexity (RSC)* and *restricted smoothness (RSS)* assumptions. While part of our results apply to this general setting, for simplicity of presentation, we focus on the case where $Q(\cdot)$ is a quadratic. More concretely, in the sparse linear regression problem where we are given a measurement matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and response $\mathbf{y} \in \mathbb{R}^n$

Algorithm 1 Orthogonal Matching Pursuit (OMP)

1: **procedure** OMP(s)
2:     $\mathbf{S}_0 = \phi$, $\mathbf{x}_0 = \mathbf{0}$, $\mathbf{r}_0 = \mathbf{y}$
3:     **for** $k = 1, 2, \ldots, s$ **do**
4:         $j \leftarrow \underset{i \notin \mathbf{S}_{k-1}}{\arg\max} |\mathbf{A}_i^T \mathbf{r}_{k-1}|$
5:         $\mathbf{S}_k \leftarrow \mathbf{S}_{k-1} \cup \{j\}$
6:         $\mathbf{x}_k \leftarrow \underset{\text{supp}(\mathbf{x}) \subseteq \mathbf{S}_k}{\arg\min} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$
7:         $\mathbf{r}_k \leftarrow \mathbf{y} - \mathbf{A}\mathbf{x}_k$
8:     **end for**
9:     **return** $\mathbf{x}_s$
10: **end procedure**
---

and we wish to solve $\min_{\|\mathbf{x}\|_0 \leq s^*} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$. We assume that the measurement matrix $\mathbf{A}$ satisfies restricted strong convexity and restricted smoothness properties [4]:

Definition 2.1 (Restricted strong convexity (RSC)). $\mathbf{A}$ *is said to be restricted strongly convex at level s with parameter ρ_s^- if for every $\mathbf{x}$ and $\mathbf{z}$ such that $\|\mathbf{x} - \mathbf{z}\|_0 \leq s$, we have*

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{z}\|_2^2 \geq \rho_s^- \|\mathbf{x} - \mathbf{z}\|_2^2.$$

Definition 2.2 (Restricted smoothness (RSS)). $\mathbf{A}$ *is said to be restricted smooth at level s with parameter ρ_s^+ if for every $\mathbf{x}$ and $\mathbf{z}$ such that $\|\mathbf{z} - \mathbf{x}\|_0 \leq s$, we have*

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{z}\|_2^2 \leq \rho_s^+ \|\mathbf{x} - \mathbf{z}\|_2^2.$$

The above definitions capture the standard strong convexity and smoothness properties but only in sparse directions. Similarly, we can define a notion of *restricted condition number*.

Definition 2.3 (Restricted condition number). *The restricted condition number at level s of a matrix $\mathbf{A}$ is defined as*

$$\widetilde{\kappa}_s(\mathbf{A}^T\mathbf{A}) = \frac{\rho_1^+}{\rho_s^-}. \tag{2.1}$$

Throughout this paper, we assume that $\mathbf{A}$ satisfies the above properties and denote the corresponding parameters as ρ_s^-, ρ_s^+, and $\widetilde{\kappa}_s$ respectively. For our lower bound matrices in Section 4 we show that these properties are satisfied.

Definition 2.4 ($\ell_\infty - norm$). *We define the $\ell_\infty - norm$ of a matrix $\mathbf{A}$:*

$$\|\mathbf{A}\|_\infty := \max_{\|\mathbf{x}\|_\infty = 1} \|\mathbf{A}\mathbf{x}\|_\infty \tag{2.2}$$

We work under the generative model where $\bar{\mathbf{x}}$ is an s^*-sparse vector supported on $\mathbf{S}^*$, that generates the data. More concretely, we assume that the measurements $\mathbf{y}$ are generated as noisy linear measurements of $\bar{\mathbf{x}}$:

$$\mathbf{y} = \mathbf{A}\bar{\mathbf{x}} + \boldsymbol{\eta}, \tag{2.3}$$

where each element of $\boldsymbol{\eta}$ is a mean zero sub-Gaussian random variable with parameter σ. This means that for some constant C, we have,

$$P\{|\eta_i| > t\} \leq C \exp\left(-t^2/2\sigma^2\right).$$

The non-zero element of $\bar{\mathbf{x}}$ with the least absolute value is denoted as $\bar{x}_{\min}$.

In this problem setting, there are two critical questions:

1. **Support recovery**: The goal here is to recover the support of $\bar{\mathbf{x}}$ after observing $\mathbf{y}$ and $\mathbf{A}$. This question can also be posed as estimating $\bar{\mathbf{x}}$ in the ℓ_∞ norm i.e., find $\hat{\mathbf{x}}$ such that $\|\hat{\mathbf{x}} - \bar{\mathbf{x}}\|_\infty$ is small.

2. **Generalization error**: Here, the goal is to compute an $\hat{\mathbf{x}}$ such that $\|\mathbf{A}(\hat{\mathbf{x}} - \bar{\mathbf{x}})\|_2$ is small. This quantity is essentially the generalization error when the learned $\hat{\mathbf{x}}$ is used to make prediction over test data generated from same distribution as training data $\mathbf{A}$ and $\mathbf{y}$.

Table 1: Comparison between our results and several prior results on support recovery for Sparse Linear Regression. HTP refers to Hard Thresholding Pursuit, PHT refers to Partial Hard Thresholding, and IHT referes to Iterative Hard Thresholding. These are all thresholding based greedy algorithms. Apart from $\widetilde{\kappa}(\cdot)$, we also use $\kappa_s(\cdot) = \rho_s^+(\cdot)/\rho_s^-(\cdot)$. All values are correct upto constants; we have skipped order notation in the interest of succinctness. Support expansion refers to the value of s in the paper. The $|\bar{x}_{\min}|$ column refers to the condition for support recovery guarantee. All support recovery happens with some probability δ, and we incur polynomial factors of $\log(d/\delta)$ in the $|\bar{x}_{\min}|$ condition. We skip these in the interest of succinctness.

Related Work	**Support expansion** (s)	$|\bar{x}_{\min}|$ **lower bound**
Yuan et al. [25] [HTP]	$\kappa_{2s}^2 s^*$	$\dfrac{\sigma\sqrt{s}}{\sqrt{\rho_{2s}^-}}$
Shen et al. [20] [HTP]	$\kappa_{2s}^2 s^*$	$\dfrac{\sigma\sqrt{\kappa_{2s}}\sqrt{\rho_1^+ s}}{\rho_{s+s^*}^-}$
Shen et al. [21] [PHT(r)]	$s^* + \kappa_{2s}^2 \min\{s^*, r\}$	$\dfrac{\sigma\sqrt{\kappa_{2s}}\sqrt{\rho_1^+ s}}{\rho_{2s}^-}$
Jain et al. [11] [IHT]	$\kappa_{2s+s^*}^2 s^*$	$-$
Zhang [28] [OMP]	$\widetilde{\kappa}_{s+s^*} s^* \log \kappa_{s+s^*}$	$-$
Theorem 3.1 [OMP]	$\widetilde{\kappa}_{s+s^*} s^* \log \kappa_{s+s^*}$	$\gamma \cdot \dfrac{\sigma\sqrt{\rho_1^+}}{\rho_{s+s^*}^-}$

We note that in both the above problems we are allowed to output $\hat{\mathbf{x}}$ that may have $s \geq s^*$ elements in the support. This is a standard and crucial relaxation needed to provide strong guarantees under weak assumptions for SLR. This work considers orthogonal matching pursuit (OMP) [16, 23] for solving both of the above problems. OMP is one of the most popular methods for sparse optimization and it is essentially a greedy method that incrementally estimates the support of $\bar{\mathbf{x}}$ by adding one element at a time. See Algorithm 1 for a pseudo-code of OMP for SLR.

In Section 3 we show our upper bounds for the performance of OMP with respect to both the problems above, under the RSS/RSC conditions. In Section 4, we provide a matching lower bound (upto logarithmic factors) which shows that there exist certain sparse linear regression problems on which OMP cannot perform significantly better than the error bounds given by our analysis. In Section 5 we show some simple simulations to ground our results.

3 Upper bounds for OMP

We first present our key contribution which is a support recovery bound for OMP under RSC/RSS.

Theorem 3.1 (Support Recovery for OMP). *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\bar{\mathbf{x}} \in \mathbb{R}^d$ be a s^*-sparse vector. Let $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}} + \boldsymbol{\eta}$ and let $\hat{\mathbf{x}}_s$ be the output of OMP after s iterations, where*

$$s \geq C_1 \widetilde{\kappa}_{s+s^*} s^* \cdot \log\left(\frac{5\rho_{s+s^*}^+}{\rho_{s+s^*}^-}\right),$$

and $\widetilde{\kappa}_{s+s^*}$ is the restricted condition number (Definition 2.3). Moreover, let $\left\|\mathbf{A}_{\mathbf{S}^*\setminus\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1}\right\|_\infty \leq \gamma$ where $\mathbf{S} = \text{supp}(\hat{\mathbf{x}}_s)$. Then, for every $\delta \in \left(0, e^{-68}\right)$, if

$$|\bar{x}_{\min}| \geq \left(1 + \sqrt{2}\,(1+\gamma)\right) \frac{\sigma}{\rho_{s+s^*}^-} \sqrt{\rho_1^+ \log \frac{d}{\delta}}, \tag{3.1}$$

and $s + s^* \geq \log\left(1/\delta\right)$, then $\mathbf{S}^* \subseteq \text{supp}(\hat{\mathbf{x}}_s)$ and $\|\hat{\mathbf{x}}_s - \bar{\mathbf{x}}\|_\infty \leq \sigma\sqrt{\frac{2}{\rho_s^-}\log\left(s/\delta\right)}$ with probability at least $1 - 7\delta$. Here $C_1 = 664$ is a universal constant.

Remark 1: $\rho_{s+s^*}^-$ is the RSC constant of the $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ objective. Hence $\rho_{s+s^*}^-$ is n times the restricted strong convexity of the normalized objective $\frac{1}{n}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ whose scale is independent of n. Similarly, $\sqrt{\rho_1^+}$ hides a $\sqrt{n}$. Thus $|\bar{x}_{\min}|$ essentially scales as $1/\sqrt{n}$.

Remark 2: The γ parameter in the above theorem is somewhat similar to the standard incoherence parameter [24], although the incoherence parameter can be significantly larger than γ. Further, existing results for OMP [26] require the incoherence parameter to be strictly less than 1 while our analysis holds for arbitrary values of γ. Thus, our results apply to more general design matrices $\mathbf{A}$.
Remark 3: Our assumption on $|\bar{x}_{\min}|$ is better at least by a factor of $\sqrt{\tilde{\kappa}}$ than corresponding assumptions made in recent work that analyzes OMP for support recovery [20, 21, 25] (see Table 1).
Remark 4: To the best of our knowledge, [13] is the only known support recovery result for LASSO under RSC, that provides strong guarantees as our result above. However, the non-convex penalty based algorithm of [13] might produce iterates which are dense, so intermediate steps can be more expensive than sparsity preserving OMP. Furthermore, while qualitatively, our bound is similar to the bound of [13], their proof requires $n \geq \|\bar{\mathbf{x}}\|_1^2 \log d$ which, naïvely, for many problems with imbalanced non-zero elements of $\bar{\mathbf{x}}$ can be as large as $(s^*)^2$.

Proof Sketch of Theorem 3.1 (see Appendix B.2 for details): Theorem 3.2 (stated below) guarantees that OMP has a very small objective value after a certain number of support expansion steps. This guarantees small generalization error (Theorem 3.3), but not support recovery. To guarantee support recovery, our proof critically exploits a novel observation (Lemma B.4 in Appendix B.2) that if at any iteration of OMP, full support recovery has not happened, then OMP decreases function value by a fixed, but small, additive constant. Theorem 3.2 allows us to say that even this small constant decrement cannot happen for too long since the objective value is already small. Overall, this means that support recovery must happen soon after we have small objective value.

Let s be the iteration index that is sufficient to satisfy the conditions for Theorem 3.2. From Theorem 3.2 we have with probability at least $1 - 2\delta$,

$$\|\mathbf{A}\mathbf{x}_s - \mathbf{y}\|_2^2 \leq \|\mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\|_2^2 + 40 \frac{\sigma^2 s \rho_1^+ \log(d/\delta)}{\rho_{s+s^*}^+}. \tag{3.2}$$

$$\leq \|\boldsymbol{\eta}\|_2^2 + 40\sigma^2 s \log(d/\delta)$$

Suppose any one of the support index has not been recovered (that is, $|\mathbf{S}^* \backslash \mathbf{S}| > 0$) then if $j \in (\mathbf{S}^* \backslash \mathbf{S})^c$ is selected by OMP in its $(s+1)^{th}$ iteration, we have by step 4 of Algorithm 1,

$$\left\|\mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T \mathbf{r}_s\right\|_\infty \leq |\mathbf{A}_j^T \mathbf{r}_s|. \tag{3.3}$$

In Lemma B.4, we lower bound the LHS of (3.3) as follows:

$$\left\|\mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T \mathbf{r}_s\right\|_\infty \geq \rho_{s+s^*}^- |\bar{x}_{\min}| - \sqrt{2}(1+\gamma)\sigma\sqrt{\rho_1^+ \log(d/\delta)}, \tag{3.4}$$

with probability at least $1 - 2\delta$. Since $|\bar{x}_{\min}| \geq \left(1 + \sqrt{2}(1+\gamma)\right) \frac{\sigma}{\rho_{s+s^*}^-}\sqrt{\rho_1^+ \log(d/\delta)}$, combining (3.3) with (3.4) gives,

$$\sigma^2 \log \frac{d}{\delta} \leq \frac{1}{\rho_1^+}\left(\mathbf{A}_j^T \mathbf{r}_s\right)^2. \tag{3.5}$$

This gives us an additive decrease in the function value:

$$\|\mathbf{A}\mathbf{x}_{s+1} - \mathbf{y}\|_2^2 \leq \min_{x_j} \|\mathbf{A}_j x_j - \mathbf{r}_s\|_2^2$$

$$= \|\mathbf{A}\mathbf{x}_s - \mathbf{y}\|_2^2 - \frac{1}{\rho_1^+}\left(\mathbf{A}_j^T \mathbf{r}_s\right)^2 \leq \|\mathbf{A}\mathbf{x}_s - \mathbf{y}\|_2^2 - \sigma^2 \log(d/\delta) \tag{3.6}$$

Suppose that for another l iterations, the full support is not recovered. Then,

$$\|\mathbf{A}\mathbf{x}_{s+l} - \mathbf{y}\|_2^2 \leq \|\mathbf{A}\mathbf{x}_s - \mathbf{y}\|_2^2 - \sigma^2 l \log(d/\delta). \tag{3.7}$$

Further it can be shown that the function value at iteration $s + l$ cannot be too small,

$$\|\mathbf{A}\mathbf{x}_{s+l} - \mathbf{y}\|_2^2 \geq \|\boldsymbol{\eta}\|_2^2 - \sigma^2(s + l + s^*) - 4\sigma^2(s + l + s^*)\sqrt{\log(d/\delta)}, \tag{3.8}$$

with probability at least $1 - \delta$. Therefore combining (3.8) and (3.2) and plugging them in (3.7), we finally get,

$$l \leq 80s + s + s^* = \mathcal{O}(s). \tag{3.9}$$

5

Therefore with good probability, OMP recovers the full support in $\mathcal{O}(s)$ iterations. See Appendix B.2 for details. $\qquad\square$

We now bound the training error for OMP after running a certain number of iterations (which are fewer than the number of iterations required for support recovery as shown in Theorem 3.1). The proof of this theorem follows via a modification of the proof of Lemma A.5 in [28]. See Appendix B.1 for the proof.

Theorem 3.2 (Training Error for OMP). *Consider the setting of Theorem 3.1. Also, let*

$$s \geq 8\widetilde{\kappa}_{s+s^*}s^* \cdot \log\left(\frac{5\rho^+_{s+s^*}}{\rho^-_{s+s^*}}\right).$$

Then with probability $1 - 2\delta$, the output $\hat{\mathbf{x}}_s$ of OMP after s steps satisfies:

$$\frac{1}{n}\left\|\mathbf{A}\hat{\mathbf{x}}_s - \mathbf{y}\right\|_2^2 \leq \frac{1}{n}\left\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\right\|_2^2 + 40\frac{\sigma^2 s \log(d/\delta)}{\rho^+_{s+s^*}} \cdot \frac{\rho^+_1}{n}. \tag{3.10}$$

Given good objective value decrease, we can show a tight generalization error on the output of OMP. While in general support recovery is the main goal of a sparse regression algorithm, in several problem scenarios one might not care about support recovery and focus only on the accuracy of the learned predictor. See Appendix B.3 for the proof.

Theorem 3.3 (Generalization Error for OMP). *Consider the setting of Theorem 3.1. Let $\hat{\mathbf{x}}_s$ be the output of OMP after s iterations. For any constant $C_1 \geq 8$, there exists a constant $C_2(\leq 9C_1)$ such that if s satisfies,*

$$C_1\widetilde{\kappa}_{s+s^*}s^* \cdot \log\left(\frac{5\rho^+_{s+s^*}}{\rho^-_{s+s^*}}\right) \geq s \geq 8\widetilde{\kappa}_{s+s^*}s^* \cdot \log\left(\frac{5\rho^+_{s+s^*}}{\rho^-_{s+s^*}}\right),$$

then with probability at least $1 - 4\delta$,

$$\frac{1}{n}\left\|\mathbf{A}(\hat{\mathbf{x}}_s^{OMP} - \bar{\mathbf{x}})\right\|_2^2 \leq C_2\frac{\sigma^2\widetilde{\kappa}_{s+s^*}s^*}{n} \cdot \log\left(\frac{5\rho^+_{s+s^*}}{\rho^-_{s+s^*}}\right) \cdot \log\frac{d}{\delta}. \tag{3.11}$$

3.1 Gaussian ensemble

Finally, we instantiate the above theorems for a Gaussian ensemble, i.e., when $\mathbf{A}$ is sampled from a Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. We denote the maximum and the minimum singular values of $\boldsymbol{\Sigma}$ as $\sigma_{\max}$ and $\sigma_{\min}$ and the condition number of $\boldsymbol{\Sigma}$ as $\kappa(\boldsymbol{\Sigma})$. To the best of our knowledge, the following is the best known generalization error guarantee in this setting in terms of the dependence on $\kappa(\boldsymbol{\Sigma})$.

Corollary 3.3.1 (Gaussian ensemble: generalization error). *Let the rows of the matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ be sampled from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\Sigma_{ii} \leq 1 \, \forall \, i \in [d]$ and $\bar{\mathbf{x}}$ be a s^*-sparse vector. Let $\hat{\mathbf{x}}_s$ be the output of OMP after s iterations and $\mathbf{S} = supp(\hat{\mathbf{x}}_s)$ be the support recovered, where,*

$$s = C_2\kappa(\boldsymbol{\Sigma}) \cdot \log\left(45\kappa(\boldsymbol{\Sigma})\right)s^*, n > 4C_1\frac{s\log d}{\sigma_{\min}(\boldsymbol{\Sigma})}, \text{ and } s + s^* \geq \log\frac{1}{\delta},$$

for any $\delta > 0$. Then with probability at least $1 - 4\delta - e^{-C_0 n}$, the following holds:

$$\frac{1}{n}\left\|\mathbf{A}(\hat{\mathbf{x}}_s^{OMP} - \bar{\mathbf{x}})\right\|_2^2 \leq C_3\frac{\sigma^2\kappa(\boldsymbol{\Sigma})s^*}{n} \cdot \log\left(45\kappa\left(\boldsymbol{\Sigma}\right)\right) \cdot \log\frac{d}{\delta}$$

Here C_0, C_1, C_3 and C_4 are universal constants independent of any problem parameters.

Note the linear dependence of generalization error on $\kappa(\boldsymbol{\Sigma})$. This matches the lower bound of [29], although technically the bound does not apply to OMP as $s > s^*$. The proof follows directly from Theorem 3.3 along with standard concentration results. See Appendix B.3 for details.

We now present support recovery result for Gaussian ensembles. For simplicity, we consider the case when $\mathbf{A}$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. This can also be extended to $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ but involves cumbersome linear algebraic computations, which we avoid for simplicity.

6

Corollary 3.3.2 (Gaussian ensemble: support recovery). *Let the rows of the matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ be sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}_{d \times d})$ and $\bar{\mathbf{x}}$ be a s^*-sparse vector. Suppose further that $|\bar{x}_{\min}| \geq 23\sigma\sqrt{\frac{\log(d/\delta)}{n}}$. Let $\hat{\mathbf{x}}_s$ be the output of OMP after s iterations and $\mathbf{S} = supp(\hat{\mathbf{x}}_s)$ be the support recovered, where,*

$$s \geq C_1 s^*, n > C_2 (s^*)^2 \log \frac{d}{\delta}, \text{ and } s + s^* \geq \log \frac{1}{\delta},$$

for any $\delta > 0$. Then $\mathbf{S}^* \subseteq supp(\hat{\mathbf{x}}_s)$ and $\|\hat{\mathbf{x}}_s - \bar{\mathbf{x}}\|_\infty \leq 2\sigma\sqrt{\frac{2\log(s/\delta)}{n}}$ with probability at least $1 - e^{-C_0 n} - 9\delta$. Here C_0, C_1 and C_2 are universal constants independent of any problem parameter.

This matches the bounds of [13] up to constants. The proof directly follows from Theorem 3.1 along with standard Gaussian concentration results. See Appendix B.3 for details.

4 Lower bounds for OMP

In this section, we provide lower bounds on the performance of OMP, both in terms of support recovery and generalization error. These bounds show that:

- The imperative quantities we make assumptions on in the upper bound section, viz: $\widetilde{\kappa}_{s+s^*}$ and γ are relevant and meaningful.
- Given bounds on these quantities, our results are tight, up to logarithmic factors.

To provide these lower bounds, we construct matrices $\mathbf{M}^{(\epsilon)}$ that are parametrized by ϵ. We fix $\bar{\mathbf{x}}$ to be an s^*-sparse vector such that:

$$\begin{cases} \bar{x}_i = \sqrt{1/s^*} & \text{if } 1 \leq i \leq s^*, \\ \bar{x}_i = 0 & \text{if } s^* < i. \end{cases} \tag{4.1}$$

Thus, $\mathbf{S}^* := supp(\bar{\mathbf{x}}) = \{1, 2, \ldots, s^*\}$. All our lower bound theorems use this fixed vector which is independent of the noise level σ. Our results are thus stronger than a typical minimax rate in which $\bar{\mathbf{x}}$ can be *scaled* based on σ. For instance, the lower bounds of [29], [30] use such a strategy. Also, the support is distributed evenly across the $\bar{x}_i$'s (4.1). Thus, we show that even large elements are not recovered.

We now define $\mathbf{M}^{(\epsilon)} \in \mathbb{R}^{n \times d}$ for a given $\epsilon \in [0, 1]$, any $s^* \leq d \leq n$ in the following manner: $\mathbf{M}^{(\epsilon)}_{1:s^*}$ are random orthogonal vectors such that $\left\| \mathbf{M}^{(\epsilon)}_i \right\|_2^2 = n, \forall i \in [s^*]$. For $i \in [d] \setminus [s^*]$, each column vector is defined as,

$$\mathbf{M}^{(\epsilon)}_i = \sqrt{\frac{1-\epsilon}{s^*}} \sum_{j=1}^{s^*} \mathbf{M}^{(\epsilon)}_j + \sqrt{\epsilon}\, \mathbf{g}_i, \tag{4.2}$$

where $\mathbf{g}_i$ is such that $\|\mathbf{g}_i\|_2^2 = n$, $\mathbf{g}_i^T \mathbf{M}^{(\epsilon)}_{1:s^*} = \mathbf{0}$ and $\mathbf{g}_i^T \mathbf{g}_j = 0$ for all $i \neq j$.

The intuition behind this construction is that OMP would prefer the *average* direction $\mathbf{M}^{(\epsilon)}_{\mathbf{S}^*} \bar{\mathbf{x}}$ over any of the correct directions $\mathbf{M}^{(\epsilon)}_i$, where $i \in \mathbf{S}^*$. Thus, we add a scaled version of $\mathbf{M}^{(\epsilon)}_{\mathbf{S}^*} \bar{\mathbf{x}}$ to each of the other orthogonal vectors of the matrix.

The parameter ϵ is set carefully to ensure that the condition number of the matrix does not increase too much, so that $\mathbf{M}^{(\epsilon)}$ satisfies the constraints of Theorem 3.3 and Theorem 3.1 (upto constants). This is captured in the next lemma:

Lemma 4.1. *The matrix $\mathbf{M}^{(\epsilon)}$ satisfies*

- $\widetilde{\kappa}_s \left(\mathbf{M}^{(\epsilon)} \right) \leq 4(1 + 2(1 - \epsilon)s) = \mathcal{O}(s)$

- $\left\| \mathbf{M}^{(\epsilon)T}_{\mathbf{S}^* \setminus \mathbf{S}} \mathbf{M}^{(\epsilon)}_{\mathbf{S}} \left(\mathbf{M}^{(\epsilon)T}_{\mathbf{S}} \mathbf{M}^{(\epsilon)}_{\mathbf{S}} \right)^{-1} \right\|_\infty \leq \frac{1}{\sqrt{s^*(1-\epsilon)}}$ *for $\mathbf{S} \cap \mathbf{S}^* = \phi$.*

We now use the above construction to show that in the noiseless case, i.e., when $\mathbf{y} = \mathbf{M}^{(\epsilon)} \bar{\mathbf{x}}$, OMP fails to recover any of the support elements in $\mathbf{S}^*$ for some ϵ. Similarly, we show that in the noisy case, support recovery fails and hence the generalization error of OMP is also large and matches the upper bound provided in Theorem 3.3. Proofs for this section can be found in Appendix C.

4.1 Noiseless case

For the deterministic noiseless case, i.e., $\sigma = 0$, we consider the matrix $\mathbf{M}^{(\epsilon)}$ for $\epsilon = (1 - 3/2s^*)$ and show that OMP requires to add *all* the elements in support to recover the correct support.

Theorem 4.2. *For every value of d, n and s^* where $s^* \leq d \leq n$, there exists a design matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ and a s^*-sparse vector $\bar{\mathbf{x}}$ (defined in (4.2), (4.1)) such that the following holds true for OMP when applied to the sparse linear regression problem with $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}}$ and when OMP is executed for $s \leq d - s^*$ iterations:*

- $\widetilde{\kappa}_s(\mathbf{A}) \leq 16(s/s^*)$ *and* $\gamma \leq \sqrt{2/3}$.

- *The support set $\mathbf{S}$ recovered by OMP after s iterations is disjoint from $\mathbf{S}^*$, i.e., $\mathbf{S}^* \cap \mathbf{S} = \phi$.*

Our support recovery result in Theorem 3.1 requires $s \geq C\widetilde{\kappa}_{s+s^*}s^*$ and one natural question is whether running OMP for s iterations is necessary for recovering the actual support. This theorem guarantees that it is indeed the case, i.e., if design matrix $\mathbf{A}$ is ill-conditioned then OMP has to work with support sets of size $s \geq \widetilde{\kappa}_{s+s^*}s^*$. This in turn implies that the number of rows in $\mathbf{A}$ (i.e., sample complexity) should also scale with $\widetilde{\kappa}_{s+s^*}$.

Note that the lower bound results of [29], [30] do not provide any insights for how the sample complexity of an algorithm should scale with κ_{s+s^*} for support recovery. In fact for this problem their results are vacuous if $|\bar{x}_{\min}|$ is reasonably large. For instance, with the $\bar{\mathbf{x}}$ defined in (4.1) and the design matrix proposed by [29], OMP can recover the true support of $\bar{\mathbf{x}}$ *exactly* after just $\mathcal{O}(s^*)$ iterations with $n = s^* \log d$ samples. Thus, a large condition number of $\mathbf{A}$ in their construction does not imply difficulty in recovery for OMP.

4.2 Noisy case

For the noisy case, i.e., $\sigma \neq 0$, we can study both support recovery as well as generalization error behavior with respect to the restricted condition number $\widetilde{\kappa}_{s+s^*}$. For this section, we consider the matrix $\mathbf{M}^{(\epsilon)}$ for $\epsilon = (1 - 1/4s^*)$. That is, we show that with high probability, OMP starts recovering the correct support only after $d^{1-\alpha}$ iterations for some constant $\alpha > 0$. This further implies that the generalization error cannot be better than the lower bound on generalization error we showed in Theorem 3.3 (upto constants).

Theorem 4.3. *For every value of d and s^*, and any constants $\alpha \in (0,1)$, $\delta \in (0,1)$, such that $8 \leq s^* \leq s \leq d^{1-\alpha}$ and $d \geq \max\left\{32 \log(1/\delta), 4^{1/\alpha}\right\}$, there exists a sparse linear regression problem with $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}} + \boldsymbol{\eta}$, $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n \times n})$, with design matrix $\mathbf{A}$, and a s^*-sparse vector $\bar{\mathbf{x}}$ defined in (4.2),(4.1) such that the following holds:*

- $\widetilde{\kappa}_s(\mathbf{A}) \leq 36(s/s^*)$ *for all s and $\gamma \leq 1/2$,*

- *With probability at least $1 - \delta$, the output $\hat{\mathbf{x}}_s$ of OMP after s steps satisfies:*

$$\frac{1}{n}\|\mathbf{A}\hat{\mathbf{x}}_s - \mathbf{A}\bar{\mathbf{x}}\|_2^2 \geq \frac{\sigma^2 \widetilde{\kappa}_{s+s^*} s^*}{18n} \cdot \log \frac{d}{\delta},$$

- *Support set $\mathbf{S}$ recovered by OMP after s iterations is disjoint from $\mathbf{S}^*$.*

Note that the dependence of the generalization error bound on $\widetilde{\kappa}_{s+s^*}$ matches our generalization error bound in Theorem 3.3. Interestingly, for our construction, noise ends up helping recovery because while Theorem 4.2 ensures that the recovery of true support elements does not occur till the very last step, noise can only help in recovering one of the true elements. However, the probability of picking up the correct element by chance is tiny as we restrict $s \leq d^{1-\alpha}$. We in fact believe that the result holds generally for any s and d. However, proving it turns out to be quite intricate since it requires finer results about the the behavior of the order statistics of independent Gaussian variables.

5 Simulations

In this section, we present simulations that verify our results. In particular, we generate a matrix $\mathbf{M}^{(\epsilon)} \in \mathbb{R}^{1000 \times 100}$, and a fixed $s^* = 10$-sparse vector $\bar{\mathbf{x}}$ by using the construction given in (4.2) and

(4.1) where $\epsilon \in (0, 1)$. We then generate $\mathbf{y} = \mathbf{M}^{(\epsilon)}\bar{\mathbf{x}} + \sigma\mathcal{N}(\mathbf{0}, \mathbf{I}_{n \times n})$ and apply OMP for recovering the support of $\bar{\mathbf{x}}$. $\hat{s}(\epsilon)$ denotes the index or support set size that is needed by OMP for fully recovering the support of $\bar{\mathbf{x}}$.

Note that we can also compute the actual value of $\widetilde{\kappa}$ for $\mathbf{M}^{(\epsilon)}$; in general the restricted condition number of $\mathbf{M}^{(\epsilon)}$ increases with decreasing ϵ, thus increasing the difficulty of the support recovery problem.
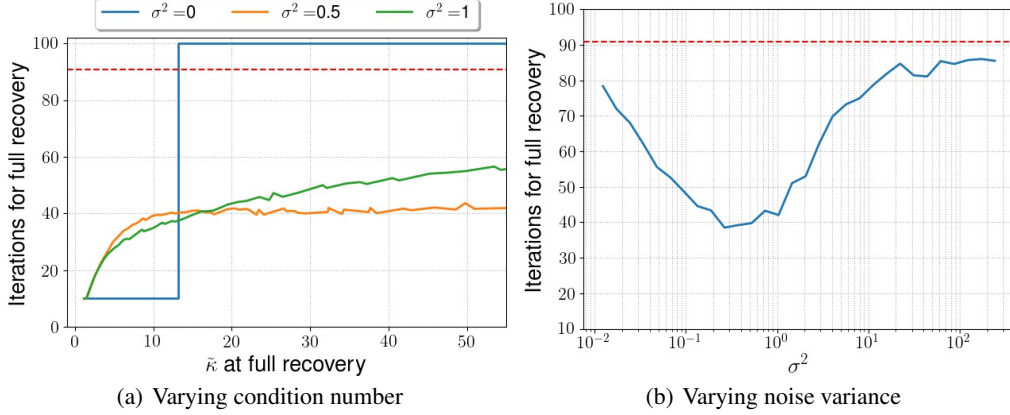

(a) Varying condition number   (b) Varying noise variance

Figure 1: Number of iterations required for recovering the full support of $\bar{\mathbf{x}}$ with respect to the restricted condition number ($\widetilde{\kappa}_{s+s^*}$) of the design matrix and the sub-Gaussian parameter of the noise term (σ^2).

Figure 1(a) plots $\hat{s}(\epsilon)$ (i.e. support size required for full recovery) against restricted condition number $\widetilde{\kappa}(\mathbf{M}^{(\epsilon)})$ of $\mathbf{M}^{(\epsilon)}$ generated by varying $\epsilon \in (0, 1)$. Theorem 4.2 claims that for $\sigma = 0$, full recovery requires $\widetilde{\kappa}_s$ to be smaller than $\mathcal{O}(d/s^*)$, which is observed in Figure 1(a). For larger variance σ^2, full recovery requires larger number of iterations for smaller $\widetilde{\kappa}$.

As mentioned in the remark below Theorem 4.3, adding noise can only help in case of large $\widetilde{\kappa}$ as our construction precludes full recovery unless $s = d$. We observe this behavior in both Figure 1(a) and 1(b), where slightly larger value of σ ends up helping support recovery, but for larger values of noise variance, OMP's performance is similar to an algorithm that simply selects each feature uniformly at random.

6 Conclusion

In this paper, we analyze OMP for the sparse regression problem under RSC/RSS assumptions. We obtain support recovery and generalization guarantees for OMP under this setting. We also provide lower bounds for OMP showing that our results are tight up to logarithmic factors. We note that our results significantly improve upon a long list of existing results for greedy methods and match the best known results for sparse regression that use nonconvex penalty based methods. In contrast to nonconvex penalty methods however, OMP guarantees the sparsity of intermediate iterates and hence can be much more efficient. We also verify our results with synthetic experiments.

References

[1] Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.*, 40(5):2452–2482, 10 2012.

[2] Thomas Blumensath and Mike E Davies. Gradient pursuit for non-linear sparse signal modelling. In *Signal Processing Conference, 2008 16th European*, pages 1–5. IEEE, 2008.

[3] Stéphane Boucheron and Maud Thomas. Concentration inequalities for order statistics. *arXiv preprint arXiv:1207.7209*, 2012.

[4] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theor.*, 51(12):4203–4215, December 2005.

[5] Emmanuel Candes and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse problems*, 23(3):969, 2007.

[6] Emmanuel Candes, Terence Tao, et al. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[7] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2006.

[8] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1.

[9] Chirag Gupta, Arun Sai Suggala, Ankit Goyal, Harsha Vardhan Simhadri, Bhargavi Paranjape, Ashish Kumar, Saurabh Goyal, Raghavendra Udupa, Manik Varma, and Prateek Jain. Protonn: Compressed and accurate knn for resource-scarce devices. In *International Conference on Machine Learning*, pages 1331–1340, 2017.

[10] D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *ArXiv e-prints*, October 2011.

[11] Prateek Jain, Ambuj Tewari, and Purushottam Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pages 685–693, 2014.

[12] Ali Jalali, Christopher C Johnson, and Pradeep K Ravikumar. On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems*, pages 1935–1943, 2011.

[13] Po-Ling Loh, Martin J Wainwright, et al. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics*, 45(6):2455–2482, 2017.

[14] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[15] Sahand Negahban and Martin J Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.

[16] Y. C. Pati, R. Rezaiifar, Y. C. Pati R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of the 27 th Annual Asilomar Conference on Signals, Systems, and Computers*, pages 40–44, 1993.

[17] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pages 40–44. IEEE, 1993.

[18] Amela Prelić, Stefan Bleuler, Philip Zimmermann, Anja Wille, Peter Bühlmann, Wilhelm Gruissem, Lars Hennig, Lothar Thiele, and Eckart Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

[19] Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[20] Jie Shen and Ping Li. On the iteration complexity of support recovery via hard thresholding pursuit. In *International Conference on Machine Learning*, pages 3115–3124, 2017.

[21] Jie Shen and Ping Li. Partial hard thresholding: Towards a principled analysis of support recovery. In *Advances in Neural Information Processing Systems*, pages 3127–3137, 2017.

[22] Joel A Tropp and Anna C Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on information theory*, 53(12):4655–4666, 2007.

[23] Joel A. Tropp and Anna C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE TRANS. INFORM. THEORY*, 53:4655–4666, 2007.

[24] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[25] Xiaotong Yuan, Ping Li, and Tong Zhang. Exact recovery of hard thresholding pursuit. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3558–3566. Curran Associates, Inc., 2016.

[26] Tong Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10(Mar):555–568, 2009.

[27] Tong Zhang. Adaptive forward-backward greedy algorithm for learning sparse representations. *IEEE transactions on information theory*, 57(7):4689–4708, 2011.

[28] Tong Zhang. Sparse recovery with orthogonal matching pursuit under rip. *IEEE Transactions on Information Theory*, 57(9):6215–6221, 2011.

[29] Yuchen Zhang, Martin J Wainwright, and Michael I Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. In *Conference on Learning Theory*, pages 921–948, 2014.

[30] Yuchen Zhang, Martin J Wainwright, Michael I Jordan, et al. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *Electronic Journal of Statistics*, 11(1):752–799, 2017.

A General properties of Gaussians

Lemma A.1. *If* $\mathbf{U} \in \mathbb{R}^{n \times k}$ *has unit norm orthogonal columns where* $k \geq \log 1/\delta$, *and* $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n \times n})$ *then*

$$\left\| \mathbf{U}^T \boldsymbol{\eta} \right\|_2^2 \leq \sigma^2 k + 4\sigma^2 \sqrt{k \log \frac{1}{\delta}} \tag{A.1}$$

holds with probability at least $1 - \delta$

Proof. Using an exponential probability tail inequality of positive semidefinite quadratic forms in a sub-gaussian random vector [10],

$$P\left\{ \frac{1}{\sigma^2} \left\| \mathbf{U}^T \boldsymbol{\eta} \right\|_2^2 > Tr(\mathbf{U}\mathbf{U}^T) + 2\sqrt{Tr(\mathbf{U}\mathbf{U}^T \mathbf{U}\mathbf{U}^T)t} + 2 \left\| \mathbf{U} \right\|_2 t \right\} \leq e^{-t}$$

$$\text{or, } P\left\{ \frac{1}{\sigma^2} \left\| \mathbf{U}^T \boldsymbol{\eta} \right\|_2^2 > k + 2\sqrt{kt} + 2t \right\} \leq e^{-t}$$

Setting $t = \log \frac{1}{\delta}$ and assuming $k \geq \log \frac{1}{\delta}$, we get the required result. $\qquad \square$

B Proofs of results in Section 3

Algorithm 2 Orthogonal Matching Pursuit (OMP) for General $Q(\cdot)$

1: **procedure** GENERAL OMP(s)
2:     $\mathbf{S}_0 = \phi, \mathbf{x}_0 = \mathbf{0}$
3:     **for** $k = 1, 2, \ldots, s$ **do**
4:         $j \leftarrow \underset{i \notin \mathbf{S}_{k-1}}{\arg \max} |\nabla Q(\mathbf{x}_{k-1})|$
5:         $\mathbf{S}_k \leftarrow \mathbf{S}_{k-1} \cup \{j\}$
6:         $\mathbf{x}_k \leftarrow \underset{\text{supp}(\mathbf{x}) \subseteq \mathbf{S}_k}{\arg \min} Q(\mathbf{x})$
7:     **end for**
8:     **return** $\mathbf{x}_s$
9: **end procedure**

B.1 Proof of Theorem 3.2

We wish to prove Theorem 3.2 in this section. We in fact prove a more general version of the theorem that holds for any function $Q(\cdot)$. Algorithm 2 is the OMP algorithm generalized to any $Q(\cdot)$. To show guarantees in this setting, we assume that $Q(\cdot)$ satisfies Restricted Smoothness (RSS) and Restricted Strong Convexity (RSC) properties, given by respective constants ρ_s^+ and ρ_s^- parametrized by the sparsity level s. We need that for all $\mathbf{x}, \mathbf{y}$ such that $\|\mathbf{y} - \mathbf{x}\|_0 \leq s$,

$$\rho_s^- \left\| \mathbf{y} - \mathbf{x} \right\|_2^2 \leq Q(\mathbf{y}) - Q(\mathbf{x}) - \langle \nabla Q(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \rho_s^+ \left\| \mathbf{y} - \mathbf{x} \right\|_2^2. \tag{B.1}$$

Note that for the sparse linear regression problem, setting the objective function to $Q(\mathbf{x}) := \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ converts Algorithm 2 to Algorithm 1. Further, this $Q(\cdot)$ indeed satisfies the RSS and RSC constants of the matrix $\mathbf{A}$ as defined in definitions 2.2 and 2.1.

In order to present this result, we need the following notation from [28]. For any sparsity level s and any vector $\mathbf{x}$, define

$$\epsilon_s(\mathbf{x}) = \sup_{\|\mathbf{u}\|_0 = s, \|\mathbf{u}\|_2 = 1} |\nabla Q(\mathbf{x})^T \mathbf{u}|, \tag{B.2}$$

$\epsilon_s(\mathbf{x})$ is essentially the ℓ_2 norm of $\nabla Q(\mathbf{x})$ restricted to its largest s coordinates (in absolute value). We are specifically interested in $\epsilon_s(\bar{\mathbf{x}})$ where $\bar{\mathbf{x}}$ is an s^*-sparse vector that we are trying to estimate. In particular for the linear regression case, $\bar{\mathbf{x}}$ is the generative parameter ($\mathbf{y} = \mathbf{A}\bar{\mathbf{x}} + \boldsymbol{\eta}$). As we shall see in B.3, $\epsilon_s(\bar{\mathbf{x}})$ determines the sub-optimality of the OMP estimator on the sparse linear regression problem. Thus, we first write a lemma quantifying how large this quantity can be.

Lemma B.1. *For the sparse linear regression problem with $Q(\mathbf{x}) = \|\mathbf{Ax} - \mathbf{y}\|_2^2$, suppose that $\max_i \|\mathbf{A}_i\|_2^2 \leq n$. Then with probability at least $1 - \delta$,*

$$\epsilon_s^2(\bar{\mathbf{x}}) \leq 8\sigma^2 \rho_1^+ s \log \frac{2d}{\delta}. \tag{B.3}$$

Proof. For the sparse linear regression problem,

$$
\begin{aligned}
\epsilon_s(\bar{\mathbf{x}}) &= \sup_{\|\mathbf{u}\|_0 = s, \|\mathbf{u}\|_2 = 1} |\nabla Q(\bar{\mathbf{x}})^T \mathbf{u}| \\
&= \sup_{\|\mathbf{u}\|_0 = s, \|\mathbf{u}\|_2 = 1} \left| 2\mathbf{u}^T \mathbf{A}^T \boldsymbol{\eta} \right| \\
&\leq \sup_{\|\mathbf{u}\|_0 = s, \|\mathbf{u}\|_2 = 1} \left(2\|\mathbf{u}\|_2 \left\| (\mathbf{A}^T \boldsymbol{\eta})_{\mathrm{supp}(\mathbf{u})} \right\|_2 \right) \\
&= 2 \sup_{|\mathbf{S}| = s} \left\| \mathbf{A}_{\mathbf{S}}^T \boldsymbol{\eta} \right\|_2 \\
&\leq 2\sqrt{s} \left\| \mathbf{A}^T \boldsymbol{\eta} \right\|_\infty \\
&\overset{\xi_1}{\leq} 2\sqrt{s}\sigma \sqrt{2\rho_1^+ \log \frac{2d}{\delta}} \\
&= 2\sqrt{2}\sigma \sqrt{\rho_1^+ s \log \frac{2d}{\delta}} \\
\implies \epsilon_s^2(\bar{\mathbf{x}}) &\leq 8\sigma^2 \rho_1^+ s \log \frac{2d}{\delta}
\end{aligned}
$$

Here ξ_1 holds with probability $(1 - \delta)$. $\qquad\square$

We now write the key technical lemma, which is a strengthened version of Lemma A.5 from [28]. We show that if OMP is run to a slightly larger (only in constants) expansion set then we get much better convergence to $Q(\bar{\mathbf{x}})$. This improved convergence is key to achieving a better generalization error (Theorem 3.3). For the statement and proof of the following lemma, we use the notation of [28]. Additionally, we define $s^* = |\bar{F}|$.

Lemma B.2 (Modified version of lemma A.5 in [28]). *Suppose OMP is run till $s \geq k + s^*$ steps, where*

$$k \geq 4|\bar{F} \setminus F^{(0)}| \frac{\rho_1^+}{\rho_s^-} \log\left(20 \left(\frac{\rho_s^+}{\rho_s^-} \right)^2 \right), \text{ then}$$

$$Q(\mathbf{x}_k) \leq Q(\bar{\mathbf{x}}) + 2.5\epsilon_s(\bar{\mathbf{x}})^2/\rho_s^+.$$

Proof. We strengthen the proof of Lemma A.5 of [28] appropriately. First, create the split on L differently, using the following μ:

$$\mu = 10 \left(\frac{\rho_s^+}{\rho_s^-} \right)^2. \tag{B.4}$$

This value of μ will be plugged in later.

Next, we use lemma A.1 from [28] at level s instead of level m as has been done in the original proof. This leads to a modified value of q_l:

$$\min_{\mathbf{x} \in \bar{F}_l} Q(\mathbf{x}) \leq Q(\bar{\mathbf{x}}) + q_l, \text{ where}$$

$$q_l = 1.5\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2 + 0.5 \frac{\epsilon_s(\bar{\mathbf{x}})^2}{\rho_s^+}.$$

Thus, equation (12) in [28] becomes

$$Q(\mathbf{x}_k) - Q(\bar{\mathbf{x}}) \leq 3\mu^{-1}\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2 + \frac{0.5}{\rho_s^+}(1 + \mu^{-1})\epsilon_s(\bar{\mathbf{x}})^2. \tag{B.5}$$

13

We have a different splitting condition compared to [28]. Now, either

$$2\mu^{-1}\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2 \le \frac{1}{\rho_s^+}(1+\mu^{-1})\epsilon_s(\bar{\mathbf{x}})^2$$

in which case the theorem follows by simply replacing the first term in the RHS of equation B.5 and using $(1+\mu^{-1}) \le 2$. We consider the other case:

$$2\mu^{-1}\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2 > \frac{1}{\rho_s^+}(1+\mu^{-1})\epsilon_s(\bar{\mathbf{x}})^2.$$

In this case, we use the same approach as in [28], but modify from the fourth inequality onwards (see below),

$$\rho_s^- \|\mathbf{x}_k - \bar{\mathbf{x}}\|_2^2 \le 2(Q(\mathbf{x}_k) - Q(\bar{\mathbf{x}})) + \frac{\epsilon_s(\bar{\mathbf{x}})^2}{\rho_s^-}$$

$$\le 6\mu^{-1}\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2 + (2+\mu^{-1})\frac{\epsilon_s(\bar{\mathbf{x}})^2}{\rho_s^-}$$

$$\le 6\mu^{-1}\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2 + 2(1+\mu^{-1})\frac{\epsilon_s(\bar{\mathbf{x}})^2}{\rho_s^-}$$

$$\le 6\mu^{-1}\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2 + \left(\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2\right)\frac{4\rho_s^+\mu^{-1}}{\rho_s^-}$$

$$= \left(\frac{4\rho_s^+\mu^{-1}}{\rho_s^-} + 6\mu^{-1}\right)\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2$$

$$\le \left(\frac{10\rho_s^+\mu^{-1}}{\rho_s^-}\right)\rho_s^+ \sum_{i=2^l}^m \bar{x}_i^2$$

$$\overset{\xi_1}{=} \rho_s^- \sum_{i=2^l}^m \bar{x}_i^2.$$

In ξ_1 we plug in the value of μ from B.4. The rest of the proof follows in the same manner as the original proof. □

Theorem B.3. *If $Q(\cdot)$ satisfies RSC and RSS and if*

$$s \ge 8\widetilde{\kappa}_{s+s^*}s^* \log 5\kappa_{s+s^*}, \tag{B.6}$$

then the output $\hat{\mathbf{x}}_s$ of OMP after s steps satisfies

$$Q(\hat{\mathbf{x}}_s) \le Q(\bar{\mathbf{x}}) + 2.5\frac{\epsilon_{s+s^*}(\bar{\mathbf{x}})^2}{\rho_{s+s^*}^+}. \tag{B.7}$$

Proof. Writing Lemma B.2 other words, if OMP is run till $\bar{s} \ge k + s^*$ steps, where

$$k \ge 8\widetilde{\kappa}_{\bar{s}}s^* \log 5\kappa_{\bar{s}}$$

then,

$$Q(\mathbf{x}_k) \le Q(\bar{\mathbf{x}}) + 2.5\frac{\epsilon_{\bar{s}}(\bar{\mathbf{x}})^2}{\rho_{\bar{s}}^+}.$$

Setting $k = s$ and $\bar{s} = s + s^*$, we have that if OMP is run till $s + s^*$ steps, where

$$s \ge 8s^*\widetilde{\kappa}_{s+s^*} \log 5\kappa_{s+s^*}$$

then,

$$Q(\mathbf{x}_s) \le Q(\bar{\mathbf{x}}) + 2.5\frac{\epsilon_{s+s^*}(\bar{\mathbf{x}})^2}{\rho_{s+s^*}^+}$$

□

14

Proof of Theorem 3.2. We set $Q(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ in Theorem B.3. Now, plugging in the value of $\epsilon_s(\bar{\mathbf{x}})^2$ from Lemma B.1, we get that with probability at least $1 - 2\delta$,

$$\|\mathbf{A}\hat{\mathbf{x}}_s - \mathbf{y}\|_2^2 \leq \|\mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\|_2^2 + 2.5 \frac{8\sigma^2 \rho_1^+ (s + s^*)}{\rho_{s+s^*}^+} \log \frac{d}{\delta}$$

$$\implies \frac{1}{n}\|\mathbf{A}\hat{\mathbf{x}}_s - \mathbf{y}\|_2^2 \leq \frac{1}{n}\|\mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\|_2^2 + 40 \frac{\sigma^2 \rho_1^+ s \log(d/\delta)}{n\rho_{s+s^*}^+}.$$

$\square$

B.2 Proof of Theorem 3.1

We first prove the following crucial lemma required for Theorem 3.1, as also discussed in the main paper.

Lemma B.4. *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ and $\bar{\mathbf{x}} \in \mathbb{R}^d$ be a s^*-sparse vector. Let $\mathbf{y} = \mathbf{A}\bar{\mathbf{x}} + \boldsymbol{\eta}$ and let $\hat{\mathbf{x}}_s$ be the output of OMP after s iterations, where*

$$s \geq 8\widetilde{\kappa}_{s+s^*} s^* \cdot \log\left(\frac{5\rho_{s+s^*}^+}{\rho_{s+s^*}^-}\right),$$

and $\widetilde{\kappa}_{s+s^*}$ is the restricted condition number (Definition 2.3). Moreover, let $\left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}(\mathbf{A}_{\mathbf{S}}^T\mathbf{A}_{\mathbf{S}})^{-1}\right\|_\infty \leq \gamma$ where $\mathbf{S} = supp(\hat{\mathbf{x}}_s)$. Then, if

$$|\bar{x}_{\min}| \geq (1 + \sqrt{2}(1+\gamma)) \frac{\sigma}{\rho_{s+s^*}^-}\sqrt{\rho_1^+ \log d/\delta}, \tag{B.8}$$

then for all $l \geq 0$, if complete support recovery hasn't happened till $s + l$ iterations of OMP, then with high probability, the function value after $s + l$ iterations satisfies,

$$\|\mathbf{A}\mathbf{x}_{s+l} - \mathbf{y}\|_2^2 \leq \|\mathbf{A}\mathbf{x}_s - \mathbf{y}\|_2^2 - \sigma^2 l \log \frac{d}{\delta} \tag{B.9}$$

Proof. If OMP at $(s+1)^{th}$ picks up an index j in $(\mathbf{S}^*)^c$, then

$$\left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp (\mathbf{A}\bar{\mathbf{x}} + \boldsymbol{\eta})\right\|_\infty \leq |\mathbf{A}_j^T \mathbf{A}_{\mathbf{S}}^\perp (\mathbf{A}\bar{\mathbf{x}} + \boldsymbol{\eta})| \tag{B.10}$$

We now compute a lower bound for the L.H.S. in (B.10):

$$\left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp (\mathbf{A}\bar{\mathbf{x}} + \boldsymbol{\eta})\right\|_\infty \geq \left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}}_{\mathbf{S}^*}\right\|_\infty - \left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \boldsymbol{\eta}\right\|_\infty$$

$$= \left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}\bar{\mathbf{x}}_{\mathbf{S}^*\backslash\mathbf{S}}\right\|_\infty - \left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \boldsymbol{\eta}\right\|_\infty$$

$$\geq \frac{\left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}\bar{\mathbf{x}}_{\mathbf{S}^*\backslash\mathbf{S}}\right\|_2}{\sqrt{|\mathbf{S}^*\backslash\mathbf{S}|}} - \left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \boldsymbol{\eta}\right\|_\infty$$

$$\geq \frac{\rho^-(\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}) \left\|\bar{\mathbf{x}}_{\mathbf{S}^*\backslash\mathbf{S}}\right\|_2}{\sqrt{|\mathbf{S}^*\backslash\mathbf{S}|}} - \left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \boldsymbol{\eta}\right\|_\infty$$

$$\geq \rho^-(\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}})|\bar{x}_{\min}| - \left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \boldsymbol{\eta}\right\|_\infty \tag{B.11}$$

We can analyze the first term in (B.11) as follows. Consider the matrix,

$$\mathbf{A}_{\mathbf{S}\cup\mathbf{S}^*}^T \mathbf{A}_{\mathbf{S}\cup\mathbf{S}^*} = \begin{bmatrix}\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}} & \mathbf{A}_{\mathbf{S}}\end{bmatrix}^T \begin{bmatrix}\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}} & \mathbf{A}_{\mathbf{S}}\end{bmatrix} = \begin{bmatrix}\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}} & \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T\mathbf{A}_{\mathbf{S}} \\ \mathbf{A}_{\mathbf{S}}^T\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}} & \mathbf{A}_{\mathbf{S}}^T\mathbf{A}_{\mathbf{S},}\end{bmatrix}$$

and its Schur complement with respect to its block $\mathbf{A}_{\mathbf{S}}^T\mathbf{A}_{\mathbf{S}}$:

$$\left[\mathbf{A}_{\mathbf{S}\cup\mathbf{S}^*}^T \mathbf{A}_{\mathbf{S}\cup\mathbf{S}^*}/\mathbf{A}_{\mathbf{S}}^T\mathbf{A}_{\mathbf{S}}\right] = \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}} - \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T\mathbf{A}_{\mathbf{S}}(\mathbf{A}_{\mathbf{S}}^T\mathbf{A}_{\mathbf{S}})^{-1}\mathbf{A}_{\mathbf{S}}^T\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}$$

$$= \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T(\mathbf{I} - \mathbf{A}_{\mathbf{S}}(\mathbf{A}_{\mathbf{S}}^T\mathbf{A}_{\mathbf{S}})^{-1}\mathbf{A}_{\mathbf{S}}^T)\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}$$

$$= \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T\mathbf{A}_{\mathbf{S}}^\perp \mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}} \tag{B.12}$$

15

Therefore from the Schur eigenvalue inequality, we have

$$\rho^- \left(\left[\mathbf{A}_{\mathbf{S} \cup \mathbf{S}^*}^T \mathbf{A}_{\mathbf{S} \cup \mathbf{S}^*} / \mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}} \right] \right) \geq \rho^- (\mathbf{A}_{\mathbf{S} \cup \mathbf{S}^*}^T \mathbf{A}_{\mathbf{S} \cup \mathbf{S}^*}) \geq \rho_{s+s^*}^- \tag{B.13}$$

We can upper bound the second term in (B.11) by using Cauchy-Schwarz and directly assuming an infinity norm upper bound as

$$
\begin{aligned}
\left\| \mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp \boldsymbol{\eta} \right\|_\infty &= \left\| \mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T (\mathbf{I} - \mathbf{A}_{\mathbf{S}} (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \mathbf{A}_{\mathbf{S}}^T) \boldsymbol{\eta} \right\|_\infty \\
&\leq \left\| \mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T \boldsymbol{\eta} \right\|_\infty + \left\| \mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T \mathbf{A}_{\mathbf{S}} (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \mathbf{A}_{\mathbf{S}}^T \boldsymbol{\eta} \right\|_\infty \\
&\leq \left\| \mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T \boldsymbol{\eta} \right\|_\infty + \left\| \mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T \mathbf{A}_{\mathbf{S}} (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \right\|_\infty \left\| \mathbf{A}_{\mathbf{S}}^T \boldsymbol{\eta} \right\|_\infty \\
&\leq (1 + \gamma) \left\| \mathbf{A}_{\mathbf{S} \cup \mathbf{S}^*}^T \boldsymbol{\eta} \right\|_\infty \quad \left(\text{Assuming} \left\| \mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T \mathbf{A}_{\mathbf{S}} (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \right\|_\infty \leq \gamma \right) \\
&\leq \sqrt{2} (1 + \gamma) \sigma \sqrt{\rho_1^+ \log \frac{2(s + s^*)}{2\delta}} \quad \text{(with probability at least } 1 - 2\delta) \\
&\leq \sqrt{2} (1 + \gamma) \sigma \sqrt{\rho_1^+ \log \frac{d}{\delta}} \quad \text{(using } d \geq s + s^*) \tag{B.14}
\end{aligned}
$$

Using (B.13) and (B.14) in (B.11), we have,

$$\left\| \mathbf{A}_{\mathbf{S}^* \backslash \mathbf{S}}^T \mathbf{A}_{\mathbf{S}}^\perp (\mathbf{A} \bar{\mathbf{x}} + \boldsymbol{\eta}) \right\|_\infty \geq \rho_{s+s^*}^- |\bar{x}_{\min}| - \sqrt{2}(1 + \gamma) \sigma \sqrt{\rho_1^+ \log \frac{d}{\delta}} \tag{B.15}$$

Since $|\bar{x}_{\min}| \geq (1 + \sqrt{2}(1 + \gamma)) \frac{\sigma}{\rho_{s+s^*}^-} \sqrt{\rho_1^+ \log d/\delta}$, from Equation (B.10) and (B.15) we get

$$
\begin{aligned}
|\mathbf{A}_j^T \mathbf{r}_s| &= |\mathbf{A}_j^T \mathbf{A}_{\mathbf{S}}^\perp (\mathbf{A} \bar{\mathbf{x}} + \boldsymbol{\eta})| \\
&\geq \rho_{s+s^*}^- |\bar{x}_{\min}| - \sqrt{2}(1 + \gamma) \sigma \sqrt{\rho_1^+ \log \frac{d}{\delta}} \\
&\geq \sigma \sqrt{\rho_1^+ \log \frac{d}{\delta}} \\
\implies \sigma^2 \log \frac{d}{\delta} &\leq \frac{1}{\rho_1^+} \left(\mathbf{A}_j^T \mathbf{r}_s \right)^2 \tag{B.16}
\end{aligned}
$$

Now, we upper bound the function value at iteration $s + 1$:

$$
\begin{aligned}
\|\mathbf{A} \mathbf{x}_{s+1} - \mathbf{y}\|_2^2 &\leq \min_{x_j} \|\mathbf{A}_j x_j - \mathbf{r}_s\|_2^2 \\
&= \left\| \mathbf{A}_j (\mathbf{A}_j^T \mathbf{A}_j)^{-1} \mathbf{A}_j^T \mathbf{r}_s - \mathbf{r}_s \right\|_2^2 \\
&= \left\| \frac{1}{\|\mathbf{A}_j\|_2^2} \mathbf{A}_j \mathbf{A}_j^T \mathbf{r}_s - \mathbf{r}_s \right\|_2^2 \\
&= \left\| \frac{1}{\|\mathbf{A}_j\|_2^2} \mathbf{A}_j \mathbf{A}_j^T \mathbf{r}_s \right\|_2^2 + \|\mathbf{r}_s\|_2^2 - 2 \left\langle \mathbf{r}_s, \frac{1}{\|\mathbf{A}_j\|_2^2} \mathbf{A}_j \mathbf{A}_j^T \mathbf{r}_s \right\rangle \\
&= \|\mathbf{A} \mathbf{x}_s - \mathbf{y}\|_2^2 - \frac{1}{\|\mathbf{A}_j\|_2^2} \left(\mathbf{A}_j^T \mathbf{r}_s \right)^2 \\
&\leq \|\mathbf{A} \mathbf{x}_s - \mathbf{y}\|_2^2 - \frac{1}{\rho_1^+} \left(\mathbf{A}_j^T \mathbf{r}_s \right)^2 \\
&\leq \|\mathbf{A} \mathbf{x}_s - \mathbf{y}\|_2^2 - \sigma^2 \log \frac{d}{\delta} \quad \text{(Using (B.16))} \tag{B.17}
\end{aligned}
$$

The bound (B.17) and the argument holds for all $t \geq s$ as long as there is at least one more element that OMP has not yet recovered. We wish to show an upper bound for the total number of steps that

16

OMP can run without recovering all the elements in the support. Suppose OMP runs for another l iterations without recovering all elements in the support. Then, telescoping (B.17), we get

$$\|\mathbf{A}\mathbf{x}_t - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}_{t+1} - \mathbf{y}\|_2^2 \geq \sigma^2 \log \frac{d}{\delta} \qquad \forall s \leq t \leq s + l - 1$$

$$\implies \|\mathbf{A}\mathbf{x}_s - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}_{s+l} - \mathbf{y}\|_2^2 \geq \sigma^2 l \log \frac{d}{\delta}.$$

$\square$

Proof of Theorem 3.1. From Theorem 3.2, for $\bar{s} \geq 8\widetilde{\kappa}_{\bar{s}+s^*} s^* \cdot \log \left(\frac{5\rho_{\bar{s}+s^*}^+}{\rho_{\bar{s}+s^*}^-} \right)$, we have with probability at least $1 - 2\delta$,

$$\|\mathbf{A}\mathbf{x}_{\bar{s}} - \mathbf{y}\|_2^2 \leq \|\boldsymbol{\eta}\|_2^2 + 40\sigma^2 \frac{\bar{s}\rho_1^+}{\rho_{\bar{s}+s^*}^+} \log \frac{d}{\delta} \tag{B.18}$$

Using Lemma B.4, after $\bar{s} + l$ OMP iterations, if there is always at least one element that is not recovered, the function value will decrease significantly, i.e.,

$$\|\mathbf{A}\mathbf{x}_{\bar{s}} - \mathbf{y}\|_2^2 - \|\mathbf{A}\mathbf{x}_{\bar{s}+l} - \mathbf{y}\|_2^2 \geq \sigma^2 l \log \frac{d}{\delta} \tag{B.19}$$

We will now lower bound $\|\mathbf{A}\mathbf{x}_{\bar{s}+l} - \mathbf{y}\|_2^2$ showing that l cannot be too large. This is basically asking how best can a sub-Gaussian noise vector be fitted in the ℓ_2 norm. That is equivalent to the problem

$$\min_{\|\mathbf{x}\|_0 = k} \|\mathbf{A}\mathbf{x} - \boldsymbol{\eta}\|_2^2 = \left\| \mathbf{A}_k^\perp \boldsymbol{\eta} \right\|_2^2$$

$$= \|\boldsymbol{\eta}\|_2^2 - \boldsymbol{\eta}^T \mathbf{A}_k (\mathbf{A}_k^T \mathbf{A}_k)^{-1} \mathbf{A}_k^T \boldsymbol{\eta} \tag{B.20}$$

$\mathbf{A}_k (\mathbf{A}_k^T \mathbf{A}_k)^{-1} \mathbf{A}_k^T$ is nothing but a projection matrix which can be written as $\mathbf{U}_k \mathbf{U}_k^T$ for some matrix $\mathbf{U}_k$ of size $n \times k$ with unit norm orthogonal columns. Consider any fixed matrix $\mathbf{U}_k'$ of size $n \times k$. Since $k \geq \log \frac{1}{\delta}$ using A.1 we have

$$\boldsymbol{\eta}^T \mathbf{U}_k' \mathbf{U}_k'^T \boldsymbol{\eta} - \sigma^2 k \leq 4\sigma^2 \sqrt{k} \sqrt{\log \frac{1}{\delta}} \tag{B.21}$$

with probability at least $1 - \delta$. Setting $\delta' = \frac{\delta}{d^k}$,

$$\boldsymbol{\eta}^T \mathbf{U}_k' \mathbf{U}_k'^T \boldsymbol{\eta} - \sigma^2 k \leq 4\sigma^2 \sqrt{k} \sqrt{\log \frac{d^k}{\delta}}$$

with probability at least $1 - \delta/d^k$. Taking a union bound over all possible projection matrices with respect to a k sized subset of the columns of $\mathbf{A}$ (of which there are about d^k), we get,

$$\max_{|\mathbf{U}_k'| = k} \left\{ \boldsymbol{\eta}^T \mathbf{U}_k' \mathbf{U}_k'^T \boldsymbol{\eta} \right\} - \sigma^2 k \leq 4\sigma^2 k \sqrt{\log \frac{d}{\delta}} \qquad \text{(with probability at least } 1 - \delta\text{)}$$

$$\implies \min_{\|\mathbf{x}\|_0 = k} \|\mathbf{A}\mathbf{x} - \boldsymbol{\eta}\|_2^2 \geq \|\boldsymbol{\eta}\|_2^2 - \sigma^2 k - 4\sigma^2 k \sqrt{\log \frac{d}{\delta}} \qquad \text{(with probability at least } 1 - \delta\text{)} \tag{B.22}$$

$$\therefore \|\mathbf{A}\mathbf{x}_{\bar{s}+l} - \mathbf{y}\|_2^2 = \|\mathbf{A}\mathbf{x}_{\bar{s}+l} - \mathbf{A}\mathbf{x}^* - \boldsymbol{\eta}\|_2^2$$

$$\geq \min_{\|\mathbf{x}\|_0 = \bar{s}+l+s^*} \|\mathbf{A}\mathbf{x} - \boldsymbol{\eta}\|_2^2$$

$$\geq \|\boldsymbol{\eta}\|_2^2 - \sigma^2 (\bar{s} + l + s^*) - 4\sigma^2 (\bar{s} + l + s^*) \sqrt{\log \frac{d}{\delta}} \tag{B.23}$$

17

Using the upper bound of $\|\mathbf{Ax}_{\bar{s}} - \mathbf{y}\|_2^2$ from (B.18) and lower bound of $\|\mathbf{Ax}_{\bar{s}+l} - \mathbf{y}\|_2^2$ from (B.23) in (B.19) we finally get

$$l\sigma^2 \log \frac{d}{\delta} \leq 40 \frac{\sigma^2 \rho_1^+}{\rho_{\bar{s}+s^*}^+} s \log \frac{d}{\delta} + \sigma^2(\bar{s}+l+s^*) + 4\sigma^2(\bar{s}+l+s^*)\sqrt{\log \frac{d}{\delta}} \quad \text{(w.p. at least } 1 - 5\delta)$$

$$\implies l \leq 40\bar{s}\frac{\rho_1^+}{\rho_{\bar{s}+s^*}^+} + (\bar{s}+l+s^*)\left[\frac{1}{\log d/\delta} + \frac{4}{\sqrt{\log d/\delta}}\right]$$

$$\implies l \leq \frac{40\bar{s}\frac{\rho_1^+}{\rho_{\bar{s}+s^*}^+}}{1-u} + (\bar{s}+s^*)\frac{u}{1-u} \tag{B.24}$$

where $u := \frac{1}{\log d/\delta} + \frac{4}{\sqrt{\log d/\delta}}$. Since $\log(d/\delta) > 68$, $\frac{1}{1-u} < 2$ and $\frac{u}{1-u} < 1$, and we have

$$l \leq 80\bar{s}\frac{\rho_1^+}{\rho_{\bar{s}+s^*}^+} + \bar{s} + s^* \leq 82\bar{s} \tag{B.25}$$

Thus OMP can run only for another l steps, after which we are ensured recovery. Along with the initial $\bar{s}$ steps, this means that after a total of $s = l + \bar{s} \leq 83\bar{s} = 664\widetilde{\kappa}_{s+s^*}s^* \cdot \log\left(\frac{5\rho_{s+s^*}^+}{\rho_{s+s^*}^-}\right)$ steps, we are ensured complete recovery. This completes the support recovery part of the proof.

We will now show that given full support recovery, i.e., $\mathbf{S}^* \subseteq \mathbf{S}$, $\|\hat{\mathbf{x}}_s - \bar{\mathbf{x}}\|_\infty$ is bounded.

$$\|\hat{\mathbf{x}}_s - \bar{\mathbf{x}}\|_\infty = \left\|\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)^{-1}\mathbf{A}_{\mathbf{S}}^T(\mathbf{A}_{\mathbf{S}\cup\mathbf{S}^*}\bar{\mathbf{x}} + \boldsymbol{\eta}) - \bar{\mathbf{x}}\right\|_\infty$$

$$= \left\|\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)^{-1}\mathbf{A}_{\mathbf{S}}^T(\mathbf{A}_{\mathbf{S}}\bar{\mathbf{x}} + \boldsymbol{\eta}) - \bar{\mathbf{x}}\right\|_\infty$$

$$= \left\|\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)^{-1}\mathbf{A}_{\mathbf{S}}^T\boldsymbol{\eta}\right\|_\infty \tag{B.26}$$

We will separately analyze the first term in (B.26). Define $\mathbf{X} := \mathbf{A}_{\mathbf{S}}\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)^{-1}$:

$$\left\|\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)^{-1}\mathbf{A}_{\mathbf{S}}^T\boldsymbol{\eta}\right\|_\infty = \|\mathbf{X}^T\boldsymbol{\eta}\|_\infty$$

We will now provide a bound on $\|\mathbf{X}^T\boldsymbol{\eta}\|_\infty$:

$$P\left\{\|\mathbf{X}^T\boldsymbol{\eta}\|_\infty > t\right\} \leq \sum_{i=1}^s P\left\{|\mathbf{X}_i^T\boldsymbol{\eta}| > t\right\}$$

$$\leq \sum_{i=1}^s 2\exp\left\{-\frac{t^2}{2\sigma^2 \|\mathbf{X}_i\|_2^2}\right\}$$

$$\leq 2s\exp\left\{-\frac{t^2}{2\sigma^2 \max_{i\leq s} \|\mathbf{X}_i\|_2^2}\right\} \tag{B.27}$$

Therefore, if $t = \sigma \max_{i\leq s} \|\mathbf{X}_i\|_2 \sqrt{2\log(s/\delta)}$, then

$$P\left\{\|\mathbf{X}^T\boldsymbol{\eta}\|_\infty \leq \sigma \max_{i\leq s} \|\mathbf{X}_i\|_2 \sqrt{2\log \frac{s}{\delta}}\right\} \geq 1 - 2\delta \tag{B.28}$$

We are now left to bound $\max_{i\leq s} \|\mathbf{X}_i\|_2$ to complete the proof. Note that $\max_{i\leq s} \|\mathbf{X}_i\|_2 \leq \left\|\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)^{-1}\mathbf{A}_{\mathbf{S}}^T\right\|_2$. Let $\mathbf{w}$ be any non-zero vector in the span of $\mathbf{A}_{\mathbf{S}}$ and define $\mathbf{v} \stackrel{\text{def}}{=}$

18

$\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)^{-1} \mathbf{A}_{\mathbf{S}}^T \mathbf{w}$. Note that

$$\mathbf{A}_{\mathbf{S}}\mathbf{v} = \mathbf{A}_{\mathbf{S}} \left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)^{-1} \mathbf{A}_{\mathbf{S}}^T \mathbf{w}$$

$$= \mathbf{w}$$

$$\implies \|\mathbf{w}\|_2 \geq \sqrt{\rho_s^-\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)} \|\mathbf{v}\|_2$$

$$\text{or, } \frac{\|\mathbf{v}\|_2}{\|\mathbf{w}\|_2} \leq \frac{1}{\sqrt{\rho_s^-\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)}}$$

$$\implies \max_{i \leq s} \|\mathbf{X}_i\|_2 \leq \frac{1}{\sqrt{\rho_s^-\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)}} \tag{B.29}$$

Using (B.29) in (B.28) we get

$$P\left\{\left\|\mathbf{X}^T\boldsymbol{\eta}\right\|_\infty \leq \sigma\sqrt{\frac{2}{\rho_s^-\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)}\log\frac{s}{\delta}}\right\} \geq 1 - 2\delta \tag{B.30}$$

Therefore, we get

$$\|\hat{\mathbf{x}}_s - \bar{\mathbf{x}}\|_\infty \leq \sigma\sqrt{\frac{2}{\rho_s^-\left(\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}}\right)}\log\frac{s}{\delta}} \tag{B.31}$$

$\square$

B.3 Proof of Theorem 3.3

Proof of Theorem 3.3. Let $\mathbf{x} = \hat{\mathbf{x}}_s^{\text{OMP}}$, and $\mathbf{S} = \text{supp}(\mathbf{x})$.

$$\|\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}\|_2^2 = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \|\mathbf{y} - \mathbf{A}\bar{\mathbf{x}}\|_2^2 - 2\langle\mathbf{A}\mathbf{x} - \mathbf{y}, \mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\rangle \tag{B.32}$$

$$= \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \|\boldsymbol{\eta}\|_2^2 + 2\langle\mathbf{A}\mathbf{x} - \mathbf{y}, \boldsymbol{\eta}\rangle$$

$$= \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \|\boldsymbol{\eta}\|_2^2 + 2\langle\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}} + \mathbf{A}\bar{\mathbf{x}} - \mathbf{y}, \boldsymbol{\eta}\rangle$$

$$= \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 - \|\boldsymbol{\eta}\|_2^2 + 2\langle\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}, \boldsymbol{\eta}\rangle$$

$$\leq \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 - \|\boldsymbol{\eta}\|_2^2 + 2\|\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}\|_2 \|P(\mathbf{A}_{\mathbf{S}\cup\mathbf{S}^*})\boldsymbol{\eta}\|_2$$

$$\leq \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 - \|\boldsymbol{\eta}\|_2^2 + 2\|\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}\|_2 \,\sigma\sqrt{s + s^*}\sqrt{\log\frac{d}{\delta}}$$

(w.p. at least $1 - 2\delta$)

$$\leq \|\mathbf{A}\bar{\mathbf{x}} - \mathbf{y}\|_2^2 + 40\frac{\rho_1^+}{\rho_{s+s^*}^+}\sigma^2 s\log\frac{d}{\delta} - \|\boldsymbol{\eta}\|_2^2 + 4\|\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}\|_2 \,\sigma\sqrt{s}\sqrt{\log\frac{d}{\delta}}$$

(w.p. at least $1 - 2\delta$, using (3.10))

$$\leq 40\sigma^2 s\log\frac{d}{\delta} + 4\|\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}\|_2 \,\sigma\sqrt{s}\sqrt{\log\frac{d}{\delta}}$$

Let us define $G := \|\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}\|_2$ and $\alpha := \sigma\sqrt{s\log\frac{d}{\delta}}$. Then, we have

$$G^2 - 4G\alpha \leq 40\alpha^2$$

$$\implies (G - 2\alpha)^2 \leq (4 + 40)\alpha^2$$

$$\implies G \leq (2 + \sqrt{44})\alpha$$

Replacing $s \leq C_1\widetilde{\kappa}_{s+s^*} s^* \log 5\rho_{s+s^*}^+/\rho_{s+s^*}^-$, we get for some constant $C_2 \leq (2 + \sqrt{44})\,C_1 \leq 9C_1$,

$$\frac{\|\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}\|_2^2}{n} \leq C_2\frac{\sigma^2 s}{n}\log\frac{d}{\delta}$$

$$\frac{\|\mathbf{A}\mathbf{x} - \mathbf{A}\bar{\mathbf{x}}\|_2^2}{n} \leq C_2\frac{\sigma^2 s^*}{n}\widetilde{\kappa}_{s+s^*}\log\kappa_{s+s^*}\log\frac{d}{\delta}$$

$\square$

Proof of Corollary 3.3.1. From [1, Lemma 6] it immediately implies that RSS and RSC at sparsity level k hold with probability at least $1 - e^{-C_0 n}$ with

$$\rho_k^- = \frac{n}{2}\sigma_{\min}(\mathbf{\Sigma}) - C_1 k \log d$$
$$\text{and, } \rho_k^+ = 2n\sigma_{\max}(\mathbf{\Sigma}) + C_1 k \log d$$

where C_0, C_1 are universal constants. For $k = s + s^*$, and since $n > 4C_1 \frac{k \log d}{\sigma_{\min}(\mathbf{\Sigma})}$ we have

$$\rho_{s+s^*}^- \geq \frac{n\sigma_{\min}(\mathbf{\Sigma})}{4}$$
$$\text{and, } \rho_{s+s^*}^+(\mathbf{\Sigma}) \leq 2.25 n\sigma_{\max}(\mathbf{\Sigma})$$

Therefore it is enough to choose $s \geq 72\kappa(\mathbf{\Sigma}) \log(45\kappa(\mathbf{\Sigma})) s^*$ to apply Theorem 3.3. $\qquad \square$

Proof of Corollary 3.3.2. Using results from the proof of 3.3.1, we can specialize for the case when $\mathbf{\Sigma} = \mathbf{I}$. To invoke Theorem 3.1 we are required to see how and when does γ concentrate as n increases for a given s. From the definition of γ, we want to upper bound

$$
\begin{aligned}
\max_{\mathbf{S}} \left\| \mathbf{A}_{\mathbf{S}^* \setminus \mathbf{S}}^T \mathbf{A}_{\mathbf{S}} (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \right\|_{\infty} &= \max_{\mathbf{S}} \max_{i \in \mathbf{S}^* \setminus \mathbf{S}} \left\| \mathbf{A}_i^T \mathbf{A}_{\mathbf{S}} (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \right\|_1 \\
&\leq \sqrt{s} \max_{\mathbf{S}} \max_{i \in \mathbf{S}^* \setminus \mathbf{S}} \left\| \mathbf{A}_i^T \mathbf{A}_{\mathbf{S}} (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \right\|_2 \\
&\leq \sqrt{s} \max_{\mathbf{S}} \max_{i \in \mathbf{S}^* \setminus \mathbf{S}} \left\| \mathbf{A}_{\mathbf{S}}^T \mathbf{A}_i \right\|_2 \left\| (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \right\|_2 \\
&\leq s \max_{\mathbf{S}} \max_{i \in \mathbf{S}^* \setminus \mathbf{S}} \left\| \mathbf{A}_{\mathbf{S}}^T \mathbf{A}_i \right\|_{\infty} \left\| (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \right\|_2 \\
&\leq s \max_{\mathbf{S}} \max_{i \in \mathbf{S}^* \setminus \mathbf{S}} \left\| \mathbf{A}_{(\mathbf{S}^* \setminus \mathbf{S})^c}^T \mathbf{A}_i \right\|_{\infty} \left\| (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \right\|_2 \\
&\leq s \max_{\mathbf{S}} \max_{i \in \mathbf{S}^* \setminus \mathbf{S}} \left\| \mathbf{A}_{(\mathbf{S}^* \setminus \mathbf{S})^c}^T \mathbf{A}_i \right\|_{\infty} \cdot \max_{\mathbf{S}} \left\| (\mathbf{A}_{\mathbf{S}}^T \mathbf{A}_{\mathbf{S}})^{-1} \right\|_2 \\
&\leq s \max_{\mathbf{S}} \max_{i \in \mathbf{S}^* \setminus \mathbf{S}} \left\| \mathbf{A}_{(\mathbf{S}^* \setminus \mathbf{S})^c}^T \mathbf{A}_i \right\|_{\infty} \frac{1}{\rho_s^-} \\
&\leq s \max_{\mathbf{S}} \max_{i \in \mathbf{S}^* \setminus \mathbf{S}} \left\| \mathbf{A}_{(\mathbf{S}^* \setminus \mathbf{S})^c}^T \mathbf{A}_i \right\|_{\infty} \frac{1}{n/4} \\
&\qquad\qquad \text{(w.p. at least } 1 - e^{-C_0 n} \text{ if } n \geq 4Cs \log d) \\
&\leq \frac{4s}{n} \max_{\mathbf{S}} \max_{i \in \mathbf{S}^* \setminus \mathbf{S}} \left\| \mathbf{A}_{(\mathbf{S}^* \setminus \mathbf{S})^c}^T \mathbf{A}_i \right\|_{\infty} \qquad\qquad\text{(B.33)}
\end{aligned}
$$

To bound the quantity in (B.33) we can use the vanilla Chernoff bound. Note that it is sufficient bounding the maximum of $\binom{d}{2}$ inner products of pair of d-length standard Gaussian vectors. For any two independent standard Gaussian vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, we will have

$$
\begin{aligned}
P\{\langle \mathbf{u}, \mathbf{v} \rangle \geq \epsilon\} = P\{\exp\{t\langle \mathbf{u}, \mathbf{v}\rangle\} \geq e^{t\epsilon}\} \\
\leq \inf_{t>0} e^{-t\epsilon} \mathbb{E}\left[e^{tZ}\right]^n \qquad (Z \text{ is a product of two standard normal random variables}) \\
\leq \inf_{t>0} \exp\left\{t^2\frac{n}{2} - t\epsilon\right\} \\
= \exp\left\{-\frac{\epsilon^2}{2n}\right\} \qquad\qquad\qquad\qquad\qquad\qquad \text{(B.34)}
\end{aligned}
$$

We need to find ϵ, so for a moment let us bound the RHS in (B.34) with some appropriate probability δ' which we will define later. This gives us

$$\exp\left\{-\frac{\epsilon^2}{2n}\right\} \leq \delta'$$

$$\implies \epsilon \geq \sqrt{2n\log\frac{1}{\delta'}}$$

$$P\left\{\langle \mathbf{u}, \mathbf{v}\rangle \geq \sqrt{2n\log\frac{1}{\delta'}}\right\} \leq \delta'$$

$$\text{or, } P\left\{|\langle \mathbf{u}, \mathbf{v}\rangle| \geq \sqrt{2n\log\frac{1}{\delta'}}\right\} \leq 2\delta'$$

Because we have $\binom{d}{2}$ such unique inner products, we set $\delta' = \delta/\binom{d}{2}$ and use a union bound to get

$$P\left\{\max_{\mathbf{S}}\max_{i\in\mathbf{S}^*\backslash\mathbf{S}}\left\|\mathbf{A}_{(\mathbf{S}^*\backslash\mathbf{S})^c}^T\mathbf{A}_i\right\|_\infty \leq 2\sqrt{n\log\frac{d}{\delta}}\right\} \geq P\left\{\max_{\mathbf{S}}\max_{i\in\mathbf{S}^*\backslash\mathbf{S}}\left\|\mathbf{A}_{(\mathbf{S}^*\backslash\mathbf{S})^c}^T\mathbf{A}_i\right\|_\infty \leq \sqrt{2n\log\left(\binom{d}{2}/\delta\right)}\right\}$$

$$\geq 1 - 2\delta \tag{B.35}$$

Using (B.35) in (B.33) we finally get

$$\max_{\mathbf{S}}\left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T\mathbf{A}_{\mathbf{S}}(\mathbf{A}_{\mathbf{S}}^T\mathbf{A}_{\mathbf{S}})^{-1}\right\|_\infty \leq 8s\sqrt{\frac{\log\left(d/\delta\right)}{n}} \qquad \text{(with probabily at least } 1 - e^{-C_0 n} - 2\delta) \tag{B.36}$$

And if further $n \geq 64s^2\log\frac{d}{\delta}$, we have $\gamma \leq 1$ with probability at least $1 - e^{-C_0 n} - 2\delta$. Ones we have γ bounded by some constant, one can easily check that the requirement on $|\bar{x}_{\min}|$ becomes

$$|\bar{x}_{\min}| \geq 23\sigma\sqrt{\log\frac{d}{\delta}} \tag{B.37}$$

Given this condition, using Theorem 3.1 we have support recovery as well as an infinity norm bound on the parameter space

$$\|\hat{\mathbf{x}}_s - \bar{\mathbf{x}}\|_\infty \leq 2\sigma\sqrt{\frac{2\log\left(s/\delta\right)}{n}} \tag{B.38}$$

which completes the proof. $\qquad\square$

C   Proofs of results in Section 4

Proof of Lemma 4.1. Consider $\mathbf{v}$ to be a sparse vector such that $\|\mathbf{v}\|_0 = s$ and $\|\mathbf{v}\|_2 = 1$. Let $\mathbf{S} := \text{supp}(\mathbf{v})$. For the rest of the proof, define $\mathbf{A} := \mathbf{M}^{(\epsilon)}$. This induces a sub-matrix $\mathbf{A}_{\mathbf{S}}$ which is the matrix of columns of $\mathbf{A}$ corresponding to the support of $\mathbf{v}$. Let us further split $\mathbf{S} = (\mathbf{S}\cap\mathbf{S}^*)\cup(\mathbf{S}\backslash\mathbf{S}^*)$ where $\mathbf{S}^* = [s^*]$. Consider a vector $\bar{\mathbf{x}}$ such that $\bar{x}_i = 1/s^*$ when $i \in [s^*]$ and 0 otherwise. Then, $\mathbf{A}_{\mathbf{S}\backslash\mathbf{S}^*} = \sqrt{1-\epsilon}\,\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}}\mathbf{1}_{|\mathbf{S}\backslash\mathbf{S}^*|}^T + \sqrt{\epsilon}\,\mathbf{G}$, where $\mathbf{G}$ is the $(n\times|\mathbf{S}\backslash\mathbf{S}^*|)$ matrix containing column vectors $\mathbf{g}_i$ for $i \in \mathbf{S}\backslash\mathbf{S}^*$.

$$\|\mathbf{A_S v}\|_2^2 = \left\|\sqrt{1-\epsilon}\mathbf{A_{S^*}}\bar{\mathbf{x}}\mathbf{1}^T\mathbf{v_{S\setminus S^*}} + \sqrt{\epsilon}\mathbf{G v_{S\setminus S^*}} + \mathbf{A_{S\cap S^*}}\mathbf{v_{S\cap S^*}}\right\|_2^2$$

$$= \left\|\sum_{i\in\mathbf{S^*}}\sum_{j\in\mathbf{S\setminus S^*}}(v_j\sqrt{1-\epsilon}\,\bar{\mathbf{x}}_i)\mathbf{A}_i + \sqrt{\epsilon}\mathbf{G v_{S\setminus S^*}} + \mathbf{A_{S\cap S^*}}\mathbf{v_{S\cap S^*}}\right\|_2^2$$

$$= \left\|\sum_{i\in\mathbf{S^*\setminus S}}\sum_{j\in\mathbf{S\setminus S^*}}(v_j\sqrt{1-\epsilon}\,\bar{\mathbf{x}}_i)\mathbf{A}_i + \sqrt{\epsilon}\mathbf{G v_{S\setminus S^*}} + \sum_{i\in\mathbf{S\cap S^*}}\left[\left(\sum_{j\in\mathbf{S\setminus S^*}}v_j\sqrt{1-\epsilon}\,\bar{\mathbf{x}}_i\right)+v_i\right]\mathbf{A}_i\right\|_2^2$$

$$= n(1-\epsilon)\sum_{i\in\mathbf{S^*\setminus S}}\left(\sum_{j\in\mathbf{S\setminus S^*}}v_j\bar{\mathbf{x}}_i\right)^2 + \epsilon\|\mathbf{G v_{S\setminus S^*}}\|_2^2 + n\sum_{i\in\mathbf{S\cap S^*}}\left[\left(\sum_{j\in\mathbf{S\setminus S^*}}v_j\sqrt{1-\epsilon}\,\bar{\mathbf{x}}_i\right)+v_i\right]^2$$

$$\text{(C.1)}$$

Define $p := \sum_{j\in\mathbf{S\setminus S^*}}v_j, q := \sum_{j\in\mathbf{S\setminus S^*}}v_j^2, r := \sum_{i\in\mathbf{S\cap S^*}}v_i, t := \sum_{i\in\mathbf{S\cap S^*}}v_i^2 = 1-q.$
Since $\bar{\mathbf{x}}_i = 1/\sqrt{s^*}$,

$$\|\mathbf{A_S v}\|_2^2 = n(1-\epsilon)p^2\frac{|\mathbf{S^*\setminus S}|}{s^*} + n\epsilon q + n\sum_{i\in\mathbf{S^*\cap S}}\left(p\sqrt{\frac{1-\epsilon}{s^*}}+v_i\right)^2$$

$$= n(1-\epsilon)p^2 + n\epsilon q + 2npr\sqrt{\frac{1-\epsilon}{s^*}} + nt. \tag{C.2}$$

We need to lower bound this quantity to get an estimate for ρ_s^-. Then,

$$\|\mathbf{A_S v}\|_2^2 = n(1-\epsilon)p^2 + n\epsilon q + 2npr\sqrt{\frac{1-\epsilon}{s^*}} + nt$$

$$\geq n(1-\epsilon)p^2 + n\epsilon q - 2np\sqrt{t}\sqrt{1-\epsilon} + nt$$

$$= n(p\sqrt{1-\epsilon}-\sqrt{t})^2 + n\epsilon q.$$

Suppose $p\sqrt{1-\epsilon} < \sqrt{t/2}$, then the bound becomes $n(t+q)/4 = n/4$. Else, $\sqrt{1-\epsilon}\sqrt{sq} \geq p\sqrt{1-\epsilon} \geq \sqrt{t/2}$. Setting $t = 1-q$, we get:

$$2(1-\epsilon)sq \geq 1-q$$

$$\implies q \geq \frac{1}{1+2(1-\epsilon)s}$$

In any case, we get,

$$\rho_s^- = \min_{\mathbf{S}\subseteq[d],|\mathbf{S}|=s}\min_{\|\mathbf{v}\|_2^2=1}\|\mathbf{A_S v}\|_2^2 \geq \min\left\{\frac{n}{4}, \frac{n}{4(1+2(1-\epsilon)s)}\right\} = \frac{n}{4(1+2(1-\epsilon)s)} \tag{C.3}$$

Therefore $\widetilde{\kappa}_s \leq 4(1+2(1-\epsilon)s)$.

Now, we wish to lower bound $\left\|\mathbf{A}_{\mathbf{S^*\setminus S}}^T\mathbf{A_S}\left(\mathbf{A_S}^T\mathbf{A_S}\right)^{-1}\right\|_\infty$ even when some correct elements might get recovered. This is essentially the max ℓ_1 norm of the rows of the matrix $\mathbf{A}_{\mathbf{S^*\setminus S}}^T\mathbf{A_S}\left(\mathbf{A_S}^T\mathbf{A_S}\right)^{-1}$, i.e.,

$$\left\|\mathbf{A}_{\mathbf{S^*\setminus S}}^T\mathbf{A_S}\left(\mathbf{A_S}^T\mathbf{A_S}\right)^{-1}\right\|_\infty = \max_{i\in\mathbf{S^*\setminus S}}\left\|\left(\mathbf{A_S}^T\mathbf{A_S}\right)^{-1}\mathbf{A_S}^T\mathbf{A}_i\right\|_1 \tag{C.4}$$

$$\mathbf{A_S}^T\mathbf{A}_i = n\begin{bmatrix}\mathbf{0}_{|\mathbf{S\cap S^*}|}\\ \mathbf{1}_{|\mathbf{S\setminus S^*}|}\sqrt{\frac{1-\epsilon}{s^*}}\end{bmatrix} \tag{C.5}$$

22

From (C.5) we see that we only requires the sub-matrix of $\left(\mathbf{A}_\mathbf{S}^T\mathbf{A}_\mathbf{S}\right)^{-1}$ with columns corresponding to the index set $\mathbf{S}\backslash\mathbf{S}^*$.

Let $p := |\mathbf{S}\cap\mathbf{S}^*|$ and $q := \mathbf{S}\backslash\mathbf{S}^*$. From block matrix inversion we have that

$$
\left[\left(\mathbf{A}_\mathbf{S}^T\mathbf{A}_\mathbf{S}\right)^{-1}\right]_{\mathbf{S}\backslash\mathbf{S}^*} = \frac{1}{n}\begin{bmatrix} \mathbf{I}_p & \sqrt{\frac{1-\epsilon}{s^*}}\mathbf{1}_p\mathbf{1}_q^T \\ \sqrt{\frac{1-\epsilon}{s^*}}\mathbf{1}_q\mathbf{1}_p^T & (1-\epsilon)\mathbf{1}_q\mathbf{1}_q^T + \epsilon\mathbf{I}_q \end{bmatrix}_{\mathbf{S}\backslash\mathbf{S}^*}^{-1}
$$

$$
= \frac{1}{n}\begin{bmatrix} -\sqrt{\frac{1-\epsilon}{s^*}}\mathbf{1}_p\mathbf{1}_q^T\left[(1-\epsilon)\mathbf{1}_q\mathbf{1}_q + \epsilon\mathbf{I}_q - \frac{1-\epsilon}{s^*}p\mathbf{1}_q\mathbf{1}_q^T\right]^{-1} \\ \left[(1-\epsilon)\mathbf{1}_q\mathbf{1}_q^T + \epsilon\mathbf{I}_q - \frac{1-\epsilon}{s^*}p\mathbf{1}_q\mathbf{1}_q^T\right]^{-1} \end{bmatrix} \tag{C.6}
$$

Let us separately compute $\left[(1-\epsilon)\mathbf{1}_q\mathbf{1}_q^T + \epsilon\mathbf{I}_q - \frac{1-\epsilon}{s^*}p\mathbf{1}_q\mathbf{1}_q^T\right]^{-1}$ first

$$
\left[(1-\epsilon)\mathbf{1}_q\mathbf{1}_q^T + \epsilon\mathbf{I}_q - \frac{1-\epsilon}{s^*}p\mathbf{1}_q\mathbf{1}_q^T\right]^{-1} = \left[\epsilon\mathbf{I}_q + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)\mathbf{1}_q\mathbf{1}_q\right]^{-1}
$$

$$
= \frac{1}{\epsilon}\mathbf{I}_q - \frac{\frac{1}{\epsilon^2}(1-\epsilon)\left(1 - \frac{p}{s^*}\right)}{1 + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)\frac{q}{\epsilon}}\mathbf{1}_q\mathbf{1}_q^T
$$

$$
= \frac{1}{\epsilon}\left[\mathbf{I}_q - \frac{(1-\epsilon)\left(1 - \frac{p}{s^*}\right)}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\mathbf{1}_q\mathbf{1}_q^T\right] \tag{C.7}
$$

Using (C.7) we have

$$
-\sqrt{\frac{1-\epsilon}{s^*}}\mathbf{1}_p\mathbf{1}_q^T\left[(1-\epsilon)\mathbf{1}_q\mathbf{1}_q + \epsilon\mathbf{I}_q - \frac{1-\epsilon}{s^*}p\mathbf{1}_q\mathbf{1}_q^T\right]^{-1} \tag{C.8}
$$

$$
= -\sqrt{\frac{1-\epsilon}{s^*}}\mathbf{1}_p\mathbf{1}_q^T\frac{1}{\epsilon}\left[\mathbf{I}_q - \frac{(1-\epsilon)\left(1 - \frac{p}{s^*}\right)}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\mathbf{1}_q\mathbf{1}_q^T\right]
$$

$$
= -\sqrt{\frac{1-\epsilon}{s^*}}\mathbf{1}_p\frac{1}{\epsilon}\left[\mathbf{1}_q^T - \frac{(1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\mathbf{1}_q^T\right]
$$

$$
= -\sqrt{\frac{1-\epsilon}{s^*}}\left[\frac{1}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\right]\mathbf{1}_p\mathbf{1}_q^T \tag{C.9}
$$

Using (C.7) and (C.9) we have

$$
\left(\mathbf{A}_\mathbf{S}^T\mathbf{A}_\mathbf{S}\right)^{-1}\mathbf{A}_\mathbf{S}^T\mathbf{A}_i = \sqrt{\frac{1-\epsilon}{s^*}}\begin{bmatrix} -\sqrt{\frac{1-\epsilon}{s^*}}\left[\frac{1}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\right]\mathbf{1}_p\mathbf{1}_q^T \\ \frac{1}{\epsilon}\left[\mathbf{I}_q - \frac{(1-\epsilon)\left(1 - \frac{p}{s^*}\right)}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\mathbf{1}_q\mathbf{1}_q^T\right] \end{bmatrix}\mathbf{1}_q
$$

$$
= \sqrt{\frac{1-\epsilon}{s^*}}\begin{bmatrix} -\sqrt{\frac{1-\epsilon}{s^*}}\left[\frac{q}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\right]\mathbf{1}_p \\ \frac{1}{\epsilon}\left[\frac{\epsilon}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\right]\mathbf{1}_q \end{bmatrix}
$$

$$
= -\sqrt{\frac{1-\epsilon}{s^*}}\frac{1}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\begin{bmatrix} -q\sqrt{\frac{1-\epsilon}{s^*}}\mathbf{1}_p \\ \mathbf{1}_q \end{bmatrix}
$$

$$
\implies \left\|\left(\mathbf{A}_\mathbf{S}^T\mathbf{A}_\mathbf{S}\right)^{-1}\mathbf{A}_\mathbf{S}^T\mathbf{A}_i\right\|_1 = \sqrt{\frac{1-\epsilon}{s^*}}\frac{q}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\left[p\sqrt{\frac{1-\epsilon}{s^*}} + 1\right] \tag{C.10}
$$

Therefore $\left\|\mathbf{A}_{\mathbf{S}^*\backslash\mathbf{S}}^T\mathbf{A}_\mathbf{S}\left(\mathbf{A}_\mathbf{S}^T\mathbf{A}_\mathbf{S}\right)^{-1}\right\|_\infty = \sqrt{\frac{1-\epsilon}{s^*}}\frac{q}{\epsilon + (1-\epsilon)\left(1 - \frac{p}{s^*}\right)q}\left[p\sqrt{\frac{1-\epsilon}{s^*}} + 1\right].$ □

Proof of Theorem 4.2. We shall construct a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$ where $n \geq d$.

Having a closer look at Algorithm 1 for sparse linear regression , we can see that $\nabla Q(\mathbf{x}_k) = \frac{2}{n}\mathbf{A}^T(\mathbf{A}\mathbf{x}_k - \mathbf{y})$. The residual vector is now $\mathbf{r}_k := \mathbf{y} - \mathbf{A}\mathbf{x}_k$ and hence the selected index is j is

$\arg\max_{i \in \mathbf{S}_k^c}|\mathbf{A}_i^T\mathbf{r}_k|$. Therefore with proper construction of the matrix $\mathbf{A}$, we can force the new selected index to be in $(\mathbf{S}^*)^c$.

We first fix the vector to be estimated, $\bar{\mathbf{x}}$, as $\bar{x}_i = 1/\sqrt{s^*}\ \forall\ i \in [s^*]$ and 0 otherwise. Thus, $\mathbf{S}^* = [s^*]$ and we take $s^* \geq 2$. We define $\mathbf{A} := \mathbf{M}^{(\epsilon)}$ and choose a constant ϵ in the range $[1 - 3/2s^*, 1 - 1/s^*)$. We assume that OMP fails to recover any support by the $(k-1)^{th}$ iterate, i.e., $\mathbf{S}_{k-1} \cap \mathbf{S} = \phi$. At the kth OMP iteration, for $i \notin (\mathbf{S}_{k-1} \cup \mathbf{S}^*)$,

$$
\begin{aligned}
|\mathbf{A}_i^T\mathbf{r}_{k-1}| &= |\sqrt{1-\epsilon}\,\mathbf{y}^T\mathbf{r}_{k-1} + \sqrt{\epsilon}\mathbf{g}_i^T\mathbf{r}_{k-1}| \\
&= \sqrt{1-\epsilon}|\mathbf{y}^T\mathbf{r}_{k-1}| \\
&= \sqrt{1-\epsilon}\sum_{j \in \mathbf{S}^*}\bar{x}_j|\mathbf{r}_{k-1}^T\mathbf{A}_j|
\end{aligned}
\tag{C.11}
$$

Now let us consider the term $\mathbf{r}_{k-1}^T\mathbf{A}_j$ for all $j \in \mathbf{S}^*$.

$$
\begin{aligned}
\mathbf{r}_{k-1}^T\mathbf{A}_j &= \mathbf{y}^T(\mathbf{I} - \mathbf{A}_{\mathbf{S}_{k-1}}(\mathbf{A}_{\mathbf{S}_{k-1}}^T\mathbf{A}_{\mathbf{S}_{k-1}})^{-1}\mathbf{A}_{\mathbf{S}_{k-1}}^T)\mathbf{A}_j \\
&\overset{\xi_1}{=} \mathbf{y}^T\mathbf{A}_j - \left[\mathbf{y}^T\mathbf{A}_{\mathbf{S}_{k-1}}(\mathbf{A}_{\mathbf{S}_{k-1}}^T\mathbf{A}_{\mathbf{S}_{k-1}})^{-1}\mathbf{1}\right](\sqrt{1-\epsilon}\,\mathbf{y}^T\mathbf{A}_j) \\
&= (\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}})^T\mathbf{A}_j - \left[\mathbf{y}^T\mathbf{A}_{\mathbf{S}_{k-1}}(\mathbf{A}_{\mathbf{S}_{k-1}}^T\mathbf{A}_{\mathbf{S}_{k-1}})^{-1}\mathbf{1}\right](\sqrt{1-\epsilon}\,(\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}})^T\mathbf{A}_j) \\
&= n\bar{x}_j - \left[\mathbf{y}^T\mathbf{A}_{\mathbf{S}_{k-1}}(\mathbf{A}_{\mathbf{S}_{k-1}}^T\mathbf{A}_{\mathbf{S}_{k-1}})^{-1}\mathbf{1}\right](\sqrt{1-\epsilon}\,(n\bar{x}_j))
\end{aligned}
$$

Here, ξ_1 holds because $\mathbf{g}_l^T\mathbf{A}_j = 0, \forall\ l \in \mathbf{S}_{k-1}, j \in \mathbf{S}^*$. Now, because $\bar{x}_j = 1/\sqrt{s^*}$ for all index $j \in \mathbf{S}^*$, the above quantity is the same irrespective of what j is. Thus, we can write (C.11) as

$$
\begin{aligned}
|\mathbf{A}_i^T\mathbf{r}_{k-1}| &= \sqrt{1-\epsilon}\sum_{j \in \mathbf{S}^*}x_j^*|\mathbf{r}_{k-1}^T\mathbf{A}_j| \\
&= \sqrt{1-\epsilon}\sqrt{s^*}\max_{j \in \mathbf{S}^*}|\mathbf{r}_{k-1}^T\mathbf{A}_j|
\end{aligned}
\tag{C.12}
$$

Therefore for ϵ such that $\sqrt{1-\epsilon}\sqrt{s^*} > 1$, or when $\epsilon < 1 - \frac{1}{s^*}$, then $|\mathbf{A}_i^T\mathbf{r}_{k-1}| > \max_{j \in \mathbf{S}^*}|\mathbf{A}_j^T\mathbf{r}_{k-1}|$ which would imply that the algorithm in the kth iteration picks an incorrect index that is not contained in $\mathbf{S}^*$.

Using Lemma 4.1, setting $\epsilon = 1 - 3/2s^*$, we get

$$
\widetilde{\kappa}_s(\mathbf{A}) \leq \frac{16s}{s^*} \text{ and } \gamma \leq \sqrt{\frac{2}{3}}
\tag{C.13}
$$

$\square$

Proof of Theorem 4.3. When the model is noisy, we consider a similar matrix $\mathbf{A} := \mathbf{M}^{(\epsilon)}$ discussed in 4.1 where we set $\epsilon = 1 - 4/s^*$. Further, we are concerned about the asymptotic failure of OMP in the presence of a large number of samples. So we set $n \geq 4\sigma^2 s^2 \log d/\delta$. Suppose that OMP has not recovered any support until k iterations. We show that even in the next (ie, $(k+1)^{th}$) iteration, an incorrect support is picked. The criteria for OMP to not select an index in $\mathbf{S}^*$ is

$$
\max_{j \notin (\mathbf{S}^* \cup \mathbf{S}_k)}\left|\mathbf{A}_j^T\mathbf{r}_k\right| \geq \max_{i \in \mathbf{S}^*}\left|\mathbf{A}_j^T\mathbf{r}_k\right|
\tag{C.14}
$$

We will show that the above inequality holds with good probability. We denote $\mathbf{1}$ as the vector with all its elements as 1. Denote the projection matrix corresponding to a matrix $\mathbf{A}_{\mathbf{S}_k}$ as $P(\mathbf{A}_{\mathbf{S}_k}) := \mathbf{A}_{\mathbf{S}_k}(\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}_k})^{-1}\mathbf{A}_{\mathbf{S}_k}^T$. Note that $\mathbf{r}_k = (\mathbf{I} - P(\mathbf{A}_{\mathbf{S}_k}))(\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} + \boldsymbol{\eta})$.

$$
\begin{aligned}
\left|\mathbf{A}_j^T\mathbf{r}_k\right| &= \left|\mathbf{A}_j^T(\mathbf{I} - P(\mathbf{A}_{\mathbf{S}_k}))(\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} + \boldsymbol{\eta})\right| \\
&= \left|\mathbf{A}_j^T\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} - \mathbf{A}_j^TP(\mathbf{A}_{\mathbf{S}_k})\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} + \mathbf{A}_j^T\boldsymbol{\eta} - \mathbf{A}_j^TP(\mathbf{A}_{\mathbf{S}_k})\boldsymbol{\eta}\right|
\end{aligned}
\tag{C.15}
$$

We can analyze these terms separately as below

$$\mathbf{A}_j^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} = (\sqrt{1-\epsilon} \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} + \sqrt{\epsilon} \mathbf{g}_j)^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}}$$

$$= n\sqrt{1-\epsilon} \tag{C.16}$$

$$\mathbf{A}_j^T P(\mathbf{A}_{\mathbf{S}_k}) \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} = \mathbf{A}_j^T \mathbf{A}_{\mathbf{S}_k} (\mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}_k})^{-1} \mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}}$$

$$= n(1-\epsilon) \mathbf{1}^T (\mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}_k})^{-1} \mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}}$$

$$= \frac{n(1-\epsilon)}{n(1+(k-1)(1-\epsilon))} \mathbf{1}^T \mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} \quad (\because \mathbf{1} \text{ is an eigenvector of } \mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}_k})$$

$$= \frac{1-\epsilon}{1+(k-1)(1-\epsilon)} \mathbf{1}^T \mathbf{1} n\sqrt{1-\epsilon}$$

$$= \frac{nk(1-\epsilon)\sqrt{1-\epsilon}}{1+(k-1)(1-\epsilon)} \tag{C.17}$$

$$\mathbf{A}_j^T P(\mathbf{A}_{\mathbf{S}_k}) \boldsymbol{\eta} = \frac{1-\epsilon}{1+(k-1)(1-\epsilon)} \mathbf{1}^T \mathbf{A}_{\mathbf{S}_k}^T \boldsymbol{\eta}$$

$$= \frac{1-\epsilon}{1+(k-1)(1-\epsilon)} \sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \boldsymbol{\eta} \tag{C.18}$$

Similarly we can analyze $\left| \mathbf{A}_i^T \mathbf{r}_k \right|$

$$\left| \mathbf{A}_i^T \mathbf{r}_k \right| = \left| \mathbf{A}_i^T (\mathbf{I} - P(\mathbf{A}_{\mathbf{S}_k}))(\mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} + \boldsymbol{\eta}) \right|$$

$$= \left| \mathbf{A}_i^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} - \mathbf{A}_i^T P(\mathbf{A}_{\mathbf{S}_k}) \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} + \mathbf{A}_i^T \boldsymbol{\eta} - \mathbf{A}_i^T P(\mathbf{A}_{\mathbf{S}_k}) \boldsymbol{\eta} \right| \tag{C.19}$$

We can analyze these terms separately as below

$$\mathbf{A}_i^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} = \frac{n}{\sqrt{s^*}} \tag{C.20}$$

$$\mathbf{A}_i^T P(\mathbf{A}_{\mathbf{S}_k}) \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} = \mathbf{A}_i^T \mathbf{A}_{\mathbf{S}_k} (\mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}_k})^{-1} \mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}}$$

$$= \frac{n\sqrt{1-\epsilon}}{\sqrt{s^*}} \mathbf{1}^T (\mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}_k})^{-1} \mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}}$$

$$= \frac{n\sqrt{1-\epsilon}}{\sqrt{s^*} n(1+(k-1)(1-\epsilon))} \mathbf{1}^T \mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} \quad (\because \mathbf{1} \text{ is an eigenvector of } \mathbf{A}_{\mathbf{S}_k}^T \mathbf{A}_{\mathbf{S}_k})$$

$$= \frac{\sqrt{1-\epsilon}}{\sqrt{s^*}(1+(k-1)(1-\epsilon))} \mathbf{1}^T \mathbf{1} n\sqrt{1-\epsilon}$$

$$= \frac{nk(1-\epsilon)}{\sqrt{s^*}(1+(k-1)(1-\epsilon))} \tag{C.21}$$

$$\mathbf{A}_i^T P(\mathbf{A}_{\mathbf{S}_k}) \boldsymbol{\eta} = \frac{\sqrt{1-\epsilon}}{\sqrt{s^*}(1+(k-1)(1-\epsilon))} \mathbf{1}^T \mathbf{A}_{\mathbf{S}_k} \boldsymbol{\eta}$$

$$= \frac{\sqrt{1-\epsilon}}{\sqrt{s^*}(1+(k-1)(1-\epsilon))} \sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \boldsymbol{\eta} \tag{C.22}$$

Thus, for OMP to not recover a correct support at any iteration $k \leq s$, using (C.15) to (C.22), we have a required condition for all $i \in \mathbf{S}^* \setminus \mathbf{S}_{k-1}$, $j \in (\mathbf{S}^* \cup \mathbf{S}_{k-1})^{\mathsf{c}}$.

$$\left| \mathbf{A}_i^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} - \mathbf{A}_i^T P(\mathbf{A}_{\mathbf{S}_k}) \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} + \mathbf{A}_i^T \boldsymbol{\eta} - \mathbf{A}_i^T P(\mathbf{A}_{\mathbf{S}_k}) \boldsymbol{\eta} \right|$$

$$\leq \left| \mathbf{A}_j^T \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} - \mathbf{A}_j^T P(\mathbf{A}_{\mathbf{S}_k}) \mathbf{A}_{\mathbf{S}^*} \bar{\mathbf{x}} + \mathbf{A}_j^T \boldsymbol{\eta} - \mathbf{A}_j^T P(\mathbf{A}_{\mathbf{S}_k}) \boldsymbol{\eta} \right|$$

$$\text{or,} \quad \left| \frac{1}{\sqrt{s^*}} \frac{1}{1+(k-1)(1-\epsilon)} \left[n\epsilon - \sqrt{1-\epsilon} \sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \boldsymbol{\eta} \right] + \mathbf{A}_i^T \boldsymbol{\eta} \right|$$

$$\leq \left| \frac{\sqrt{1-\epsilon}}{1+(k-1)(1-\epsilon)} \left[n\epsilon - \sqrt{1-\epsilon} \sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \boldsymbol{\eta} \right] + \mathbf{A}_j^T \boldsymbol{\eta} \right| \tag{C.23}$$

25

Now, using triangle inequality on the LHS and RHS of C.23, we have a sufficient condition for failure:

$$\frac{1}{\sqrt{s^*}}\frac{1}{1+(k-1)(1-\epsilon)}\left|n\epsilon - \sqrt{1-\epsilon}\sum_{l\in\mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta}\right| + \left|\mathbf{A}_i^T\boldsymbol{\eta}\right|$$

$$\leq \frac{\sqrt{1-\epsilon}}{1+(k-1)(1-\epsilon)}\left|n\epsilon - \sqrt{1-\epsilon}\sum_{l\in\mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta}\right| - \left|\mathbf{A}_j^T\boldsymbol{\eta}\right|$$

Rearranging terms we get

$$\left|\mathbf{A}_i^T\boldsymbol{\eta}\right| + \left|\mathbf{A}_j^T\boldsymbol{\eta}\right| \leq \left(\sqrt{1-\epsilon} - \frac{1}{\sqrt{s^*}}\right)\frac{1}{1+(k-1)(1-\epsilon)}\left|n\epsilon - \sqrt{1-\epsilon}\sum_{l\in\mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta}\right| \quad \text{(C.24)}$$

Suppose $k \leq \frac{1}{9}\sigma\sqrt{n/\log\frac{d}{\delta}}$. Because $\left\|\mathbf{A}^T\boldsymbol{\eta}\right\|_\infty \leq \sigma\sqrt{2n\log\frac{d}{\delta}}$ with probability at least $1-\delta$,

$$n\epsilon \geq \sqrt{n}\epsilon\sqrt{n} \quad \text{(C.25)}$$

$$\geq 9k\sigma\epsilon\sqrt{n\log\frac{d}{\delta}}$$

$$\overset{\xi_1}{\geq} 9\sqrt{1-\epsilon}\,k\sigma\sqrt{2n\log\frac{d}{\delta}}$$

$$\geq 9\sqrt{1-\epsilon}\sum_{l\in\mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta}$$

Here, ξ_1 holds because $s^* \geq 8$. Then, a sufficient condition for (C.24) to be satisfied is,

$$2\sigma\sqrt{2n\log\frac{d}{\delta}} \leq \left(\sqrt{1-\epsilon} - \frac{1}{\sqrt{s^*}}\right)\frac{1}{1+(k-1)(1-\epsilon)}8\sqrt{1-\epsilon}\left|\sum_{l\in\mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta}\right|$$

$$= \left(\sqrt{1-\epsilon} - \frac{1}{\sqrt{s^*}}\right)\frac{1}{1+(k-1)(1-\epsilon)}8\sqrt{1-\epsilon}\sum_{l\in\mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta} \quad \text{(C.26)}$$

In the last equality (ξ_1), we remove absolute value signs by instantiating Lemma C.1 (with $\mathrm{T} \leftarrow$ k). The equality holds with probability larger than $1-\delta$. We will now lower bound $\sum_{l\in\mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta}$. Consider the random variables $\mathbf{g}_j^T\boldsymbol{\eta}$ for $j \in [d]\setminus\mathbf{S}^*$ and $\mathbf{A}_i^T\boldsymbol{\eta}$ for $i \in \mathbf{S}^*$. Since $\mathbf{g}_j$'s and $\mathbf{A}_i$'s are orthogonal to each other, these can be looked at as independent Gaussian variables with variance $\sigma\sqrt{n}$. Now, define

$$X_i := \begin{cases} \frac{\mathbf{g}_i^T\boldsymbol{\eta}}{\sigma\sqrt{n}} & \text{for } i \in [d]\setminus\mathbf{S}^* \\ \frac{\mathbf{A}_i^T\boldsymbol{\eta}}{\sigma\sqrt{n}} & \text{for } i \in \mathbf{S}^* \end{cases} \quad \text{(C.27)}$$

Thus, X_i are now independent standard Gaussian variables. We wish to show high probability bounds for the k-th order statistic of X_i's. Suppose that the kth largest element of the set $\{X_1, X_2, \ldots, X_d\}$ be denoted as $\mathbf{X}^{(k)}$, then for all $t \geq 0$

$$P\left\{X^{(k+1)} \leq t\right\} = \sum_{l=0}^{k}\binom{d}{l}\left[P\left\{X < t\right\}\right]^{d-l}\left[P\left\{X \geq t\right\}\right]^l$$

$$= \sum_{l=0}^{k}\binom{d}{l}\left[1 - \Phi(t)\right]^{d-l}\Phi(t)^l \quad \text{(C.28)}$$

The term on the right is the same as the probability that a binomial random variable X, with parameters d and $\Phi(t)$ is less than k, and from [3] we know that it is sub-gamma on the left tail with variance factor less than $d\Phi(t)$ and scale factor 0. Also a random variable X is said to be sub-gamma on the right tail with variance v and scale c if

$$\log\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \frac{\lambda^2 v}{2(1-c\lambda)} \quad \forall\,\lambda > 0 \quad \text{(C.29)}$$

Also, X is said to be sub-gamma on the left tail with variance factor v and scale parameter c, if $-X$ is sub-gamma on the right tail with the same variance and scale parameters. Define $z := k - \mathbb{E}[X]$. From Chernoff bounds we have

$$
\begin{aligned}
P\left\{X^{(k+1)} \leq t\right\} &= P\left\{X \geq k\right\} \\
&= P\left\{X - \mathbb{E}[X] \geq k - \mathbb{E}[X]\right\} \\
&= P\left\{e^{\lambda(\mathbb{E}[X]-X)} \geq e^{\lambda z}\right\} \\
&\leq e^{-\lambda z}\mathbb{E}\left[e^{\lambda(\mathbb{E}[X]-X)}\right] \\
&\leq \exp\left\{-\lambda z + \frac{\lambda^2 v}{2}\right\} \\
&= \exp\left\{-\frac{z^2}{2v}\right\}
\end{aligned}
$$

Now, since $d \geq 4^{\frac{1}{\alpha}}$, $k \leq s \leq d^{1-\alpha} \leq d/4 \leq d/2 \leq d\Phi(t) = \mathbb{E}[X]$. Thus, $|z| \geq \left|d^{1-\alpha} - d\Phi(t)\right| \geq \left|d^{1-\alpha} - d/2\right|$. Further, we saw that the variance factor $v \leq d\Phi(t)$. Now, set $t = \sqrt{\log \frac{d}{k}}$. This gives,

$$
\begin{aligned}
P\left\{X^{(k+1)} \leq \sqrt{\log \frac{d}{k}}\right\} &\leq \exp\left\{-\frac{\left(d^{1-\alpha} - d\Phi\left(\left(\sqrt{\log \frac{d}{k}}\right)\right)\right)^2}{2d\Phi\left(\sqrt{\log \frac{d}{k}}\right)}\right\} \\
&\leq \exp\left\{-\frac{\left(d^{1-\alpha} - \frac{d}{2}\right)^2}{2d}\right\} \\
&\leq \exp\left\{-\frac{d\left(d^{-\alpha} - \frac{1}{2}\right)^2}{2}\right\} \\
&\leq \exp\left\{-\frac{d\left(\frac{1}{4} - \frac{1}{2}\right)}{2}\right\} \\
&\leq \exp\left\{-\frac{d}{32}\right\} \\
&\overset{\xi_1}{\leq} \delta
\end{aligned}
$$
(C.30)

Here, ξ_1 holds because $d \geq 32\log(1/\delta)$. Thus, for the top k $\mathbf{g}_l$'s that maximize $\mathbf{g}_l^T \boldsymbol{\eta}$,

$$
P\left\{\mathbf{g}_l^T \boldsymbol{\eta} \geq \sigma\sqrt{n \log \frac{d}{k}}\right\} \geq 1 - \delta
$$
(C.31)

Now, with probability at least $1 - 2\delta$

$$
\begin{aligned}
\left|\mathbf{A}_l^T \boldsymbol{\eta}\right| &= \left|\sqrt{1-\epsilon}\,\mathbf{A}_{\mathbf{S}}^* \bar{\mathbf{x}} + \sqrt{\epsilon}\mathbf{g}_l^T \boldsymbol{\eta}\right| \\
&\geq \sqrt{1-\epsilon}\,\sigma\sqrt{2n \log\left(\frac{d}{\delta}\right)} + \sqrt{\epsilon}\,\sigma\sqrt{n \log\left(\frac{d}{k}\right)} \\
&\geq \sqrt{1-\epsilon}\,\sigma\sqrt{2n \log\left(\frac{d}{k}\right)} + \sqrt{\epsilon}\,\sigma\sqrt{n \log\left(\frac{d}{k}\right)} \\
&\geq \sigma\sqrt{n \log\left(\frac{d}{k}\right)} \\
&\geq \sigma\sqrt{\alpha n \log d}
\end{aligned}
$$
(C.32)

Therefore $\left|\sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \boldsymbol{\eta}\right| \geq \sqrt{\alpha} \sigma k \sqrt{n \log d}$ with probability at least $1 - 2\delta$. Since, $\epsilon = 1 - {}^4/{s^*}$ and $s^* \geq 8$, then (C.26) reduces to,

$$k \geq \frac{(s^* - 4)\sqrt{\log {}^d/\delta}}{2\sqrt{2\alpha \log d} - 4\sqrt{\log {}^d/\delta}} \tag{C.33}$$

which is trivially satisfies because $s^* \geq 8$ and $\alpha < 1$. Therefore from the above argument, we have that since $k \leq \frac{1}{9\sigma}\sqrt{\frac{n}{\log {}^d/\delta}}$, then with probability at least $1 - 2\delta$, an incorrect support is recovered at the kth OMP iterate.

Let us now have a look at the generalization error when none of the correct support is recovered, i.e., $\mathbf{S}_k \cap \mathbf{S}^* = \phi$.

$$\frac{1}{n}\left\|\mathbf{A}_{\mathbf{S}_k}\hat{\mathbf{x}}_k^{\text{OMP}} - \mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}}\right\|_2^2 = \frac{1}{n}\left\|\mathbf{A}_{\mathbf{S}_k}(\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}_k})^{-1}\mathbf{A}_{\mathbf{S}_k}^T(\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} + \boldsymbol{\eta}) - \mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}}\right\|_2^2$$

$$= \frac{1}{n}\left\|\mathbf{A}_{\mathbf{S}_k}(\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}_k})^{-1}\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} + P(\mathbf{A}_{\mathbf{S}_k})\boldsymbol{\eta} - \mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}}\right\|_2^2 \tag{C.34}$$

$$\mathbf{A}_{\mathbf{S}_k}(\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}_k})^{-1}\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} = \mathbf{A}_{\mathbf{S}_k}(\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}_k})^{-1}\mathbf{1}n\sqrt{1-\epsilon}$$

$$= \mathbf{A}_{\mathbf{S}_k}\mathbf{1}\frac{n\sqrt{1-\epsilon}}{n(1 + (k-1)(1-\epsilon))}$$

$$= \frac{1}{1 + (k-1)(1-\epsilon)}\left[k(1-\epsilon)\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} + \sqrt{\epsilon(1-\epsilon)}\sum_{l \in \mathbf{S}_k}\mathbf{g}_i\right]$$

$$\implies \mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} - \mathbf{A}_{\mathbf{S}_k}(\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}_k})^{-1}\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} = \frac{1}{1 + (k-1)(1-\epsilon)}\left[\epsilon\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} + \sqrt{\epsilon(1-\epsilon)}\sum_{l \in \mathbf{S}_k}\mathbf{g}_i\right]$$

$$\implies \left\|\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}} - \mathbf{A}_{\mathbf{S}_k}(\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}_k})^{-1}\mathbf{A}_{\mathbf{S}_k}^T\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}}\right\|_2^2 = \frac{n\left[\epsilon^2 + k\epsilon(1-\epsilon)\right]}{(1 + (k-1)(1-\epsilon))^2} \tag{C.35}$$

Using (C.34) and (C.35) we get

$$\frac{1}{n}\left\|\mathbf{A}_{\mathbf{S}_k}\hat{\mathbf{x}}_k^{\text{OMP}} - \mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}}\right\|_2^2 = \frac{1}{n}\left\|P(\mathbf{A}_{\mathbf{S}_k})\boldsymbol{\eta}\right\|_2^2 + \frac{1}{n}\frac{n\left[\epsilon^2 + k\epsilon(1-\epsilon)\right]}{(1 + (k-1)(1-\epsilon))^2}$$

$$- \frac{2}{n}\langle P(\mathbf{A}_{\mathbf{S}_k})\boldsymbol{\eta}, (\mathbf{I} - P(\mathbf{A}_{\mathbf{S}_k})\mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}})\rangle$$

$$= \frac{1}{n}\left\|P(\mathbf{A}_{\mathbf{S}_k})\boldsymbol{\eta}\right\|_2^2 + \frac{\left[\epsilon^2 + k\epsilon(1-\epsilon)\right]}{(1 + (k-1)(1-\epsilon))^2}$$

$$= \frac{1}{n}\left\|P(\mathbf{A}_{\mathbf{S}_k})\boldsymbol{\eta}\right\|_2^2 + \frac{\epsilon}{1 + (k-1)(1-\epsilon)} \tag{C.36}$$

When $\epsilon = 1 - {}^4/{s^*}$ and $s^* \geq 8$, we have

$$\frac{1}{n}\left\|\mathbf{A}_{\mathbf{S}_k}\hat{\mathbf{x}}_k^{\text{OMP}} - \mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}}\right\|_2^2 = \frac{1}{n}\left\|P(\mathbf{A}_{\mathbf{S}_k})\boldsymbol{\eta}\right\|_2^2 + \frac{1 - \frac{4}{s^*}}{1 - \frac{4}{s^*} + \frac{4k}{s^*}}$$

$$\geq \frac{\frac{1}{2}}{\frac{1}{2} + \frac{4k}{s^*}}$$

$$\geq \frac{s^*}{9k} \tag{C.37}$$

28

Since $s \le \frac{1}{2\sigma}\sqrt{\frac{n}{\log d/\delta}}$,

$$\frac{1}{n}\left\|\mathbf{A}_{\mathbf{S}_k}\hat{\mathbf{x}}_k^{\mathrm{OMP}} - \mathbf{A}_{\mathbf{S}^*}\bar{\mathbf{x}}\right\|_2^2 \ge \frac{2\sigma s^*}{9}\sqrt{\frac{\log d/\delta}{n}} \tag{C.38}$$

$$\ge \sigma\sqrt{\frac{\log d/\delta}{n}}$$

$$\ge \sigma\frac{\sqrt{\log d/\delta}}{n}\sqrt{n}$$

$$\ge \frac{2\sigma^2 s\log(d/\delta)}{n}$$

$$\ge \frac{\sigma^2\widetilde{\kappa}_s s^*\log(d/\delta)}{18n}$$

In the last line, we use $s \ge \frac{1}{36}\widetilde{\kappa}_s s^*$.

$\square$

Lemma C.1. *Suppose the same conditions as Theorem 4.3 hold. For the lower bound matrix defined in the proof of 4.3, suppose that a correct support has not been recovered until step $T \le s$, then for all $k \in \mathbf{S}_t$, $\mathbf{A}_k^T\boldsymbol{\eta} \ge 0$.*

Proof. With probability $1 - \delta$, for all $i \in [d]$, $\left|\mathbf{A}_i^T\boldsymbol{\eta}\right| \le \sigma\sqrt{2n\log\frac{d}{\delta}}$. Thus, for iteration $k + 1$,

$$\sum_{l \in \mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta} \le k\sigma\sqrt{2n\log\frac{d}{\delta}}$$

$$\le s\sigma\sqrt{2n\log\frac{d}{\delta}}$$

$$\le \left(\sqrt{\frac{n}{32\sigma^2\log\frac{d}{\delta}}}\right)\sigma\sqrt{2n\log\frac{d}{\delta}}$$

$$\le \frac{n}{4}$$

Further, since $s^* \ge 8$, $\epsilon \ge 2/3 \ge \sqrt{1-\epsilon}$. Thus, $\left[n\epsilon - \sqrt{1-\epsilon}\sum_{l\in\mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta}\right] \ge 3n\epsilon/4 \ge n/2$. Now,

$$\frac{\sqrt{1-\epsilon}}{1 + (k-1)(1-\epsilon)} \ge \frac{\sqrt{1-\epsilon}}{1 + (s-1)(1-\epsilon)}$$

$$\ge \frac{\sqrt{1-\epsilon}}{2s(1-\epsilon)}$$

$$\ge \frac{1}{2s}$$

$$\implies \frac{\sqrt{1-\epsilon}}{1 + (k-1)(1-\epsilon)}\left[n\epsilon - \sqrt{1-\epsilon}\sum_{l\in\mathbf{S}_k}\mathbf{A}_l^T\boldsymbol{\eta}\right] \ge \frac{n}{4s}$$

$$\ge \frac{\sqrt{n}\left(\sqrt{32}\sigma s\sqrt{\log\frac{d}{\delta}}\right)}{4s}$$

$$\ge \sigma\sqrt{2n\log\frac{d}{\delta}}$$

$$\ge \left|\mathbf{A}_j^T\boldsymbol{\eta}\right| \ge 0 \qquad (\forall j \in (\mathbf{S}^*\cup\mathbf{S}_{k-1})^c) \tag{C.39}$$

From C.23, if a $j \in (\mathbf{S}^* \cup \mathbf{S}_{k-1})^{\mathrm{c}}$ is picked, then it maximizes:

$$\left| \frac{\sqrt{1-\epsilon}}{1+(k-1)(1-\epsilon)} \left[n\epsilon - \sqrt{1-\epsilon} \sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \eta \right] + \mathbf{A}_j^T \eta \right|$$

From C.39 we can see that for any j, j' such that $\mathbf{A}_j^T \eta \geq 0$ and $\mathbf{A}_{j'}^{\prime T} \eta \leq 0$,

$$\left| \frac{\sqrt{1-\epsilon}}{1+(k-1)(1-\epsilon)} \left[n\epsilon - \sqrt{1-\epsilon} \sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \eta \right] + \mathbf{A}_j^T \eta \right|$$

$$= \frac{\sqrt{1-\epsilon}}{1+(k-1)(1-\epsilon)} \left[n\epsilon - \sqrt{1-\epsilon} \sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \eta \right] + \mathbf{A}_j^T \eta$$

$$\geq \frac{\sqrt{1-\epsilon}}{1+(k-1)(1-\epsilon)} \left[n\epsilon - \sqrt{1-\epsilon} \sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \eta \right]$$

$$\geq \frac{\sqrt{1-\epsilon}}{1+(k-1)(1-\epsilon)} \left[n\epsilon - \sqrt{1-\epsilon} \sum_{l \in \mathbf{S}_k} \mathbf{A}_l^T \eta \right] + \mathbf{A}_{j'}^{\prime T} \eta$$

Thus, OMP will indeed pick up an index j such that $\mathbf{A}_j^T \eta \geq 0$. This proof holds for all steps $k < T$. Thus we can say that for all $k \leq T$, the index that is picked up (say $\mathbf{A}_l$) satisfies $\mathbf{A}_l^T \eta \geq 0$. $\qquad\square$