Coherence-Based Performance Guarantees for Estimating a Sparse Vector Under Random Noise

Zvika Ben-Haim, Yonina C. Eldar, and Michael Elad

***Abstract*—We consider the problem of estimating a deterministic sparse vector $\mathbf{x}_0$ from underdetermined measurements $\mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\mathbf{w}$ represents white Gaussian noise and $\mathbf{A}$ is a given deterministic dictionary. We analyze the performance of three sparse estimation algorithms: basis pursuit denoising (BPDN), orthogonal matching pursuit (OMP), and thresholding. These algorithms are shown to achieve near-oracle performance with high probability, assuming that $\mathbf{x}_0$ is sufficiently sparse. Our results are non-asymptotic and are based only on the coherence of $\mathbf{A}$, so that they are applicable to arbitrary dictionaries. Differences in the precise conditions required for the performance guarantees of each algorithm are manifested in the observed performance at high and low signal-to-noise ratios. This provides insight on the advantages and drawbacks of ℓ_1 relaxation techniques such as BPDN as opposed to greedy approaches such as OMP and thresholding.**

EDICS Topics: SSP-PARE, SSP-PERF.

Index terms: Sparse estimation, basis pursuit, matching pursuit, thresholding algorithm, oracle.

I. Introduction

Estimation problems with sparsity constraints have attracted considerable attention in recent years because of their potential use in numerous signal processing applications, such as denoising, compression and sampling. In a typical setup, an unknown deterministic parameter $\mathbf{x}_0 \in \mathbb{R}^m$ is to be estimated from measurements $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\mathbf{A} \in \mathbb{R}^{n\times m}$ is a deterministic matrix and $\mathbf{w}$ is a noise vector. Typically, the dictionary $\mathbf{A}$ consists of more columns than rows (i.e., $m > n$), so that without further assumptions, $\mathbf{x}_0$ is unidentifiable from $\mathbf{b}$. The impassé is resolved by assuming that the parameter vector is sparse, i.e., that most elements of $\mathbf{x}_0$ are zero. Under the assumption of sparsity, several estimation approaches can be used. These include greedy algorithms, such as thresholding and orthogonal matching pursuit (OMP) [1], and ℓ_1 relaxation methods, such as the Dantzig selector [2] and basis pursuit denoising (BPDN) [3] (also known as the Lasso). A comparative analysis of these techniques is crucial for determining the appropriate strategy in a given situation.

There are two standard approaches to modeling the noise $\mathbf{w}$ in the sparse estimation problem. The first is to assume that $\mathbf{w}$ is deterministic and bounded [4]–[6]. This leads to a worst-case analysis in which an estimator must perform adequately even when the noise maximally damages the measurements. The noise in this case is thus called adversarial. By contrast, if one assumes that the noise is random, then the analysis aims to describe estimator behavior for typical noise values [2], [7], [8]. The random noise scenario is the main focus of this paper. As one might expect, stronger performance guarantees can be obtained in this setting.

It is common to judge the quality of an estimator by comparing its mean-squared error (MSE) with the Cramér–Rao bound (CRB) [9]. In the case of sparse estimation under Gaussian noise, it has recently been shown that the unbiased CRB is identical (for almost all values of $\mathbf{x}_0$) to the MSE of the "oracle" estimator, which knows the locations of the nonzero elements of $\mathbf{x}_0$ [10]. Thus, a gold standard for estimator performance is the MSE of the oracle. Indeed, it can be shown that ℓ_1 relaxation algorithms come close to the oracle when the noise is Gaussian. Results of this type are sometimes referred to as "oracle inequalities." Specifically, Candès and Tao [2] have shown that, with high probability, the ℓ_2 distance between $\mathbf{x}_0$ and the Dantzig estimate is within a constant times $\log m$ of the performance of the oracle. Recently, Bickel et al. [8] have demonstrated that the performance of BPDN is similarly bounded, with high probability, by $C \log m$ times the oracle performance, for a constant C. However, the constant involved in this analysis is considerably larger than that of the Dantzig selector. Interestingly, it turns out that the $\log m$ gap between the oracle and practical estimators is an unavoidable consequence of the fact that the nonzero locations in $\mathbf{x}_0$ are unknown [11].

The contributions [2], [8] state their results using the restricted isometry constants (RICs). These measures of the dictionary quality can be efficiently approximated in specific cases, e.g., when the dictionary is selected randomly from an appropriate ensemble. However, in general it is NP-hard to evaluate the RICs for a given matrix $\mathbf{A}$, and they must then be bounded by efficiently computable properties of $\mathbf{A}$, such as the mutual coherence [12]. In this respect, coherence-based results are appealing since they can be used with arbitrary dictionaries [13], [14].

In this paper, we seek performance guarantees for sparse estimators based directly on the mutual coherence of the matrix $\mathbf{A}$ [15]. While such results are suboptimal when the RICs of $\mathbf{A}$ are known, the proposed approach yields tighter bounds than those obtained by applying coherence bounds to RIC-based results. Specifically, we demonstrate that BPDN, OMP and thresholding all achieve performance within a constant times $\log m$ of the oracle estimator, under suitable conditions.

Z. Ben-Haim and Y. C. Eldar are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: {zvikabh@tx, yonina@ee}.technion.ac.il). M. Elad is with the Department of Computer Science, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: elad@cs.technion.ac.il). Contact information for Z. Ben-Haim: phone +972-4-8294700, fax +972-4-8295757.

This work was supported in part by the Israel Science Foundation under Grants 1081/07 and 599/08, and by the European Commission's FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (grant agreement no. 216715) and FP7-FET program, SMALL project (grant agreement no. 225913).

In the case of BPDN, our result provides a tighter guarantee than the coherence-based implications of the work of Bickel et al. [8]. To the best of our knowledge, there are no prior performance guarantees for greedy approaches such as OMP and thresholding when the noise is random.

It is important to distinguish the present work from Bayesian performance analysis, as practiced in [13], [16]–[18], where on top of the assumption of stochastic noise, a probabilistic model for $\mathbf{x}_0$ is also used. Our results hold for any specific value of $\mathbf{x}_0$ (satisfying appropriate conditions), rather than providing results on average over realizations of $\mathbf{x}_0$; this necessarily leads to weaker guarantees. It also bears repeating that our results apply to a fixed, finite-sized matrix $\mathbf{A}$; this distinguishes our work from asymptotic performance guarantees for large m and n, such as [19].

The rest of this paper is organized as follows. We begin in Section II by comparing dictionary quality measures and reviewing standard estimation techniques. In Section III, we analyze the limitations of estimator performance under adversarial noise. This motivates the introduction of random noise, for which substantially better guarantees are obtained in Section IV. Finally, the validity of these results is examined by simulation in practical estimation scenarios in Section V.

The following notation is used throughout the paper. Vectors and matrices are denoted, respectively, by boldface lowercase and boldface uppercase letters. The set of indices of the nonzero entries of a vector $\mathbf{x}$ is called the support of $\mathbf{x}$ and denoted $\operatorname{supp}(\mathbf{x})$. Given an index set Λ and a matrix $\mathbf{A}$, the notation $\mathbf{A}_\Lambda$ refers to the submatrix formed from the columns of $\mathbf{A}$ indexed by Λ. The ℓ_p norm of a vector $\mathbf{x}$, for $1 \leq p \leq \infty$, is denoted $\|\mathbf{x}\|_p$, while $\|\mathbf{x}\|_0$ denotes the number of nonzero elements in $\mathbf{x}$.

II. Preliminaries

A. Characterizing the Dictionary

Let $\mathbf{x}_0 \in \mathbb{R}^m$ be an unknown deterministic vector, and denote its support set by $\Lambda_0 = \operatorname{supp}(\mathbf{x}_0)$. Let $s = \|\mathbf{x}_0\|_0$ be the number of nonzero entries in $\mathbf{x}_0$. In our setting, it is typically assumed that s is much smaller than m, i.e., that most elements in $\mathbf{x}_0$ are zero. Suppose we obtain noisy measurements

$$\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times m}$ is a known overcomplete dictionary ($m > n$). We refer to the columns $\mathbf{a}_i$ of $\mathbf{A}$ as the *atoms* of the dictionary, and assume throughout our work that the atoms are normalized, $\|\mathbf{a}_i\|_2 = 1$. We will consider primarily the situation in which the noise $\mathbf{w}$ is random, though for comparison we will also examine the case of a bounded deterministic noise vector; a precise definition of $\mathbf{w}$ is deferred to subsequent sections.

For $\mathbf{x}_0$ to be identifiable, one must guarantee that different values of $\mathbf{x}_0$ produce significantly different values of $\mathbf{b}$. One way to ensure this is to examine all possible *subdictionaries*, or s-element sets of atoms, and verify that the subspaces spanned by these subdictionaries differ substantially from one another.

More specifically, several methods have been proposed to formalize the notion of the suitability of a dictionary for sparse estimation. These include the mutual coherence [12], the cumulative coherence [7], the exact recovery coefficient (ERC) [7], the spark [4], and the RICs [2], [5]. Except for the mutual coherence and cumulative coherence, none of these measures can be efficiently calculated for an arbitrary given dictionary $\mathbf{A}$. Since the values of the cumulative and mutual coherence are quite close, our focus in this paper will be on the mutual coherence $\mu = \mu(\mathbf{A})$, which is defined as

$$\mu \triangleq \max_{i \neq j} \left| \mathbf{a}_i^T \mathbf{a}_j \right|. \tag{2}$$

While the mutual coherence can be efficiently calculated directly from (2), it is not immediately clear in what way μ is related to the requirement that subdictionaries must span different subspaces. Indeed, μ ensures a lack of correlation between single atoms, while we require a distinction between s-element subdictionaries. To explore this relation, let us recall the definitions of the RICs, which are more directly related to the subdictionaries of $\mathbf{A}$. We will then show that the mutual coherence can be used to bound the constants involved in the RICs, a fact which will also prove useful in our subsequent analysis. This strategy is inspired by earlier works, which have used the mutual coherence to bound the ERC [7] and the spark [4]. Thus, the coherence can be viewed as a tractable proxy for more accurate measures of the quality of a dictionary, which cannot themselves be calculated efficiently.

By the RICs we refer to two properties describing "good" dictionaries, namely, the restricted isometry property (RIP) and the restricted orthogonality property (ROP), which we now define. A dictionary $\mathbf{A}$ is said to satisfy the RIP [5] of order s with parameter δ_s if, for every index set Λ of size s, we have

$$(1 - \delta_s)\|\mathbf{y}\|_2^2 \leq \|\mathbf{A}_\Lambda \mathbf{y}\|_2^2 \leq (1 + \delta_s)\|\mathbf{y}\|_2^2 \tag{3}$$

for all $\mathbf{y} \in \mathbb{R}^s$. Thus, when δ_s is small, the RIP ensures that any s-atom subdictionary is nearly orthogonal, which in turn implies that any two disjoint $(s/2)$-atom subdictionaries are well-separated.

Similarly, $\mathbf{A}$ is said to satisfy the ROP [2] of order (s_1, s_2) with parameter θ_{s_1,s_2} if, for every pair of disjoint index sets Λ_1 and Λ_2 having cardinalities s_1 and s_2, respectively, we have

$$\left| \mathbf{y}_1^T \mathbf{A}_{\Lambda_1}^T \mathbf{A}_{\Lambda_2} \mathbf{y}_2 \right| \leq \theta_{s_1,s_2} \|\mathbf{y}_1\|_2 \|\mathbf{y}_2\|_2 \tag{4}$$

for all $\mathbf{y}_1 \in \mathbb{R}^{s_1}$ and for all $\mathbf{y}_2 \in \mathbb{R}^{s_2}$. In words, the ROP requires any two disjoint subdictionaries containing s_1 and s_2 elements, respectively, to be nearly orthogonal to each other. These two properties are therefore closely related to the requirement that distinct subdictionaries of $\mathbf{A}$ behave dissimilarly.

In recent years, it has been demonstrated that various practical estimation techniques successfully approximate $\mathbf{x}_0$ from $\mathbf{b}$, if the constants δ_s and θ_{s_1,s_2} are sufficiently small [2], [5], [20]. This occurs, for example, when the entries in $\mathbf{A}$ are chosen randomly according to an independent, identically distributed Gaussian law, as well as in some specific deterministic dictionary constructions.

Unfortunately, in the standard estimation setting, one cannot design the system matrix $\mathbf{A}$ according to these specific rules. In general, if one is given a particular dictionary $\mathbf{A}$, then

there is no known algorithm for efficiently determining its RICs. Indeed, the very nature of the RICs seems to require enumerating over an exponential number of index sets in order to find the "worst" subdictionary. While the mutual coherence μ of (2) tends to be far less accurate in capturing the accuracy of a dictionary, it is still useful to be able to say something about the RICs based only on μ. Such a result is given in the following lemma.

Lemma 1: For any matrix $\mathbf{A}$, the RIP constant δ_s of (3) and the ROP constant θ_{s_1,s_2} of (4) satisfy the bounds

$$\delta_s \leq (s-1)\mu, \tag{5}$$

$$\theta_{s_1,s_2} \leq \mu\sqrt{s_1 s_2} \tag{6}$$

where μ is the mutual coherence (2).

The proof of Lemma 1 can be found in Appendix A. We will apply this lemma in Section IV, when examining the performance of the Dantzig selector. This tool can also be used in conjunction with other results that rely on the RIP and ROP.

B. Estimation Techniques

To fix notation, we now briefly review several approaches for estimating $\mathbf{x}_0$ from noisy measurements $\mathbf{b}$ given by (1). The two main strategies for efficiently estimating a sparse vector are ℓ_1 relaxation and greedy methods. The first of these involves solving an optimization problem wherein the nonconvex constraint $\|\mathbf{x}_0\|_0 = s$ is relaxed to a constraint on the ℓ_1 norm of the estimated vector $\mathbf{x}_0$. Specifically, we consider the ℓ_1-penalty version of BPDN, which estimates $\mathbf{x}_0$ as a solution $\hat{\mathbf{x}}_{\mathrm{BP}}$ to the quadratic program

$$\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 + \gamma\|\mathbf{x}\|_1 \tag{7}$$

for some regularization parameter γ. We refer to the optimization problem (7) as BPDN, although it should be noted that some authors reserve this term for the related optimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 \leq \delta \tag{8}$$

where δ is a given constant.

Another estimator based on the idea of ℓ_1 relaxation is the Dantzig selector [2], defined as a solution $\hat{\mathbf{x}}_{\mathrm{DS}}$ to the optimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{A}^T(\mathbf{b}-\mathbf{A}\mathbf{x})\|_\infty \leq \tau \tag{9}$$

where τ is again a user-selected parameter. The Dantzig selector, like BPDN, is a convex relaxation method, but rather than penalizing the ℓ_2 norm of the residual $\mathbf{b}-\mathbf{A}\mathbf{x}$, the Dantzig selector ensures that the residual is weakly correlated with all dictionary atoms.

Instead of solving an optimization problem, greedy approaches estimate the support set Λ_0 from the measurements $\mathbf{b}$. Once a support set Λ is chosen, the parameter vector $\mathbf{x}_0$ can be estimated using least-squares (LS) to obtain

$$\widehat{\mathbf{x}} = \begin{cases} \mathbf{A}_\Lambda^\dagger \mathbf{b} & \text{on the support set } \Lambda, \\ \mathbf{0} & \text{elsewhere.} \end{cases} \tag{10}$$

Greedy techniques differ in the method by which the support set is selected. The simplest method is known as the thresholding algorithm. This technique computes the correlation of the measured signal $\mathbf{b}$ with each of the atoms $\mathbf{a}_i$ and defines Λ as the set of indices of the s atoms having the highest correlation. Subsequently, the LS technique (10) is applied to obtain the thresholding estimate $\widehat{\mathbf{x}}_{\mathrm{th}}$.

A somewhat more sophisticated greedy algorithm is OMP [1]. This iterative approach begins by initializing the estimated support set Λ^0 to the empty set and setting a residual vector $\mathbf{r}^0$ to $\mathbf{b}$. Subsequently, at each iteration $i = 1, \ldots, s$, the algorithm finds the single atom which is most highly correlated with $\mathbf{r}^{i-1}$. The index of this atom, say k_i, is added to the support set, so that $\Lambda^i = \Lambda^{i-1} \cup \{k_i\}$. The estimate $\widehat{\mathbf{x}}^i_{\mathrm{OMP}}$ at the ith iteration is then defined by the LS solution (10) using the support set Λ^i. Next, the residual is updated using the formula

$$\mathbf{r}^i = \mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}^i_{\mathrm{OMP}}. \tag{11}$$

The residual thus describes the part of $\mathbf{b}$ which has yet to be accounted for by the estimate. The counter i is now incremented, and s iterations are performed, after which the OMP estimate $\widehat{\mathbf{x}}_{\mathrm{OMP}}$ is defined as the estimate at the final iteration, $\widehat{\mathbf{x}}^s_{\mathrm{OMP}}$. A well-known property of OMP is that the algorithm never chooses the same atom twice [4]. Consequently, stopping after s iterations guarantees that $\|\widehat{\mathbf{x}}_{\mathrm{OMP}}\|_0 = s$.

Finally, we also mention the so-called oracle estimator, which is based both on $\mathbf{b}$ and on the true support set Λ_0 of $\mathbf{x}_0$; the support set is assumed to have been provided by an "oracle". The oracle estimator $\hat{\mathbf{x}}_{\mathrm{or}}$ calculates the LS solution (10) for Λ_0. In the case of white Gaussian noise, the MSE obtained using this technique equals that of the CRB [10]. Thus, it makes sense to use the oracle estimator as a gold standard against which the performance of practical algorithms can be compared.

III. Performance under Adversarial Noise

In this section, we briefly discuss the case in which the noise $\mathbf{w}$ is an unknown deterministic vector which satisfies $\|\mathbf{w}\|_2 \leq \varepsilon$. As we will see, performance guarantees in this case are rather weak, and indeed no denoising capability can be ensured for any known algorithm. In Section IV, we will compare this setting with the results which can be obtained when $\mathbf{w}$ is random.

Typical "stability" results under adversarial noise guarantee that if the mutual coherence μ of $\mathbf{A}$ is sufficiently small, and if $\mathbf{x}_0$ is sufficiently sparse, then the distance between $\mathbf{x}_0$ and its estimate is on the order of the noise magnitude. Such results can be derived for algorithms including BPDN, OMP, and thresholding. Consider, for example, the following theorem, which is based on the work of Tropp [7, §IV-C].[1]

Theorem 1 (Tropp): Let $\mathbf{x}_0$ be an unknown deterministic vector with known sparsity $\|\mathbf{x}_0\|_0 = s$, and let $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\|\mathbf{w}\|_2 \leq \varepsilon$. Suppose the mutual coherence μ of the dictionary $\mathbf{A}$ satisfies $s < 1/(3\mu)$. Let $\hat{\mathbf{x}}_{\mathrm{BP}}$ denote a solution

[1] Tropp considers only the case in which the entries of $\mathbf{x}_0$ belong to the set $\{0, \pm 1\}$. However, since the analysis performed in [7, §IV-C] can readily be applied to the general setting considered here, we omit the proof of Theorem 1.

of BPDN (7) with regularization parameter $\gamma = 2\varepsilon$. Then, $\hat{\mathbf{x}}_{\mathrm{BP}}$ is unique, the support of $\hat{\mathbf{x}}_{\mathrm{BP}}$ is a subset of the support of $\mathbf{x}_0$, and

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_\infty < \left(3 + \sqrt{\tfrac{3}{2}}\right)\varepsilon \approx 4.22\varepsilon. \tag{12}$$

Results similar to Theorem 1 have also been obtained [4], [5], [14], [20] for the related ℓ_1-error estimation approach (8), as well as for the OMP algorithm [4]. Furthermore, the technique used in the proof for the OMP [4] can also be applied to demonstrate a (slightly weaker) performance guarantee for the thresholding algorithm.

In all of the aforementioned results, the only guarantee is that the distance between $\hat{\mathbf{x}}_{\mathrm{BP}}$ and $\mathbf{x}_0$ is on the order of the noise power ε. Such results are somewhat disappointing, because one would expect the knowledge that $\mathbf{x}_0$ is sparse to assist in denoising; yet Theorem 1 promises only that the ℓ_∞ distance between $\hat{\mathbf{x}}_{\mathrm{BP}}$ and $\mathbf{x}_0$ is less than about four times the maximum noise level. However, the fact that no denoising has occurred is a consequence of the problem setting itself, rather than a limitation of the algorithms proposed above. In the adversarial case, even the oracle estimator can only guarantee an estimation error on the order of ε. This is because $\mathbf{w}$ can be chosen so that $\mathbf{w} \in \mathrm{span}(\mathbf{A}_{\Lambda_0})$, in which case projection onto $\mathrm{span}(\mathbf{A}_{\Lambda_0})$, as performed by the oracle estimator, does not remove any portion of the noise.

In conclusion, results in this adversarial context must take into account values of $\mathbf{w}$ which are chosen so as to cause maximal damage to the estimation algorithm. In many practical situations, such a scenario is overly pessimistic. Thus, it is interesting to ask what guarantees can be made about the performance of practical estimators under the assumption of random (and thus non-adversarial) noise. This scenario is considered in the next section.

IV. Performance under Random Noise

We now turn to the setting in which the noise $\mathbf{w}$ is a Gaussian random vector with mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}$. In this case, it can be shown [10] that the MSE of any unbiased estimator of $\mathbf{x}_0$ satisfies the Cramér–Rao bound

$$\mathrm{MSE}(\hat{\mathbf{x}}) \geq \mathrm{CRB} = \sigma^2 \,\mathrm{Tr}((\mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0})^{-1}) \tag{13}$$

whenever $\|\mathbf{x}_0\|_0 = s$. Interestingly, CRB is also the MSE of the oracle estimator [2].

It follows from the Gershgorin disc theorem [21] that all eigenvalues of $\mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0}$ are between $1 - (s-1)\mu$ and $1 + (s+1)\mu$. Therefore, for reasonable sparsity levels, $\mathrm{Tr}((\mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0})^{-1})$ is not much larger than s; for example, if we assume, as in Theorem 1, that $s < 1/(3\mu)$, then CRB of (13) is no larger than $\frac{3}{2}s\sigma^2$. Considering that the mean power of $\mathbf{w}$ is $n\sigma^2$, it is evident that the oracle estimator has substantially reduced the noise level. In this section, we will demonstrate that comparable performance gains are achievable using practical methods, which do not have access to the oracle.

A. ℓ_1-Relaxation Approaches

Historically, performance guarantees under random noise were first obtained for the Dantzig selector (9). The result, due to Candès and Tao [2], is derived using the RICs (3)–(4). Using the bounds of Lemma 1 yields the following coherence-based result.

Theorem 2 (Candès and Tao): Let $\mathbf{x}_0$ be an unknown deterministic vector such that $\|\mathbf{x}_0\|_0 = s$, and let $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is a random noise vector. Assume that

$$s < 1 + \frac{1}{(1+\sqrt{2})\mu} \tag{14}$$

and consider the Dantzig selector (9) with parameter

$$\tau = \sigma\sqrt{2(1+\alpha)\log m} \tag{15}$$

for some constant $\alpha > 0$. Then, with probability exceeding

$$1 - \frac{1}{m^\alpha \sqrt{\pi \log m}}, \tag{16}$$

the Dantzig selector $\hat{\mathbf{x}}_{\mathrm{DS}}$ satisfies

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{DS}}\|_2^2 \leq 2c_1^2(1+\alpha)s\sigma^2 \log m \tag{17}$$

where

$$c_1 = \frac{4}{1 - \left((1+\sqrt{2})s - 1\right)\mu}. \tag{18}$$

This theorem is significant because it demonstrates that, while $\hat{\mathbf{x}}_{\mathrm{DS}}$ does not quite reach the performance of the oracle estimator, it does come within a constant factor multiplied by $\log m$, with high probability. Interestingly, the $\log m$ factor is an unavoidable result of the fact that the locations of the nonzero elements in $\mathbf{x}_0$ are unknown (see [11, §7.4] and the references therein).

It is clearly of interest to determine whether results similar to Theorem 2 can be obtained for other sparse estimation algorithms [22], [23]. In this context, Bickel et al. [8] have recently shown that, with high probability, BPDN also comes within a factor of $C \log m$ of the oracle performance, for a constant C. In fact, their analysis is quite versatile, and simultaneously provides a result for both the Dantzig selector and BPDN. However, the constant C obtained in this BPDN guarantee is always larger than 128, often substantially so; this is considerably weaker than the result of Theorem 2. Furthermore, while the necessary conditions for the results of Bickel et al. are not directly comparable with those of Candès and Tao, an application of Lemma 1 indicates that coherence-based conditions stronger than (14) are required for the results of Bickel et al. to hold.

In the following, we obtain a coherence-based performance guarantee for BPDN. In particular, we demonstrate that, for an appropriate choice of the regularization parameter γ, the squared error of the BPDN estimate is bounded, with high probability, by a small constant times $s\sigma^2 \log(m-s)$, and that this constant is lower than that of Theorem 2. We begin by stating the following somewhat more general result, whose proof is found in Appendix B.

Theorem 3: Let $\mathbf{x}_0$ be an unknown deterministic vector with known sparsity $\|\mathbf{x}_0\|_0 = s$, and let $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is a random noise vector. Suppose that[2]

$$s < \frac{1}{3\mu}. \tag{19}$$

Then, with probability exceeding

$$\left(1 - (m-s)\exp\left(-\frac{\gamma^2}{8\sigma^2}\right)\right)\left(1 - e^{-s/7}\right), \tag{20}$$

the solution $\hat{\mathbf{x}}_{\mathrm{BP}}$ of BPDN (7) is unique and satisfies

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq \left(\sigma\sqrt{3} + \tfrac{3}{2}\gamma\right)^2 s. \tag{21}$$

To compare the results for BPDN and the Dantzig selector, we now derive from Theorem 3 a result which holds with a probability on the order of (16). Observe that in order for (20) to be a high probability, we require $\exp(-\gamma^2/(8\sigma^2))$ to be substantially smaller than $1/(m-s)$. This requirement can be used to select a value for the regularization parameter γ. In particular, one requires γ to be at least on the order of $\sqrt{8\sigma^2\log(m-s)}$. However, γ should not be much larger than this value, as this will increase the error bound (21). We propose to use

$$\gamma = \sqrt{8\sigma^2(1+\alpha)\log(m-s)} \tag{22}$$

for some fairly small $\alpha > 0$. Substituting this value of γ into Theorem 3 yields the following result.

Corollary 1: Under the conditions of Theorem 3, let $\hat{\mathbf{x}}_{\mathrm{BP}}$ be a solution of BPDN (7) with γ given by (22). Then, with probability exceeding

$$\left(1 - \frac{1}{(m-s)^{\alpha}}\right)\left(1 - e^{-s/7}\right) \tag{23}$$

we have

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq \left(\sqrt{3} + 3\sqrt{2(1+\alpha)\log(m-s)}\right)^2 s\sigma^2. \tag{24}$$

Let us examine the probability (23) with which Corollary 1 holds, to verify that it is indeed roughly equal to (16). The expression (23) consists of a product of two terms, both of which converge to 1 as the problem dimensions increase. The right-hand term may seem odd because it appears to favor non-sparse signals; however, this is an artifact of the method of proof, which requires a sufficient number of nonzero coefficients for large number approximations to hold. This right-hand term converges to 1 exponentially and therefore typically has a negligible effect on the overall probability of success; for example, for $s \geq 50$ this term is larger than 0.999.

The left-hand term in (23) tends to 1 polynomially as $m-s$ increases. This is a slightly lower rate than the probability (16) with which the Dantzig selector bound holds; however, this difference is compensated for by a correspondingly lower multiplicative factor of $\log(m-s)$ in the BPDN error bound (24), as opposed to the $\log m$ factor in the Dantzig selector. In any case, for both theorems to hold, m must increase much more quickly than s, so that these differences are negligible.

For large s and $m-s$, Corollary 1 ensures that, with high probability, $\|\hat{\mathbf{x}}_{\mathrm{BP}} - \mathbf{x}_0\|_2^2$ is no larger than a constant multiplied by $s\sigma^2\log(m-s)$. Up to a multiplicative constant, this error bound is essentially identical to the result (17) for the Dantzig selector. As we have seen, the probabilities with which these bounds hold are likewise almost identical. However, the constants involved in the BPDN, as demonstrated by Corollary 1, are substantially lower than those previously known for the Dantzig selector. To see this, consider a situation in which $s = 1/(4\mu)$. In this case, for large s, the bound (17) on the Dantzig selector rapidly converges to

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{DS}}\|_2^2 \leq 203.6(1+\alpha)\cdot\log m\cdot s\sigma^2. \tag{25}$$

By comparison, the performance of BPDN in the same setting, as bounded by Corollary 1, is

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq 18(1+\alpha)\cdot\log(m-s)\cdot s\sigma^2 \tag{26}$$

which is over 10 times lower. This improvement is not merely a result of the particular choice of s or μ. Indeed, the multiplicative factor of 18 which appeared in the BPDN bound (26) holds for large s with any value of μ, as long as $s < 1/(3\mu)$; whereas it can be seen from (17)–(18) that the multiplicative factor of the Dantzig selector is always larger than 32. Further comparison between these guarantees will be presented in Section V.

B. Greedy Approaches

The performance guarantees obtained for the ℓ_1-relaxation techniques required only the assumption that $\mathbf{x}_0$ is sufficiently sparse. By contrast, for greedy algorithms, successful estimation can only be guaranteed if one further assumes that all nonzero components of $\mathbf{x}_0$ are somewhat larger than the noise level. The reason is that greedy techniques are based on a LS solution for an estimated support, an approach whose efficacy is poor unless the support is correctly identified. Indeed, when using the LS technique (10), even a single incorrectly identified support element may cause the entire estimate to be severely incorrect. To ensure support recovery, all nonzero elements must be large enough to overcome the noise.

To formalize this notion, denote $\mathbf{x}_0 = (x_{0,1}, \ldots, x_{0,m})^T$ and define

$$\begin{aligned} |x_{\min}| &= \min_{i\in\Lambda_0} |x_{0,i}|, \\ |x_{\max}| &= \max_{i\in\Lambda_0} |x_{0,i}|. \end{aligned} \tag{27}$$

A performance guarantee for both OMP and the thresholding algorithm is then given by the following theorem, whose proof can be found in Appendix C.

Theorem 4: Let $\mathbf{x}_0$ be an unknown deterministic vector with known sparsity $\|\mathbf{x}_0\|_0 = s$, and let $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is a random noise vector. Suppose that

$$|x_{\min}| - (2s-1)\mu|x_{\min}| \geq 2\sigma\sqrt{2(1+\alpha)\log m} \tag{28}$$

for some constant $\alpha > 0$. Then, with probability at least

$$1 - \frac{1}{m^{\alpha}\sqrt{\pi(1+\alpha)\log m}}, \tag{29}$$

[2] As in [7], analogous findings can also be obtained under the weaker requirement $s < 1/(2\mu)$, but the resulting expressions are somewhat more involved.

the OMP estimate $\widehat{\mathbf{x}}_{\mathrm{OMP}}$ identifies the correct support Λ_0 of $\mathbf{x}_0$ and, furthermore, satisfies

$$\|\widehat{\mathbf{x}}_{\mathrm{OMP}} - \mathbf{x}_0\|_2^2 \leq \frac{2(1+\alpha)}{(1-(s-1)\mu)^2} s\sigma^2 \log m \tag{30a}$$

$$\leq 8(1+\alpha) s\sigma^2 \log m. \tag{30b}$$

If the stronger condition

$$|x_{\min}| - (2s-1)\mu |x_{\max}| \geq 2\sigma\sqrt{2(1+\alpha)\log m} \tag{31}$$

holds, then with probability exceeding (29), the thresholding algorithm also correctly identifies Λ_0 and satisfies (30).

The performance guarantee (30) is better than that provided by Theorem 2 and Corollary 1. However, this result comes at the expense of requirements on the magnitude of the entries of $\mathbf{x}_0$. Our analysis thus suggests that greedy approaches may outperform ℓ_1-based methods when the entries of $\mathbf{x}_0$ are large compared with the noise, but that the greedy approaches will deteriorate when the noise level increases. As we will see in Section V, simulations also appear to support this conclusion.

It is interesting to compare the success conditions (28) and (31) of the OMP and thresholding algorithms. For given problem dimensions, the OMP algorithm requires $|x_{\min}|$, the smallest nonzero element of $\mathbf{x}_0$, to be larger than a constant multiple of the noise standard deviation σ. This is required in order to ensure that all elements of the support of $\mathbf{x}_0$ will be identified with high probability. The requirement of the thresholding algorithm is stronger, as befits a simpler approach: In this case $|x_{\min}|$ must be larger than the noise standard deviation plus a constant times $|x_{\max}|$. In other words, one must be able to separate $|x_{\min}|$ from the combined effect of noise and interference caused by the other nonzero components of $\mathbf{x}_0$. This results from the thresholding technique, in which the entire support is identified simultaneously from the measurements. By comparison, the iterative approach used by OMP identifies and removes the large elements in $\mathbf{x}_0$ first, thus facilitating the identification of the smaller elements in later iterations.

V. Numerical Results

In this section, we describe a number of numerical experiments comparing the performance of various estimators to the guarantees of Section IV. Our first experiment measured the median estimation error, i.e., the median of the ℓ_2 distance between $\mathbf{x}_0$ and its estimate. The median error is intuitively appealing as it characterizes the "typical" estimation error, and it can be readily bounded by the performance guarantees of Section IV.

Specifically, we chose the two-ortho dictionary $\mathbf{A} = [\mathbf{I} \;\; \mathbf{H}]$, where $\mathbf{I}$ is the 512×512 identity matrix and $\mathbf{H}$ is the 512×512 Hadamard matrix with normalized columns. The RICs of this dictionary are unknown, but the coherence can be readily calculated and is given by $\mu = 1/\sqrt{512}$. Consequently, the theorems of Section IV can be used to obtain performance guarantees for sufficiently sparse vectors. In particular, in our simulations we chose parameters $\mathbf{x}_0$ having a support of size $s = 7$. The smallest nonzero entry in $\mathbf{x}_0$ was $|x_{\min}| = 0.1$ and the largest entry was $|x_{\max}| = 1$. Under these conditions, applying the theorems of Section IV yields the bounds[3]

$$\begin{aligned} \|\mathbf{x}_0 - \widehat{\mathbf{x}}_{\mathrm{OMP}}\|_2^2 &\leq 3.7 s\sigma^2 \log m \quad \text{w.p. } \tfrac{3}{4}, \text{ if } \sigma \leq 0.057; \\ \|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 &\leq 22.1 s\sigma^2 \log m \quad \text{w.p. } \tfrac{1}{2}; \\ \|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{DS}}\|_2^2 &\leq 361.8 s\sigma^2 \log m \quad \text{w.p. } \tfrac{3}{4}. \end{aligned} \tag{32}$$

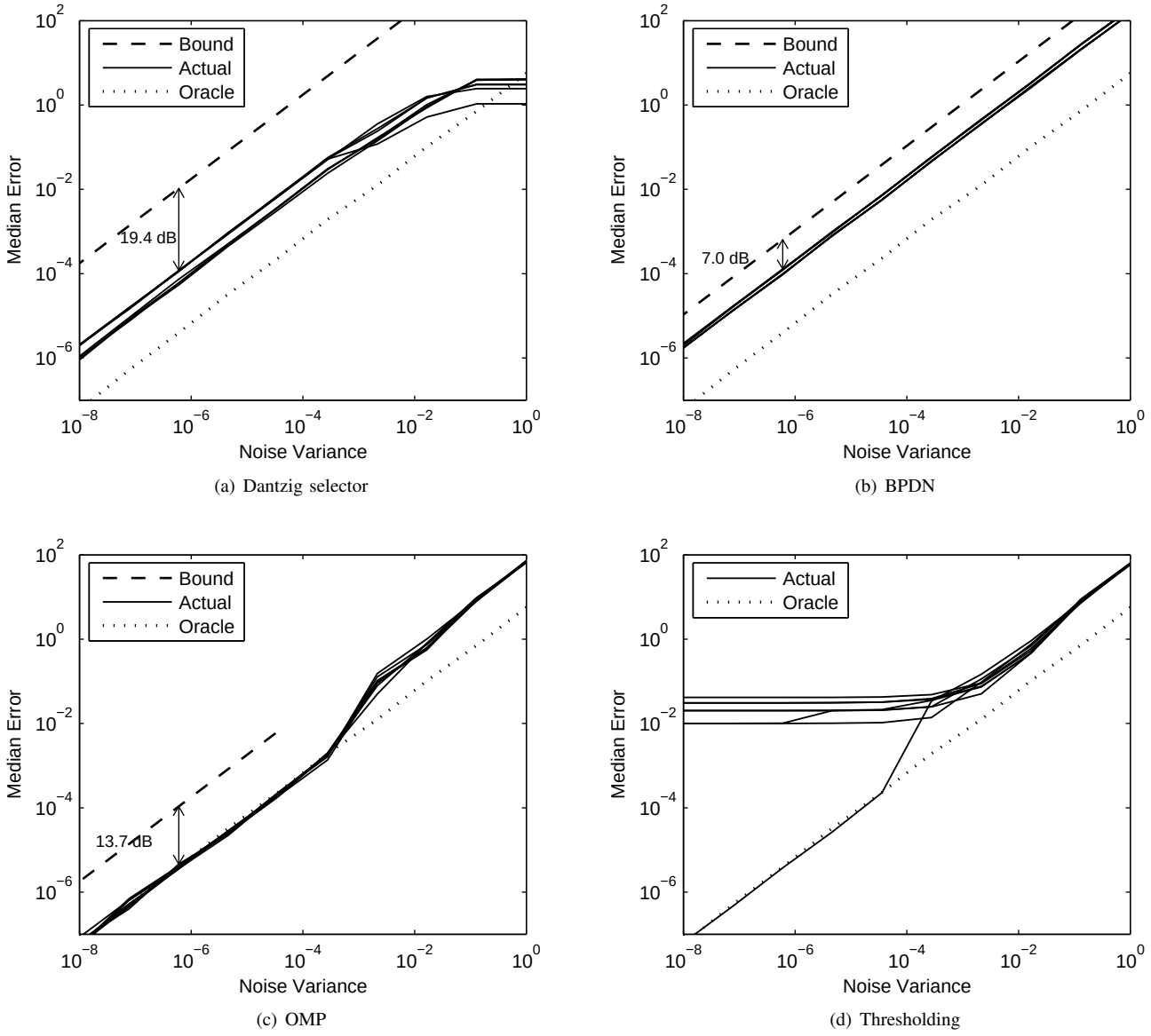
We have thus obtained guarantees for the median estimation error of the Dantzig selector, BPDN, and OMP. Under these settings, no guarantee can be made for the performance of the thresholding algorithm. Indeed, as we will see, for some choices of $\mathbf{x}_0$ satisfying the above requirements, the performance of the thresholding algorithm is not proportional to $s\sigma^2 \log m$. To obtain thresholding guarantees, one requires a narrower range between $|x_{\min}|$ and $|x_{\max}|$.

To measure the actual median error obtained by various estimators, 8 different parameter vectors $\mathbf{x}_0$ were selected. These differed in the distribution of the magnitudes of the nonzero components within the range $[|x_{\min}|, |x_{\max}|]$ and in the locations of the nonzero elements. For each parameter $\mathbf{x}_0$, a set of measurement vectors $\mathbf{b}$ were obtained from (1) by adding white Gaussian noise. The estimation algorithms of Section II-B were then applied to each measurement realization; for the Dantzig selector and BPDN, the parameters τ and γ were chosen as the smallest values such that the probabilities of success (16) and (23), respectively, would exceed $1/2$. The median over noise realizations of the distance $\|\mathbf{x}_0 - \widehat{\mathbf{x}}\|_2^2$ was then computed for each estimator. This process was repeated for 10 values of the noise variance σ^2 in the range $10^{-8} \leq \sigma^2 \leq 1$. The results are plotted in Fig. 1 as a function of σ^2.

It is evident from Fig. 1 that some parameter vectors are more difficult to estimate than others. Indeed, there is a large variety of parameters $\mathbf{x}_0$ satisfying the problem requirements, and it is likely that some of them come closer to the theoretical limits than the parameters chosen in our experiment. This highlights the importance of performance guarantees in ensuring adequate performance for *all* parameter values. On the other hand, it is quite possible that further improvements of the constants in the performance bounds are possible. For example, the Dantzig selector guarantee, which is obtained by applying coherence bounds to RIC-based results [2], is almost 100 times higher than the worst of the examined parameter values. It should also be noted that applying coherence bounds to RIC-based BPDN guarantees [8] yields a bound which applies to the aforementioned matrix $\mathbf{A}$ only when $s \leq 3$, and thus cannot be used in the present setting. Therefore, it appears that when dealing with dictionaries for which only the coherence μ is known, guarantees based directly on μ are tighter than RIC-based results.

In practice, it is more common to measure the MSE of an estimator than its median error. Our next goal is to determine whether the behavior predicted by our theoretical analysis is also manifested in the MSE of the various estimators. To this

[3] In the current setting, the results for the Dantzig selector (Theorem 2) and OMP (Theorem 4) can only be used to yield guarantees holding with probabilities of approximately $3/4$ and higher. These are, of course, also bounds on the median error.



(a) Dantzig selector

(b) BPDN

(c) OMP

(d) Thresholding

Fig. 1. Median estimation error for practical estimators (solid line) compared with the performance guarantees (dashed line) and the oracle estimator (dotted line). The solid lines report performance for 8 different values of the unknown parameter vector $\mathbf{x}_0$. For OMP, performance is only guaranteed for $\sigma \leq 0.057$, while for thresholding, nothing can be guaranteed for the given problem dimensions.

end, we conducted an experiment in which the MSEs of the estimators of Section II-B were compared. In this simulation, we chose the two-ortho dictionary $\mathbf{A} = [\mathbf{I} \;\; \mathbf{H}]$, where $\mathbf{I}$ is the 256×256 identity matrix and $\mathbf{H}$ is the 256×256 Hadamard matrix with normalized columns.[4] Once again, the RICs of this dictionary are unknown. However, the coherence in this case is given by $\mu = 1/16$, and consequently, the ℓ_1 relaxation guarantees of Section IV-A hold for $s \leq 5$.

We obtained the parameter vector $\mathbf{x}_0$ for this experiment by selecting a 5-element support at random, choosing the nonzero entries from a white Gaussian distribution, and then normalizing the resulting vector so that $\|\mathbf{x}_0\|_2 = 1$. The regularization parameters τ and γ of the Dantzig selector and BPDN were chosen as recommended by Theorem 2 and Corollary 1, respectively; for both estimators a value of $\alpha = 1$ was chosen, so that the guaranteed probability of success for the two algorithms has the same order of magnitude. The MSE of each estimate was then calculated by averaging over repeated realizations of $\mathbf{x}_0$ and the noise. The experiment was conducted for 10 values of the noise variance σ^2 and the results are plotted in Fig. 2 as a function of the signal-to-noise ratio (SNR), which is defined by

$$\text{SNR} = \frac{\|\mathbf{x}_0\|_2^2}{n\sigma^2} = \frac{1}{n\sigma^2}. \tag{33}$$

To compare this plot with the theoretical results of Section IV, observe first the situation at high SNR. In this case, OMP, BPDN, and the Dantzig selector all achieve performance

[4]Similar experiments were performed on a variety of other dictionaries, including an overcomplete DCT [24] and a matrix containing Gaussian random entries. The different dictionaries yielded comparable results, which are not reported here.

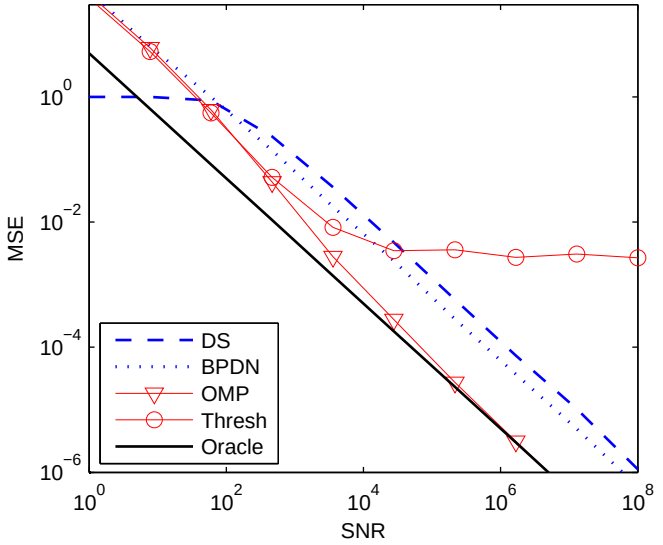


Fig. 2. MSE of various estimators as a function of the SNR. The sparsity level is $s = 5$ and the dictionary is a 256×512 two-ortho matrix.

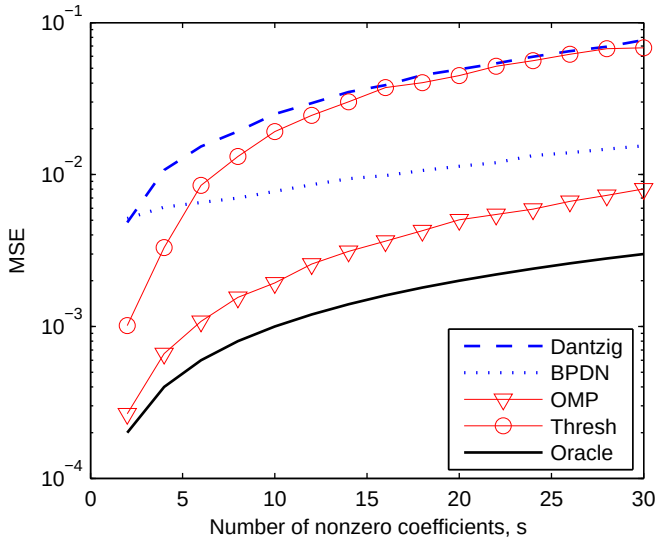


Fig. 3. MSE of various estimators as a function of the support size s. The noise standard deviation is $\sigma = 0.01$ and the dictionary is a 256×512 two-ortho matrix.

which is proportional to the oracle MSE (or CRB) given by (13). Among these, OMP is closest to the CRB, followed by BPDN and, finally, the Dantzig selector. This behavior matches the proportionality constants given in the theorems of Section IV. Indeed, for small σ, the condition (28) holds even for large α, and thus Theorem 4 guarantees that OMP will recover the correct support of $\mathbf{x}_0$ with high probability, explaining the convergence of this estimator to the oracle. By contrast, the performance of the thresholding algorithm levels off at high SNR; this is again predicted by Theorem 4, since, even when $\sigma = 0$, the condition (31) does not always hold, unless $|x_{\min}|$ is not much smaller than $|x_{\max}|$. Thus, for our choice of $\mathbf{x}_0$, Theorem 4 does not guarantee near-oracle performance for the thresholding algorithm, even at high SNR.

With increasing noise, Theorem 4 requires a corresponding increase in $|x_{\min}|$ to guarantee the success of the greedy algorithms. Consequently, Fig. 2 demonstrates a deterioration of these algorithms when the SNR is low. On the other hand, the theorems for the relaxation algorithms make no such assumptions, and indeed these approaches continue to perform well, compared with the oracle estimator, even when the noise level is high. In particular, the Dantzig selector outperforms the CRB at low SNR; this is because the CRB is a bound on unbiased techniques, whereas when the noise is large, biased techniques such as an ℓ_1 penalty become very effective. Robustness to noise is thus an important advantage of ℓ_1-relaxation techniques.

It is also interesting to examine the effect of the support size s on the performance of the various estimators. To this end, 15 support sizes in the range $2 \leq s \leq 30$ were tested. For each value of s, random vectors $\mathbf{x}_0$ having s nonzero entries were selected as in the previous simulation. The dictionary $\mathbf{A}$ was the 256×512 two-ortho matrix defined above; as in the previous experiment, other matrices were also tested and provided similar results. The standard deviation of the noise for this experiment was $\sigma = 0.01$. The results are plotted in Fig. 3.

As mentioned above, the mutual coherence of the dictionary $\mathbf{A}$ is $1/16$, so that the proposed performance guarantees apply only when $\mathbf{x}_0$ is quite sparse ($s \leq 5$). Nevertheless, Fig. 3 demonstrates that the estimation algorithms (with the exception of the thresholding approach) exhibit a graceful degradation as the support of $\mathbf{x}_0$ increases. At first sight this would appear to mean that the performance guarantees provided are overly pessimistic. For example, it is possible that the RICs in the present setting, while unknown, are fairly low and permit a stronger analysis than that of Section IV. It is also quite reasonable to expect, as mentioned above, that some improvement in the theoretical guarantees is possible. However, it is worth recalling that the performance guarantees proposed in this paper apply to all sparse vectors, while the numerical results describe the performance averaged over different values of $\mathbf{x}_0$. Thus it is possible that there exist particular parameter values for which the performance is considerably poorer than that reported in Fig. 3. Indeed, there exist values of $\mathbf{A}$ and $\mathbf{x}_0$ for which BPDN yields grossly incorrect results even when $\|\mathbf{x}_0\|_0$ is on the order of $1/\mu$ [13]. However, identifying such worst-case parameters numerically is quite difficult; this is doubtlessly at least part of the reason for the apparent pessimism of the performance guarantees.

VI. Conclusion

The performance of an estimator depends on the problem setting under consideration. As we have seen, under the adversarial noise scenario of Section III, the estimation error of any algorithm can be as high as the noise power; in other words, the assumption of sparsity has not yielded any denoising effect. On the other hand, in the Bayesian regime in which both $\mathbf{x}_0$ and the noise vector are random, practical estimators come close to the performance of the oracle estimator [13]. In Section IV, we examined a middle ground between these two extremes, namely the setting in which $\mathbf{x}_0$ is deterministic but the noise is random. As we have shown, despite the fact

that less is known about $\mathbf{x}_0$ in this case than in the Bayesian scenario, a variety of estimation techniques are still guaranteed to achieve performance close to that of the oracle estimator.

Our theoretical and numerical results suggest some conclusions concerning the choice of an estimator. In particular, at high SNR values, it appears that the greedy OMP algorithm has an advantage over the other algorithms considered herein. In this case the support set of $\mathbf{x}_0$ can be recovered accurately and OMP thus converges to the oracle estimator; by contrast, ℓ_1 relaxations have a shrinkage effect which causes a loss of accuracy at high SNR. This is of particular interest since greedy algorithms are also computationally more efficient than relaxation methods. On the other hand, the ℓ_1 relaxation techniques, and particularly the Dantzig selector, appear to be more effective than the greedy algorithms when the noise level is significant: in this case, shrinkage is a highly effective denoising technique. Indeed, as a result of the bias introduced by the shrinkage, ℓ_1-based approaches can even perform better than the oracle estimator and the Cramér–Rao bound.

Appendix A
Proof of Lemma 1

By Gershgorin's disc theorem [21], all eigenvalues of $\mathbf{A}_\Lambda^T\mathbf{A}_\Lambda$ are between $1-(s-1)\mu$ and $1+(s-1)\mu$. Combining this with the fact that, for all $\mathbf{y}$,

$$\lambda_{\min}(\mathbf{A}_\Lambda^T\mathbf{A}_\Lambda)\|\mathbf{y}\|_2^2 \le \|\mathbf{A}_\Lambda\mathbf{y}\|_2^2 \le \lambda_{\max}(\mathbf{A}_\Lambda^T\mathbf{A}_\Lambda)\|\mathbf{y}\|_2^2, \quad (34)$$

we obtain (5). Next, to demonstrate (6), observe that

$$\left|\mathbf{y}_1^T\mathbf{A}_{\Lambda_1}^T\mathbf{A}_{\Lambda_2}\mathbf{y}_2\right| \le \left|\mathbf{y}_1^T\right| \cdot \left|\mathbf{A}_{\Lambda_1}^T\mathbf{A}_{\Lambda_2}\right| \cdot |\mathbf{y}_2| \quad (35)$$

where the absolute value of a matrix or vector is taken elementwise. Since $\mathbf{A}_{\Lambda_1}^T\mathbf{A}_{\Lambda_2}$ is a submatrix of $\mathbf{A}^T\mathbf{A}$ which does not contain any of the diagonal elements of $\mathbf{A}^T\mathbf{A}$, it follows that each element in $\mathbf{A}_{\Lambda_1}^T\mathbf{A}_{\Lambda_2}$ is smaller in absolute value than μ. Thus

$$\left|\mathbf{y}_1^T\mathbf{A}_{\Lambda_1}^T\mathbf{A}_{\Lambda_2}\mathbf{y}_2\right| \le \mu\left|\mathbf{y}_1^T\right|\mathbb{1}\mathbb{1}^T|\mathbf{y}_2| = \mu\|\mathbf{y}_1\|_1\|\mathbf{y}_2\|_1 \quad (36)$$

where $\mathbb{1}$ indicates a vector of ones. Using the fact that $\|\mathbf{y}\|_1 \le \sqrt{s}\|\mathbf{y}\|_2$ for any s-vector $\mathbf{y}$, we obtain

$$\left|\mathbf{y}_1^T\mathbf{A}_{\Lambda_1}^T\mathbf{A}_{\Lambda_2}\mathbf{y}_2\right| \le \mu\sqrt{s_1 s_2}\|\mathbf{y}_1\|_2\|\mathbf{y}_2\|_2, \quad (37)$$

which implies that θ_{s_1,s_2} satisfies (6).

Appendix B
Proof of Theorem 3

The proof is based closely on the work of Tropp [7]. From the triangle inequality,

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm BP}\|_2 \le \|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm or}\|_2 + \|\hat{\mathbf{x}}_{\rm or} - \hat{\mathbf{x}}_{\rm BP}\|_2 \quad (38)$$

where $\hat{\mathbf{x}}_{\rm or}$ is the oracle estimator. Our goal is to separately bound the two terms on the right-hand side of (38). Indeed, as we will see, the two constants $\sigma\sqrt{3}$ and $\frac{3}{2}\gamma$ in (21) arise, respectively, from the two terms in (38).

Beginning with the term $\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm or}\|_2$, let $\mathbf{x}_{0,\Lambda}$ denote the s-vector containing the elements of $\mathbf{x}_0$ indexed by Λ_0, and similarly, let $\hat{\mathbf{x}}_{{\rm or},\Lambda}$ denote the corresponding subvector of $\hat{\mathbf{x}}_{\rm or}$. We then have

$$\begin{aligned}\mathbf{x}_{0,\Lambda} - \hat{\mathbf{x}}_{{\rm or},\Lambda} &= \mathbf{x}_{0,\Lambda} - \mathbf{A}_{\Lambda_0}^\dagger(\mathbf{A}\mathbf{x}_0 + \mathbf{w}) \\ &= \mathbf{x}_{0,\Lambda} - \mathbf{A}_{\Lambda_0}^\dagger(\mathbf{A}_{\Lambda_0}\mathbf{x}_{0,\Lambda} + \mathbf{w}) \\ &= -\mathbf{A}_{\Lambda_0}^\dagger\mathbf{w}, \end{aligned} \quad (39)$$

where we have used the fact that $\mathbf{A}_{\Lambda_0}$ has full column rank, which is a consequence [25] of the condition (19). Thus, $\mathbf{x}_{0,\Lambda} - \hat{\mathbf{x}}_{{\rm or},\Lambda}$ is a Gaussian random vector with mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{A}_{\Lambda_0}^\dagger\mathbf{A}_{\Lambda_0}^{\dagger T} = \sigma^2(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0})^{-1}$.

For future use, we note that the cross-correlation between $\mathbf{A}_{\Lambda_0}^\dagger\mathbf{w}$ and $(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^\dagger)\mathbf{w}$ is

$$\begin{aligned}E\left\{\mathbf{A}_{\Lambda_0}^\dagger\mathbf{w}\mathbf{w}^T(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^\dagger)^T\right\} &= \sigma^2\mathbf{A}_{\Lambda_0}^\dagger(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^\dagger)^T \\ &= \mathbf{0}, \end{aligned} \quad (40)$$

where we have used the fact [26, Th. 1.2.1] that for any matrix $\mathbf{M}$

$$\mathbf{M}^\dagger\mathbf{M}^{\dagger T}\mathbf{M}^T = (\mathbf{M}^T\mathbf{M})^\dagger\mathbf{M}^T = \mathbf{M}^\dagger. \quad (41)$$

Since $\mathbf{w}$ is Gaussian, it follows that $\mathbf{A}_{\Lambda_0}^\dagger\mathbf{w}$ and $(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^\dagger)\mathbf{w}$ are statistically independent. Furthermore, because $\mathbf{x}_{0,\Lambda} - \hat{\mathbf{x}}_{{\rm or},\Lambda}$ depends on $\mathbf{w}$ only through $\mathbf{A}_{\Lambda_0}^\dagger\mathbf{w}$, we conclude that

$$\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm or} \text{ is statistically independent of } (\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^\dagger)\mathbf{w}. \quad (42)$$

We now wish to bound the probability that $\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm or}\|_2^2 > 3s\sigma^2$. Let $\mathbf{z}$ be a normalized Gaussian random variable, $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I}_s)$. Then

$$\begin{aligned}&\Pr\left\{\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm or}\|_2^2 > 3s\sigma^2\right\} \\ &= \Pr\left\{\left\|\sigma(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0})^{-1/2}\mathbf{z}\right\|_2^2 \ge 3s\sigma^2\right\} \\ &\le \Pr\left\{\left\|(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0})^{-1/2}\right\|^2\|\mathbf{z}\|_2^2 \ge 3s\right\} \end{aligned} \quad (43)$$

where $\|\mathbf{M}\|$ denotes the maximum singular value of the matrix $\mathbf{M}$. Thus, $\|(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0})^{-1/2}\| = 1/s_{\min}$, where $s_{\min}$ is the minimum singular value of $\mathbf{A}_{\Lambda_0}$. From the Gershgorin disc theorem [21, p. 320], it follows that $s_{\min} \ge \sqrt{1-(s-1)\mu}$. Using (19), this can be simplified to $s_{\min} \ge \sqrt{2/3}$, and therefore

$$\left\|(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0})^{-1/2}\right\| \le \sqrt{\frac{3}{2}}. \quad (44)$$

Combining with (43) yields

$$\Pr\left\{\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm or}\|_2^2 > 3s\sigma^2\right\} \le \Pr\left\{\|\mathbf{z}\|_2^2 \ge 2s\right\}. \quad (45)$$

Observe that $\|\mathbf{z}\|_2^2$ is the sum of s independent normalized Gaussian random variables. The right-hand side of (45) is therefore $1-F_{\chi_s^2}(2s)$, where $F_{\chi_s^2}(\cdot)$ is the cumulative distribution function of the χ^2 distribution with s degrees of freedom. Using the formula [27, §16.3] for $F_{\chi_s^2}(\cdot)$, we have

$$\Pr\left\{\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm or}\|_2^2 > 3s\sigma^2\right\} \le Q\left(\tfrac{1}{2}s, s\right) \quad (46)$$

where $Q(a,z)$ is the regularized Gamma function

$$Q(a,z) \triangleq \frac{\int_z^\infty t^{a-1}e^{-t}dt}{\int_0^\infty t^{a-1}e^{-t}dt}. \quad (47)$$

$Q\big(\frac{1}{2}s, s\big)$ decays exponentially as $s \to \infty$, and it can be seen that

$$Q\big(\tfrac{1}{2}s, s\big) < e^{-s/7} \quad \text{for all } s. \tag{48}$$

We thus conclude that the event

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm or}\|_2^2 \le 3s\sigma^2 \tag{49}$$

occurs with probability no smaller than $1 - e^{-s/7}$. Note that the same technique can be applied to obtain bounds on the probability that $\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\rm or}\|_2^2 > \alpha s\sigma^2$, for any $\alpha > \frac{2}{3}$. The only difference will be the rate of exponential decay in (48). However, the distance between $\mathbf{x}_0$ and $\hat{\mathbf{x}}_{\rm or}$ is usually small compared with the distance between $\hat{\mathbf{x}}_{\rm or}$ and $\hat{\mathbf{x}}_{\rm BP}$, so that such an approach does not significantly affect the overall result.

The above calculations provided a bound on the first term in (38). To address the second term $\|\hat{\mathbf{x}}_{\rm or} - \hat{\mathbf{x}}_{\rm BP}\|_2$, define the random event

$$G: \ \max_i \left|\mathbf{a}_i^T(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^{\dagger})\mathbf{b}\right| \le \tfrac{1}{2}\gamma \tag{50}$$

where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$. It is shown in [7, App. IV-A] that

$$\Pr\{G\} \ge 1 - (m-s)\exp\left(-\frac{\gamma^2}{8\sigma^2}\right). \tag{51}$$

If G indeed occurs, then the portion of the measurements $\mathbf{b}$ which do not belong to the range space of $\mathbf{A}_{\Lambda_0}$ are small, and consequently it has been shown [7, Cor. 9] that, in this case, the solution $\hat{\mathbf{x}}_{\rm BP}$ to (7) is unique, the support of $\hat{\mathbf{x}}_{\rm BP}$ is a subset of Λ_0, and

$$\|\hat{\mathbf{x}}_{\rm BP} - \hat{\mathbf{x}}_{\rm or}\|_\infty \le \tfrac{3}{2}\gamma. \tag{52}$$

Since both $\hat{\mathbf{x}}_{\rm BP}$ and $\hat{\mathbf{x}}_{\rm or}$ are nonzero only in Λ_0, this implies that

$$\|\hat{\mathbf{x}}_{\rm BP} - \hat{\mathbf{x}}_{\rm or}\|_2 \le \tfrac{3}{2}\gamma\sqrt{s}. \tag{53}$$

The event G depends on the random variable $\mathbf{w}$ only through $(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^{\dagger})\mathbf{w}$. Thus, it follows from (42) that G is statistically independent of the event (43). The probability that both events occur simultaneously is therefore given by the product of their respective probabilities. In other words, with probability exceeding (20), both (53) and (49) hold. Using (38) completes the proof of the theorem.

Appendix C
Proof of Theorem 4

The claims concerning both algorithms are closely related. To emphasize this similarity, we first provide several lemmas which will be used to prove both results. These lemmas are all based on an analysis of the random event

$$B = \left\{\max_{1\le i\le m} |\mathbf{a}_i^T\mathbf{w}| < \tau\right\} \tag{54}$$

where

$$\tau \triangleq \sigma\sqrt{2(1+\alpha)\log m} \tag{55}$$

and $\alpha > 0$. Our proof will be based on demonstrating that B occurs with high probability, and that when B does occur, both thresholding and OMP achieve near-oracle performance.

Lemma 2: Suppose that $\mathbf{w} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Then, the event B of (54) occurs with a probability of at least (29).

Proof: The random variables $\{\mathbf{a}_i^T\mathbf{w}\}_{i=1}^m$ are jointly Gaussian. Therefore, by Šidák's lemma [28, Th. 1]

$$\Pr\{B\} = \Pr\left\{\max_{1\le i\le m} |\mathbf{a}_i^T\mathbf{w}| < \tau\right\} \ge \prod_{i=1}^m \Pr\left\{|\mathbf{a}_i^T\mathbf{w}| \le \tau\right\}. \tag{56}$$

Since $\|\mathbf{a}_i\|_2 = 1$, each random variable $\mathbf{a}_i^T\mathbf{w}$ has mean zero and variance σ^2. Consequently,

$$\Pr\left\{|\mathbf{a}_i^T\mathbf{w}| < \tau\right\} = 1 - 2Q\Big(\frac{\tau}{\sigma}\Big) \tag{57}$$

where $Q(x) = (1/\sqrt{2\pi})\int_x^\infty e^{-z^2/2}dz$ is the Gaussian tail probability. Using the bound

$$Q(x) \le \frac{1}{x\sqrt{2\pi}}e^{-x^2/2} \tag{58}$$

we obtain from (57)

$$\Pr\left\{|\mathbf{a}_i^T\mathbf{w}| < \tau\right\} \ge 1 - \eta \tag{59}$$

where

$$\eta \triangleq \sqrt{\frac{2}{\pi}}\cdot\frac{\sigma}{\tau}e^{-\tau^2/2\sigma^2}. \tag{60}$$

When $\eta > 1$, the bound (29) is meaningless and the theorem holds vacuously. Otherwise, when $\eta \le 1$, we have from (56) and (59)

$$\Pr\{B\} \ge (1-\eta)^m \ge 1 - m\eta \tag{61}$$

where the final inequality holds for any $\eta \le 1$ and $m \ge 1$. Substituting the values of η and τ and simplifying, we obtain that B holds with a probability no lower than (29), as required. ■

The next lemma demonstrates that, under suitable conditions, correlating $\mathbf{b}$ with the dictionary atoms $\mathbf{a}_i$ is an effective method of identifying the atoms participating in the support of $\mathbf{x}_0$.

Lemma 3: Let $\mathbf{x}_0$ be a vector with support $\Lambda_0 = \mathrm{supp}(\mathbf{x}_0)$ of size $s = |\Lambda_0|$, and let $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$ for some noise vector $\mathbf{w}$. Define $|x_{\min}|$ and $|x_{\max}|$ as in (27), and suppose that

$$|x_{\max}| - (2s-1)\mu|x_{\max}| \ge 2\tau. \tag{62}$$

Then, if the event B of (54) holds, we have

$$\max_{j\in\Lambda_0} |\mathbf{a}_j^T\mathbf{b}| > \max_{j\notin\Lambda_0} |\mathbf{a}_j^T\mathbf{b}|. \tag{63}$$

If, rather than (62), the stronger condition

$$|x_{\min}| - (2s-1)\mu|x_{\max}| \ge 2\tau \tag{64}$$

is given, then, under the event B, we have

$$\min_{j\in\Lambda_0} |\mathbf{a}_j^T\mathbf{b}| > \max_{j\notin\Lambda_0} |\mathbf{a}_j^T\mathbf{b}|. \tag{65}$$

Proof: The proof is an adaptation of [4, Lemma 5.2]. Beginning with the term $\max_{j\notin\Lambda_0} |\mathbf{a}_j^T\mathbf{b}|$, we have, under the

event B,

$$\begin{aligned}\max_{j\notin\Lambda_0}|\mathbf{a}_j^T\mathbf{b}| &= \max_{j\notin\Lambda_0}\left|\mathbf{a}_j^T\mathbf{w}+\sum_{i\in\Lambda_0}x_i\mathbf{a}_j^T\mathbf{a}_i\right| \\ &\le \max_{j\notin\Lambda_0}|\mathbf{a}_j^T\mathbf{w}|+\max_{j\notin\Lambda_0}\sum_{i\in\Lambda_0}\left|x_i\mathbf{a}_j^T\mathbf{a}_i\right| \\ &< \tau+s\mu|x_{\max}|. \end{aligned} \tag{66}$$

On the other hand, when B holds,

$$\begin{aligned}\max_{j\in\Lambda_0}|\mathbf{a}_j^T\mathbf{b}| &= \max_{j\in\Lambda_0}\left|x_j+\mathbf{a}_j^T\mathbf{w}+\sum_{i\in\Lambda_0\setminus\{j\}}x_i\mathbf{a}_j^T\mathbf{a}_i\right| \\ &\ge |x_{\max}|-\max_{j\in\Lambda_0}\left|\mathbf{a}_j^T\mathbf{w}+\sum_{i\in\Lambda_0\setminus\{j\}}x_i\mathbf{a}_j^T\mathbf{a}_i\right| \\ &> |x_{\max}|-\tau-(s-1)\mu|x_{\max}| \\ &= |x_{\max}|-(2s-1)\mu|x_{\max}|-\tau+s\mu|x_{\max}|. \end{aligned} \tag{67}$$

Together with (66), this yields

$$\max_{j\in\Lambda_0}|\mathbf{a}_j^T\mathbf{b}| > |x_{\max}|-(2s-1)\mu|x_{\max}|-2\tau+\max_{j\notin\Lambda_0}|\mathbf{a}_j^T\mathbf{b}|. \tag{68}$$

Thus, under the condition (62), we obtain (63). Similarly, when B holds, we have

$$\begin{aligned}\min_{j\in\Lambda_0}\left|\mathbf{a}_j^T\mathbf{b}\right| &= \min_{j\in\Lambda_0}\left|x_j+\mathbf{a}_j^T\mathbf{w}+\sum_{i\in\Lambda_0\setminus\{j\}}x_i\mathbf{a}_j^T\mathbf{a}_i\right| \\ &> |x_{\min}|-\tau-(s-1)\mu|x_{\max}| \\ &= |x_{\min}|-(2s-1)\mu|x_{\max}|-\tau+s\mu|x_{\max}|. \end{aligned} \tag{69}$$

Again using (66), we obtain

$$\min_{j\in\Lambda_0}\left|\mathbf{a}_j^T\mathbf{b}\right| > |x_{\min}|-(2s-1)\mu|x_{\max}|-2\tau+\max_{j\notin\Lambda_0}|\mathbf{a}_j^T\mathbf{b}|. \tag{70}$$

Consequently, under the assumption (64), we conclude that (65) holds, as required. ∎

The following lemma bounds the performance of the oracle estimator under the event B. The usefulness of this lemma stems from the fact that, if either OMP or the thresholding algorithm correctly identify the support of $\mathbf{x}_0$, then their estimate is identical to that of the oracle.

Lemma 4: Let $\mathbf{x}_0$ be a vector with support $\Lambda_0=\mathrm{supp}(\mathbf{x}_0)$, and let $\mathbf{b}=\mathbf{A}\mathbf{x}_0+\mathbf{w}$ for some noise vector $\mathbf{w}$. If the event B of (54) occurs, then

$$\|\hat{\mathbf{x}}_{\mathrm{or}}-\mathbf{x}_0\|_2^2 \le 2s\sigma^2(1+\alpha)\log m\frac{1}{(1-(s-1)\mu)^2}. \tag{71}$$

Proof: Note that both $\hat{\mathbf{x}}_{\mathrm{or}}$ and $\mathbf{x}_0$ are supported on Λ_0, and therefore

$$\|\hat{\mathbf{x}}_{\mathrm{or}}-\mathbf{x}_0\|_2^2 = \|\mathbf{A}_{\Lambda_0}^{\dagger}\mathbf{b}-\mathbf{x}_{0,\Lambda_0}\|_2^2 \tag{72}$$

where $\mathbf{x}_{0,\Lambda_0}$ is the subvector of nonzero entries of $\mathbf{x}_0$. We thus have, under the event B,

$$\begin{aligned}\|\hat{\mathbf{x}}_{\mathrm{or}}-\mathbf{x}_0\|_2^2 &= \|\mathbf{A}_{\Lambda_0}^{\dagger}\mathbf{A}_{\Lambda_0}\mathbf{x}_{0,\Lambda_0}+\mathbf{A}_{\Lambda_0}^{\dagger}\mathbf{w}-\mathbf{x}_{0,\Lambda_0}\|_2^2 \\ &= \|\mathbf{A}_{\Lambda_0}^{\dagger}\mathbf{w}\|_2^2 \\ &= \left\|(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0})^{-1}\mathbf{A}_{\Lambda_0}^T\mathbf{w}\right\|_2^2 \\ &\le \left\|(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0})^{-1}\right\|^2\sum_{i\in\Lambda_0}(\mathbf{a}_i^T\mathbf{w})^2 \\ &\le \frac{1}{(1-(s-1)\mu)^2}s\sigma^2 2(1+\alpha)\log m \end{aligned} \tag{73}$$

where, in the last step, we used the definition (54) of B and the fact that $\|\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0}\|\ge 1-(s-1)\mu$, which was demonstrated in Appendix B. This completes the proof the lemma. ∎

We are now ready to prove Theorem 4. The proof for the thresholding algorithm is obtained by combining the three lemmas presented above. Indeed, Lemma 2 ensures that the event B occurs with probability at least as high as the required probability of success (29). Whenever B occurs, we have by Lemma 3 that the atoms corresponding to Λ_0 all have strictly higher correlation with $\mathbf{b}$ than the off-support atoms, so that the thresholding algorithm identifies the correct support Λ_0, and is thus equivalent to the oracle estimator $\hat{\mathbf{x}}_{\mathrm{or}}$ as long as B holds. Finally, by Lemma 4, identification of the true support Λ_0 guarantees the required error (30).

We now prove the OMP performance guarantee. Our aim is to show that when B occurs, OMP correctly identifies the support of $\mathbf{x}_0$; the result then follows by Lemmas 2 and 4. To this end we employ the technique used in the proof of [4, Th. 5.1]. We begin by examining the first iteration of the OMP algorithm, in which one identifies the atom $\mathbf{a}_i$ whose correlation with $\mathbf{b}$ is maximal. Note that (28) implies (62), and therefore, by Lemma 3, the atom having the highest correlation with $\mathbf{b}$ corresponds to an element in the support Λ_0 of $\mathbf{x}_0$. Consequently, the first step of the OMP algorithm correctly identifies an element in Λ_0.

The proof now continues by induction. Suppose we are currently in the ith iteration of OMP, with $1<i\le s$, and assume that atoms from the correct support were identified in all $i-1$ previous steps. Referring to the notation used in the definition of OMP in Section II-B, this implies that $\mathrm{supp}(\widehat{\mathbf{x}}_{\mathrm{OMP}}^{i-1})=\Lambda^{i-1}\subset\Lambda_0$. The ith step consists of identifying the atom $\mathbf{a}_i$ which is maximally correlated with the residual $\mathbf{r}^i$. By the definition of $\mathbf{r}^i$, we have

$$\mathbf{r}^i = \mathbf{A}\tilde{\mathbf{x}}^{i-1}+\mathbf{w} \tag{74}$$

where $\tilde{\mathbf{x}}^{i-1}=\mathbf{x}_0-\widehat{\mathbf{x}}_{\mathrm{OMP}}^{i-1}$. Thus $\mathrm{supp}(\tilde{\mathbf{x}}^{i-1})\subseteq\Lambda_0$, so that $\mathbf{r}^i$ is a noisy measurement of the vector $\mathbf{A}\tilde{\mathbf{x}}^{i-1}$, which has a sparse representation consisting of no more than s atoms. Now, since

$$\|\widehat{\mathbf{x}}_{\mathrm{OMP}}^{i-1}\|_0 = i-1 < s = \|x_0\|_0, \tag{75}$$

it follows that at least one nonzero entry in $\tilde{\mathbf{x}}^{i-1}$ is equal to the corresponding entry in $\mathbf{x}_0$. Consequently

$$\max_i|\tilde{x}_i^{i-1}|\ge|x_{\min}|. \tag{76}$$

Note that the model (74) is precisely of the form (1), with $\mathbf{r}^i$ taking the place of the measurements $\mathbf{b}$ and $\tilde{\mathbf{x}}^{i-1}$ taking the place of the sparse vector $\mathbf{x}_0$. It follows from (76) and (28) that this model satisfies the requirement (62). Consequently, by Lemma 3, we have that under the event B,

$$\max_{i\in\Lambda_0} |\mathbf{a}_i^T \mathbf{r}^i| > \max_{i\notin\Lambda_0} |\mathbf{a}_i^T \mathbf{r}^i|. \tag{77}$$

Therefore, the ith iteration of OMP will choose an element within Λ_0 to add to the support. By induction it follows that the first s steps of OMP all identify elements in Λ_0, and since OMP never chooses the same element twice, the entire support Λ_0 will be identified after s iterations. This completes the proof of Theorem 4.

REFERENCES

[1] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Asilomar Conf. Signals, Systems, and Computers*, Nov. 1993, pp. 40–44.

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

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, pp. 33–61, 1998.

[4] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, 2006.

[5] E. J. Candès, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Comm. Pure Appl. Math.*, vol. LIX, pp. 1207–1223, 2006.

[6] J. J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Trans. Inf. Theory*, vol. 51, no. 10, pp. 3601–3608, 2005.

[7] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals in noise," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, 2006.

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

[9] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice Hall, 1993.

[10] Z. Ben-Haim and Y. C. Eldar, "The Cramér–Rao bound for sparse estimation," *IEEE Trans. Signal Process.*, submitted. [Online]. Available: http://arxiv.org/abs/0905.4378

[11] E. J. Candès, "Modern statistical estimation via oracle inequalities," *Acta Numerica*, pp. 1–69, 2006.

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

[13] E. J. Candès and Y. Plan, "Near-ideal model selection by ℓ_1 minimization," *Ann. Statist.*, vol. 37, no. 5A, pp. 2145–2177, Oct. 2009.

[14] T. Cai, L. Wang, and G. Xu, "Stable recovery of sparse signals and an oracle inequality," U. Penn., Tech. Rep., 2009. [Online]. Available: http://www-stat.wharton.upenn.edu/~tcai/paper/Stable-Recovery-MIP.pdf

[15] Z. Ben-Haim, Y. C. Eldar, and M. Elad, "Coherence-based near-oracle performance guarantees for sparse estimation under Gaussian noise," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, Mar. 2010, submitted.

[16] K. Schnass and P. Vandergheynst, "Average performance analysis for thresholding," *IEEE Signal Process. Lett.*, vol. 14, no. 11, pp. 828–831, Nov. 2007.

[17] J. A. Tropp, "Norms of random submatrices and sparse approximation," *C. R. Acad. Sci. Paris, Ser. I*, vol. 346, no. 23-24, pp. 1271–1274, 2008.

[18] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, to appear. [Online]. Available: http://arxiv.org/abs/0904.0494

[19] M. J. Wainwright, "Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[20] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, vol. 346, pp. 589–592, 2008. [Online]. Available: http://www.acm.caltech.edu/~emmanuel/papers/RIP.pdf

[21] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins University Press, 1996.

[22] B. Efron, T. Hastie, and R. Tibshirani, "Discussion: The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2358–2364, 2007.

[23] E. Candes and T. Tao, "Rejoinder: The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2392–2404, 2007.

[24] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.

[25] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization," *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 5, pp. 2197–2202, March 4, 2003.

[26] S. L. Campbell and C. D. Meyer, Jr., *Generalized Inverses of Linear Transformations*. London, UK: Pitman, 1979.

[27] A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics*, 6th ed. London: Edward Arnold, 1994, vol. 1.

[28] Z. Šidák, "Rectangular confidence regions for the means of multivariate normal distributions," *J. Amer. Statist. Assoc.*, vol. 62, no. 318, pp. 626–633, Jun. 1967.