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

Zvika Ben-Haim, *Student Member, IEEE*, Yonina C. Eldar, *Senior Member, IEEE*, and
Michael Elad, *Senior Member, IEEE*

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 provide theoretical performance guarantees for three sparse estimation algorithms: basis pursuit denoising (BPDN), orthogonal matching pursuit (OMP), and thresholding. The performance of these techniques is quantified as the ℓ_2 distance between the estimate and the true value of $\mathbf{x}_0$. We demonstrate that, with high probability, the analyzed algorithms come close to the behavior of the oracle estimator, which knows the locations of the nonzero elements in $\mathbf{x}_0$. Our results are non-asymptotic and are based only on the coherence of $\mathbf{A}$, so that they are applicable to arbitrary dictionaries. This provides insight on the advantages and drawbacks of ℓ_1 relaxation techniques such as BPDN and the Dantzig selector, as opposed to greedy approaches such as OMP and thresholding.

Index Terms—Basis pursuit, Dantzig selector, matching pursuit, oracle, sparse estimation, thresholding algorithm.

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 [1]. 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 impasse 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) [2], and ℓ_1 relaxation methods, such as the Dantzig selector [3] and basis pursuit denoising (BPDN) [4], [5] (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 [6]–[8]. 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 [3], [9], [10]. In this paper, we focus on the random noise scenario. 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) [11]. 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 $\mathbf{x}_0$ [12]. 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 [3] 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.* [10] 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 [13].

The contributions [3], [10] 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 [14]. In this respect, coherence-based results are appealing since they can be used with arbitrary dictionaries [15]–[17].

In this paper, we seek performance guarantees for sparse estimators based directly on the mutual coherence of the matrix $\mathbf{A}$ [18]. 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. In the case of BPDN, our result provides a tighter guarantee than the coherence-based implications of the work of Bickel *et al.* [10]. 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 [19]–[23], 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 [24], [25].

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, our main results are presented in the form of performance guarantees for various sparse estimation algorithms. Subsequently, in Section IV, these results are compared with previous findings for different sparse models, including the case of adversarial noise and the random design matrix setting. Finally, in Section V, the validity of our results is examined by simulation in practical estimation scenarios.

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 $\mathrm{supp}(\mathbf{x})$. 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 identity matrix is denoted by $\mathbf{I}$. Furthermore, for any matrix $\mathbf{A}$, $\mathbf{A}^T$ refers to the matrix transpose, $\mathbf{A}^\dagger$ is the Moore-Penrose pseudoinverse, $\mathrm{Tr}(\mathbf{A})$ is the trace, and $\mathrm{span}(\mathbf{A})$ is the column span, while $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ denote the smallest and largest eigenvalues of $\mathbf{A}$. Finally, 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 $\Lambda_0 = \mathrm{supp}(\mathbf{x}_0)$. 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

$$\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w} \qquad (1)$$

where $\mathbf{A} \in \mathbb{R}^{n \times m}$ is a known deterministic 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 such that

$$\|\mathbf{a}_i\|_2 = 1. \qquad (2)$$

We further assume that $\mathbf{w}$ is zero-mean white Gaussian noise with covariance $E\left\{\mathbf{w}\mathbf{w}^T\right\} = \sigma^2 \mathbf{I}_{m \times m}$.

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 [14], the cumulative coherence [9], the exact recovery coefficient (ERC) [9], the spark [6], and the RICs [3], [7]. 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

$$\mu \triangleq \max_{i \neq j} \left|\mathbf{a}_i^T \mathbf{a}_j\right|. \qquad (3)$$

While the mutual coherence can be efficiently calculated directly from (3), 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 [9] and the spark [6], [17], [22]. 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 [7] of order s with parameter δ_s if, for every index set Λ of size s, we have

$$(1 - \delta_s)\|\mathbf{y}\|_2^2 \leq \|\mathbf{A}_\Lambda \mathbf{y}\|_2^2 \leq (1 + \delta_s)\|\mathbf{y}\|_2^2 \qquad (4)$$

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 [3] 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

$$\left| \mathbf{y}_1^T \mathbf{A}_{\Lambda_1}^T \mathbf{A}_{\Lambda_2} \mathbf{y}_2 \right| \leq \theta_{s_1,s_2} \|\mathbf{y}_1\|_2 \|\mathbf{y}_2\|_2 \qquad (5)$$

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 [3], [7], [26], [27]. 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 (3) 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, whose proof can be found in [28] and [29].

Lemma 1 (Cai, Xu, and Zhang): For any matrix $\mathbf{A}$, the RIP constant δ_s of (4) and the ROP constant θ_{s_1,s_2} of (5) satisfy the bounds

$$\delta_s \leq (s-1)\mu, \qquad (6)$$
$$\theta_{s_1,s_2} \leq \mu\sqrt{s_1 s_2} \qquad (7)$$

where μ is the mutual coherence (3).

We will apply this lemma in Section III in order to obtain coherence-based settings from results which use the RICs.

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 $\widehat{\mathbf{x}}_{\mathrm{BP}}$ to the quadratic program

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \gamma\|\mathbf{x}\|_1 \qquad (8)$$

for some regularization parameter γ. We refer to the optimization problem (8) as BPDN, although it should be noted that some authors reserve this term for the related optimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \leq \nu \qquad (9)$$

where ν is a given constant.

Another estimator based on the idea of ℓ_1 relaxation is the Dantzig selector [3], defined as a solution $\widehat{\mathbf{x}}_{\mathrm{DS}}$ to the optimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x})\|_\infty \leq \tau \qquad (10)$$

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

$$\widehat{\mathbf{x}} = \begin{cases} \mathbf{A}_\Lambda^\dagger \mathbf{b}, & \text{on the support set } \Lambda \\ \mathbf{0}, & \text{elsewhere.} \end{cases} \qquad (11)$$

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 (11) is applied to obtain the thresholding estimate $\widehat{\mathbf{x}}_{\mathrm{th}}$.

A somewhat more sophisticated greedy algorithm is OMP [2]. 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 $\widehat{\mathbf{x}}_{\mathrm{OMP}}^i$ at the ith iteration is then defined by the LS solution (11) using the support set Λ^i. Next, the residual is updated using the formula

$$\mathbf{r}^i = \mathbf{b} - \mathbf{A}\widehat{\mathbf{x}}_{\mathrm{OMP}}^i. \qquad (12)$$

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, $\widehat{\mathbf{x}}_{\mathrm{OMP}}^s$. A well-known property of OMP is that the algorithm never chooses the same atom twice [6]. 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 $\widehat{\mathbf{x}}_{\mathrm{or}}$ calculates the LS solution (11) for Λ_0, and is often used as a gold standard against which the performance of practical algorithms can be compared.

III. PERFORMANCE GUARANTEES

Under the setting (1), it can be shown [12] that the MSE of any unbiased estimator $\widehat{\mathbf{x}}$ of $\mathbf{x}_0$ satisfies the Cramér-Rao bound [30]

$$\mathrm{MSE}(\widehat{\mathbf{x}}) \triangleq E\left\{\|\widehat{\mathbf{x}} - \mathbf{x}_0\|_2^2\right\}$$
$$\geq \sigma^2 \mathrm{Tr}\left(\left(\mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0}\right)^{-1}\right) \triangleq \mathrm{CRB}. \qquad (13)$$

Interestingly, CRB is also the MSE of the oracle estimator mentioned above [3]. It follows from the Gershgorin disc theorem [31] 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, $\mathrm{Tr}\big(\big(\mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0} \big)^{-1} \big)$ is not much larger than s, and consequently (13) is on the order of $s\sigma^2$. 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

Review of Previous Results: Historically, performance guarantees under random noise were first obtained for the Dantzig selector (10) by Candès and Tao [3]. Their result is derived using the RICs (4)–(5); by applying the bounds of Lemma 1, it is possible to obtain from their work a coherence-based guarantee. More recently, Cai, Wang, and Xu [15] have shown that one can obtain a tighter performance guarantee for the Dantzig selector directly from the mutual coherence, as shown in the following theorem.

Theorem 1 (Cai, Wang, and Xu): Under the setting (1), assume that

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

and consider the Dantzig selector (10) with parameter

$$\tau = \sigma\sqrt{2\log m} + \frac{3}{2}\sigma. \tag{15}$$

Then, with probability exceeding

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

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

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{DS}}\|_2^2 \leq \frac{8\left(\frac{3}{2} + \sqrt{2\log m}\right)^2}{(1 - (2s-1)\mu)^2}(s+1)\sigma^2. \tag{17}$$

Thus, 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 [13, Sec. 7.4, and the references therein]).

Recently, a performance guarantee has also been demonstrated for BPDN [10]. Once again, this result is based on RIC-like properties. Its translation to a coherence-based guarantee is given in the following theorem, whose derivation from [10] is described in Appendix A.

Theorem 2 (Bickel, Ritov, and Tsybakov): Under the setting (1), assume that

$$s < \frac{1+\mu}{(2 + 3\sqrt{2})\mu} \tag{18}$$

and consider the BPDN estimator (8) with parameter

$$\gamma = \sqrt{8\sigma^2(1+\alpha)\log m} \tag{19}$$

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

$$1 - \frac{1}{m^\alpha} \tag{20}$$

BPDN satisfies

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq 2048(1+\alpha)\frac{1}{\kappa^4}s\sigma^2\log m \tag{21}$$

where

$$\kappa = \sqrt{1 - (2s-1)\mu} - \frac{s\mu 3\sqrt{2}}{\sqrt{1 - (2s-1)\mu}} < 1. \tag{22}$$

Coherence-Based Guarantee for BPDN: The constant in the BPDN performance guarantee (21) is typically much larger than that given in (17) for the Dantzig selector. The necessary condition (18) is also more stringent than the requirements for the Dantzig selector. However, following the experience in the case of the Dantzig selector, one may hope for substantially better guarantees to be obtained by directly relying on the mutual coherence. This approach indeed bears fruit, as we now show. 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[1]

$$s < \frac{1}{3\mu}. \tag{23}$$

Then, with probability exceeding

$$\left(1 - (m-s)\exp\left(-\frac{\gamma^2}{8\sigma^2}\right)\right)\left(1 - e^{-s/7}\right) \tag{24}$$

the solution $\hat{\mathbf{x}}_{\mathrm{BP}}$ of BPDN (8) is unique, its support is contained in the true support Λ_0, and

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq \left(\sigma\sqrt{3} + \frac{3}{2}\gamma\right)^2 s. \tag{25}$$

To compare this result with the previous theorems, we now derive from Theorem 3 a result which holds with a probability on the order of (20). Observe that for (24) 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 (25). Therefore, it is natural to use

$$\gamma = \sqrt{8\sigma^2(1+\alpha)\log(m-s)} \tag{26}$$

for some fairly small $\alpha > 0$; note the encouraging similarity of this value of γ to (19). Substituting γ of (26) 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 (8) with γ given by (26). Then, with probability exceeding

$$\left(1 - \frac{1}{(m-s)^\alpha}\right)\left(1 - e^{-s/7}\right) \tag{27}$$

---

[1] As in [9], analogous findings can also be obtained under the weaker requirement $s < 1/(2\mu)$, but the resulting expressions are somewhat more involved.

the solution $\hat{\mathbf{x}}_{\text{BP}}$ is unique, its support is a subset of Λ_0, and

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\text{BP}}\|_2^2 \leq \left(\sqrt{3} + 3\sqrt{2(1+\alpha)\log(m-s)}\right)^2 s\sigma^2. \quad (28)$$

To compare the BPDN guarantees of Corollary 1 and Theorem 2, we first examine the probability (27). This expression 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 side of (27) tends to 1 polynomially as $m - s$ increases. This is a slightly lower rate than that of Theorem 2; however, this difference is compensated for by a correspondingly lower multiplicative factor of $\log(m-s)$ in the error bound (28), as opposed to the $\log m$ factor in previous results. In any case, for any of the theorems to hold, m must increase much more quickly than s, so that these differences are negligible.

Whereas the probability of success of Corollary 1 is comparable to that of Theorem 2, the required sparsity level (23) in the proposed guarantee is substantially better than that of previous results. Furthermore, the constant (25) in the new result is also much smaller than that of the previous BPDN guarantee (21). Thus, it appears that the direct application of the mutual coherence is successful in obtaining tighter performance guarantees.

What can be learned by comparing the guarantees for BPDN and the Dantzig selector? In some respects, the BPDN result appears stronger; in particular, the probability of success in Corollary 1 is better than that of Theorem 1, and the resulting guarantee has somewhat smaller constants. On the other hand, the sparsity requirements of Theorem 1 are somewhat less stringent. Choosing the most accurate guarantee (and, consequently, the preferred estimator) will thus depend on the specifics of the setting under consideration. One such specific example 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 (11), 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

$$|x_{\min}| = \min_{i \in \Lambda_0} |x_{0,i}|$$
$$|x_{\max}| = \max_{i \in \Lambda_0} |x_{0,i}|. \quad (29)$$

A performance guarantee for both OMP and the thresholding algorithm is then given by the following theorem.

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

$$|x_{\min}| - (2s-1)\mu|x_{\min}| \geq 2\sigma\sqrt{2(1+\alpha)\log m} \quad (30)$$

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

$$1 - \frac{1}{m^\alpha\sqrt{\pi(1+\alpha)\log m}} \quad (31)$$

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

$$\|\hat{\mathbf{x}}_{\text{OMP}} - \mathbf{x}_0\|_2^2 \leq \frac{2(1+\alpha)}{(1-(s-1)\mu)^2} s\sigma^2 \log m \quad (32a)$$
$$\leq 8(1+\alpha)s\sigma^2 \log m. \quad (32b)$$

If the stronger condition

$$|x_{\min}| - (2s-1)\mu|x_{\max}| \geq 2\sigma\sqrt{2(1+\alpha)\log m} \quad (33)$$

holds, then with probability exceeding (31), the thresholding algorithm also correctly identifies Λ_0 and satisfies (32).

The performance guarantee (32) is better than that provided by Theorem 1 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 (30) and (33) 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.

IV. COMPARISON WITH RELATED ESTIMATION SETTINGS

The difficulty of an estimation problem naturally depends on the strength of the assumptions in the underlying model. We now compare the results of the previous section with prior performance guarantees for sparse estimation algorithms in different settings.

The equation $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$ describes a variety of situations, depending on what is assumed to be known about $\mathbf{A}$, $\mathbf{x}_0$, and

w. The focus in this paper is on the frequentist estimation setting, wherein the known matrix $\mathbf{A}$ and the unknown parameter vector $\mathbf{x}_0$ are both deterministic, and the noise $\mathbf{w}$ is Gaussian and white. In the following, we compare the performance guarantees obtained for this setting with known results for other scenarios. This comparison emphasizes that better performance can be guaranteed if a more detailed model is assumed. We note that many of the results cited in this section are obtained up to an unknown numerical constant, and consequently the discussion herein likewise performs qualitative comparisons based on asymptotic rates.

Adversarial Noise: In the adversarial or bounded noise scenario, the noise $\mathbf{w}$ is assumed to be deterministic and unknown, and to have a bounded ℓ_2 norm, $\|\mathbf{w}\|_2 \leq \varepsilon$. The matrix $\mathbf{A}$ and the unknown vector $\mathbf{x}_0$ are also deterministic. Since no probabilistic information is assumed about $\mathbf{w}$, one must perform a worst-case analysis, namely, in this case guarantees for an estimator $\hat{\mathbf{x}}$ ensure that $\|\hat{\mathbf{x}} - \mathbf{x}_0\|$ is small for all noise realizations and all feasible values of $\mathbf{x}_0$. Consequently, adversarial performance guarantees are relatively weak, and indeed no denoising capability can be ensured for any known algorithm.

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 [9, Sec. IV-C].[2]

Theorem 5 (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 of BPDN (8) 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

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_\infty < \left(3 + \sqrt{\frac{3}{2}}\right) \varepsilon \approx 4.22\varepsilon. \tag{34}$$

Results similar to Theorem 5 have also been obtained [6], [7], [15], [26] for the related ℓ_1-error estimation approach (9), as well as for the OMP algorithm [6]. Furthermore, the technique used in the proof for the OMP [6] 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. As expected, these results are much weaker than those obtained in Section III for the frequentist setting. Indeed, in the frequentist setting, which differs only in the adoption of the Gaussian distribution for the noise, typical performance guarantees ensure that $\|\hat{\mathbf{x}} - \mathbf{x}_0\|_2$ is on the order of $s\sigma^2 \log m$, while the average noise power is $n\sigma^2$, which is much larger. This difference is due to the fact that results in the adversarial context

must take into account values of $\mathbf{w}$ which are chosen so as to cause maximal damage to the estimation algorithm.

Random Design Matrix: Whereas the adversarial setting assumes less about the signal and consequently provides weak guarantees, incorporating further probabilistic assumptions can provide much more optimistic assurances of the achievable performance than those obtained in our setting. For example, a common setting in the compressed sensing literature assumes that the matrix $\mathbf{A}$ is known, but chosen randomly from an appropriate ensemble. The primary advantage of such an assumption is that with high probability, the RICs of the resulting matrix will be low, and performance guarantees obtained based on the use of the RICs are typically sharper than those based on the mutual coherence.

Specifically, suppose that the entries of $\mathbf{A}$ are chosen from a white Gaussian distribution and are then normalized to satisfy (2). Also suppose that $\mathbf{x}_0$ is deterministic and that $\mathbf{w}$ is white Gaussian noise, as in our setting. Then it can be shown that, with high probability, the Dantzig selector achieves an error on the order of $s\sigma^2 \log m$ as long as the sparsity level s is on the order of $n/\log(m/n)$ [3]. Comparable results can also be demonstrated for OMP,[3] and in this case near-oracle performance is asymptotically guaranteed with high probability when s is no larger than about $n/\log n$ [25], [32].

The guarantees of Section III are much weaker than these random-matrix results. Indeed, to obtain near-oracle performance guarantees, the results of Section III all require the sparsity level s to be on the order of $1/\mu$. However, for any matrix $\mathbf{A}$ we have [33, Thm.2.3]

$$\mu \geq \sqrt{\frac{m-n}{n(m-1)}}. \tag{35}$$

Making the reasonable assumption that $m \geq 2n$, it follows that, at best, our frequentist results hold with sparsity levels s on the order of $\sqrt{n}$. This is not nearly as strong as the random matrix results, for which the number of nonzero entries in $\mathbf{x}_0$ is allowed to come within a log factor of the number of measurements n.

The difference between the random and deterministic matrix guarantees should not, however, be interpreted to mean that the proposed performance guarantees can be substantially improved. Rather, it indicates that truly better performance should be expected in the random matrix setting. Indeed, there exist suitably chosen deterministic matrices $\mathbf{A}$ and vectors $\mathbf{x}_0$ for which s is on the order of $1/\mu$, wherein reasonable algorithms completely fail to recover $\mathbf{x}_0$ [21]. In these constructions, the value of s fails to meet the requirements for the theorems of Section III by no more than a small multiplicative constant. It follows that any performance guarantee based solely on the mutual coherence of $\mathbf{A}$ cannot guarantee performance for sparsity levels higher than $O(1/\mu)$.

V. NUMERICAL RESULTS

In this section, we describe a number of numerical experiments comparing the performance of various estimators to the

---

[2]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 [9, Sec. IV-C], can readily be applied to the general setting considered here, we omit the proof of Theorem 5.

[3]As in any analysis of a greedy algorithm, the OMP result requires some further assumptions on the SNR and, in particular, on the value of $|x_{\min}|$ relative to the noise power.
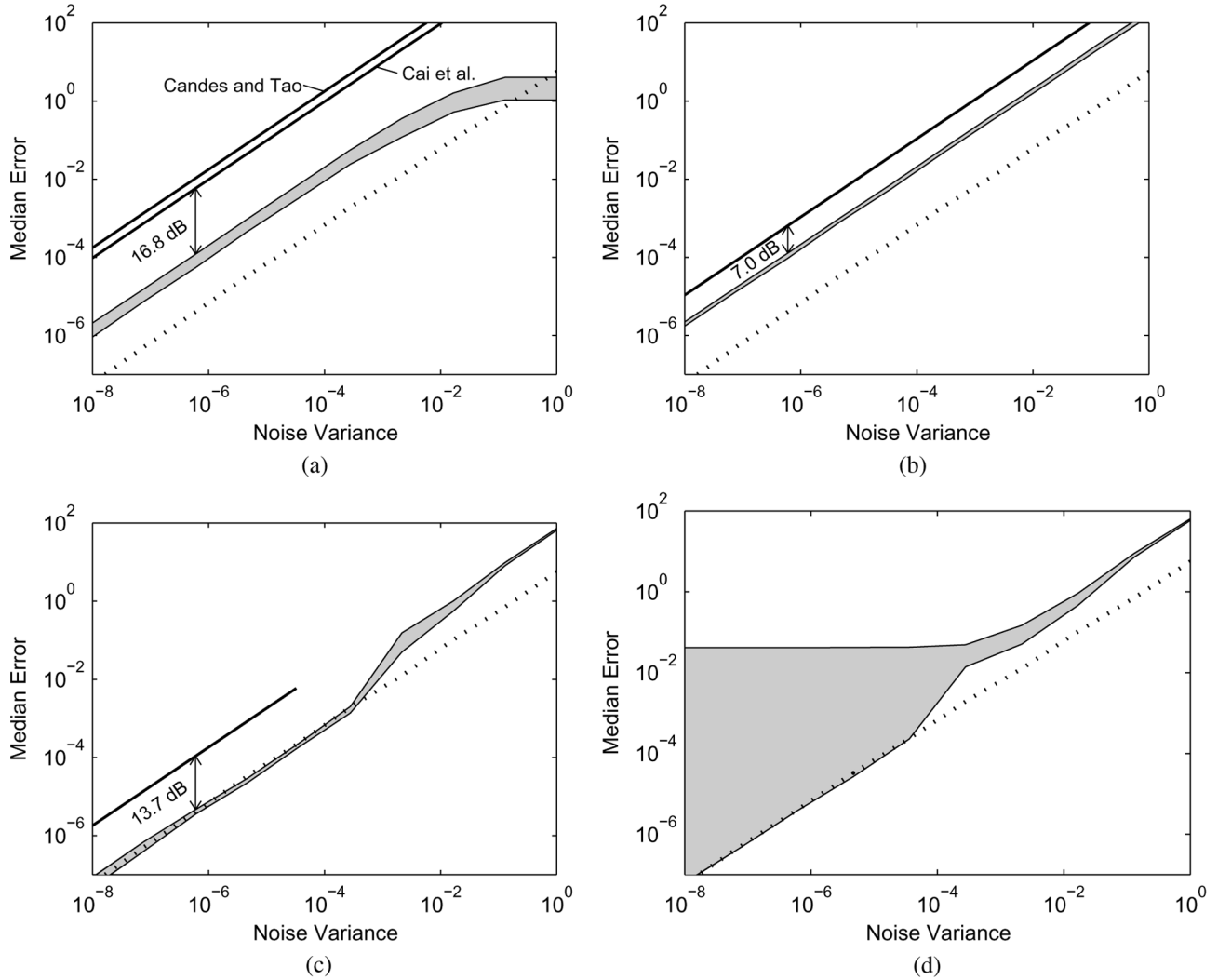
Fig. 1.   Median estimation error for practical estimators (gray regions) compared with performance guarantees (solid lines) and the oracle estimator (dotted line). The gray regions report the range of performances observed for 8 different values of the unknown parameter vector $\mathbf{x}_0$. For the Dantzig selector, both the guarantee of Candes and Tao [3] and that of Cai, Wang, and Xu (Theorem 1) are plotted. For OMP, performance is only guaranteed for $\sigma \leq 0.057$, while for thresholding, nothing can be guaranteed for the given problem dimensions. (a) Dantzig selector. (b) BPDN. (c) OMP. (d) Thresholding.

guarantees of Section III. Our first experiment measured the median estimation error, i.e., the median of the squared 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 III.

Specifically, we chose the two-ortho dictionary $\mathbf{A} = [\mathbf{I} \quad \mathbf{H}]$, 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 III 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 $|x_{\max}| = 1$. To obtain guarantees on the median error, for each of the theorems of Section III a value of α was chosen such that the resulting error bound holds with

probability 1/2 or greater.[4] Under these conditions, applying the theorems of Section III yields the bounds

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{OMP}}\|_2^2 \leq 3.7 s\sigma^2 \log m \qquad \text{w.p. } \frac{3}{4}, \text{ if } \sigma \leq 0.057;$$

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq 22.1 s\sigma^2 \log m \qquad \text{w.p. } \frac{1}{2};$$

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{DS}}\|_2^2 \leq 198.4 s\sigma^2 \log m \quad \text{w.p. } \frac{3}{4}. \qquad (36)$$

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

[4]In particular, the results for the Dantzig selector (Theorem 1) 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.

thresholding algorithm is not proportional to $s\sigma^2 \log m$. To obtain thresholding guarantees, one requires a narrower range between $|x_{\min}|$ and $|x_{\max}|$.

The RIC-based BPDN guarantee of Theorem 2 is not applicable in the present setting; however, by comparing the BPDN guarantees of Theorems 2 and 3, it is evident that the latter is substantially tighter, even when the conditions for both theorems hold. It is also interesting to note that by applying Lemma 1 to the (RIC-based) Dantzig selector guarantee of Candès and Tao [3], one can obtain for the current setting

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{DS}}\|_2^2 \le 361.8 s\sigma^2 \log m \quad \text{w.p.} \; \frac{3}{4}. \tag{37}$$

This result is again looser than (36), once again demonstrating that considerable improvement in performance guarantees is possible by direct use of the mutual coherence.

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). The estimation algorithms of Section II-B were then applied to each measurement realization. For the Dantzig selector, τ was selected using (15), and for BPDN, γ was chosen as the smallest value such that the probability of success (27) would exceed 1/2. The median over noise realizations of the distance $\|\mathbf{x}_0 - \hat{\mathbf{x}}\|_2^2$ was then computed for each estimator. This process was repeated for 10 values of the noise variance σ^2 in the range $10^{-8} \le \sigma^2 \le 1$. The results are plotted in Fig. 1 as a function of σ^2. The performance guarantees (36)–(37) are also plotted.[5]

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 is almost 50 times higher than the worst of the examined parameter values.

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 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 $\mathbf{A} = [\mathbf{I} \quad \mathbf{H}]$, where $\mathbf{I}$ is the 256×256 identity matrix and $\mathbf{H}$ is the 256×256 Hadamard matrix with normalized columns.[6] Once again, the RICs of this dictionary are unknown. However, the coherence in this case is given by
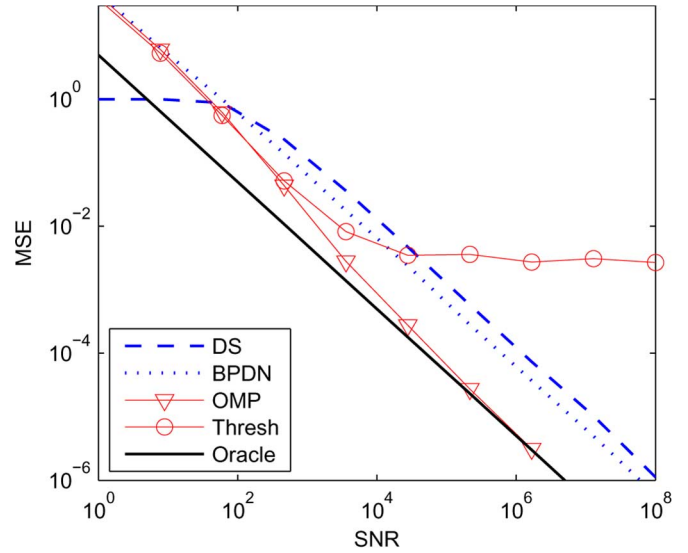

Fig. 2. MSE of various estimators as a function of the SNR. The sparsity level is $s = 5$ and the dictionary is a 256×512 two-ortho matrix.

$\mu = 1/16$, and consequently, the ℓ_1 relaxation guarantees of Section III-A hold for $s \le 5$.

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 1 and Corollary 1, respectively; for the latter, a value of $\alpha = 1$ was chosen. 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

$$\mathrm{SNR} = \frac{\|\mathbf{x}_0\|_2^2}{n\sigma^2} = \frac{1}{n\sigma^2}. \tag{38}$$

To compare this plot with the theoretical results of Section III, observe first the situation at high SNR. In this case, OMP, BPDN, and the Dantzig selector all achieve performance 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 III. Indeed, for small σ, the condition (30) 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 (33) 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

---

[5] The guarantee (37) actually requires a slightly different value of τ, but this difference has a negligible effect on performance and is ignored.

[6] Similar experiments were performed on a variety of other dictionaries, including an overcomplete DCT and a matrix containing Gaussian random entries. The different dictionaries yielded comparable results, which are not reported here.
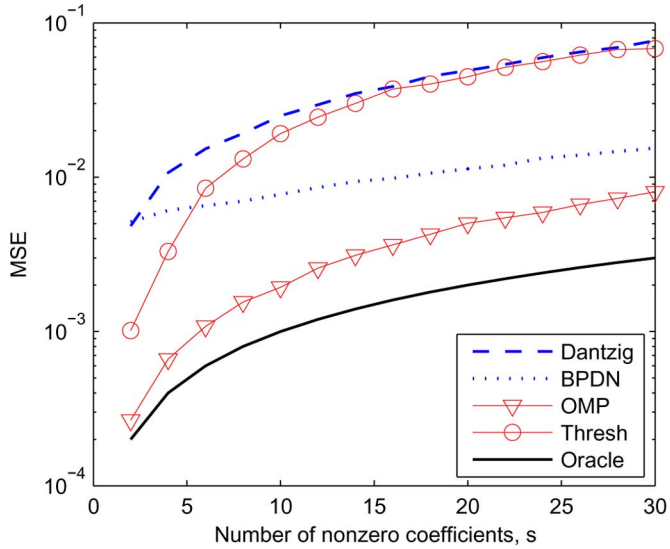
Fig. 3. MSE of various estimators as a function of the support size s. The noise standard deviation is $\sigma = 0.01$ and the dictionary is a 256×512 two-ortho matrix.

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.

Aforementioned, 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 III. 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 $1/\mu$ [21]. 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 in Section IV, under the adversarial noise scenario, 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, when both the noise and the design matrix $\mathbf{A}$ are random, practical estimators come close to the performance of the oracle estimator. In this paper, we examined a middle ground between these two cases, namely the setting in which $\mathbf{x}_0$ and $\mathbf{A}$ are deterministic but the noise is random. As we have shown, despite the fact that less information is available in this case, 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 THEOREM 2

To prove the theorem, we will adapt [10, Theorem 7.2 and Lemma 4.1] to the present setting, and then apply the coherence bounds of Lemma 1. Note that the normalization of the matrix $\mathbf{A}$ differs between the present paper and that of [10], and consequently the results stated herein differs by a factor of n from the original.

We begin by recalling the definition of the "restricted eigenvalue condition" $RE(s, q, 3)$ of [10]. Let

$$\kappa = \min_{J_0 \subseteq \{1,\ldots,m\}:|J_0| \leq s} \min_{\mathbf{x} \neq 0: \left\|\mathbf{x}_{J_0^c}\right\|_1 \leq 3\|\mathbf{x}_{J_0}\|_1} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}_{J_0 \cup J_1}\|_2} \quad (39)$$

where $\mathbf{x}_J$ denotes the subvector of $\mathbf{x}$ indexed by the elements of the set J, J_0^c is the complement of the set J_0, and J_1 contains the indices of the q largest elements in $\mathbf{x}$ which are not contained in J_0. If $\kappa > 0$, then the matrix $\mathbf{A}$ is said to satisfy the restricted eigenvalue condition $RE(s, q, 3)$ with parameter κ. Note that $\kappa \leq 1$, since one could, for example, choose $J_0 = \{1\}$ and $\mathbf{x} =$

$[1, 0, \ldots, 0]^T$, whereupon it follows from (2) that the objective function in (39) equals 1.

Using this definition, the following result follows directly from [10, Theorem 7.2].

Theorem 6 (Bickel, Ritov, and Tsybakov): Under the setting (1), assume that the restricted eigenvalue condition $RE(s, q, 3)$ holds for some q. Consider the BPDN estimator (8) with γ given by (19) and choose $a > 2\sqrt{2}$. Then, with probability exceeding $1 - m^{1-a^2/8}$, we have

$$\|\hat{\mathbf{x}}_{\mathrm{BP}} - \mathbf{x}_0\|_2^2 \leq 16 \left(1 + 3\sqrt{\frac{s}{q}}\right)^2 \frac{a^2}{\kappa^4} s\sigma^2 \log m. \quad (40)$$

From [10, Lemma 4.1], it follows that a sufficient condition for $RE(s, s, 3)$ to hold is $\delta_{2s} + 3\theta_{s,2s} < 1$, where δ_{2s} and $\theta_{s,2s}$ are the RICs defined in Section II-A. Applying the bounds of Lemma 1, it follows that $RE(s, s, 3)$ holds if (18) holds. Furthermore, from [10, Lemma 4.1], we have that in this case

$$\kappa \geq \sqrt{1 - \delta_{2s}} - \frac{3\theta_{s,2s}}{\sqrt{1 - \delta_{2s}}} \quad (41)$$

from which, again by applying Lemma 1, we obtain

$$\kappa \geq \sqrt{1 - (2s-1)\mu} - \frac{s\mu 3\sqrt{2}}{\sqrt{1 - (2s-1)\mu}}. \quad (42)$$

Since decreasing κ does not violate the bound (40), our coherence-based bound must choose κ given by (22). Defining $1 + \alpha = a^2/8$ and choosing $q = s$, we obtain the probability of success (20) and the bound (21). Thus we have translated the results of Bickel *et al.* to the coherence-based Theorem 2.

APPENDIX B
PROOF OF THEOREM 3

The proof is based closely on the work of Tropp [9]. From the triangle inequality

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2 \leq \|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{or}}\|_2 + \|\hat{\mathbf{x}}_{\mathrm{or}} - \hat{\mathbf{x}}_{\mathrm{BP}}\|_2 \quad (43)$$

where $\hat{\mathbf{x}}_{\mathrm{or}}$ is the oracle estimator. Our goal is to separately bound the two terms on the right-hand side of (43). Indeed, as we will see, the two constants $\sigma\sqrt{3}$ and $(3/2)\gamma$ in (25) arise, respectively, from the two terms in (43).

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

$$\begin{aligned}
\mathbf{x}_{0,\Lambda} - \hat{\mathbf{x}}_{\mathrm{or},\Lambda} &= \mathbf{x}_{0,\Lambda} - \mathbf{A}_{\Lambda_0}^\dagger (\mathbf{A}\mathbf{x}_0 + \mathbf{w}) \\
&= \mathbf{x}_{0,\Lambda} - \mathbf{A}_{\Lambda_0}^\dagger (\mathbf{A}_{\Lambda_0}\mathbf{x}_{0,\Lambda} + \mathbf{w}) \\
&= -\mathbf{A}_{\Lambda_0}^\dagger \mathbf{w} \quad (44)
\end{aligned}$$

where we have used the fact that $\mathbf{A}_{\Lambda_0}$ has full column rank, which is a consequence [34] of the condition (23). Thus, $\mathbf{x}_{0,\Lambda} - \hat{\mathbf{x}}_{\mathrm{or},\Lambda}$ is a Gaussian random vector with mean $\mathbf{0}$ and covariance $\sigma^2 \mathbf{A}_{\Lambda_0}^\dagger \mathbf{A}_{\Lambda_0}^{\dagger T} = \sigma^2 \left(\mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0}\right)^{-1}$.

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

$$\begin{aligned}
E\left\{\mathbf{A}_{\Lambda_0}^\dagger \mathbf{w}\mathbf{w}^T(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^\dagger)^T\right\} &= \sigma^2 \mathbf{A}_{\Lambda_0}^\dagger \left(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^\dagger\right)^T \\
&= \mathbf{0} \quad (45)
\end{aligned}$$

where we have used the fact [35, Th. 1.2.1] that for any matrix $\mathbf{M}$

$$\mathbf{M}^\dagger \mathbf{M}^{\dagger T} \mathbf{M}^T = (\mathbf{M}^T \mathbf{M})^\dagger \mathbf{M}^T = \mathbf{M}^\dagger. \quad (46)$$

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

$$\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{or}} \text{ is statistically independent of } \left(\mathbf{I} - \mathbf{A}_{\Lambda_0}\mathbf{A}_{\Lambda_0}^\dagger\right)\mathbf{w}. \quad (47)$$

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

$$\begin{aligned}
&\mathrm{Pr}\left\{\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{or}}\|_2^2 > 3s\sigma^2\right\} \\
&= \mathrm{Pr}\left\{\left\|\sigma(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0})^{-1/2}\mathbf{z}\right\|_2^2 \geq 3s\sigma^2\right\} \\
&\leq \mathrm{Pr}\left\{\left\|(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0})^{-1/2}\right\|^2 \|\mathbf{z}\|_2^2 \geq 3s\right\} \quad (48)
\end{aligned}$$

where $\|\mathbf{M}\|$ denotes the maximum singular value of the matrix $\mathbf{M}$. Thus, $\|\left(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0}\right)^{-1/2}\| = 1/s_{\min}$, where $s_{\min}$ is the minimum singular value of $\mathbf{A}_{\Lambda_0}$. From the Gershgorin disc theorem [31, p.320], it follows that $s_{\min} \geq \sqrt{1 - (s-1)\mu}$. Using (23), this can be simplified to $s_{\min} \geq \sqrt{2/3}$, and therefore

$$\left\|\left(\mathbf{A}_{\Lambda_0}^T\mathbf{A}_{\Lambda_0}\right)^{-1/2}\right\| \leq \sqrt{\frac{3}{2}}. \quad (49)$$

Combining with (48) yields

$$\mathrm{Pr}\left\{\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{or}}\|_2^2 > 3s\sigma^2\right\} \leq \mathrm{Pr}\left\{\|\mathbf{z}\|_2^2 \geq 2s\right\}. \quad (50)$$

Observe that $\|\mathbf{z}\|_2^2$ is the sum of s independent normalized Gaussian random variables. The right-hand side of (50) 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 [36, Sec. 16.3] for $F_{\chi_s^2}(\cdot)$, we have

$$\mathrm{Pr}\left\{\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{or}}\|_2^2 > 3s\sigma^2\right\} \leq Q\left(\frac{1}{2}s, s\right) \quad (51)$$

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

$$Q(a, z) \triangleq \frac{\int_z^\infty t^{a-1}e^{-t}dt}{\int_0^\infty t^{a-1}e^{-t}dt}. \quad (52)$$

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

$$Q\left(\frac{1}{2}s, s\right) < e^{-s/7} \quad \text{for all } s. \quad (53)$$

We thus conclude that the event

$$\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\text{or}}\|_2^2 \leq 3s\sigma^2 \tag{54}$$

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 $\|\mathbf{x}_0 - \hat{\mathbf{x}}_{\text{or}}\|_2^2 > \alpha s\sigma^2$, for any $\alpha > (2/3)$. The only difference will be the rate of exponential decay in (53). However, the distance between $\mathbf{x}_0$ and $\hat{\mathbf{x}}_{\text{or}}$ is usually small compared with the distance between $\hat{\mathbf{x}}_{\text{or}}$ and $\hat{\mathbf{x}}_{\text{BP}}$, so that such an approach does not significantly affect the overall result.

The above calculations provided a bound on the first term in (43). To address the second term $\|\hat{\mathbf{x}}_{\text{or}} - \hat{\mathbf{x}}_{\text{BP}}\|_2$, define the random event

$$G : \max_i \left| \mathbf{a}_i^T \left(\mathbf{I} - \mathbf{A}_{\Lambda_0} \mathbf{A}_{\Lambda_0}^\dagger \right) \mathbf{b} \right| \leq \frac{1}{2}\gamma \tag{55}$$

where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$. It is shown in [9, App. IV-A] that

$$\Pr\{G\} \geq 1 - (m - s) \exp\left(-\frac{\gamma^2}{8\sigma^2} \right). \tag{56}$$

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 [9, Cor. 9] that, in this case, the solution $\hat{\mathbf{x}}_{\text{BP}}$ to (8) is unique, the support of $\hat{\mathbf{x}}_{\text{BP}}$ is a subset of Λ_0, and

$$\|\hat{\mathbf{x}}_{\text{BP}} - \hat{\mathbf{x}}_{\text{or}}\|_\infty \leq \frac{3}{2}\gamma. \tag{57}$$

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

$$\|\hat{\mathbf{x}}_{\text{BP}} - \hat{\mathbf{x}}_{\text{or}}\|_2 \leq \frac{3}{2}\gamma\sqrt{s}. \tag{58}$$

The event G depends on the random variable $\mathbf{w}$ only through $\left(\mathbf{I} - \mathbf{A}_{\Lambda_0} \mathbf{A}_{\Lambda_0}^\dagger \right) \mathbf{w}$. Thus, it follows from (47) that G is statistically independent of the event (48). The probability that both events occur simultaneously is therefore given by the product of their respective probabilities. In other words, with probability exceeding (24), both (58) and (54) hold. Using (43) 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

$$B = \left\{ \max_{1 \leq i \leq m} \left| \mathbf{a}_i^T \mathbf{w} \right| < \tau \right\} \tag{59}$$

where

$$\tau \triangleq \sigma \sqrt{2(1 + \alpha)\log m} \tag{60}$$

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 (59) occurs with a probability of at least (31).

Proof: The random variables $\{\mathbf{a}_i^T\mathbf{w}\}_{i=1}^m$ are jointly Gaussian. Therefore, by 'idák's lemma [37, Th. 1]

$$\Pr\{B\} = \Pr\left\{ \max_{1 \leq i \leq m} \left| \mathbf{a}_i^T\mathbf{w} \right| < \tau \right\} \geq \prod_{i=1}^m \Pr\left\{ \left| \mathbf{a}_i^T\mathbf{w} \right| \leq \tau \right\}. \tag{61}$$

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

$$\Pr\left\{ \left| \mathbf{a}_i^T\mathbf{w} \right| < \tau \right\} = 1 - 2Q\left(\frac{\tau}{\sigma} \right) \tag{62}$$

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

$$Q(x) \leq \frac{1}{x\sqrt{2\pi}} e^{-x^2/2} \tag{63}$$

we obtain from (62)

$$\Pr\left\{ \left| \mathbf{a}_i^T\mathbf{w} \right| < \tau \right\} \geq 1 - \eta \tag{64}$$

where

$$\eta \triangleq \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma}{\tau} e^{-\tau^2/2\sigma^2}. \tag{65}$$

When $\eta > 1$, the bound (31) is meaningless and the theorem holds vacuously. Otherwise, when $\eta \leq 1$, we have from (61) and (64)

$$\Pr\{B\} \geq (1 - \eta)^m \geq 1 - m\eta \tag{66}$$

where the final inequality holds for any $\eta \leq 1$ and $m \geq 1$. Substituting the values of η and τ and simplifying, we obtain that B holds with a probability no lower than (31), 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 $\Lambda_0 = \text{supp}(\mathbf{x}_0)$ 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 (29), and suppose that

$$|x_{\max}| - (2s - 1)\mu|x_{\max}| \geq 2\tau. \tag{67}$$

Then, if the event B of (59) holds, we have

$$\max_{j \in \Lambda_0} \left| \mathbf{a}_j^T\mathbf{b} \right| > \max_{j \notin \Lambda_0} \left| \mathbf{a}_j^T\mathbf{b} \right|. \tag{68}$$

If, rather than (67), the stronger condition

$$|x_{\min}| - (2s - 1)\mu|x_{\max}| \geq 2\tau \tag{69}$$

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

$$\min_{j \in \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right| > \max_{j \notin \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right|. \tag{70}$$

Proof: The proof is an adaptation of [6, Lemma 5.2]. Beginning with the term $\max_{j \notin \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right|$, we have, under the event B

$$
\begin{aligned}
\max_{j \notin \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right| &= \max_{j \notin \Lambda_0} \left| \mathbf{a}_j^T \mathbf{w} + \sum_{i \in \Lambda_0} x_i \mathbf{a}_j^T \mathbf{a}_i \right| \\
&\leq \max_{j \notin \Lambda_0} \left| \mathbf{a}_j^T \mathbf{w} \right| + \max_{j \notin \Lambda_0} \sum_{i \in \Lambda_0} \left| x_i \mathbf{a}_j^T \mathbf{a}_i \right| \\
&< \tau + s\mu |x_{\max}|.
\end{aligned}
\tag{71}
$$

On the other hand, when B holds,

$$
\begin{aligned}
\max_{j \in \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right| &= \max_{j \in \Lambda_0} \left| x_j + \mathbf{a}_j^T \mathbf{w} + \sum_{i \in \Lambda_0 \setminus \{j\}} x_i \mathbf{a}_j^T \mathbf{a}_i \right| \\
&\geq |x_{\max}| - \max_{j \in \Lambda_0} \left| \mathbf{a}_j^T \mathbf{w} + \sum_{i \in \Lambda_0 \setminus \{j\}} x_i \mathbf{a}_j^T \mathbf{a}_i \right| \\
&> |x_{\max}| - \tau - (s-1)\mu |x_{\max}| \\
&= |x_{\max}| - (2s-1)\mu |x_{\max}| \\
&\quad - \tau + s\mu |x_{\max}|.
\end{aligned}
\tag{72}
$$

Together with (71), this yields

$$\max_{j \in \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right| > |x_{\max}| - (2s-1)\mu |x_{\max}| - 2\tau + \max_{j \notin \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right|. \tag{73}$$

Thus, under the condition (67), we obtain (68). Similarly, when B holds, we have

$$
\begin{aligned}
\min_{j \in \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right| &= \min_{j \in \Lambda_0} \left| x_j + \mathbf{a}_j^T \mathbf{w} + \sum_{i \in \Lambda_0 \setminus \{j\}} x_i \mathbf{a}_j^T \mathbf{a}_i \right| \\
&> |x_{\min}| - \tau - (s-1)\mu |x_{\max}| \\
&= |x_{\min}| - (2s-1)\mu |x_{\max}| \\
&\quad - \tau + s\mu |x_{\max}|.
\end{aligned}
\tag{74}
$$

Again using (71), we obtain

$$\min_{j \in \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right| > |x_{\min}| - (2s-1)\mu |x_{\max}| - 2\tau + \max_{j \notin \Lambda_0} \left| \mathbf{a}_j^T \mathbf{b} \right|. \tag{75}$$

Consequently, under the assumption (69), we conclude that (70) 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 (59) occurs, then

$$\|\hat{\mathbf{x}}_{\mathrm{or}} - \mathbf{x}_0\|_2^2 \leq 2s\sigma^2 (1+\alpha) \log m \frac{1}{(1-(s-1)\mu)^2}. \tag{76}$$

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

$$\|\hat{\mathbf{x}}_{\mathrm{or}} - \mathbf{x}_0\|_2^2 = \left\| \mathbf{A}_{\Lambda_0}^\dagger \mathbf{b} - \mathbf{x}_{0,\Lambda_0} \right\|_2^2 \tag{77}$$

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

$$
\begin{aligned}
\|\hat{\mathbf{x}}_{\mathrm{or}} - \mathbf{x}_0\|_2^2 &= \left\| \mathbf{A}_{\Lambda_0}^\dagger \mathbf{A}_{\Lambda_0} \mathbf{x}_{0,\Lambda_0} + \mathbf{A}_{\Lambda_0}^\dagger \mathbf{w} - \mathbf{x}_{0,\Lambda_0} \right\|_2^2 \\
&= \left\| \mathbf{A}_{\Lambda_0}^\dagger \mathbf{w} \right\|_2^2 \\
&= \left\| (\mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0})^{-1} \mathbf{A}_{\Lambda_0}^T \mathbf{w} \right\|_2^2 \\
&\leq \left\| (\mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0})^{-1} \right\|^2 \sum_{i \in \Lambda_0} (\mathbf{a}_i^T \mathbf{w})^2 \\
&\leq \frac{1}{(1-(s-1)\mu)^2} s\sigma^2 2(1+\alpha) \log m
\end{aligned}
\tag{78}
$$

where, in the last step, we used the definition (59) of B and the fact that $\left\| \mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0} \right\| \geq 1 - (s-1)\mu$, 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 (31). 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 (32).

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 [6, 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 (30) implies (67), 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 $1 < i \leq s$, 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 $\mathrm{supp}\left(\hat{\mathbf{x}}_{\mathrm{OMP}}^{i-1}\right) = \Lambda^{i-1} \subset \Lambda_0$. 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

$$\mathbf{r}^i = \mathbf{A}\tilde{\mathbf{x}}^{i-1} + \mathbf{w} \tag{79}$$

where $\tilde{\mathbf{x}}^{i-1} = \mathbf{x}_0 - \hat{\mathbf{x}}_{\mathrm{OMP}}^{i-1}$. 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

$$\left\| \hat{\mathbf{x}}_{\mathrm{OMP}}^{i-1} \right\|_0 = i - 1 < s = \|x_0\|_0 \tag{80}$$

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

$$\max_i \left| \tilde{x}_i^{i-1} \right| \geq |x_{\min}|. \tag{81}$$

Note that the model (79) 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 (81) and (30) that this model satisfies the requirement (67). Consequently, by Lemma 3, we have that under the event B,

$$\max_{i \in \Lambda_0} \left| \mathbf{a}_i^T \mathbf{r}^i \right| > \max_{i \notin \Lambda_0} \left| \mathbf{a}_i^T \mathbf{r}^i \right|. \tag{82}$$

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] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, Feb. 2009.

[2] 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, Syst., Comput.*, Nov. 1993, pp. 40–44.

[3] 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.

[4] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] Z. Ben-Haim and Y. C. Eldar, "The Cramér-Rao bound for estimating a sparse parameter vector," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3384–3389, Jun. 2010.

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

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

[15] T. T. Cai, L. Wang, and G. Xu, "Stable recovery of sparse signals and an oracle inequality," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3516–3522, Jul. 2010.

[16] Y. C. Eldar, "Uncertainty relations for shift-invariant analog signals," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5742–5757, Dec. 2009.

[17] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.

[18] 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. Acoust., Speech, Signal Processing (ICASSP 2010)*, Dallas, TX, Mar. 2010, pp. 3590–3593.

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

[20] J. A. Tropp, "On the conditioning of random subdictionaries," *Appl. Computat. Harmon. Anal.*, vol. 25, no. 1, pp. 1–24, 2008.

[21] 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.

[22] Y. C. Eldar and H. Rauhut, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 505–519, Jan. 2010.

[23] J. S. Turek, I. Yavneh, M. Protter, and M. Elad, "On MMSE and MAP denoising under sparse representation modeling over a unitary dictionary," *Appl. Computat. Harmon. Anal.*, 2010, submitted for publication.

[24] 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.

[25] A. K. Fletcher and S. Rangan, "Orthogonal matching pursuit from noisy measurements: A new analysis," in *Proc. Neural Inf. Processing Syst. Conf. (NIPS)*, Vancouver, BC, Canada, Dec. 2009.

[26] 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.

[27] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.

[28] T. T. Cai, G. Xu, and J. Zhang, "On recovery of sparse signals via ℓ_1 minimization," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3388–3397, Jul. 2009.

[29] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. New York: Springer, 2010.

[30] Y. C. Eldar, "Rethinking biased estimation: Improving maximum likelihood and the Cramér-Rao bound," *Foundations and Trends in Signal Processing*, vol. 1, no. 4, pp. 305–449, 2008.

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

[32] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[33] T. Strohmer and R. W. Heath, Jr, "Grassmannian frames with applications to coding and communication," *Appl. Computat. Harmon. Anal.*, vol. 14, no. 3, pp. 257–275, May 2003.

[34] 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, Mar. 2003.

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

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

[37] 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.

Zvika Ben-Haim (S'05) received the B.Sc. degree in electrical engineering and the B.A. degree in physics, both in 2000, and the M.Sc. degree in electrical engineering in 2005, all *summa cum laude* from the Technion—Israel Institute of Technology, Haifa, Israel. From 2000 to 2006, he served as a researcher in the Israel Defense Forces. He is currently working towards the Ph.D. degree in the Department of Electrical Engineering at the Technion.

From 1996 to 2000, he was a member of the Technion Excellence Program. His research is in the field of statistical signal processing, with a particular interest in theoretical performance limits of estimation techniques.

Mr. Ben-Haim was awarded the Irwin and Joan Jacobs Excellence Scholarship in 2008. In 2009, he received the Andrew and Erna Finci Viterbi Fellowship for Ph.D. Students, and in 2010, he was awarded the Wolf Foundation Prize for Outstanding Ph.D. Students.

Yonina C. Eldar (S'98–M'02–SM'07) received the B.Sc. degree in physics in 1995 and the B.Sc. degree in electrical engineering in 1996 both from Tel-Aviv University (TAU), Tel-Aviv, Israel, and the Ph.D. degree in electrical engineering and computer science in 2001 from the Massachusetts Institute of Technology (MIT), Cambridge.

From January 2002 to July 2002, she was a Postdoctoral Fellow at the Digital Signal Processing Group at MIT. She is currently a Professor in the Department of Electrical Engineering at the Technion—Israel Institute of Technology, Haifa. She is also a Research Affiliate with the Research Laboratory of Electronics at MIT and a Visiting Professor at the Statistics and Electrical Engineering Departments at Stanford. Her research interests are in the general areas of sampling theory, statistical signal processing, optimization methods and their applications to biology and optics.

Dr. Eldar was in the program for outstanding students at TAU from 1992 to 1996. In 1998, she held the Rosenblith Fellowship for study in Electrical Engineering at MIT, and in 2000, she held an IBM Research Fellowship. From 2002–2005 she was a Horev Fellow of the Leaders in Science and Technology program at the Technion and an Alon Fellow. In 2004, she was awarded the Wolf Foundation Krill Prize for Excellence in Scientific Research, in 2005 the Andre and Bella Meyer Lectureship, in 2007 the Henry Taub Prize for Excellence in Research, in 2008 the Hershel Rich Innovation Award, the Award for Women with Distinguished Contributions, the Muriel & David Jacknow Award for Excellence in Teaching, and the Technion Outstanding Lecture Award, in 2009 the Technion's Award for Excellence in Teaching, and in 2010 the Michael Bruno Memorial Award from the Rothschild Foundation. She is a member of the IEEE Signal Processing Theory and Methods technical committee and the Bio Imaging Signal Processing technical committee, an Associate Editor for the IEEE Transactions on Signal Processing, the *EURASIP Journal of Signal Processing*, the *SIAM Journal on Matrix Analysis and Applications*, and the *SIAM Journal on Imaging Science*s, and on the Editorial Board of *Foundations and Trends in Signal Processing*.

Michael Elad (SM'05) received the B.Sc., M.Sc., and D.Sc. degrees from the Department of Electrical Engineering at the Technion—Israel Institute of Technology, Haifa.

From 1988 to 1993, he served in the Israeli Air Force. From 1997 to 2000, he worked at Hewlett-Packard Laboratories Israel as an R&D engineer. During 2000 to 2001, he headed the research division at Jigami Corporation, Israel. During 2001–2003, he was a postdoctoral researcher at Stanford University. Since late 2003, he has been a faculty member at the Computer Science Department at the Technion. In May 2007, he was tenured to an associate professorship. He works in the field of signal and image processing, specializing in particular on inverse problems, sparse representations, and overcomplete transforms.

Dr. Elad received the Technion's best lecturer award six times, he is the recipient of the Solomon Simon Mani award for excellence in teaching in 2007, and he is also the recipient of the Henri Taub Prize for academic excellence (2008) and the Hershel-Rich prize (2010) for innovation. He is a senior IEEE member, and he is currently serving as an associate editor for *SIAM Journal on Imaging Sciences*.