COHERENCE-BASED NEAR-ORACLE PERFORMANCE GUARANTEES FOR SPARSE ESTIMATION UNDER GAUSSIAN NOISE

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

Technion—Israel Institute of Technology, Haifa 32000, Israel
{zvikabh@tx, yonina@ee, elad@cs}.technion.ac.il

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, orthogonal matching pursuit, and thresholding. These approaches 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.

Index Terms— Sparse estimation, basis pursuit, matching pursuit, thresholding algorithm, oracle

1. INTRODUCTION

Consider the setting in which 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 dictionary and $\mathbf{w}$ is white Gaussian noise. It is assumed that $\mathbf{x}_0$ is sparse, i.e., that most elements of $\mathbf{x}_0$ are zero. Several estimation approaches have been proposed for this setting. These include greedy algorithms, such as thresholding and orthogonal matching pursuit (OMP) [1], and ℓ_1-based methods, such as the Dantzig selector [2] and basis pursuit denoising (BPDN) [3] (also known as the Lasso).

Candès and Tao [2] have shown that the Dantzig selector is close to optimal under the assumption of Gaussian random noise. Specifically, they proved that, with high probability, the ℓ_2 distance between $\mathbf{x}_0$ and the Dantzig estimate is within $O(\log m)$ of the performance of an ideal "oracle" estimator, which knows the locations of the nonzero elements of $\mathbf{x}_0$. Recently, Bickel et al. [4] have shown that BPDN also shares this property. To the best of our knowledge, there are no such guarantees for greedy algorithms under random noise.

The contributions [2, 4] state their results using the restricted isometry property (RIP). This measure is useful when information about the RIP constants is available (e.g., when $\mathbf{A}$ is selected randomly from an appropriate ensemble). However, in general it is NP-hard to evaluate the RIP constants. These constants must then be bounded by efficiently computable properties of $\mathbf{A}$, such as the mutual coherence. Although weaker than RIP results, coherence-based bounds are appealing since they can be used with arbitrary dictionaries.

In this paper, we seek performance guarantees for sparse estimators based directly on the mutual coherence of the matrix $\mathbf{A}$. While these results are suboptimal when the RIP constants of $\mathbf{A}$ are known, the coherence approach allows us to derive tighter bounds than those obtained by applying coherence bounds to RIP-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. Furthermore, for BPDN, our result is tighter than previous performance guarantees [4].

2. BACKGROUND

Let $\mathbf{x}_0 \in \mathbb{R}^m$ be an unknown vector, and denote its support set by Λ_0. Let $s = \|\mathbf{x}_0\|_0$ be the number of nonzero entries in $\mathbf{x}_0$. We assume that s is known to be substantially smaller than m, i.e., most elements in $\mathbf{x}_0$ are zero. Suppose we obtain the measurements

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

where $\mathbf{A} \in \mathbb{R}^{n\times m}$ is a known dictionary and $\mathbf{w}$ is white Gaussian noise with variance σ^2. We assume that the columns ("atoms") $\mathbf{a}_i$ of $\mathbf{A}$ are normalized such that $\|\mathbf{a}_i\|_2 = 1$.

Given an index set Λ, denote by $\mathbf{A}_\Lambda$ the submatrix formed from the columns of $\mathbf{A}$ referenced by Λ. Using this notation, we briefly recall some characteristics of the matrix $\mathbf{A}$. First, $\mathbf{A}$ is said to satisfy the RIP of order s with parameter δ_s [2] if, for every index set Λ of size s, we have

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

for all $\mathbf{y} \in \mathbb{R}^s$. Similarly, $\mathbf{A}$ is said to satisfy the restricted orthogonality property (ROP) of order (s_1, s_2) with parameter θ_{s_1,s_2} [2] if, for every pair of disjoint index sets Λ_1 and Λ_2 having cardinalities s_1 and s_2, respectively, we have

$$\left|\mathbf{y}_1^T \mathbf{A}_{\Lambda_1}^T \mathbf{A}_{\Lambda_2} \mathbf{y}_2\right| \le \theta_{s_1,s_2}\|\mathbf{y}_1\|_2\|\mathbf{y}_2\|_2 \tag{3}$$

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).

for all $\mathbf{y}_1 \in \mathbb{R}^{s_1}$ and for all $\mathbf{y}_2 \in \mathbb{R}^{s_2}$.

While the RIP and ROP are accurate characterizations of $\mathbf{A}$, their computation for a predetermined dictionary is NP-hard in general. By contrast, the mutual coherence

$$\mu \triangleq \max_{i \neq j} \left| \mathbf{a}_i^T \mathbf{a}_j \right| \tag{4}$$

can be computed efficiently directly from (4). The RIP and ROP constants can be bounded by μ, as demonstrated by the following lemma [5]. Thus, RIP-based guarantees [2, 4] can be applied to cases in which only the coherence is known.

Lemma 1 *For any matrix $\mathbf{A}$ and any s, s_1, and s_2,*

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

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

To fix notation, we now briefly describe techniques for approximating $\mathbf{x}_0$ from measurements given by (1). Two main strategies are available to this end: ℓ_1 relaxation methods and greedy techniques. Relaxation approaches include BPDN, which is a solution $\widehat{\mathbf{x}}_{\mathrm{BP}}$ to the quadratic program

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

and the Dantzig selector, given by a solution $\widehat{\mathbf{x}}_{\mathrm{DS}}$ to

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{A}^T(\mathbf{b} - \mathbf{A}\mathbf{x})\|_\infty \leq \tau. \tag{8}$$

In (7) and (8), the constants γ and τ control a tradeoff between sparsity and conformance to the measurements.

Rather than solving an optimization problem, greedy approaches estimate the support set Λ_0 from the measurements $\mathbf{b}$. The simplest of these approaches is the thresholding algorithm, which selects a support containing the s atoms most highly correlated with $\mathbf{b}$. The estimate itself is then chosen by finding the least-squares (LS) solution within this support.

A somewhat more sophisticated greedy approach is OMP [1]. This algorithm keeps track of a "residual" $\mathbf{r}$, which contains the portion of $\mathbf{b}$ yet to be accounted for by the estimate. The algorithm iteratively finds the single atom which is most highly correlated with $\mathbf{r}$, adds it to the support set, and recalculates the residual. We consider a version of this algorithm in which s iterations are performed, yielding an estimate with s nonzero entries.

3. PERFORMANCE GUARANTEES

A common technique for assessing the quality of an estimator is to compare its performance to the lowest mean-squared error (MSE) obtainable in the given setting. One way to achieve this is through the Cramér–Rao bound (CRB), a limit on the MSE of any unbiased estimator. It has recently been shown [6] that the CRB for our setting (1) is given by

$$\sigma^2 \operatorname{Tr}((\mathbf{A}_{\Lambda_0}^T \mathbf{A}_{\Lambda_0})^{-1}). \tag{9}$$

Thus, no unbiased estimator can achieve an MSE below (9). A different strategy for appraising estimators is to compare practical techniques with the "oracle" estimator, which is the LS solution among vectors $\mathbf{x}$ whose support is Λ_0 [2]. The MSE of the oracle estimator is also given by (9), so that the CRB and oracle approaches are equivalent in this respect.

For reasonable dictionaries and sparsity levels, (9) is on the order of $s\sigma^2$. Thus, the oracle estimator has substantially reduced the effect of the noise, whose input power is $E\{\|\mathbf{w}\|_2^2\} = n\sigma^2$. Our goal in this paper is to determine whether comparable performance gains can be achieved by practical methods.

This question was first addressed in the context of the Dantzig selector (8) by Candès and Tao [2]. Their result is derived using the RIP and ROP constants (2)–(3). For a given dictionary $\mathbf{A}$, one can obtain the following coherence-based guarantee on the performance of the Dantzig selector by applying Lemma 1 to [2, Th. 1.1].

Theorem 1 (Candès and Tao) *Assume that*

$$s < 1 + \frac{1}{(1+\sqrt{2})\mu} \tag{10}$$

and consider the Dantzig selector (8) *with parameter $\tau = \sigma\sqrt{2(1+\alpha)\log m}$, for some constant $\alpha > 0$. Then, with probability exceeding*

$$1 - \frac{1}{m^\alpha \sqrt{\pi \log m}}, \tag{11}$$

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

$$\|\mathbf{x}_0 - \widehat{\mathbf{x}}_{\mathrm{DS}}\|_2^2 \leq \frac{32(1+\alpha)}{[1-((1+\sqrt{2})s-1)\mu]^2} s\sigma^2 \log m. \tag{12}$$

This theorem is significant because it demonstrates that, while $\widehat{\mathbf{x}}_{\mathrm{DS}}$ does not quite reach the performance of the CRB (9), it does come within a constant factor multiplied by $\log m$, with high probability. It turns out that the $\log m$ factor is an unavoidable result of the fact that the locations of the nonzero elements in $\mathbf{x}_0$ are unknown [7, §7.4].

Bickel et al. [4] have recently demonstrated similar results for BPDN; they showed that, with high probability, BPDN 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 their BPDN guarantee is always larger than 128, which is considerably weaker than that of Theorem 1.

By directly using the mutual coherence of $\mathbf{A}$, an improved coherence-based performance guarantee can be obtained for BPDN. This is demonstrated in the following theorem [5].

Theorem 2 *Suppose that*[1] *$s < 1/(3\mu)$, and let $\widehat{\mathbf{x}}_{\mathrm{BP}}$ be a solution of BPDN* (7) *with $\gamma = \sqrt{8\sigma^2(1+\alpha)\log(m-k)}$.*

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

Then, with probability exceeding

$$\left(1 - \frac{1}{(m-s)^{\alpha}}\right)\left(1 - e^{-s/7}\right), \tag{13}$$

we have

$$\|\mathbf{x}_0 - \widehat{\mathbf{x}}_{\mathrm{BP}}\|_2^2 \leq \left(\sqrt{3} + 3\sqrt{2(1+\alpha)\log(m-s)}\right)^2 s\sigma^2. \tag{14}$$

Let us compare Theorems 1 and 2 in terms of probabilities of success. Eq. (13) is 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, but this is only an artifact of the method of proof (see [5] for details). This term converges to 1 exponentially and typically has a negligible effect on the overall probability of success. The left-hand term in (13), which is the significant factor in the probability of success, tends to 1 polynomially as $m-s$ increases, while in (11) the probability tends to 1 polynomially in m. However, for both theorems to hold, we must have $m \gg s$, so this difference is negligible. We also note that for $m \to \infty$ and constant s, (11) tends to 1, while (13) does not. On the other hand, if both $m-s$ and s tend to infinity, then both probabilities converge to 1.

We next compare the error bounds provided by the two theorems. For large s and $m-s$, the result (14) for BPDN is on the order of $18(1+\alpha)s\sigma^2\log(m-s)$, while the result (12) for the Dantzig selector is $C(1+\alpha)s\sigma^2\log m$, where $C \geq 32$ and, depending on the values of s and μ, may be much larger.

The above performance guarantees assumed only 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 all support elements are correctly identified. To ensure support recovery, all nonzero elements must be large enough to overcome the noise.

To formalize this notion, we denote by $|x_{\min}|$ and $|x_{\max}|$, respectively, the smallest and largest values of $|x_{0,i}|$ among $i \in \Lambda_0$. A performance guarantee for both OMP and the thresholding algorithm can then be stated as follows [5].

Theorem 3 *Suppose that*

$$|x_{\min}| - (2s-1)\mu|x_{\min}| \geq 2\sigma\sqrt{2(1+\alpha)\log m} \tag{15}$$

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

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

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

$$\|\widehat{\mathbf{x}}_{\mathrm{OMP}} - \mathbf{x}_0\|_2^2 \leq \frac{2(1+\alpha)}{(1-(s-1)\mu)^2} s\sigma^2\log m \tag{17a}$$

$$\leq 8(1+\alpha)s\sigma^2\log m. \tag{17b}$$

If the stronger condition

$$|x_{\min}| - (2s-1)\mu|x_{\max}| \geq 2\sigma\sqrt{2(1+\alpha)\log m} \tag{18}$$

holds, then with probability exceeding (16)*, the thresholding algorithm also correctly identifies Λ_0 and satisfies* (17)*.*

The performance guarantee (17) is better than that provided by Theorems 1 and 2. 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 fail when the noise level increases. As we will see in Section 4, simulations also appear to support this conclusion.

It is interesting to compare the success conditions (15) and (18) 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. 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.

4. NUMERICAL RESULTS

The results of Section 3 guarantee that, with high probability, a specified distance between $\mathbf{x}_0$ and its estimate will be achieved. In other words, these are "nearly worst-case" guarantees in that they hold for all but a very unlikely set of noise realizations. In practice, however, it is more common to measure the MSE of an estimator. The MSE is likely to be somewhat lower than the theoretical results of Section 3, as it averages the errors of all noise realizations. Thus, our next goal is to determine whether the behavior predicted by Theorems 1–3 is also manifested in the MSE of the various estimators.

To this end, we conducted a series of simulations in which the MSEs of the estimators of Section 2 were compared. The regularization parameters τ and γ of the Dantzig selector and BPDN were chosen as recommended by Theorems 1 and 2, respectively. For these estimators a value of $\alpha = 1$ was chosen; thus, the guaranteed probability of success for the two algorithms has the same order of magnitude.

We chose the two-ortho dictionary $\mathbf{A} = [\mathbf{I} \;\; \mathbf{H}]$, where $\mathbf{I}$ is the 128×128 identity matrix and $\mathbf{H}$ is the 128×128 Hadamard matrix with normalized columns. The RIP and

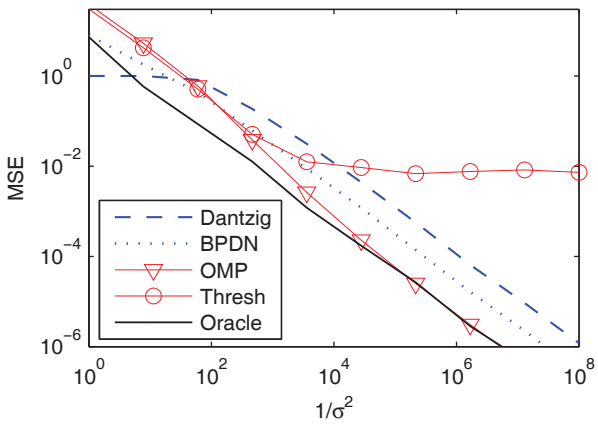

Fig. 1. MSE of various estimators as a function of the SNR.

ROP constants of this dictionary are unknown, but the coherence can be readily calculated and is given by $\mu \approx 0.088$. Similar experiments were performed on a variety of other dictionaries and yielded comparable results (not reported here).

The parameter vector $\mathbf{x}_0$ was obtained 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 measurements $\mathbf{b}$ were obtained from (1), and $\mathbf{x}_0$ was estimated using the algorithms defined in Section 2. The MSE of each estimate was then calculated by averaging over repeated realizations of $\mathbf{x}_0$ and the noise. The experiment was repeated for 10 values of the noise variance σ^2 and the results are plotted in Fig. 1.

To compare this plot with the theoretical results of Section 3, observe first the situation at high signal-to-noise ratio (SNR). Here, OMP, BPDN, and the Dantzig selector all achieve performance which is proportional to the oracle MSE (or CRB) of (9). Among these, OMP is closest to the CRB, followed by BPDN and the Dantzig selector. This behavior matches the proportionality constants given in the theorems of Section 3. Furthermore, for small σ, the condition (15) holds even for large α, and thus Theorem 3 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 3, since, even when $\sigma = 0$, the condition (18) does not always hold, unless $|x_{\min}|$ is not much smaller than $|x_{\max}|$. Thus, for our choice of $\mathbf{x}_0$, Theorem 3 does not guarantee near-oracle performance for the thresholding algorithm, even at high SNR.

With increasing noise, Theorem 3 requires a corresponding increase in $|x_{\min}|$ to guarantee the success of the greedy algorithms. Consequently, Fig. 1 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 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.

5. DISCUSSION

The performance of an estimator depends on the problem setting under consideration. For example, suppose the parameter $\mathbf{x}_0$ and the noise $\mathbf{w}$ are both deterministic and assume that $\|\mathbf{w}\|_2 \leq \varepsilon$. In this case, the estimation error of any algorithm can be as high as ε; in other words, the assumption of sparsity has not yielded any reduction in noise power [8]. 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 [9]. In this paper, we have examined a middle ground between these two extremes, namely the frequentist 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 guaranteed to achieve performance close to that of the oracle estimator.

6. 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] 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.

[5] Z. Ben-Haim, Y. C. Eldar, and M. Elad, "Coherence-based performance guarantees for estimating a sparse vector under random noise," *IEEE Trans. Signal Process.*, submitted. [Online]. Available: http://arxiv.org/abs/0903.4579

[6] 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

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

[8] 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.

[9] 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.