PERFORMANCE BOUNDS FOR SPARSE ESTIMATION WITH RANDOM NOISE

Zvika Ben-Haim and Yonina C. Eldar

Technion—Israel Institute of Technology
Haifa 32000, Israel

ABSTRACT

The problem considered in this paper is to estimate a deterministic vector representing elements in an overcomplete dictionary. The vector is assumed to be sparse and is to be estimated from measurements corrupted by Gaussian noise. Our goal is to derive a lower bound on the mean-squared error (MSE) achievable in this setting. To this end, an appropriate definition of unbiasedness in the sparse setting is developed, and the unbiased Cramér–Rao bound (CRB) is derived. The resulting bound is shown to be identical to the MSE of the oracle estimator. Combined with the fact that the CRB is achieved at high signal-to-noise ratios by the maximum likelihood technique, our result provides a new interpretation for the common practice of using the oracle estimator as a gold standard against which practical approaches are compared.

Index Terms— Sparse estimation, Cramér–Rao bound

1. INTRODUCTION

The problem of estimating a sparse vector has been analyzed intensively in recent years, and its applications span diverse fields in signal processing and statistics [1–4]. We consider the setting in which a deterministic sparse vector $\mathbf{x}_0 \in \mathbb{R}^m$ is to be estimated from measurements $\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w}$, where $\mathbf{w}$ is white Gaussian noise and $\mathbf{A} \in \mathbb{R}^{n \times m}$ is a deterministic dictionary for which $m > n$. The maximum-likelihood (ML) estimator for this setting cannot be calculated efficiently by any known algorithm. However, several alternative techniques are surprisingly successful in estimating $\mathbf{x}_0$; these include the Dantzig selector (DS) [4] and basis pursuit denoising (BPDN), which is also referred to as the Lasso [1, 2, 5].

In this paper, we characterize the best achievable mean-squared error (MSE) of estimators of $\mathbf{x}_0$, and compare this lower bound with the actual performance obtained by practical techniques. A common approach for obtaining lower bounds on the MSE is to derive the Cramér–Rao bound (CRB) for unbiased estimators. Among other reasons, the CRB is a meaningful lower bound because it is typically achieved by the ML estimator when the noise variance is small.

To apply these concepts to the sparse setting, we first define an appropriate extension of the notion of unbiasedness. The unbiased CRB is then derived for the sparse estimation scenario, based on the concept of a constrained CRB [6]. As we show, the unbiased CRB equals the MSE of the "oracle estimator" which knows the locations of the nonzero entries of $\mathbf{x}_0$. The CRB can thus be viewed as an alternative justification for the common use of the oracle estimator as a baseline against which practical estimators are compared. This gives further merit to recent results, which demonstrate that BPDN and the DS both achieve near-oracle performance [4, 7].

2. PROBLEM SETTING

Let $\mathbf{x}_0 \in \mathbb{R}^m$ be an unknown deterministic vector satisfying

$$\mathbf{x}_0 \in \mathcal{S} \triangleq \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x}\|_0 \le k\} \tag{1}$$

for some known integer $k \ll m$. Here, $\|\mathbf{x}\|_0$ denotes the number of nonzero entries in $\mathbf{x}$. We refer to the indices of these nonzero components as the support set, and denote the support set of the true parameter $\mathbf{x}_0$ by Λ_0.

Our goal is to reconstruct $\mathbf{x}_0$ from the measurements

$$\mathbf{b} = \mathbf{A}\mathbf{x}_0 + \mathbf{w} \tag{2}$$

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

To estimate $\mathbf{x}_0$, one might consider the ML technique

$$\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \quad \text{s.t. } \|\mathbf{x}\|_0 \le k. \tag{3}$$

Unfortunately, solving (3) is NP-hard, meaning that an efficient algorithm providing the ML estimator for general $\mathbf{A}$ is unlikely to exist. Consequently, several practical alternatives have been proposed for estimating $\mathbf{x}_0$. One of these is the ℓ_1-penalty version of BPDN [1], which is obtained by solving

$$\min_{\mathbf{x}} \tfrac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \gamma\|\mathbf{x}\|_1 \tag{4}$$

with some regularization parameter γ. More recently, an alternative known as the DS was proposed [4]; this approach

This work was supported in part by the Israel Science Foundation under Grant 1081/07, by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (contract no. 216715), and by the Goldstein UAV and Satellite Center.

estimates $\mathbf{x}_0$ as a solution $\widehat{\mathbf{x}}_{\mathrm{DS}}$ to the optimization problem

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{A}^*(\mathbf{b}-\mathbf{A}\mathbf{x})\|_\infty \le \tau \tag{5}$$

where τ is again a user-selected parameter. A modification of the DS, known as the Gauss–Dantzig selector (GDS) [4], is to use $\widehat{\mathbf{x}}_{\mathrm{DS}}$ only to estimate the support set Λ_0. In this approach, one solves (5) and determines the support set of $\widehat{\mathbf{x}}_{\mathrm{DS}}$. The GDS estimate is then obtained as

$$\widehat{\mathbf{x}}_{\mathrm{GDS}} = \begin{cases} \mathbf{A}_\Lambda^\dagger \mathbf{b} & \text{on the support set of } \widehat{\mathbf{x}}_{\mathrm{DS}} \\ \mathbf{0} & \text{elsewhere} \end{cases} \tag{6}$$

where the submatrix $\mathbf{A}_\Lambda$ consists of the columns of $\mathbf{A}$ corresponding to the support of $\widehat{\mathbf{x}}_{\mathrm{DS}}$.

3. PERFORMANCE BOUNDS

3.1. Background

The MSE of an estimator is, in general, a function of the unknown parameter $\mathbf{x}_0$; an estimator might be better suited for some values of $\mathbf{x}_0$ than others. Performance bounds must thus describe the achievable MSE as a function of $\mathbf{x}_0$.

Previous research has examined the performance of estimation techniques in terms of their worst-case MSE among all possible values $\mathbf{x}_0 \in \mathcal{S}$. Specifically, it has been shown [4] that, for sufficiently small k and for an appropriate choice of the regularization parameter τ, the DS of (5) satisfies

$$\|\mathbf{x}_0 - \widehat{\mathbf{x}}_{\mathrm{DS}}\|_2^2 \le Ck\sigma^2 \log m \quad \text{with high probability} \tag{7}$$

for some constant C. More recently, an identical property was demonstrated for BPDN (4) with an appropriate choice of γ [7]. Conversely, the worst-case error of *any* estimator is at least a constant times $k\sigma^2 \log m$ [8, §7.4]. Thus, both BPDN and DS are optimal, up to a constant, in terms of worst-case error. Nevertheless, the MSE of these approaches for specific values of $\mathbf{x}_0$, even for a vast majority of such values, might be much lower. Our goal is therefore to characterize the *pointwise* performance of an estimator, i.e., the MSE for specific values of $\mathbf{x}_0$.

One baseline with which practical techniques are often compared is the oracle estimator, given by

$$\widehat{\mathbf{x}}_{\mathrm{or}} = \begin{cases} \mathbf{A}_{\Lambda_0}^\dagger \mathbf{b} & \text{on the support set } \Lambda_0 \\ \mathbf{0} & \text{elsewhere} \end{cases} \tag{8}$$

where $\mathbf{A}_{\Lambda_0}$ is the submatrix constructed from the columns of $\mathbf{A}$ corresponding to the nonzero entries of $\mathbf{x}_0$. In other words, $\widehat{\mathbf{x}}_{\mathrm{or}}$ is the least-squares (LS) solution among vectors whose support coincides with Λ_0, which is assumed to have been provided by an "oracle." Of course, in practice Λ_0 is unknown, so that $\widehat{\mathbf{x}}_{\mathrm{or}}$ cannot actually be implemented. Nevertheless, one often compares the performance of true estimators with $\widehat{\mathbf{x}}_{\mathrm{or}}$, whose MSE is given by [4]

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

While $\widehat{\mathbf{x}}_{\mathrm{or}}$ is a reasonable technique to adopt if Λ_0 is known, this does not imply that (9) is a lower bound on the performance of practical estimators. Indeed, as shown in Section 4, when the SNR is low, both BPDN and the DS outperform $\widehat{\mathbf{x}}_{\mathrm{or}}$, thanks to the use of shrinkage in these estimators. If Λ_0 is known, then there are even techniques which are better than $\widehat{\mathbf{x}}_{\mathrm{or}}$ for *all* values of $\mathbf{x}_0$ [9]. In the sequel, we demonstrate that one can indeed interpret (9) as a lower bound on the achievable MSE, but such a result requires a certain restriction of the class of estimators under consideration.

3.2. Unbiasedness in the Sparse Setting

To obtain a meaningful pointwise bound on the MSE, one must exclude some estimators from consideration; otherwise, the bound will be tarnished by estimators such as $\widehat{\mathbf{x}} = \mathbf{x}_{\mathrm{u}}$, for some constant $\mathbf{x}_{\mathrm{u}}$, which achieve an MSE of 0 at the specific point $\mathbf{x}_0 = \mathbf{x}_{\mathrm{u}}$. Thus, the only pointwise lower bound on the MSE of all estimators is 0. This is clearly not a useful result.

A standard method of working around this difficulty is to restrict attention to unbiased estimators, i.e., techniques $\widehat{\mathbf{x}}$ whose bias $\mathbf{b}(\widehat{\mathbf{x}}, \mathbf{x}_0) \triangleq E\{\widehat{\mathbf{x}} - \mathbf{x}_0\}$ is zero. Thus, while the estimate is not always accurate, it yields the correct value $\mathbf{x}_0$ "on average." Furthermore, for high SNR, it can be shown that biased estimators are suboptimal. Admittedly, there are situations in which bias is productive (see Section 4). However, in this paper we focus on the unbiasedness approach, and attempt to adapt it to the sparse estimation framework.

Observe first that since $\mathbf{A}$ is overcomplete, no estimator can be unbiased for all $\mathbf{x}_0 \in \mathbb{R}^m$. Indeed, all values of $\mathbf{x}_0$ in the nullspace of $\mathbf{A}$ yield an identical distribution of $\mathbf{b}$, so an estimator can be unbiased for one of these values at most. This is a consequence of the underdetermined nature of (2).

The question is whether it is possible to construct estimators which are unbiased over a subset of $\mathbb{R}^m$. For example, it is always possible to construct a technique which is unbiased at a single point, say $\mathbf{x}_{\mathrm{u}}$: this is again $\widehat{\mathbf{x}} = \mathbf{x}_{\mathrm{u}}$, which is unbiased at $\mathbf{x}_{\mathrm{u}}$ but nowhere else. To avoid this loophole, one can require an estimator to be unbiased in the neighborhood

$$\mathcal{B}_\varepsilon(\mathbf{x}_0) = \{\mathbf{x} \in \mathbb{R}^m : \|\mathbf{x} - \mathbf{x}_0\|_2 \le \varepsilon\} \tag{10}$$

of $\mathbf{x}_0$, for some small ε. It follows that both the bias $\mathbf{b}(\mathbf{x})$ and the bias gradient $\partial\mathbf{b}/\partial\mathbf{x}$ vanish at $\mathbf{x} = \mathbf{x}_0$. This formulation is the basis of the CRB, a lower bound on the MSE at $\mathbf{x}_0$ which applies to all estimators whose bias and bias gradient are zero at $\mathbf{x}_0$.

However, it turns out that even this requirement is too stringent: the CRB for the estimation problem (2) is infinite, implying that no finite-variance estimator is unbiased in any ε-neighborhood of any $\mathbf{x}_0$ [10]. The reason is related to the fact that unbiasedness is required over the m-dimensional set $\mathcal{B}_\varepsilon(\mathbf{x}_0)$, whereas only $n < m$ measurements are available.

A reasonable compromise is to require unbiasedness over

$\mathcal{B}_\varepsilon(\mathbf{x}_0) \cap \mathcal{S}$, i.e., over the neighborhood of $\mathbf{x}_0$ restricted[1] to the feasible set $\mathcal{S}$ of (1). This leads to a weaker requirement on the bias gradient: namely, for any vector $\mathbf{v} \in \mathbb{R}^m$ such that $(\mathbf{x}_0 + \varepsilon\mathbf{v}) \in \mathcal{S}$ for all sufficiently small ε, it is required that

$$\left.\frac{\partial \mathbf{b}}{\partial \mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_0} \cdot \mathbf{v} = \mathbf{0}. \tag{11}$$

The vector $\mathbf{v}$ represents a direction by which $\mathbf{x}_0$ can change without violating the constraint set $\mathcal{S}$. Thus, we relax the demand $\partial\mathbf{b}/\partial\mathbf{x} = \mathbf{0}$ by requiring that the bias gradient vanish only in feasible directions. We refer to this requirement as $\mathcal{S}$-unbiasedness at $\mathbf{x}_0$. Under this definition, one can construct a CRB based on the methodology of [6], as shown below.

3.3. The Cramér–Rao Bound

The following theorem demonstrates that, under the definition of $\mathcal{S}$-unbiasedness obtained above, the CRB is finite for most, but not all, values of $\mathbf{x}_0$ in $\mathcal{S}$. The proof of Theorem 1 follows the lines of [11] and appears in full in [10].

Theorem 1 *Let $\widehat{\mathbf{x}}$ be a finite-variance estimator of a parameter $\mathbf{x}_0 \in \mathcal{S}$ from observations $\mathbf{b}$ given by (2). Then, $\widehat{\mathbf{x}}$ cannot be $\mathcal{S}$-unbiased at any point for which $\|\mathbf{x}_0\|_0 < k$. Also, if $\widehat{\mathbf{x}}$ is $\mathcal{S}$-unbiased at a point for which $\|\mathbf{x}_0\|_0 = k$, then*

$$E\left\{\|\widehat{\mathbf{x}} - \mathbf{x}_0\|_2^2\right\} \geq \sigma^2 \operatorname{Tr}((\mathbf{A}_{\Lambda_0}^* \mathbf{A}_{\Lambda_0})^{-1}). \tag{12}$$

The most striking feature of Theorem 1 is the fact that for $\|\mathbf{x}_0\|_0 = k$, the CRB (12) is identical to the oracle MSE (9). However, the CRB is of additional importance because of the fact that the ML estimator achieves the CRB in the limit when a large number of independent measurements are available, a situation which is equivalent in our setting to the limit $\sigma \to 0$. In other words, an MSE of (9) is achieved at high SNR by the ML approach (3). While the ML approach is computationally intractable in the sparse estimation setting, it is still implementable in principle, as opposed to $\widehat{\mathbf{x}}_{\text{or}}$, which relies on unavailable information (namely, the support set of $\mathbf{x}_0$). Thus, Theorem 1 gives an alternative interpretation to comparisons of estimator performance with the oracle.

Next, suppose that $\|\mathbf{x}_0\|_0 < k$. In this case, changing any single entry in $\mathbf{x}_0$, even an entry not in Λ_0, will not violate the constraint set $\mathcal{S}$. Thus, $\mathcal{B}_\varepsilon(\mathbf{x}_0) \cap \mathcal{S}$ is an m-dimensional set at such points. Theorem 1 then states that no estimator can be expected to be unbiased for such a high-dimensional set, just as unbiased estimation is impossible in the m-dimensional set $\mathcal{B}_\varepsilon(\mathbf{x}_0)$. However, it is still possible to obtain a finite CRB in this setting by further restricting the definition of unbiasedness, as follows. If it is known that $\|\mathbf{x}_0\|_0 = \tilde{k} < k$, then one can redefine $\mathcal{S}$ in (1) by replacing k with $\tilde{k}$. This will reduce the class of estimators considered $\mathcal{S}$-unbiased, and Theorem 1 would then provide a finite lower bound on those estimators.

Observe that the bound (12) depends on the value of $\mathbf{x}_0$ (through its support set Λ_0), which implies that some values of $\mathbf{x}_0$ are more difficult to estimate than others. However, for sufficiently incoherent dictionaries, $\operatorname{Tr}((\mathbf{A}_{\Lambda_0}^* \mathbf{A}_{\Lambda_0})^{-1})$ is bounded above and below by a small constant times k, so that in this case the CRB is similar for all values of $\mathbf{x}_0$. To see this, let μ be the coherence of $\mathbf{A}$, defined as

$$\mu \triangleq \max_{i \neq j} |\mathbf{a}_i^* \mathbf{a}_j| \,. \tag{13}$$

By the Gershgorin disc theorem, the eigenvalues of $\mathbf{A}_{\Lambda_0}^* \mathbf{A}_{\Lambda_0}$ are in the range $[1 - k\mu, 1 + k\mu]$. It follows that the CRB (12) is bounded above and below by

$$\frac{k\sigma^2}{1 + k\mu} \leq \sigma^2 \operatorname{Tr}((\mathbf{A}_{\Lambda_0}^* \mathbf{A}_{\Lambda_0})^{-1}) \leq \frac{k\sigma^2}{1 - k\mu}. \tag{14}$$

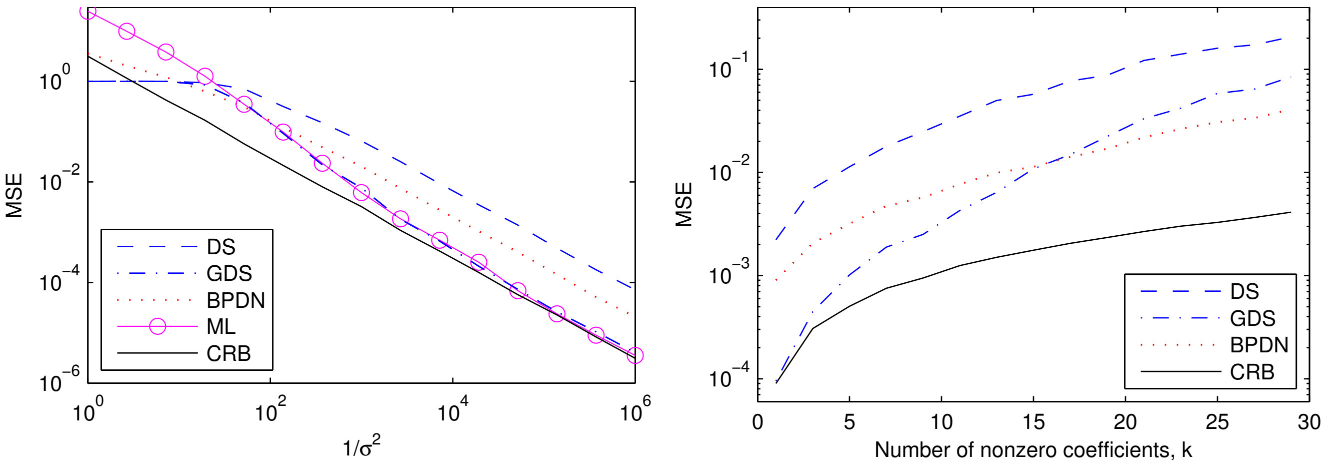
In other words, when k is somewhat smaller than $1/\mu$, the CRB is roughly equal to $k\sigma^2$ for all values of $\mathbf{x}_0$. As we have seen in Section 3.1, for sufficiently small k, the worst-case MSE of practical estimators, such as BPDN and the DS, is $O(k\sigma^2 \log m)$. Thus, practical estimators come almost within a constant of the unbiased CRB, implying that they are close to optimal for all values of $\mathbf{x}_0$, at least when compared with unbiased techniques.

4. NUMERICAL RESULTS

To demonstrate the use of the CRB for measuring the achievable MSE in a sparse estimation problem, a series of simulations was performed. Specifically, a random 100×200 dictionary $\mathbf{A}$ was constructed from a zero-mean Gaussian IID distribution, whose columns were normalized so that $\|\mathbf{a}_i\|_2 = 1$. A random parameter $\mathbf{x}_0$ was selected by choosing a support uniformly at random and selecting the nonzero elements as Gaussian IID variables with mean 0 and variance 1. Noisy measurements $\mathbf{b}$ were obtained from (2), and $\mathbf{x}_0$ was then estimated using BPDN (4), the DS (5), and the GDS (6). The regularization parameters were chosen as $\tau = 2\sigma\sqrt{\log m}$ and $\gamma = 4\sigma\sqrt{\log(m-k)}$, rules of thumb which are motivated by a theoretical analysis [7]. The MSE of each estimate was then calculated by repeating this process with different realizations of the random variables. The CRB was calculated from (12).

We first examined the CRB at various SNR levels. Here, the ML estimator was also computed, in order to verify its convergence to the CRB at high SNR. This necessitated the selection of the rather low support size, $k = 3$. The MSE and CRB were calculated for 15 different SNRs by choosing values of σ in the range 1 to 10^{-3}. The MSE of the ML approach, as well as the other estimators of Section 2, is compared with the CRB in Fig. 1(a). The convergence of the ML estimator to the CRB is clearly visible in this figure. The performance of the GDS is also impressive, being as good or better than the ML approach in this setting. Apparently, at high SNR, the DS tends to correctly recover the true support set Λ_0, and thus the

[1] We assume that $\mathbf{x}_0 \in \mathcal{S}$, since otherwise $\mathcal{B}_\varepsilon(\mathbf{x}_0) \cap \mathcal{S} = \varnothing$ for small ε.



(a)

(b)

Fig. 1. MSE of various estimators compared with the CRB (12), for (a) varying SNR and (b) varying sparsity levels.

GDS (6) approaches the performance of the oracle (8). Perhaps surprisingly, applying a LS estimate on the support set obtained by BPDN (not plotted in Fig. 1) does not work well at all, and in fact results in higher MSE than a direct application of BPDN.

While the CRB of Theorem 1 provides a useful lower bound, we emphasize that it is applicable only to unbiased estimators. The bias of most estimators tends to be negligible in low-noise settings, but often increases with σ^2. Indeed, when σ^2 is as large as $\|\mathbf{x}_0\|_2^2$, the measurements carry very little useful information about $\mathbf{x}_0$, and an estimator can improve performance by shrinkage. Such a strategy, while clearly biased, yields lower MSE than a naive reliance on the noisy measurements. This is indeed the behavior of the DS and BPDN, since for large σ^2, the ℓ_1 regularization becomes the dominant term, resulting in heavy shrinkage. Consequently, the unbiased CRB no longer applies to these estimators. This is seen from the fact that some of the estimators outperform the CRB when the SNR is exceedingly low.

The performance of the estimators of Section 2, excluding the ML method, was also compared for varying sparsity levels. To this end, the simulation was repeated for 15 support sizes $1 \leq k \leq 30$, with $\sigma = 0.01$. The results are plotted in Fig. 1(b). While a substantial gap exists between the CRB and the MSE of the practical estimators in this case, a similar rate of increase is obtained as k grows. Interestingly, a drawback of the GDS approach is visible in this setting: as k increases, correct support recovery becomes more difficult, and shrinkage becomes a valuable asset for reducing the sensitivity of the estimate to random measurement fluctuations. The LS approach practiced by the GDS, which does not perform shrinkage, leads to gradual performance deterioration.

Similar results were obtained with several deterministic dictionaries $\mathbf{A}$. These are omitted due to space restrictions.

5. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[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] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimation with constraints," *IEEE Trans. Inf. Theory*, vol. 26, no. 6, pp. 1285–1301, Nov. 1990.

[7] Z. Ben-Haim, Y. C. Eldar, and M. Elad, "Near-oracle performance of basis pursuit under random noise," *IEEE Trans. Signal Process.*, Mar. 2009, submitted. [Online]. Available: http://arxiv.org/abs/0903.4579

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

[9] Z. Ben-Haim and Y. C. Eldar, "Blind minimax estimation," *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3145–3157, Sep. 2007.

[10] ——, "The Cramér–Rao bound for sparse estimation," *IEEE Trans. Signal Process.*, May 2009, submitted. [Online]. Available: http://arxiv.org/abs/0905.4378

[11] ——, "On the constrained Cramér-Rao bound with a singular Fisher information matrix," *IEEE Signal Process. Lett.*, vol. 16, no. 6, pp. 453–456, Jun. 2009.