Letter to the Editor

How well can we estimate a sparse vector? ☆

Emmanuel J. Candès [a], Mark A. Davenport [b],*

[a] Departments of Mathematics and Statistics, Stanford University, Stanford, CA 94035, United States
[b] School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, United States

ARTICLE INFO

ABSTRACT

The estimation of a sparse vector in the linear model is a fundamental problem in signal processing, statistics, and compressive sensing. This paper establishes a lower bound on the mean-squared error, which holds regardless of the sensing/design matrix being used and regardless of the estimation procedure. This lower bound very nearly matches the known upper bound one gets by taking a random projection of the sparse vector followed by an ℓ_1 estimation procedure such as the Dantzig selector. In this sense, compressive sensing techniques cannot essentially be improved.

© 2012 Elsevier Inc. All rights reserved.

1. Introduction

The estimation of a sparse vector from noisy observations is a fundamental problem in signal processing and statistics, and lies at the heart of the growing field of compressive sensing [4,5,8]. At its most basic level, we are interested in accurately estimating a vector $\boldsymbol{x} \in \mathbb{R}^n$ that has at most k nonzeros from a set of noisy linear measurements

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}, \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$. We are often interested in the underdetermined setting where m may be much smaller than n. In general, one would not expect to be able to accurately recover $\boldsymbol{x}$ when $m < n$ since there are more unknowns than observations. However it is by now well known that by exploiting sparsity, it is possible to accurately estimate $\boldsymbol{x}$.

As an example, consider what is known concerning ℓ_1 minimization techniques, which are among the most powerful and well understood with respect to their performance in noise. Specifically, if we suppose that the entries of the matrix $\boldsymbol{A}$ are i.i.d. $\mathcal{N}(0, 1/n)$, then one can show that for any $\boldsymbol{x} \in \Sigma_k := \{\boldsymbol{x}: \|\boldsymbol{x}\|_0 \leqslant k\}$, ℓ_1 minimization techniques such as the Lasso or the Dantzig selector produce a recovery $\hat{\boldsymbol{x}}$ such that

$$\frac{1}{n}\big\|(\hat{\boldsymbol{x}} - \boldsymbol{x})\big\|_2^2 \leqslant C_0 \frac{k\sigma^2}{m} \log n \tag{2}$$

holds with high probability provided that $m = \Omega(k \log(n/k))$ [6]. We refer to [3] and [9] for further results.

1.1. Criticism

A noteworthy aspect of the bound in (2) is that the recovery error increases linearly as we decrease m, and thus we pay a penalty for taking a small number of measurements. Although this effect is sometimes cited as a drawback of the compressive sensing framework, it should not be surprising — we fully expect that if each measurement has a constant SNR, then taking more measurements should reduce our estimation error.

However, there is another somewhat more troubling aspect of (2). Specifically, by filling the rows of A with i.i.d. random variables, we are ensuring that our "sensing vectors" are almost orthogonal to our signal of interest, leading to a tremendous SNR loss. To quantify this loss, suppose that we had access to an oracle that knows *a priori* the locations of the nonzero entries of x and could instead construct A with vectors localized to the support of x. For example, if m is an integer multiple of k then we could simply measure sample each coefficient directly m/k times and then average these samples. One can check that this procedure would yield an estimate obeying

$$\mathbb{E}\left[\frac{1}{n}\|\hat{x} - x\|_2^2\right] = \left(\frac{k\sigma^2}{m}\right)\left(\frac{k}{n}\right). \tag{3}$$

Thus, the performance in (2) is worse than what would be possible with an oracle by a factor of $(n/k)\log n$. When k is small, this is very large! Of course, we won't have access to an oracle in practice, but the substantial difference between (2) and (3) naturally leads one to question whether (2) can be improved upon.

1.2. Can we do better?

In this paper we will approach this question from the viewpoint of compressive sensing and/or of experimental design. Specifically, we assume that we are free to choose *both* the matrix A *and* the sparse recovery algorithm. Our results will have implications for the case where A is determined by factors beyond our control, but our primary interest will be in considering the performance obtained by the best possible choice of A. In this setting, our fundamental question is:

Can we ever hope to do better than (2)? Is there a more intelligent choice for the matrix A? Is there a more effective recovery algorithm?

In this paper we show that the answer is *no*, and that there exists no choice of A or recovery algorithm that can significantly improve upon the guarantee in (2). Specifically, we consider the worst-case error over all $x \in \Sigma_k$, i.e.,

$$M^*(A) = \inf_{\hat{x}} \sup_{x \in \Sigma_k} \mathbb{E}\left[\frac{1}{n}\|\hat{x}(y) - x\|_2^2\right]. \tag{4}$$

Our main result consists of the following bound, which establishes a fundamental limit on the minimax risk which holds for any matrix A and any possible recovery algorithm.

Theorem 1. *Suppose that we observe $y = Ax + z$ where x is a k-sparse vector, A is an $m \times n$ matrix with $m \geqslant k$, and $z \sim \mathcal{N}(0, \sigma^2 I)$. Then there exists a constant $C_1 > 0$ such that for all A,*

$$M^*(A) \geqslant C_1 \frac{k\sigma^2}{\|A\|_F^2} \log(n/k). \tag{5}$$

We also have that for all A

$$M^*(A) \geqslant \frac{k\sigma^2}{\|A\|_F^2}. \tag{6}$$

This theorem says that *there is no A* and *no recovery algorithm* that does fundamentally better than the Dantzig selector (2) up to a constant[1]; that is, ignoring the difference in the factors $\log n/k$ and $\log n$. In this sense, the results of compressive sensing are at the limit.

Although the noise model in (1) is fairly common, in some settings (such as the estimation of a signal transmitted over a noisy channel) it is more natural to consider noise that has been added directly to the signal prior to the acquisition of the measurements. In this case we can directly apply Theorem 1 to obtain the following corollary.

---

[1] Our analysis shows that asymptotically C_1 can be taken as $1/128$. We have made no effort to optimize this constant, and it is probably far from sharp. This is why we give the simpler bound (6) which is proven by considering the error we would incur even if we knew the support of x *a priori*. However, our main result is (5). We leave the calculation of an improved constant to future work.

Corollary 1. *Suppose that we observe $y = A(x + w)$ where x is a k-sparse vector, A is an $m \times n$ matrix with $k \leqslant m \leqslant n$, and $w \sim \mathcal{N}(0, \sigma^2 I)$. Then for all A*

$$M^*(A) \geqslant C_1 \frac{k\sigma^2}{m} \log(n/k) \quad and \quad M^*(A) \geqslant \frac{k\sigma^2}{m}. \tag{7}$$

Proof. We assume that A has rank $m' \leqslant m$. Let $U\Sigma V^*$ be the reduced SVD of A, where U is $m \times m'$, Σ is $m' \times m'$, and V is $n \times m'$. Applying the matrix $\Sigma^{-1} U^*$ to y preserves all the information about x, and so we can equivalently assume that the data is given by

$$y' = \Sigma^{-1} U^* y = V^* x + V^* w. \tag{8}$$

Note that $V^* w$ is a Gaussian vector with covariance matrix $\sigma^2 V^* V = \sigma^2 I$. Moreover, V^* has unit-norm rows, so that $\|V^*\|_F \leqslant m' \leqslant m$. We then apply Theorem 1 to establish (7). □

The intuition behind this result is that when noise is added to the measurements, we can boost the SNR by rescaling A to have higher norm. When we instead add noise to the signal, the noise is also scaled by A, and so no matter how A is designed there will always be a penalty of $1/m$.

1.3. Related work

There have been a number of prior works that have established lower bounds on $M^*(A)$ or related quantities under varying assumptions [1,13–17,19]. In [1,17], techniques from information theory similar to the ones that we use below are used to establish rather general lower bounds under the assumption that the entries of x are generated i.i.d. according to some distribution. For an appropriate choice of distribution, x will be approximately sparse and [1,17] will yield asymptotic lower bounds of a similar flavor to ours.

The prior work most closely related to our results is that of Ye and Zhang [19] and Raskutti, Wainwright, and Yu [15]. In [19] Ye and Zhang establish a bound similar to (5) in Theorem 1. While the resulting bounds are substantially the same, the bounds in [19] hold only in the asymptotic regime where $k \to \infty$, $n \to \infty$, and $\frac{k}{n} \to 0$, whereas our bounds hold for arbitrary finite values of k and n, including the case where k is relatively large compared to n. In [15] Raskutti et al. reach a somewhat similar conclusion to our Theorem 1 via a similar argument, but where it is assumed that A satisfies $\|Ax\|_2^2 \leqslant (1 + \delta)\|x\|_2^2$ for all $x \in \Sigma_{2k}$ (i.e., the upper bound of the *restricted isometry property* (RIP)). In this case the authors show that[2]

$$M^*(A) \geqslant C \frac{k\sigma^2}{(1+\delta)n} \log(n/k).$$

Our primary aim, however, is to challenge the use of the RIP and/or random matrices and to determine whether we can do better via a different choice in A. Our approach relies on standard tools from information theory such as Fano's inequality, and as such is very similar in spirit to the approaches in [1,15,17]. The proof of Theorem 1 begins by following a similar path to that taken in [15]. As in the results of [15], we rely on the construction of a packing set of sparse vectors. However, we place no assumptions whatsoever on the matrix A. To do this we must instead consider a random construction of this set, allowing us to apply the recently established matrix-version of Bernstein's inequality due to Ahlswede and Winter [2] to bound the empirical covariance matrix of the packing set. Our analysis is divided into two parts. In Section 2 we provide the proof of Theorem 1, and in Section 3 we provide the construction of the necessary packing set.

1.4. Notation

We now provide a brief summary of the notations used throughout the paper. If A is an $m \times n$ matrix and $T \subset \{1, \ldots, n\}$, then A_T denotes the $m \times |T|$ submatrix with columns indexed by T. Similarly, for a vector $x \in \mathbb{R}^n$ we let $x|_T$ denote the restriction of x to T. We will use $\|x\|_p$ to denote the standard ℓ_p norm of a vector, and for a matrix A, we will use $\|A\|$ and $\|A\|_F$ to denote the operator and Frobenius norms respectively.

---

[2] Note that it is possible to remove the assumption that A satisfies the upper bound of the RIP, but with a rather unsatisfying result. Specifically, for an arbitrary matrix A with a fixed Frobenius norm, we have that $\|A\|_2^2 \leqslant \|A\|_F^2$, so that $(1 + \delta) \leqslant \|A\|_F^2$. This bound can be shown to be tight by considering a matrix A with only one nonzero column. However, applying this bound underestimates $M^*(A)$ by a factor of n. Of course, the bounds coincide for "good" matrices (such as random matrices) which will have a significantly smaller value of δ [14]. However, the random matrix framework is precisely that which we wish to challenge.

2. Proof of main result

In this section we establish the lower bound (5) in Theorem 1. The proof of (6) is provided in Appendix A. In the proofs of both (5) and (6), we will assume that $\sigma = 1$ since the proof for arbitrary σ follows by a simple rescaling. To obtain the bound in (5) we begin by following a similar course as in [15]. Specifically, we will suppose that $\boldsymbol{x}$ is distributed uniformly on a finite set of points $\mathcal{X} \subset \Sigma_k$, where $\mathcal{X}$ is constructed so that the elements of $\mathcal{X}$ are well separated. This allows us to apply the following lemma which follows from Fano's inequality combined with the convexity of the Kullback–Leibler (KL) divergence. We provide a proof of the lemma in Appendix A.

Lemma 1. *Consider the measurement model where $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{z}$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Suppose that there exists set of points $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^{|\mathcal{X}|} \subset \Sigma_k$ such that for any $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$, $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \geqslant 8nM^*(\boldsymbol{A})$, where $M^*(\boldsymbol{A})$ is defined as in (4). Then*

$$\frac{1}{2}\log|\mathcal{X}| - 1 \leqslant \frac{1}{2|\mathcal{X}|^2} \sum_{i,j=1}^{|\mathcal{X}|} \|\boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{x}_j\|_2^2. \tag{9}$$

By taking the set $\mathcal{X}$ in Lemma 2 below and rescaling these points by $4\sqrt{nM^*(\boldsymbol{A})}$, we have that there exists a set $\mathcal{X}$ satisfying the assumptions of Lemma 1 with

$$|\mathcal{X}| = (n/k)^{k/4},$$

and hence from (9) we obtain

$$\frac{k}{4}\log(n/k) - 2 \leqslant \frac{1}{|\mathcal{X}|^2} \sum_{i,j=1}^{|\mathcal{X}|} \|\boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{x}_j\|_2^2 = \mathrm{Tr}\left(\boldsymbol{A}^*\boldsymbol{A}\left(\frac{1}{|\mathcal{X}|^2} \sum_{i,j=1}^{|\mathcal{X}|} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^*\right)\right). \tag{10}$$

If we set

$$\boldsymbol{\mu} = \frac{1}{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \boldsymbol{x}_i \quad \text{and} \quad \boldsymbol{Q} = \frac{1}{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \boldsymbol{x}_i\boldsymbol{x}_i^*,$$

then one can show that

$$\frac{1}{|\mathcal{X}|^2} \sum_{i,j=1}^{|\mathcal{X}|} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^* = 2(\boldsymbol{Q} - \boldsymbol{\mu}\boldsymbol{\mu}^*).$$

Thus, we can bound (10) by

$$2\,\mathrm{Tr}(\boldsymbol{A}^*\boldsymbol{A}(\boldsymbol{Q} - \boldsymbol{\mu}\boldsymbol{\mu}^*)) \leqslant 2\,\mathrm{Tr}(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{Q}),$$

where the inequality follows since $\mathrm{Tr}(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{\mu}\boldsymbol{\mu}^*) = \|\boldsymbol{A}\boldsymbol{\mu}\|_2^2 \geqslant 0$. Moreover, since $\boldsymbol{A}^*\boldsymbol{A}$ and $\boldsymbol{Q}$ are positive semidefinite,

$$\mathrm{Tr}(\boldsymbol{A}^*\boldsymbol{A}\boldsymbol{Q}) \leqslant \mathrm{Tr}(\boldsymbol{A}^*\boldsymbol{A})\|\boldsymbol{Q}\| = \|\boldsymbol{A}\|_F^2\|\boldsymbol{Q}\|.$$

Combining this with (10) and applying Lemma 2 to bound the norm of $\boldsymbol{Q}$ — recalling that it has been appropriately rescaled — we obtain

$$\frac{k}{4}\log(n/k) - 2 \leqslant (1 + \beta)32M^*(\boldsymbol{A})\|\boldsymbol{A}\|_F^2,$$

where β is a constant that can be arbitrarily close to 0. This yields the desired result.

3. Packing set construction

We now return to the problem of constructing the packing set $\mathcal{X}$. As noted above, our construction exploits the following matrix Bernstein inequality of Ahlswede and Winter [2]. See also [18].

Theorem 2 *(Matrix Bernstein inequality). Let $\{\boldsymbol{X}_i\}$ be a finite sequence of independent zero-mean random self-adjoint matrices of dimension $n \times n$. Suppose that $\|\boldsymbol{X}_i\| \leqslant 1$ almost surely for all i and set $\rho^2 = \sum_i \|\mathbb{E}[\boldsymbol{X}_i^2]\|$. Then for all $t \in [0, 2\rho^2]$,*

$$\mathbb{P}\left[\left\|\sum_i \boldsymbol{X}_i\right\| \geqslant t\right] \leqslant 2n\exp\left(-\frac{t^2}{4\rho^2}\right). \tag{11}$$

We construct the set $\mathcal{X}$ by choosing points at random, which allows us to apply Theorem 2 to establish a bound on the empirical covariance matrix. In bounding the size of $\mathcal{X}$ we follow a similar course as in [15] and rely on techniques from [11].

Lemma 2. *Let n and k be given, and suppose for simplicity that k is even and $k < n/2$. There exists a set $\mathcal{X} = \{x_i\}_{i=1}^{|\mathcal{X}|} \subset \Sigma_k$ of size*

$$|\mathcal{X}| = (n/k)^{k/4} \tag{12}$$

such that

(i) $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 \geqslant 1/2$ *for all* $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ *with* $i \neq j$; *and*
(ii) $\|\frac{1}{|\mathcal{X}|} \sum_{i=1}^{|\mathcal{X}|} \boldsymbol{x}_i \boldsymbol{x}_i^* - \frac{1}{n}\boldsymbol{I}\| \leqslant \beta/n$,

where β can be made arbitrarily close to 0 as $n \to \infty$.

Proof. We will show that such a set $\mathcal{X}$ exists via the probabilistic method. Specifically, we will show that if we draw $|\mathcal{X}|$ independent k-sparse vectors at random, then the set will satisfy both (i) and (ii) with probability strictly greater than 0. We will begin by considering the set

$$\mathcal{U} = \left\{ \boldsymbol{x} \in \left\{0, +\sqrt{1/k}, -\sqrt{1/k}\right\}^n \colon \|\boldsymbol{x}\|_0 = k \right\}.$$

Clearly, $|\mathcal{U}| = \binom{n}{k} 2^k$. Next, note that for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{U}$, $\frac{1}{k}\|\boldsymbol{x}' - \boldsymbol{x}\|_0 \leqslant \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2$, and thus if $\|\boldsymbol{x}' - \boldsymbol{x}\|_2^2 \leqslant 1/2$ then $\|\boldsymbol{x}' - \boldsymbol{x}\|_0 \leqslant k/2$. From this we observe that for any fixed $\boldsymbol{x} \in \mathcal{U}$,

$$\left| \left\{ \boldsymbol{x}' \in \mathcal{U} \colon \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2 \leqslant 1/2 \right\} \right| \leqslant \left| \left\{ \boldsymbol{x}' \in \mathcal{U} \colon \|\boldsymbol{x}' - \boldsymbol{x}\|_0 \leqslant k/2 \right\} \right| \leqslant \binom{n}{k/2} 3^{k/2}.$$

Suppose that we construct $\mathcal{X}$ by picking elements of $\mathcal{U}$ uniformly at random. When adding the jth point to $\mathcal{X}$, the probability that $\boldsymbol{x}_j$ violates (i) with respect to the previously added points is bounded by

$$\frac{(j-1)\binom{n}{k/2}3^{k/2}}{\binom{n}{k}2^k}.$$

Thus, using the union bound, we can bound the total probability that $\mathcal{X}$ will fail to satisfy (i), denoted P_1, by

$$P_1 \leqslant \sum_{j=1}^{|\mathcal{X}|} \frac{(j-1)\binom{n}{k/2}3^{k/2}}{\binom{n}{k}2^k} \leqslant \frac{|\mathcal{X}|^2}{2} \frac{\binom{n}{k/2}}{\binom{n}{k}} \left(\frac{\sqrt{3}}{2}\right)^k.$$

Next, observe that

$$\frac{\binom{n}{k}}{\binom{n}{k/2}} = \frac{(k/2)!(n-k/2)!}{k!(n-k)!} = \prod_{i=1}^{k/2} \frac{n-k+i}{k/2+i} \geqslant \left(\frac{n-k+k/2}{k/2+k/2}\right)^{k/2} = \left(\frac{n}{k} - \frac{1}{2}\right)^{k/2},$$

where the inequality follows since $(n-k+i)/(k/2+i)$ is decreasing as a function of i provided that $n-k > k/2$. Also,

$$\left(\frac{n}{k}\right)^{k/2}\left(\frac{\sqrt{3}}{2}\right)^k = \left(\frac{3n}{4k}\right)^{k/2} \leqslant \left(\frac{n}{k} - \frac{1}{2}\right)^{k/2}$$

with the proviso $k \leqslant n/2$. Thus, for $|\mathcal{X}|$ of size given in (12),

$$P_1 \leqslant \frac{1}{2}\left(\frac{n}{k}\right)^{k/2}\frac{\binom{n}{k/2}}{\binom{n}{k}}\left(\frac{\sqrt{3}}{2}\right)^k \leqslant \frac{1}{2}\left(\frac{n}{k} - \frac{1}{2}\right)^{k/2}\frac{\binom{n}{k/2}}{\binom{n}{k}} \leqslant \frac{1}{2}\frac{\binom{n}{k}}{\binom{n}{k/2}}\frac{\binom{n}{k/2}}{\binom{n}{k}} \leqslant \frac{1}{2}. \tag{13}$$

Next, we consider (ii). We begin by letting

$$\boldsymbol{X}_i = \boldsymbol{x}_i \boldsymbol{x}_i^* - \frac{\boldsymbol{I}}{n}.$$

Since $\boldsymbol{x}_i$ is drawn uniformly at random from $\mathcal{U}$, it is straightforward to show that $\|\boldsymbol{X}_i\| \leqslant 1$ and that $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^*] = \boldsymbol{I}/n$, which implies that $\mathbb{E}[\boldsymbol{X}_i] = 0$. Moreover,

$$\mathbb{E}[\boldsymbol{X}_i^2] = \mathbb{E}[(\boldsymbol{x}_i \boldsymbol{x}_i^*)^2] - \left(\frac{1}{n}\boldsymbol{I}\right)^2 = \frac{(n-1)}{n^2}\boldsymbol{I}.$$

Thus we obtain $\rho^2 = \sum_{i=1}^{|\mathcal{X}|} \|\mathbb{E}[\boldsymbol{X}_i^2]\| = |\mathcal{X}|(n-1)/n^2 \leqslant |\mathcal{X}|/n$. Hence, we can apply Theorem 2 to obtain

$$\mathbb{P}\left[\left\|\sum_{i=1}^{|\mathcal{X}|} \boldsymbol{X}_i\right\| \geqslant t\right] \leqslant 2n\exp\left(-\frac{t^2 n}{4|\mathcal{X}|}\right).$$

Setting $t = |\mathcal{X}|\beta/n$, this reduces to show that the probability that $\mathcal{X}$ will fail to satisfy (ii), denoted P_2, is bounded by

$$P_2 \leqslant 2n\exp\left(-\frac{\beta^2 |\mathcal{X}|}{4n}\right).$$

For the lemma to hold we require that $P_1 + P_2 < 1$, and since $P_1 < \frac{1}{2}$ it is sufficient to show that $P_2 < \frac{1}{2}$. This will occur provided that

$$\beta^2 > \frac{4n\log(4n)}{|\mathcal{X}|}.$$

Since $|\mathcal{X}| = \Theta((n/k)^k)$, β can be made arbitrarily small as $n \to \infty$. □

Appendix A

Proof of (6) in Theorem 1.
We begin by noting that

$$M^*(\boldsymbol{A}) = \inf_{\hat{\boldsymbol{x}}} \sup_{T:|T|\leqslant k} \sup_{\boldsymbol{x}:\mathrm{supp}(\boldsymbol{x})=T} \mathbb{E}\left[\frac{1}{n}\|\hat{\boldsymbol{x}}(\boldsymbol{y})-\boldsymbol{x}\|_2^2\right] \geqslant \sup_{T:|T|\leqslant k} \inf_{\hat{\boldsymbol{x}}} \sup_{\boldsymbol{x}:\mathrm{supp}(\boldsymbol{x})=T} \mathbb{E}\left[\frac{1}{n}\|\hat{\boldsymbol{x}}(\boldsymbol{y})-\boldsymbol{x}\|_2^2\right].$$

Thus for the moment we restrict our attention to the subproblem of bounding

$$M^*(\boldsymbol{A}_T) = \inf_{\hat{\boldsymbol{x}}} \sup_{\boldsymbol{x}:\mathrm{supp}(\boldsymbol{x})=T} \mathbb{E}\left[\frac{1}{n}\|\hat{\boldsymbol{x}}(\boldsymbol{y})-\boldsymbol{x}\|_2^2\right] = \inf_{\hat{\boldsymbol{x}}} \sup_{\boldsymbol{x}\in\mathbb{R}^k} \mathbb{E}\left[\frac{1}{n}\|\hat{\boldsymbol{x}}(\boldsymbol{A}_T\boldsymbol{x}+\boldsymbol{z})-\boldsymbol{x}\|_2^2\right], \tag{14}$$

where $\hat{\boldsymbol{x}}(\cdot)$ takes values in $\mathbb{R}^k$. The last equality of (14) follows since if $\mathrm{supp}(\boldsymbol{x}) = T$ then

$$\|\hat{\boldsymbol{x}}(\boldsymbol{y})-\boldsymbol{x}\|_2^2 = \|\hat{\boldsymbol{x}}(\boldsymbol{y})|_T - \boldsymbol{x}|_T\|_2^2 + \|\hat{\boldsymbol{x}}(\boldsymbol{y})|_{T^c}\|_2^2,$$

so that the risk can always be decreased by setting $\hat{\boldsymbol{x}}(\boldsymbol{y})|_{T^c} = 0$. This subproblem (14) has a well-known solution (see Exercise 5.8 on p. 403 of [12]). Specifically, let $\lambda_i(\boldsymbol{A}_T^*\boldsymbol{A}_T)$ denote the eigenvalues of the matrix $\boldsymbol{A}_T^*\boldsymbol{A}_T$. Then

$$M^*(\boldsymbol{A}_T) = \frac{1}{n}\sum_{i=1}^{k}\frac{1}{\lambda_i(\boldsymbol{A}_T^*\boldsymbol{A}_T)}. \tag{15}$$

Thus we obtain

$$M^*(\boldsymbol{A}) \geqslant \sup_{T:|T|\leqslant k} M^*(\boldsymbol{A}_T) = \sup_{T:|T|\leqslant k} \frac{1}{n}\sum_{i=1}^{k}\frac{1}{\lambda_i(\boldsymbol{A}_T^*\boldsymbol{A}_T)}. \tag{16}$$

Note that if there exists a subset T for which $\boldsymbol{A}_T$ is not full rank, then at least one of the eigenvalues $\lambda_i(\boldsymbol{A}_T^*\boldsymbol{A}_T)$ will vanish and the minimax risk will be unbounded. This also shows that the minimax risk is always unbounded when $m < k$.

Thus, we now assume that $\boldsymbol{A}_T$ is full rank for any choice of T. Since $f(x) = 1/x$ is a convex function for $x > 0$, we have that

$$\sum_{i=1}^{k}\frac{1}{\lambda_i(\boldsymbol{A}_T^*\boldsymbol{A}_T)} \geqslant \frac{k^2}{\sum_{i=1}^{k}\lambda_i(\boldsymbol{A}_T^*\boldsymbol{A}_T)} = \frac{k^2}{\|\boldsymbol{A}_T\|_F^2}.$$

Since there always exists a set of k columns T_0 such that $\|\boldsymbol{A}_{T_0}\|_F^2 \leqslant (k/n)\|\boldsymbol{A}\|_F^2$, (16) reduces to yield the desired result. □

Proof of Lemma 1. To begin, note that if $\boldsymbol{x}$ is uniformly distributed on the set of points in $\mathcal{X}$, then there exists an estimator $\hat{\boldsymbol{x}}(\boldsymbol{y})$ such that

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}\left[\frac{1}{n}\|\hat{\boldsymbol{x}}(\boldsymbol{y})-\boldsymbol{x}\|_2^2\right] \leqslant M^*(\boldsymbol{A}), \tag{17}$$

where the expectation is now taken with respect to both the signal and the noise. We next consider the problem of deciding which $\boldsymbol{x}_i \in \mathcal{X}$ generated the observations $\boldsymbol{y}$. Towards this end, set

$$T\big(\hat{\boldsymbol{x}}(\boldsymbol{y})\big) = \arg\min_{\boldsymbol{x}_i \in \mathcal{X}} \big\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}_i\big\|_2.$$

Define $P_e = \mathbb{P}[T(\hat{\boldsymbol{x}}(\boldsymbol{y})) \neq \boldsymbol{x}]$. From Fano's inequality [7] we have that

$$H(\boldsymbol{x}|\boldsymbol{y}) \leqslant 1 + P_e \log |\mathcal{X}|. \tag{18}$$

We now aim to bound P_e. We begin by noting that for any $\boldsymbol{x}_i \in \mathcal{X}$ and any $\hat{\boldsymbol{x}}(\boldsymbol{y})$, $T(\hat{\boldsymbol{x}}(\boldsymbol{y})) \neq \boldsymbol{x}_i$ if and only if there exists an $\boldsymbol{x}_j \in \mathcal{X}$ with $j \neq i$ such that

$$\big\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}_i\big\|_2 \geqslant \big\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}_j\big\|_2 \geqslant \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 - \big\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}_i\big\|_2.$$

This would imply that

$$2\big\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}_i\big\|_2 \geqslant \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \geqslant \sqrt{8nM^*(\boldsymbol{A})}.$$

Thus, we can bound P_e using Markov's inequality as follows:

$$P_e \leqslant \mathbb{P}\big[\big\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}_i\big\|_2^2 \geqslant 8nM^*(\boldsymbol{A})/4\big] \leqslant \frac{\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\|\hat{\boldsymbol{x}}(\boldsymbol{y}) - \boldsymbol{x}_i\|_2^2]}{2nM^*(\boldsymbol{A})} \leqslant \frac{nM^*(\boldsymbol{A})}{2nM^*(\boldsymbol{A})} = \frac{1}{2}.$$

Combining this with (18) and the fact that $H(\boldsymbol{x}) = \log |\mathcal{X}|$, we obtain

$$I(\boldsymbol{x}, \boldsymbol{y}) = H(\boldsymbol{x}) - H(\boldsymbol{x}|\boldsymbol{y}) \geqslant \frac{1}{2}\log |\mathcal{X}| - 1.$$

From the convexity of KL divergence (see [10] for details), we have that

$$I(\boldsymbol{x}, \boldsymbol{y}) \leqslant \frac{1}{|\mathcal{X}|^2} \sum_{j,k=1}^{|\mathcal{X}|} D(\mathcal{P}_i, \mathcal{P}_j),$$

where $D(\mathcal{P}_i, \mathcal{P}_j)$ represents the KL divergence from $\mathcal{P}_i$ to $\mathcal{P}_j$ where $\mathcal{P}_i$ denotes the distribution of $\boldsymbol{y}$ conditioned on $\boldsymbol{x} = \boldsymbol{x}_i$. Since $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, $\mathcal{P}_i$ is simply given by $\mathcal{N}(\boldsymbol{A}\boldsymbol{x}_i, \boldsymbol{I})$. Standard calculations demonstrate that $D(\mathcal{P}_i, \mathcal{P}_j) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{x}_j\|_2^2$, establishing (9). $\quad\square$

References

[1] S. Aeron, V. Saligrama, M. Zhao, Information theoretic bounds for compressed sensing, IEEE Trans. Inform. Theory 56 (10) (2010) 5111–5130.
[2] R. Ahlswede, A. Winter, Strong converse for identification via quantum channels, IEEE Trans. Inform. Theory 48 (3) (2002) 569–579.
[3] P. Bickel, Y. Ritov, A. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist. 37 (4) (2009) 1705–1732.
[4] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inform. Theory 52 (2) (2006) 489–509.
[5] E. Candès, T. Tao, Near optimal signal recovery from random projections: Universal encoding strategies?, IEEE Trans. Inform. Theory 52 (12) (2006) 5406–5425.
[6] E. Candès, T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, Ann. Statist. 35 (6) (2007) 2313–2351.
[7] T. Cover, J. Thomas, Elements of Information Theory, Wiley–Interscience, New York, 1991.
[8] D. Donoho, Compressed sensing, IEEE Trans. Inform. Theory 52 (4) (2006) 1289–1306.
[9] D. Donoho, I. Johnstone, A. Maleki, A. Montanari, Compressed sensing over ℓ_p balls: Minimax mean square error, arXiv:1103.1943, 2011.
[10] T. Han, S. Verdú, Generalizing the Fano inequality, IEEE Trans. Inform. Theory 40 (4) (1994) 1247–1251.
[11] T. Kühn, A lower estimate for entropy numbers, J. Approx. Theory 110 (1) (2001) 120–124.
[12] E. Lehman, G. Casella, Theory of Point Estimation, Springer-Verlag, New York, 1998.
[13] K. Lounici, Generalized mirror averaging and d-convex aggregation, Math. Methods Statist. 16 (3) (2007) 246–259.
[14] G. Raskutti, M. Wainwright, B. Yu, Lower bounds on minimax rates for nonparametric regression with additive sparsity and smoothness, in: Proc. Adv. in Neural Proc. Systems (NIPS), Vancouver, BC, December 2009.
[15] G. Raskutti, M.J. Wainwright, B. Yu, Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls, IEEE Trans. Inform. Theory 57 (10) (2011) 6976–6994.
[16] P. Rigollet, A. Tsybakov, Exponential screening and optimal rates of sparse estimation, Ann. Statist. 39 (2) (2011) 731–771.
[17] S. Sarvotham, D. Baron, R. Baraniuk, Measurements vs. bits: Compressed sensing meets information theory, in: Proc. Allerton Conf. Communication, Control, and Computing, Monticello, IL, September 2006.
[18] J. Tropp, User-friendly tail bounds for sums of random matrices, Found. Comput. Math. 12 (4) (2012) 389–434.
[19] F. Ye, C. Zhang, Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls, J. Mach. Learn. Res. 11 (2010) 3519–3540.