

Stable Recovery of Sparse Signals and an Oracle Inequality

Tony Tony Cai, Lie Wang, and Guangwu Xu

Abstract—This article considers sparse signal recovery in the presence of noise. A mutual incoherence condition which was previously used for exact recovery in the noiseless case is shown to be sufficient for stable recovery in the noisy case. Furthermore, the condition is proved to be sharp. A specific counterexample is given. In addition, an oracle inequality is derived under the mutual incoherence condition in the case of Gaussian noise.

Index Terms— ℓ_1 minimization, compressed sensing, mutual incoherence, oracle inequality, sparse recovery.

I. INTRODUCTION

COMPRESSED sensing has received much recent attention in signal and imaging processing, applied mathematics, and statistics. The central goal is to accurately reconstruct a high dimensional sparse signal based on a small number linear measurements. Specifically one considers the following linear model:

$$y = \Phi\beta + z \quad (1)$$

where Φ is an $n \times p$ matrix (with $n \ll p$) and $z \in \mathbb{R}^n$ is a vector of measurement errors. The goal is to recover the unknown signal $\beta \in \mathbb{R}^p$ based on y and Φ . Throughout the paper we shall assume that the columns of Φ are standardized to have unit ℓ_2 norm.

It is now well understood that the method of ℓ_1 minimization provides an effective way for reconstructing a sparse signal in many settings. The ℓ_1 minimization problem in this context is

$$(P_{\mathcal{B}}) \quad \min_{\gamma \in \mathbb{R}^p} \|\gamma\|_1 \quad \text{subject to} \quad y - \Phi\gamma \in \mathcal{B}. \quad (2)$$

Here $\mathcal{B} \subset \mathbb{R}^n$ is a bounded set. For example, \mathcal{B} is taken to be $\{0\}$ in the noiseless case and can be $\mathcal{B}^{\ell_2} = \{z : \|z\|_2 \leq \eta\}$ or $\mathcal{B}^{DS} = \{z : \|\Phi'z\|_{\infty} \leq \eta\}$ in the noisy case.

It is clear that with $n \ll p$ the linear system (1) is underdetermined and regularity conditions are needed. A widely used

condition for sparse signal recovery is the mutual incoherence property (MIP) introduced in [10]. The MIP requires the pairwise correlations among the column vectors of Φ to be small. Let

$$\mu = \max_{i \neq j} |\langle \Phi_i, \Phi_j \rangle|. \quad (3)$$

It was first shown by Donoho and Huo [10], in the noiseless case for the setting where Φ is a concatenation of two square orthogonal matrices, that

$$k < \frac{1}{2} \left(\frac{1}{\mu} + 1 \right) \quad (4)$$

ensures the exact recovery of β when β has at most k nonzero entries (such a signal is called k -sparse). This result was then extended in the noiseless case in [14] and [16] to a general dictionary Φ .

Stronger MIP conditions have been used in the literature to guarantee stable recovery of sparse signals in the noisy case. When noise is assumed to be bounded in ℓ_2 norm, i.e., $\mathcal{B} = \mathcal{B}^{\ell_2}$, [9] showed that sparse signals can be recovered approximately through ℓ_1 minimization, with the error at worst proportional to the input noise level, when

$$k < \frac{1}{4} \left(\frac{1}{\mu} + 1 \right).$$

The results in [3] imply that

$$k < \frac{2}{3 + \sqrt{6}} \left(\frac{1}{\mu} + 1 \right) \approx 0.368 \left(\frac{1}{\mu} + 1 \right)$$

is sufficient for stable recovery. And Tseng [19] used

$$k < \left(\frac{1}{2} - O(\mu) \right) \frac{1}{\mu} + 1.$$

However, to the best of our knowledge, the natural question whether (4), namely

$$k < \frac{1}{2} \left(\frac{1}{\mu} + 1 \right)$$

is sufficient for stable recovery in the noisy case remains open.

In this paper, we consider stable recovery of sparse signals under the MIP framework. Our results show that not only the condition (4) is indeed sufficient for stable recovery in the noisy case, it is also sharp in the sense that there exist dictionaries such that it is not possible to recover certain k -sparse signals with $k = \frac{1}{2} \left(\frac{1}{\mu} + 1 \right)$. A specific counterexample is constructed in Section III.

Manuscript received June 17, 2009; revised October 19, 2009. Current version published June 16, 2010. The work of T. T. Cai was supported in part by the NSF by Grant DMS-0604954. The work of G. Xu was supported in part by the National 973 Project of China (No. 2007CB807902).

T. T. Cai is with the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: tcai@wharton.upenn.edu).

L. Wang is with Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: liewang@math.mit.edu).

G. Xu is with the Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee, Milwaukee, WI 53211 USA (e-mail: gxu4uwm@uwm.edu).

Communicated by J. Romberg, Associate Editor for Signal Processing.

Digital Object Identifier 10.1109/TIT.2010.2048506

In addition to bounded noise in an ℓ_2 ball, we also consider bounded error in $\mathcal{B}^{DS} = \{z : \|\Phi'z\|_\infty \leq \eta\}$. This case is closely connected to the Dantzig Selector introduced in Candès and Tao [7] in the framework of the restricted isometry property. The results for bounded error can be extended directly to the Gaussian noise case, which is of particular interest in signal processing and in statistics.

Oracle inequality is a powerful decision-theoretic tool that provides great insight into the performance of a procedure as compared to that of an ideal estimator. It was first introduced by Donoho and Johnstone [11] to demonstrate the optimalities of certain wavelet thresholding rules in statistical signal processing problems. Candès and Tao [7] established an oracle inequality in the setting of compressed sensing for the Dantzig Selector under the restrictive isometry property. In this paper, we present an oracle inequality for sparse signal recovery under the MIP condition (4) in the Gaussian noise case. It is worth noting that our proof is simple and elementary.

The rest of the paper is organized as follows. Section II considers stable recovery of sparse signals under the MIP. We shall show that the condition $k < \frac{1}{2}(\frac{1}{\mu} + 1)$ is sufficient for stable recovery with bounded noise as well as Gaussian noise. We then show in Section III that $k < \frac{1}{2}(\frac{1}{\mu} + 1)$ is not only sufficient but in fact sharp by providing a counterexample when $k = \frac{1}{2}(\frac{1}{\mu} + 1)$. Section IV establishes the oracle inequality in the case of Gaussian noise under the MIP condition. Section V discusses some relations to the restricted isometry property.

II. STABLE RECOVERY OF SPARSE SIGNALS

As aforementioned, $k < \frac{1}{2}(\frac{1}{\mu} + 1)$ has been proved to guarantee the exact recovery of k -sparse signals in noiseless case. We shall show in this section that this condition is also sufficient for stable reconstruction of k -sparse signals in the noisy case. We shall consider the case where the error is bounded. Two specific bounded sets are considered: the ℓ_2 ball $\mathcal{B}^{\ell_2} = \{z : \|z\|_2 \leq \epsilon\}$ and the set $\mathcal{B}^{DS} = \{z : \|\Phi'z\|_\infty \leq \epsilon\}$. These results can be extended directly to the Gaussian noise case since Gaussian noise is “essentially bounded.” The Gaussian noise case is of particular interest in statistics. See, for example, Efron, *et al.* [13], and Candès and Tao [7]. We will then derive an oracle inequality for the Gaussian noise case in Section IV.

We begin by introducing basic notation and definitions as well as some elementary results that will be needed later.

The support of a vector $v = (v_1, v_2, \dots, v_p) \in \mathbb{R}^p$ is defined to be $\text{supp}(v) = \{i : v_i \neq 0\}$. A vector v is said to be k -sparse if $|\text{supp}(v)| \leq k$. We use the standard notation $\|v\|_q = (\sum_{i=1}^q |v_i|^q)^{1/q}$ to denote the ℓ_q -norm of the vector v . We shall also treat a vector $v = (v_i) \in \mathbb{R}^p$ as a function $v : \{1, 2, \dots, p\} \rightarrow \mathbb{R}$ by assigning $v(i) = v_i$.

Consider the ℓ_1 minimization problem $(P_{\mathcal{B}})$. Let β be a feasible solution to $(P_{\mathcal{B}})$, i.e., $y - \Phi\beta \in \mathcal{B}$. Let $\hat{\beta}$ be a solution to the minimization problem $(P_{\mathcal{B}})$. Then by definition $\|\hat{\beta}\|_1 \leq \|\beta\|_1$. Let $h = \hat{\beta} - \beta$ and $h_0 = h|_{\text{supp}(\beta)}$. Here $\mathbb{1}_A$ denotes the indicator function of a set $A \subseteq \{1, 2, \dots, p\}$, i.e., $\mathbb{1}_A(j) = 1$ if $j \in A$ and 0 if $j \notin A$.

The following is a widely used fact (see, for example, [3], [5], [7], and [10])

$$\|h - h_0\|_1 \leq \|h_0\|_1. \quad (5)$$

This follows directly from the fact that

$$\begin{aligned} \|\beta\|_1 &\geq \|\hat{\beta}\|_1 = \|\beta + h\|_1 = \|\beta + h_0\|_1 + \|h - h_0\|_1 \\ &\geq \|\beta\|_1 - \|h_0\|_1 + \|h - h_0\|_1. \end{aligned}$$

The following fact is well known. Let α be any k -sparse signal, then

$$(1 - (k - 1)\mu)\|\alpha\|_2^2 \leq \|\Phi\alpha\|_2^2 \leq (1 + (k - 1)\mu)\|\alpha\|_2^2. \quad (6)$$

See, e.g., [3], [9], [18], and [19].

We now consider stable recovery of sparse signals with error in the ℓ_2 ball \mathcal{B}^{ℓ_2} . The following result shows that $k < \frac{1}{2}(\frac{1}{\mu} + 1)$ is a sufficient condition for stable recovery.

Theorem 2.1: Consider the model (1) with $\|z\|_2 \leq \epsilon$. Suppose β is k -sparse with $k < \frac{1}{2}(\frac{1}{\mu} + 1)$. Let $\hat{\beta}$ be the minimizer of $(P_{\mathcal{B}})$ with $\mathcal{B} = \{z : \|z\|_2 \leq \eta\}$, and $\epsilon \leq \eta$. Then $\hat{\beta}$ satisfies

$$\|\hat{\beta} - \beta\|_2 \leq \frac{\sqrt{3(1 + \mu)}}{1 - (2k - 1)\mu} \cdot (\eta + \epsilon).$$

Remark 2.1: This theorem improves several stable recovery results in the literature. In particular, our condition $k < \frac{1}{2}(\frac{1}{\mu} + 1)$ is weaker than $k < \frac{1}{4}(\frac{1}{\mu} + 1)$ [9], $k < \frac{2}{3 + \sqrt{6}}(\frac{1}{\mu} + 1)$ [3], and $k < (\frac{1}{2} - O(\mu))\frac{1}{\mu} + 1$ [19].

Proof: First we note that

$$\|\Phi h\|_2 \leq \eta + \epsilon.$$

This simply follows from

$$\begin{aligned} \|\Phi h\|_2 &= \|\Phi\hat{\beta} - y - (\Phi\beta - y)\|_2 \\ &\leq \|\Phi\hat{\beta} - y\|_2 + \|\Phi\beta - y\|_2 \leq \eta + \epsilon. \end{aligned}$$

It is also noted that $k \leq \frac{1}{2}(\frac{1}{\mu} + 1)$ is equivalent to

$$(2k - 1)\mu < 1 \quad \text{or} \quad 1 - (2k - 1)\mu > 0.$$

Without loss of generality, we shall assume that $\text{supp}(\beta) = \text{supp}(h_0) \subseteq \{1, 2, \dots, k\}$. It follows from the facts $\|\Phi_i\|_2 = 1$, $|\langle \Phi_i, \Phi_j \rangle| \leq \mu$ for $i \neq j$, and (6) that

$$\begin{aligned} |\langle \Phi h, \Phi h_0 \rangle| &\geq |\langle \Phi h_0, \Phi h_0 \rangle| - |\langle \Phi(h - h_0), \Phi h_0 \rangle| \\ &\geq (1 - (k - 1)\mu)\|h_0\|_2^2 \\ &\quad - \left| \sum_{i=0}^k \sum_{j=k+1}^p \langle \Phi_i, \Phi_j \rangle h_0(i)h(j) \right| \\ &\geq (1 - (k - 1)\mu)\|h_0\|_2^2 - \mu\|h_0\|_1\|h - h_0\|_1 \\ &\geq (1 - (k - 1)\mu)\|h_0\|_2^2 - \mu\|h_0\|_1^2 \\ &\geq (1 - (k - 1)\mu)\|h_0\|_2^2 - k\mu\|h_0\|_2^2 \\ &= (1 - (2k - 1)\mu)\|h_0\|_2^2. \end{aligned}$$

On the other hand, it follows from (6) that

$$\|\Phi h_0\|_2^2 \leq (1 + (k-1)\mu) \|h_0\|_2^2.$$

Therefore, we have¹

$$\begin{aligned} \|h_0\|_2 &\leq \frac{|\langle \Phi h, \Phi h_0 \rangle|}{(1 - (2k-1)\mu) \|h_0\|_2} \\ &\leq \frac{\|\Phi h\|_2 \|\Phi h_0\|_2}{(1 - (2k-1)\mu) \|h_0\|_2} \\ &\leq \frac{\sqrt{1 + (k-1)\mu}}{1 - (2k-1)\mu} (\eta + \epsilon). \end{aligned}$$

Finally

$$\begin{aligned} \|\Phi h\|_2^2 &= \langle \Phi h, \Phi h \rangle = \sum_{i,j} \langle \Phi_i h_i, \Phi_j h_j \rangle \\ &= \sum_i \|\Phi_i\|_2^2 |h_i|^2 + \sum_{i \neq j} \langle \Phi_i, \Phi_j \rangle h_i h_j \\ &\geq \|h\|_2^2 - \mu \sum_{i \neq j} |h_i h_j| \\ &= \|h\|_2^2 + \mu \sum_i |h_i|^2 - \mu \sum_{i,j} |h_i h_j| \\ &= (1 + \mu) \|h\|_2^2 - \mu \|h\|_1^2 \\ &\geq (1 + \mu) \|h\|_2^2 - \mu (2 \|h_0\|_1)^2 \\ &\geq (1 + \mu) \|h\|_2^2 - 4\mu k \|h_0\|_2^2. \end{aligned}$$

Note that $2\mu k < 1 + \mu$. So

$$\begin{aligned} \|h\|_2^2 &\leq \frac{(\eta + \epsilon)^2 + 4k\mu \|h_0\|_2^2}{1 + \mu} \\ &\leq \frac{(1 - (2k-1)\mu)^2 + 4k\mu(1 + (k-1)\mu)}{(1 + \mu)(1 - (2k-1)\mu)^2} (\eta + \epsilon)^2 \\ &= \frac{1 + 2\mu + (8k^2 - 8k + 1)\mu^2}{(1 + \mu)(1 - (2k-1)\mu)^2} (\eta + \epsilon)^2 \\ &\leq \frac{1 + 2\mu + (8k^2 - 7)\mu^2}{(1 + \mu)(1 - (2k-1)\mu)^2} (\eta + \epsilon)^2 \\ &= \frac{1 + 2\mu + 2(2k\mu)^2 - 7\mu^2}{(1 + \mu)(1 - (2k-1)\mu)^2} (\eta + \epsilon)^2 \\ &\leq \frac{1 + 2\mu + 2(1 + \mu)^2 - 7\mu^2}{(1 + \mu)(1 - (2k-1)\mu)^2} (\eta + \epsilon)^2 \\ &= \frac{3 + 6\mu - 5\mu^2}{(1 + \mu)(1 - (2k-1)\mu)^2} (\eta + \epsilon)^2 \\ &\leq \frac{3(1 + \mu)}{(1 - (2k-1)\mu)^2} (\eta + \epsilon)^2. \end{aligned}$$

We now turn to stable recovery of k -sparse signals with error in the bounded set $\mathcal{B}^{DS} = \{z : \|\Phi'z\|_\infty \leq \epsilon\}$ under the MIP framework. Candès and Tao [7] treated the sparse signal recovery problem by solving $(P_{\mathcal{B}})$ with $\mathcal{B} = \mathcal{B}^{DS} = \{z : \|\Phi'z\|_\infty \leq \eta\}$, in the framework of restricted isometry property, and referred the solution as the Dantzig Selector. We shall

¹Here we assume that $h_0 \neq 0$ as the case for $h_0 = 0$ is trivial.

show here and in Section IV the Dantzig Selector can be analyzed easily using elementary tools under the MIP condition $k < \frac{1}{2}(\frac{1}{\mu} + 1)$.

Theorem 2.2: Consider the model (1) with z satisfying $\|\Phi'z\|_\infty \leq \epsilon$. Suppose β is k -sparse with $k < \frac{1}{2}(\frac{1}{\mu} + 1)$. Let $\hat{\beta}$ be the minimizer of $(P_{\mathcal{B}})$ with $\mathcal{B} = \{z : \|\Phi'z\|_\infty \leq \eta\}$, and $\epsilon \leq \eta$. Then $\hat{\beta}$ satisfies

$$\|\hat{\beta} - \beta\|_2 \leq \frac{\sqrt{2k}}{1 - (2k-1)\mu} \cdot (\eta + \epsilon).$$

Proof: Note that from the fact $\|h - h_0\|_1 \leq \|h_0\|_1$ and the first part of the proof of Theorem 2.1, we have

$$|\langle \Phi h, \Phi h_0 \rangle| \geq (1 - (2k-1)\mu) \|h_0\|_2^2.$$

Using the fact $\|\Phi'_T(\Phi h)\|_2 \leq \sqrt{k}(\eta + \epsilon)$ where $T = \text{supp}(\beta)$, we get

$$\begin{aligned} \|h_0\|_2 &\leq \frac{|\langle \Phi h, \Phi h_0 \rangle|}{(1 - (2k-1)\mu) \|h_0\|_2} \\ &= \frac{|\langle \Phi h, \Phi_T h_0 \rangle|}{(1 - (2k-1)\mu) \|h_0\|_2} \\ &= \frac{|\langle \Phi'_T \Phi h, h_0 \rangle|}{(1 - (2k-1)\mu) \|h_0\|_2} \\ &\leq \frac{\sqrt{k}}{(1 - (2k-1)\mu)} (\eta + \epsilon). \end{aligned}$$

Again, as in the proof of Theorem 2.1

$$\|\Phi h\|_2^2 \geq (1 + \mu) \|h\|_2^2 - 4\mu k \|h_0\|_2^2$$

so we conclude

$$\begin{aligned} \|h\|_2^2 &\leq \frac{1}{1 + \mu} (\|\Phi h\|_2^2 + 4\mu k \|h_0\|_2^2) \\ &= \frac{1}{1 + \mu} (\langle \Phi' \Phi h, h \rangle + 4k\mu \|h_0\|_2^2) \\ &\leq \frac{1}{1 + \mu} ((\eta + \epsilon) \|h\|_1 + 4k\mu \|h_0\|_2^2) \\ &\leq \frac{1}{1 + \mu} ((\eta + \epsilon) 2 \|h_0\|_1 + 4k\mu \|h_0\|_2^2) \\ &\leq \frac{1}{1 + \mu} ((\eta + \epsilon) 2\sqrt{k} \|h_0\|_2 + 4k\mu \|h_0\|_2^2) \\ &\leq \frac{(\eta + \epsilon)^2}{1 + \mu} \left(\frac{2k}{(1 - (2k-1)\mu)} + \frac{4k^2\mu}{(1 - (2k-1)\mu)^2} \right) \\ &\leq \frac{2k(\eta + \epsilon)^2}{(1 + \mu)(1 - (2k-1)\mu)^2} (1 - (2k-1)\mu + 2k\mu) \\ &\leq \frac{2k}{(1 - (2k-1)\mu)^2} \cdot (\eta + \epsilon)^2. \end{aligned}$$

Remark 2.2: For simplicity, we have focused on recovering k -sparse signals in the present paper. When β is not k -sparse, ℓ_1 minimization can also recover β with accuracy if β has good k -term approximation. For a general vector $\beta \in \mathbb{R}^p$, denote

$\beta_{-\max(k)}$ the vector β with the k -largest entries (in absolute value) set to zero. Then similar to [2] and [3], our results can be extended to the general setting. Under $k < \frac{1}{2}(\frac{1}{\mu} + 1)$, Theorem 2.1 holds with the error bound

$$\|\hat{\beta} - \beta\|_2 \leq C_1(\eta + \epsilon) + C_2\sqrt{\mu}\|\beta_{-\max(k)}\|_1$$

for some constants C_1 and C_2 , and Theorem 2.2 holds with the error bound

$$\|\hat{\beta} - \beta\|_2 \leq C_3\sqrt{k}(\eta + \epsilon) + C_4\sqrt{\mu}\|\beta_{-\max(k)}\|_1$$

for some constants C_3 and C_4 .

We have so far focused on stable recovery with bounded error. The results can be extended directly to the Gaussian noise case. This is due to the fact that Gaussian noise is “essentially bounded”. See, for example, [3] and [2].

III. SHARPNESS OF $k < \frac{1}{2}(\frac{1}{\mu} + 1)$

We have just shown in Section II that the condition $k < \frac{1}{2}(\frac{1}{\mu} + 1)$ is sufficient for stable recovery of k -sparse signals in both bounded noise and Gaussian noise settings. An inspection of our proof shows that this condition arises naturally. It is interesting to ask whether the condition is sharp. Our next result implies that this condition is indeed sharp.

Theorem 3.1: Let k be a positive integer. Let $t > 0$ be any integer, and set $n = (2k - 1)t$ and $p = 2kt$. Then there exists an $n \times p$ matrix Φ with mutual incoherence $\mu = \frac{1}{2k-1}$ (or equivalently $k = \frac{1}{2}(\frac{1}{\mu} + 1)$), and two nonzero k -sparse vectors β_1 and β_2 with disjoint supports such that

$$\Phi\beta_1 = \Phi\beta_2.$$

Remark 3.1: This result implies that the model (1) is not identifiable in general under the condition $k = \frac{1}{2}(\frac{1}{\mu} + 1)$ and therefore not all k -sparse signals can be recovered exactly in the noiseless case. In the noisy case, it is easy to see that both Theorems 2.1 and 2.2 fail because no estimator $\hat{\beta}$ can be close to both β_1 and β_2 when ϵ and η are sufficiently small. The proof reveals that one can actually get at least t pairs of such k -sparse signals (β_1, β_2) .

Remark 3.2: The specific counterexample constructed in the proof shows that the bound $k < \frac{1}{2}(\frac{1}{\mu} + 1)$ is tight when $\frac{n}{p} = \frac{2kt-t}{2kt} = 1 - \frac{1}{2k}$. It should be noted that some necessary and sufficient sparse recovery conditions for more general (n, p) have been discussed in the Grassmannian angle framework. See, for example, [12] and [20].

Proof: Let Γ be a $2k \times 2k$ matrix such that each diagonal element of Γ is 1 and each off diagonal element equals $-\frac{1}{2k-1}$. Then it is easy to see that Γ is a positive-semidefinite matrix with rank $2k - 1$.

Note that the symmetric matrix Γ can be decomposed as $\Gamma = A'A$ where A is a $(2k - 1) \times 2k$ matrix with rank $2k - 1$. More precisely, since Γ has two distinct eigenvalues $\frac{2k}{2k-1}$ and

0, with the multiplicities of $2k - 1$ and 1, respectively, there is an orthogonal matrix U such that

$$\Gamma = U \text{Diag} \left\{ \underbrace{\frac{2k}{2k-1}, \frac{2k}{2k-1}, \dots, \frac{2k}{2k-1}}_{2k-1}, 0 \right\} U'.$$

It suffices to take A as

$$A = \begin{pmatrix} \sqrt{\frac{2k}{2k-1}} & 0 & \dots & 0 & 0 \\ 0 & \sqrt{\frac{2k}{2k-1}} & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & \sqrt{\frac{2k}{2k-1}} & 0 \end{pmatrix} U'.$$

Write $U = (u_1, u_2, \dots, u_{2k})$ where $u_1, u_2, \dots, u_{2k-1}$ are unit eigenvectors corresponding to eigenvalue $\frac{2k}{2k-1}$ and $u_{2k} = (\frac{1}{\sqrt{2k}}, \frac{1}{\sqrt{2k}}, \dots, \frac{1}{\sqrt{2k}})'$ is the eigenvector corresponding to eigenvalue 0. It is verified that each column of A is of length 1 in ℓ_2 norm. This is because $A = \sqrt{\frac{2k}{2k-1}}(u_1, u_2, \dots, u_{2k-1})'$.

Since the rank of A is $2k - 1$, there exists some $\gamma \in \mathbb{R}^{2k}$ such that $\gamma \neq 0$ and $A\gamma = 0$. Now define an $n \times p$ matrix Φ by

$$\Phi = \text{Diag} \left\{ \underbrace{A, A, \dots, A}_t \right\}.$$

It can be seen easily that the mutual incoherence μ for Φ satisfies $\mu = \frac{1}{2k-1}$. Suppose $\beta_1, \beta_2 \in \mathbb{R}^p$ are given by

$$\beta_1 = (\gamma(1), \gamma(2), \dots, \gamma(k), 0, 0, \dots, 0)',$$

and

$$\beta_2 = (\underbrace{0, 0, \dots, 0}_k, -\gamma(k+1), -\gamma(k+2), \dots, -\gamma(2k), 0, 0, \dots, 0)'$$

Then both β_1 and β_2 are k -sparse vectors but $\Phi\beta_1 = \Phi\beta_2$. This means the model is not identifiable within the class of k -sparse signals. ■

IV. AN ORACLE INEQUALITY

As aforementioned, oracle inequality was first introduced by Donoho and Johnstone [11] in the context of signal denoising using wavelet thresholding. The oracle inequality approach provides an effective tool for studying the optimalities of a procedure. An oracle inequality compares the properties of a given procedure to that of an ideal estimator with the aid of an oracle. The ideal risk is used as a benchmark to evaluate the performance of the procedure of interest. This approach has since been extended to study many other problems. In particular, [7] developed an oracle inequality for the Dantzig Selector $\hat{\beta}^{DS}$ in the Gaussian noise setting in the framework of the restrictive isometry property. In this section, we derive an oracle inequality for sparse signal recovery under the MIP condition (4). We should note that our proof is particularly simple and elementary.

Consider the Gaussian noise problem where we observe (Φ, y) with

$$y = \Phi\beta + z, \quad z \sim N(0, \sigma^2 I_n). \quad (7)$$

We shall assume that the noise level σ is known. We wish to reconstruct the signal β accurately based on Φ and y .

We first briefly describe the main ideas behind the oracle inequality approach in the context of sparse signal recovery. For more details, see [11] and [7].

Denote the support of the signal β by J_* . If J_* were known to us, β could be simply recovered by $\hat{\beta}^{LS}$ using the least squares estimator

$$\hat{\beta}_{J_*}^{LS} = (\Phi'_{J_*} \Phi_{J_*})^{-1} \Phi'_{J_*} y$$

with $\hat{\beta}^{LS}(i) = \hat{\beta}_{J_*}^{LS}(i)$ for $i \in J_*$ and $\hat{\beta}^{LS}(i) = 0$ otherwise. Note that $\hat{\beta}_{J_*}^{LS}$ minimizes the squared error $\|y - \Phi_{J_*} \gamma\|_2^2$ among all possible choices of γ .

One can be even more ambitious by asking for the best k -sparse ‘‘estimator’’ of β . More specifically, denote by \mathcal{J} the collection of all subsets of $\{1, 2, \dots, p\}$ with cardinality less than or equal to k . Then for a given index set $J \in \mathcal{J}$, we can estimate β by $\hat{\beta}^J$ using the least squares estimator $\hat{\beta}_J^{LS} = (\Phi'_J \Phi_J)^{-1} \Phi'_J y$, with $\hat{\beta}^J(i) = \hat{\beta}_J^{LS}(i)$ for $i \in J$ and $\hat{\beta}^J(i) = 0$ otherwise. The ideal choice of the index set J_{oracle} is the one that minimizes the risk over all choices in \mathcal{J} , that is

$$J_{\text{oracle}} = \arg \min_{J \in \mathcal{J}} E \|\hat{\beta}^J - \beta\|_2^2$$

and the oracle risk R_{oracle} is the minimum mean squared error achievable over \mathcal{J}

$$R_{\text{oracle}} = E \|\hat{\beta}^{J_{\text{oracle}}} - \beta\|_2^2 = \min_{J \in \mathcal{J}} E \|\hat{\beta}^J - \beta\|_2^2.$$

Note that for a given index set J , the mean squared error of $\hat{\beta}_J^{LS}$ can be easily calculated as

$$\begin{aligned} E \|\hat{\beta}^J - \beta\|_2^2 &= E \|\hat{\beta}_J^{LS} - \beta_J\|_2^2 + \|\beta_{J^c}\|_2^2 \\ &= \|(\Phi'_J \Phi_J)^{-1} \Phi'_J \Phi \beta_{J^c}\|_2^2 \\ &\quad + \sigma^2 \text{Tr}((\Phi'_J \Phi_J)^{-1}) + \|\beta_{J^c}\|_2^2 \\ &\geq \frac{1}{1 + (|J| - 1)\mu} (\|\beta_{J^c}\|_2^2 + \sigma^2 |J|). \end{aligned} \quad (8)$$

See, for example, [7]. Note that $(|J| - 1)\mu \leq (k - 1)\mu < \frac{1}{2}$. Hence, the ideal risk is bounded by

$$\begin{aligned} R_{\text{oracle}} &\geq \frac{2}{3} \min_{J \in \mathcal{J}} (\|\beta_{J^c}\|_2^2 + \sigma^2 |J|) \\ &\geq \frac{2}{3} \sum_{i=1}^p \min(\beta^2(i), \sigma^2). \end{aligned}$$

Note that the ideal risk R_{oracle} is neither known nor attainable as it requires the knowledge of the ideal k -sparse subset J_{oracle} . So an interesting question is: Can the oracle risk R_{oracle} be (almost) attained by a purely data-driven procedure without the knowledge of J_{oracle} ?

The following result shows that the answer is affirmative: under the sufficient and sharp MIP condition $k < \frac{1}{2}(\frac{1}{\mu} + 1)$, ℓ_1 minimization nearly attains the oracle risk without the need of knowing the optimal subset J_{oracle} or the value of the ideal risk R_{oracle} .

Theorem 4.1: Consider the Gaussian model (7). Suppose β is k -sparse with $k < \frac{1}{2}(\frac{1}{\mu} + 1)$. Set $\lambda_* = \sigma(\sqrt{2 \log p} + \frac{3}{2})$. Let $\hat{\beta}$ be the minimizer of the problem

$$\min \|\gamma\|_1 \quad \text{subject to} \quad \|\Phi'(y - \Phi\gamma)\|_\infty < \lambda_*. \quad (9)$$

Then with probability at least $1 - \frac{1}{\sqrt{\pi \log p}}$, $\hat{\beta}$ satisfies

$$\|\hat{\beta} - \beta\|_2^2 \leq \frac{8(\sqrt{2 \log p} + 2)^2}{(1 - (2k - 1)\mu)^2} \left(\sigma^2 + \sum_{i=1}^p \min(\beta^2(i), \sigma^2) \right). \quad (10)$$

Remark 4.1: As aforementioned, a similar oracle inequality was derived in [7, Th. 1.2] in the context of restrictive isometry property. In that setting, it was required to have the ℓ_1 minimization parameter $\lambda_* \geq (1 + c)\lambda$ where $\lambda = \sigma\sqrt{2 \log p}$ and $c > 0$ is a constant. In comparison, under the MIP condition we require a weaker condition on λ_* as the λ_* is essentially the same as λ in the sense that $\lambda_*/\lambda \rightarrow 1$ as $p \rightarrow \infty$. In addition, as we shall see later, under the MIP condition, the proof of the oracle inequality (10) is particularly simple.

Proof: Without loss of generality, we assume $\sigma = 1$ and $|\beta(1)| \geq |\beta(2)| \geq \dots \geq |\beta(k)|$ and $\beta(i) = 0$ for $i > k$. Set $\lambda = \sqrt{2 \log p}$ and define the event $A = \{z : \|\Phi'z\|_\infty \leq \lambda\}$. Standard bound on Gaussian tail probability shows that

$$P\left(\|\Phi'z\|_\infty \leq \sqrt{2 \log p}\right) \geq 1 - \frac{1}{\sqrt{\pi \log p}}.$$

That is, the event A occurs with probability at least $1 - \frac{1}{\sqrt{\pi \log p}}$.

In the following we shall assume that the event A occurs, i.e., $\|\Phi'z\|_\infty \leq \lambda$. Let $k_0 = \lceil \sum_{i=1}^p \min(\beta^2(i), 1) \rceil$ and write $\alpha_1 = \beta \mathbb{1}_{\{1, 2, \dots, k_0\}}$ and $\alpha_2 = \beta \mathbb{1}_{\{k_0+1, \dots, k\}}$. Then

- 1) $\beta = \alpha_1 + \alpha_2$.
- 2) By the definition of k_0 , it is clear that $\{|i : |\beta(i)| \geq 1|\} \leq k_0$. This implies that $|\beta(i)| < 1$ if $i > k_0$. Therefore

$$\begin{aligned} \|\alpha_2\|_1 &= \sum_{i=k_0+1}^k |\beta(i)| < k - k_0; \\ &\quad \text{and} \\ \|\alpha_2\|_2 &= \sqrt{\sum_{i=k_0+1}^k \beta^2(i)} \\ &= \sqrt{\sum_{i=k_0+1}^k \min(\beta^2(i), 1)} \leq \sqrt{k_0}. \end{aligned}$$

First we verify that α_1 is a feasible solution to (9). In fact, for any i with $1 \leq i \leq p$

$$\begin{aligned} & |\langle \Phi_i, \Phi \alpha_1 - y \rangle| \\ &= |\langle \Phi_i, \Phi \beta - y \rangle - \langle \Phi_i, \Phi \alpha_2 \rangle| \\ &\leq |\langle \Phi_i, z \rangle| + |\langle \Phi_i, \Phi \alpha_2 \rangle| \\ &\leq \lambda + \sum_{j=k_0+1}^k |\langle \Phi_i, \Phi_j \rangle| |\alpha_2(j)| \\ &\leq \begin{cases} \lambda + (1 - \mu) |\alpha_2(i)| + \mu \|\alpha_2\|_1 & \text{if } k_0 + 1 \leq i \leq k \\ \lambda + \mu \|\alpha_2\|_1 & \text{otherwise} \end{cases} \\ &\leq \lambda + 1 + (k - k_0 - 1)\mu \\ &\leq \lambda + \frac{3}{2} = \lambda_*. \end{aligned}$$

From Theorem 2.2, we have

$$\begin{aligned} \|\hat{\beta} - \alpha_1\|_2 &\leq \frac{\sqrt{2k_0}}{1 - (2k_0 - 1)\mu} (\lambda_* + \lambda_*) \\ &= \frac{2\sqrt{2k_0}}{1 - (2k_0 - 1)\mu} (\sqrt{2 \log p} + 2) \\ &\quad - \frac{\sqrt{2k_0}}{1 - (2k_0 - 1)\mu} \\ &\leq \frac{2\sqrt{2}}{1 - (2k_0 - 1)\mu} (\sqrt{2 \log p} + 2) \sqrt{k_0} - \sqrt{k_0} \end{aligned}$$

and, hence

$$\begin{aligned} \|\hat{\beta} - \beta\|_2 &\leq \|\hat{\beta} - \alpha_1\|_2 + \|\alpha_2\|_2 \\ &\leq \frac{2\sqrt{2}}{1 - (2k_0 - 1)\mu} (\sqrt{2 \log p} + 2) \sqrt{k_0}. \end{aligned}$$

Consequently

$$\begin{aligned} \|\hat{\beta} - \beta\|_2^2 &\leq \frac{8}{(1 - (2k - 1)\mu)^2} (\sqrt{2 \log p} + 2)^2 k_0 \\ &\leq \frac{8(\sqrt{2 \log p} + 2)^2}{(1 - (2k - 1)\mu)^2} \left(1 + \sum_{i=1}^p \min(\beta^2(i), 1) \right). \end{aligned}$$

■

V. RELATIONS TO THE RESTRICTED ISOMETRY PROPERTY

We have shown in the earlier sections that the MIP condition $k \leq \frac{1}{2}(\frac{1}{\mu} + 1)$ is sufficient and sharp for stable recovery of k -sparse signals in the presence of noise. Besides MIP, the sparse signal recovery problem has also been well studied in the framework of the restricted isometry property (RIP) introduced by [6]. For an $n \times p$ matrix Φ and an integer k , $1 \leq k \leq p$, the k -restricted isometry constant δ_k is the smallest constant such that

$$\sqrt{1 - \delta_k} \|c\|_2 \leq \|\Phi c\|_2 \leq \sqrt{1 + \delta_k} \|c\|_2 \quad (11)$$

for every k -sparse vector c . If $k + k' \leq p$, the k, k' -restricted orthogonality constant $\theta_{k,k'}$, is the smallest number that satisfies

$$|\langle \Phi c, \Phi c' \rangle| \leq \theta_{k,k'} \|c\|_2 \|c'\|_2 \quad (12)$$

for all c and c' such that c and c' are k -sparse and k' -sparse respectively, and have disjoint supports.

It has been shown that ℓ_1 minimization can recover a sparse signal with a small or zero error under various conditions on δ_k and $\theta_{k,k'}$. For example, $\delta_k + \theta_{k,k} + \theta_{k,2k} < 1$ was used in [6], $\delta_{3k} + 3\delta_{4k} < 2$ in [5], and $\delta_{2k} + \theta_{k,2k} < 1$ in [7], $\delta_{1.5k} + \theta_{k,1.5k} < 1$ in [3] and $\delta_{1.25k} + \theta_{k,1.25k} < 1$ in [2].

Simple conditions involving only the isometry constant δ have also been given in the literature on sparse signal recovery, for example, $\delta_{2k} < \sqrt{2} - 1$ was used in [4] and $\delta_{2k} < 0.472$ was given by [2]. Davies and Gribonval [8] constructed examples which demonstrate that stable recovery cannot be guaranteed if $\delta_{2k} \geq \frac{1}{\sqrt{2}} + \epsilon$ for $\epsilon > 0$. This indicates that the upper bound for δ_{2k} is likely less than $\frac{1}{\sqrt{2}}$ in order to ensure stable recovery for all k -sparse signals. The exact sharp bound for the RIP constant δ_{2k} remains unknown. This is an interesting topic for further research.

An advantage of MIP is that it can be used to deterministically verify whether a given matrix satisfy (4). In contrast, calculating the RIP parameters is typically computationally infeasible and it is thus not possible to verify RIP conditions for a given matrix. However, RIP constants can be used to probabilistically construct compressed sensing matrices which achieves the nice effect that the number of rows can be considerably less than the number of columns. For example, as long as $n = O(k \log \frac{p}{k})$, then with positive probability (when n is large), the $n \times p$ matrix Φ with entries drawn independently according to Gaussian distribution, $\Phi_{i,j} \sim N(0, \frac{1}{n})$, is a compressed sensing matrix, see [1], [6], and [17].

The connections between RIP and MIP have been noted in the literature. It can be seen that $\mu = \theta_{1,1} = \delta_2$, and the following relations can be found in [3] (the first inequality can be also found in [18])

$$\delta_k \leq (k - 1)\mu, \quad \text{and} \quad \theta_{k,k'} \leq \sqrt{kk'}\mu. \quad (13)$$

With the above relations, the results in RIP can also be stated in terms of MIP. However, current RIP conditions are not enough for achieving the (sharp) MIP condition (4)

$$k < \frac{1}{2} \left(\frac{1}{\mu} + 1 \right).$$

For example, to the best of our knowledge, one of the best RIP conditions in the literature so far is

$$\delta_{1.25k} + \theta_{k,1.25k} < 1.$$

Using (13), this yields the corresponding MIP condition

$$k < \frac{1}{1.25 + \sqrt{1.25}} \left(\frac{1}{\mu} + 1 \right) \approx \frac{1}{2.368} \left(\frac{1}{\mu} + 1 \right),$$

which is stronger than (4).

It should be noted that RIP and MIP conditions, although strongly connected, are different conditions. As indicated by the

above example, the RIP framework does not contain the MIP framework, or vice versa. Methods and techniques for treating MIP and RIP are different as well.

ACKNOWLEDGMENT

The authors would like to thank the referees for thorough and useful comments which have helped to improve the presentation of the paper.

REFERENCES

- [1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constr. Approx.*, vol. 28, no. 2008, pp. 253–263.
- [2] T. Cai, L. Wang, and G. Xu, "Shifting inequality and recovery of sparse signals," *IEEE Trans. Signal Process.*, vol. 58, pp. 1300–1308.
- [3] T. Cai, G. Xu, and Zhang, "On recovery of sparse signals via ℓ_1 minimization," *IEEE Trans. Inf. Theory*, vol. 55, no. 2009, pp. 3388–3397.
- [4] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l' Acad. Sci., Paris*, ser. I, vol. 346, no. 2008, pp. 589–592.
- [5] E. J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Commun. Pure Appl. Math.*, vol. 59, no. 2006, pp. 1207–1223.
- [6] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 2005, pp. 4203–4215.
- [7] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n (with discussion)," *Ann. Statist.*, vol. 35, no. 2007, pp. 2313–2351.
- [8] M. E. Davies and R. Gribonval, "Restricted isometry constants where ℓ_p sparse recovery can fail for $0 < p \leq 1$," *IEEE Trans. Inf. Theory*, vol. 55, no. 2009, pp. 2203–2214.
- [9] 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. 2006, pp. 6–18.
- [10] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 2001, pp. 2845–2862.
- [11] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 1994, pp. 425–455.
- [12] D. L. Donoho and J. Tanner, "Neighborliness of randomly projected simplices in high dimensions," *Proc. Nat. Acad. Sci.*, vol. 102, no. 27, pp. 9452–9457.
- [13] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression (with discussion)," *Ann. Statist.*, vol. 32, no. 2004, pp. 407–451.
- [14] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 2004, pp. 1341–1344.
- [15] J.-J. Fuchs, "Recovery of exact sparse representations in the presence of bounded noise," *IEEE Trans. Inf. Theory*, vol. 51, no. 2005, pp. 3601–3608.
- [16] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 2003, pp. 3320–3325.
- [17] M. Rudelson and R. Vershynin, "Sparse reconstruction by convex relaxation: Fourier and Gaussian measurements," in *Proc. CISS 2006 (40th Ann. Conf. Inf. Sci. Syst.)*, 2006.
- [18] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 2004, pp. 2231–2242.
- [19] P. Tseng, "Further results on a stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 55, no. 2009, pp. 888–899.
- [20] W. Xu and B. Hassibi, "Compressive sensing over the Grassmann manifold: A unified analytical approach," in *Proc. Allerton Conf. Commun., Contr. Comput.*, Sep. 2008, pp. 562–567.

Tony Tony Cai received the Ph.D. degree from Cornell University, Ithaca, NY, in 1996.

His research interests include high-dimensional inference, large-scale multiple testing, nonparametric function estimation, functional data analysis, and statistical decision theory. He is the Dorothy Silberberg Professor of Statistics at the Wharton School of the University of Pennsylvania, Philadelphia.

Dr. Cai is the recipient of the 2008 COPSS Presidents' Award and a fellow of the Institute of Mathematical Statistics. He is the current editor of the *Annals of Statistics*.

Lie Wang received the Ph.D. degree in statistics from the University of Pennsylvania, Philadelphia, in 2008.

He is now with the Department of Mathematics, Massachusetts Institute of Technology, Cambridge. His research interests include nonparametric function estimation, high-dimensional sparse regression, semiparametric models, and functional data regression.

Guangwu Xu received the Ph.D. degree in mathematics from the State University of New York, Buffalo.

He is now with the Department of Electrical Engineering and Computer Science, University of Wisconsin-Milwaukee. His research interests include cryptography and information security, computational number theory, algorithms, and functional analysis.