Support recovery in compressed sensing: An estimation theoretic approach

Amin Karbasi, Ali Hormati, Soheil Mohajer, Martin Vetterli

School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne, Switzerland

February 9, 2009

Abstract

Compressed sensing (CS) deals with the reconstruction of sparse signals from a small number of linear measurements. One of the main challenges in CS is to find the support of a sparse signal from a set of noisy observations. In the CS literature, several information-theoretic bounds on the scaling law of the required number of measurements for exact support recovery have been derived, where the focus is mainly on random measurement matrices.

In this paper, we investigate the support recovery problem from an estimation theory point of view, where no specific assumption is made on the underlying measurement matrix. By using the Hammersley-Chapman-Robbins (HCR) bound, we derive a fundamental lower bound on the performance of any unbiased estimator which provides necessary conditions for reliable ℓ_2-norm support recovery. We then analyze the optimal decoder to provide conditions under which the HCR bound is achievable. This leads to a set of sufficient conditions for reliable ℓ_2-norm support recovery.

1 Introduction

Linear sampling of sparse signals, with a number of samples close to their sparsity level, has recently received great attention under the name of Compressed Sensing or Compressive Sampling (CS) [1, 2]. A k-sparse signal $\boldsymbol{\theta} \in \mathbb{R}^p$ is defined as a signal with $k \ll p$ nonzero expansion coefficients in some orthonormal basis or frame. The goal of compressed sensing is to find

measurement matrices $\mathbf{\Phi}_{m\times p}$, followed by reconstruction algorithms which allow robust recovery of sparse signals using the least number of measurements m, and low computational complexity.

In practice, however, all the measurements are noisy, and thus the exact recovery of $\boldsymbol{\theta}$ is impossible. Support recovery refers to the problem of correctly estimating the position of the non-zero entries based on a set of noisy observations. A large body of recent work (e.g., [3], [4], [5], [6]) has established information theoretic limits for *exact* support recovery based on the $\{0,1\}-$valued loss function. This work mainly focuses on the standard Gaussian measurement ensemble where the elements of the measurement matrix are drawn i.i.d from the Gaussian distribution $\mathcal{N}(0,1)$.

In this paper, we look at the support recovery problem from an estimation theory point of view, where the error metric between the true and the estimated support is the ℓ_2-norm. The positions of the nonzero entries of $\boldsymbol{\theta}$ forms a set of k integers between 1 and p. Consequently, the support recovery in a discrete setup can be regarded as estimating restricted parameters. This leads us to use the Hammersley-Chapman-Robbins (HCR) bound which provides a lower bound on the variance of any unbiased estimator of a set of restricted parameters [7, 8].

The organization of this paper is as follows. In Section 2, we provide a more precise formulation of the problem. We derive the HCR bound for the support recovery problem in Section 3, where no assumption is made on the measurement matrix. We then apply the obtained bound on random measurement matrices, in order to determine a lower bound on the number of measurements for reliable ℓ_2-norm support recovery. Of equal interest are the conditions under which the derived HCR bound is achievable. To this end, in Section 4, we study the performance of the Maximum-Likelihood (ML) decoder and derive conditions under which it becomes unbiased and achieves the HCR bound. Again, no assumption is made on the measurement matrix. Using the Gaussian measurement ensemble, as an example, we can then identify the sufficient number of measurements for reliable ℓ_2-norm support recovery.

2 Problem statement

In this paper, we consider a deterministic signal model in which $\boldsymbol{\theta} \in \mathbb{R}^p$ is a fixed but unknown vector with exactly k non-zero entries. We refer to k as the signal sparsity, p as the signal dimension, and define the support vector

as the positions of the non-zero elements of $\boldsymbol{\theta}$. More precisely,

$$\boldsymbol{s}(\boldsymbol{\theta}) \triangleq (n_1, n_2, \ldots, n_k), \tag{1}$$

where the corresponding non-zero entries of $\boldsymbol{\theta}$ are

$$\boldsymbol{\theta_s} \triangleq (\theta_{n_1}, \theta_{n_2}, \ldots, \theta_{n_k}). \tag{2}$$

We assume that $n_1 < n_2 < \cdots < n_k$. Suppose we are given a vector of m noisy observations $\boldsymbol{y} \in \mathbb{R}^m$ of the form

$$\boldsymbol{y} = \boldsymbol{\Phi\theta} + \boldsymbol{\epsilon}, \tag{3}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{m\times p}$ is the measurement matrix, and $\boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma^2 I_{m\times m}\right)$ is additive Gaussian noise. Throughout this paper, we assume w.l.o.g that σ^2 is fixed, since any scaling of σ^2 can be acounted for in the scaling of $\boldsymbol{\theta}$. Let $\boldsymbol{x} = \boldsymbol{\Phi\theta}$, and $\boldsymbol{\Phi}_s$ denote the subspace spanned by the columns of $\boldsymbol{\Phi}$ at positions indexed by $\boldsymbol{s}(\boldsymbol{\theta})$. Since there are $N = \binom{p}{k}$ subspaces of dimension k, a number from 1 to N can be assigned to them and w.l.o.g., we assume that $\boldsymbol{x}$ belongs to the first subspace $\boldsymbol{s}_1 = \boldsymbol{s}$.

Due to the presence of noise, $\boldsymbol{\theta}$ cannot be recovered exactly. However, a sparse-recovery algorithm outputs an estimate $\boldsymbol{\theta'}$. In the support recovery problem, we are only interested in estimating the support. To that end, we can consider different performance metrics for the estimate. In [6], the measure of error between the estimate and the true signal is a $\{0,1\}-$valued loss function:

$$\rho_1(\boldsymbol{\theta}, \boldsymbol{\theta'}) = \mathbb{I}\left(\boldsymbol{s}(\boldsymbol{\theta}) \neq \boldsymbol{s}(\boldsymbol{\theta'})\right), \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This metric is appropriate for the exact support recovery. In this work, we are interested in an approximate support recovery. For this purpose, we consider the following ℓ_2-norm error metric

$$\rho_2(\boldsymbol{\theta}, \boldsymbol{\theta'}) = \|\boldsymbol{s}(\boldsymbol{\theta}) - \boldsymbol{s}(\boldsymbol{\theta'})\|_2^2. \tag{5}$$

Note that $\rho_2(\boldsymbol{\theta}, \boldsymbol{\theta'}) = 0$ implies $\rho_1(\boldsymbol{\theta}, \boldsymbol{\theta'}) = 0$ and vice-versa.

As was mentioned in [6], the SNR is not suitable for the support recovery problem. It is possible to generate problem instances for which the support recovery is arbitrarily difficult, in particular, by sending the smallest coefficient to zero (assuming that $k > 1$) at an arbitrarily rapid rate, even as the SNR becomes arbitrarily large by increasing the rest. Hence, we also define

$$\theta_{\min} = \min_{i\in s} |\theta_i|. \tag{6}$$

In particular, our results apply to any unbiased decoder that operates over the signal class

$$C(\theta_{\min}) = \{\boldsymbol{\theta} \in \mathbb{R}^p : |\theta_i| \geq \theta_{\min} \ \forall i \in \boldsymbol{s}\}. \tag{7}$$

With this setup, our goal is to find conditions for any unbiased estimator, based on the parameters p, m, k and $\theta_{\min}$, under which the variance of error for *any* signal picked from the signal class $C(\theta_{\min})$ goes to zero as the signal dimension increases. Our analysis is high dimensional in nature, in the sense that the signal dimension p goes to infinity. More precisely, we say the ℓ_2-norm support recovery is reliable if

$$\lim_{p\to\infty} \rho_2(\boldsymbol{\theta}, \boldsymbol{\theta}') = 0, \tag{8}$$

for any $\boldsymbol{\theta} \in C(\theta_{\min})$, under some scaling of $(\theta_{\min}, k, m)$ as a function of p. For unbiased estimators, (8) is equivalent to

$$\lim_{p\to\infty} \mathrm{tr}[\mathrm{cov}(\hat{\boldsymbol{s}}(\boldsymbol{\theta}))] = 0, \tag{9}$$

where $\hat{\boldsymbol{s}}(\boldsymbol{\theta})$ is the estimated support of $\boldsymbol{\theta}$. Since the support estimation is based on $\boldsymbol{y}$, with abuse of notation, we also denote it by $\hat{\boldsymbol{s}}(\boldsymbol{y})$. Throughout this paper, we only consider unbiased estimators.

3 Hammersley-Chapman-Robbins Bound

The Cramer-Rao (CR) bound is a well-known tool in statistics which provides a lower bound on the variance of the error of any unbiased estimator of an unknown deterministic parameter δ, from a set of measurements $\boldsymbol{y}$ [9]. More specifically, in a single parameter scenario, the estimated value $\hat{\delta}$ satisfies

$$\mathrm{var}(\hat{\delta}) \geq \frac{1}{-\int_{-\infty}^{\infty} \frac{\partial^2 \ln \mathbb{P}(y;\delta)}{\partial \delta^2} \mathbb{P}(y;\delta) dy}, \tag{10}$$

where $\mathbb{P}(y;\delta)$ is the pdf of the measurements which depends on the parameter δ. As (10) suggests, the CR bound is typically derived for estimating a continuous parameter.

In many cases, there is *a priori* information on the estimated parameter which restricts it to take values from a pre determined set. An example is the estimation of the mean of a normal distribution when one knows that the true mean is an integer. In such scenarios, the Hammersley-Chapman-Robbins (HCR) bound provides a stronger lower bound on the variance of

any unbiased estimator [7, 8]. More specifically, let us assume that the set of independent observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$ is drawn according to a probability distribution with density function $\mathbb{P}(\boldsymbol{y}; \boldsymbol{\delta})$ where δ is a parameter belonging to some parameter set Δ (e.g., the set of integer numbers) and completely characterizes the pdf. In addition, the sequence $\boldsymbol{\delta}$ is partitioned into two subsequences $\boldsymbol{\delta} = (\boldsymbol{\delta}_1, \boldsymbol{\delta}_2)$ where we are only interested in estimating the parameters included in subsequence $\boldsymbol{\delta}_1$. Let $\hat{\boldsymbol{\delta}}_1(\boldsymbol{y})$ denote an unbiased estimator of $\boldsymbol{\delta}_1$. The HCR bound on the trace of the covariance matrix of any unbiased estimator of $\boldsymbol{\delta}_1$ is given by

$$\mathrm{tr}[\mathrm{cov}(\hat{\boldsymbol{\delta}}_{\mathbf{1}})] \geq \sup_{\boldsymbol{\delta}' \neq \boldsymbol{\delta}} \frac{\|\boldsymbol{\delta}_1 - \boldsymbol{\delta}'_1\|_2^2}{\int_{\mathbb{R}^m} \frac{\mathbb{P}^2(\boldsymbol{y};\boldsymbol{\delta}')}{\mathbb{P}(\boldsymbol{y};\boldsymbol{\delta})} d\boldsymbol{y} - 1}, \tag{11}$$

in which $\boldsymbol{\delta}' = (\boldsymbol{\delta}'_1, \boldsymbol{\delta}'_2) \in \Delta$. The set Δ is chosen so that $\boldsymbol{\delta}'$ takes values according to the *a priori* information.

Example 3.1. *For clarity, let us consider the performance of an unbiased estimator of the mean of a normal distribution based on independent samples of size m, i.e. $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$. In this case, $\boldsymbol{\delta} = (\mu, \sigma^2)$, $\delta_1 = \mu$, $\delta_2 = \sigma^2$ and*

$$\mathbb{P}(\boldsymbol{y}; \boldsymbol{\delta}) = (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \mu)^2}. \tag{12}$$

Let $\hat{\mu}(\boldsymbol{y})$ denote an unbiased estimator of μ, the parameter we want to estimate. When there is no prior information on μ, it follows from the CR bound that

$$var(\hat{\mu}) \geq \sigma^2/m. \tag{13}$$

Once the mean is restricted to be an integer, we may write $\delta_1 = \mu$ and $\delta'_1 = \mu + \alpha$ where α is a non-zero integer. Then upon integration we get

$$\begin{aligned} var(\hat{\mu}) &\geq \max_{\alpha \neq 0} \frac{\alpha^2}{e^{m\alpha^2/\sigma^2} - 1} && (14) \\ &= \frac{1}{e^{m/\sigma^2} - 1}, && (15) \end{aligned}$$

where the maximum is attained for $\alpha = \pm 1$. A point worth mentioning is the role of prior information. While (13) drops linearly, (15) decreases exponentially with respect to the number of observations. It is also interesting to note that (14) applies as well to the case in which the parameter is not restricted. We then have to deal with the maximization in (14) for variations in α where α may take any value (not necessarily integral) except $\alpha = 0$. Since the RHS of (14) is a decreasing function of α^2, we let $\alpha \to 0$ and we deduce (13).

In the support recovery problem, we know a priori that each entry of the support vector takes values from the restricted set $\Delta = \{1, 2, \ldots, p\}$. Hence the HCR bound can provide us with a lower bound on the performance of any unbiased estimator.

Theorem 3.2. *Assume $\hat{\boldsymbol{s}}(\boldsymbol{y})$ to be an unbiased estimator of the support $\boldsymbol{s}$. The HCR lower bound on the variance of $\hat{\boldsymbol{s}}(\boldsymbol{y})$ is given by*

$$tr[cov(\hat{\boldsymbol{s}})] \geq \max_{i \in \{2, \cdots, N\}} \frac{\|\boldsymbol{s} - \boldsymbol{s}_i\|^2}{e^{\|\boldsymbol{x} - p_{\boldsymbol{s}_i}\boldsymbol{x}\|^2/\sigma^2} - 1}, \tag{16}$$

in which $p_{\boldsymbol{s}_i}\boldsymbol{x}$ denotes the projection of $\boldsymbol{x}$ onto the subspace spanned by $\boldsymbol{\Phi}_{\boldsymbol{s}_i}$.

Proof. Since our observations are of the form $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{\epsilon}$, the set of unknown parameters $\boldsymbol{\delta}$ consists of the support vector $\boldsymbol{s}(\boldsymbol{\theta}) = (n_1, n_2, \ldots, n_k)$ and the corresponding coefficients $\boldsymbol{\theta}_{\boldsymbol{s}} = (\theta_{n_1}, \theta_{n_2}, \ldots, \theta_{n_k})$. We are only interested in estimating the support, hence, $\boldsymbol{\delta}_1 = \boldsymbol{s}(\boldsymbol{\theta})$ and $\boldsymbol{\delta}_2 = \boldsymbol{\theta}_{\boldsymbol{s}}$. Then

$$\frac{\mathbb{P}^2(\boldsymbol{y}; \boldsymbol{\delta}')}{\mathbb{P}(\boldsymbol{y}; \boldsymbol{\delta})} = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - 2x_i' + x_i)^2 - 2(x_i' - x_i)^2}{2\sigma^2}}, \tag{17}$$

where $\boldsymbol{x}' = \boldsymbol{\Phi}\boldsymbol{\theta}'$. Upon integration we get

$$\int_{\mathbb{R}^m} \frac{\mathbb{P}^2(\boldsymbol{y}; \boldsymbol{\delta}')}{\mathbb{P}(\boldsymbol{y}; \boldsymbol{\delta})} d\boldsymbol{y} - 1 = e^{\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{\sigma^2}} - 1. \tag{18}$$

Using the HCR bound

$$\mathrm{tr}[\mathrm{cov}(\hat{\boldsymbol{s}})] \geq \sup_{\boldsymbol{\delta}' \neq \boldsymbol{\delta}} \frac{\|\boldsymbol{s} - \boldsymbol{s}'\|^2}{e^{\|\boldsymbol{x} - \boldsymbol{x}'\|^2/\sigma^2} - 1}. \tag{19}$$

If $\boldsymbol{x}$ and $\boldsymbol{x}'$ live in the same subspace, i.e., $\boldsymbol{s} = \boldsymbol{s}'$, the RHS of (19) will be zero. Therefore, in order to find the supremum, we can restrict our attention to all the signals which do not live in the same subspace as $\boldsymbol{x}$ does:

$$\mathrm{tr}[\mathrm{cov}(\hat{\boldsymbol{s}})] \geq \sup_{\{\boldsymbol{\theta}' : \boldsymbol{s}(\boldsymbol{\theta}') \neq \boldsymbol{s}(\boldsymbol{\theta})\}} \frac{\|\boldsymbol{s} - \boldsymbol{s}'\|^2}{e^{\|\boldsymbol{x} - \boldsymbol{x}'\|^2/\sigma^2} - 1}. \tag{20}$$

For each sequence $\boldsymbol{s}'$, the numerator of (20) is fixed (it is the ℓ_2 distance between the supports and does not depend on the coefficients) while the denominator is minimized by setting $\boldsymbol{x}' = p_{\boldsymbol{s}'}\boldsymbol{x}$. This leads to (16). □

In the following, we see how Theorem 3.2 helps us find a lower bound on the number of measurements for reliable ℓ_2-norm support recovery.

3.1 Necessary Conditions

Using the HCR bound, Theorem 3.2 provides a lower bound on the performance of any unbiased estimator for the ℓ_2-norm support recovery problem. In words, the ℓ_2-norm support recovery is unreliable if the RHS of (16) is bounded away from zero which yields to a lower bound on the minimum number of measurements. The following example illustrates how this bound can be used when the Gaussian measurement matrices $\mathbf{\Phi}$ are deployed.

Random Matrices: As an example, we obtain the necessary conditions on the number of measurements required for reliable ℓ_2-norm support recovery, when each entry Φ_{ij} is drawn i.i.d. from a Gaussian distribution $\mathcal{N}(0,1)$.

Theorem 3.3. *Let the measurement matrix $\mathbf{\Phi} \in \mathbb{R}^{m\times p}$ be drawn with i.i.d. elements from a Gaussian distribution with zero-mean and variance one. Then the ℓ_2-norm support recovery over the signal class $C(\theta_{\min})$ is unreliable if*

$$m < \max\left\{k, \frac{\sigma^2 \log(p-k)}{\theta_{\min}^2}\right\}. \tag{21}$$

Proof. From Theorem 3.2 we know that for any $\boldsymbol{x}' \in \boldsymbol{s}'$ we have

$$\mathrm{tr}[\mathrm{cov}(\hat{\boldsymbol{s}})] \geq \frac{\|\boldsymbol{s}-\boldsymbol{s}'\|^2}{e^{\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\sigma^2}-1}. \tag{22}$$

The ℓ_2-norm support recovery is reliable if (8) holds for any $\boldsymbol{\theta} \in C(\theta_{\min})$. In particular, when $\boldsymbol{s}(\boldsymbol{\theta}) = (1, 2, \ldots, k)$ and it takes on $\theta_{\min}$ as its last non-zero entry, i.e., $\theta_k = \theta_{\min}$. Moreover, assume that $\boldsymbol{\theta}'$ is equal to $\boldsymbol{\theta}$ on all the positions but the smallest non-zero value. Note that one can find a $\boldsymbol{\theta}'$ such that $\|\boldsymbol{s}-\boldsymbol{s}'\|^2$ be at least $(p-k)^2$ by simply choosing $\boldsymbol{s}(\boldsymbol{\theta}') = (1, 2, \ldots, k-1, p)$, i.e., putting the smallest non-zero entry of $\boldsymbol{\theta}'$ in the last position. Now

$$\boldsymbol{x} - \boldsymbol{x}' = \mathbf{\Phi}(\boldsymbol{\theta} - \boldsymbol{\theta}'). \tag{23}$$

This implies that

$$\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{\sigma^2} = \frac{\theta_{\min}^2}{\sigma^2} Z, \tag{24}$$

where $Z \sim \chi^2(m)$. Note that $\mathrm{tr}[\mathrm{cov}(\hat{s})]$ is bounded away from zero if $\|\boldsymbol{s} - \boldsymbol{s}'\|^2/(e^{\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\sigma^2} - 1)$ does not go to zero. This will happen if

$$\Pr\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{\sigma^2} \dot{<} \log(p-k)\right) \to 1, \tag{25}$$

as $p \to \infty$, where by $A \dot{<} B$ we mean multiplicatively less than B in asymptote, i.e., there exists a *constant* $\delta > 0$ such that $A \leq (1+\delta)B$. The expression (25) is equivalent to

$$\Pr\left(Z \dot{>} \frac{\sigma^2 \log(p-k)}{\theta_{\min}^2}\right) \to 0, \tag{26}$$

as $p \to \infty$. It is known that a centralized χ^2 variate with m degrees of freedom satisfies

$$\Pr\left[Z - m \geq 2\sqrt{mt}\right] \leq e^{-t}, \tag{27}$$

for all $t \geq 0$ [10]. Combining (26) and (27) leads to

$$\Pr\left(Z \dot{>} \frac{\sigma^2 \log(p-k)}{\theta_{\min}^2}\right) \leq e^{-\left(\frac{\sigma^2 \log(p-k)}{\theta_{\min}^2} - m\right)^2 / 4m}, \tag{28}$$

provided that

$$m < (1+C)\frac{\sigma^2 \log(p-k)}{\theta_{\min}^2}, \tag{29}$$

for some constant $C > 0$ (note that (27) is only valid for $t \geq 0$). Clearly, under the condition (29), the right hand side of (28) tends to zero as p grows. $\square$

Table 1 demonstrates the necessary conditions for different scalings of k and $\theta_{\min}$ as a function of p.

Up to this point we have discussed the HCR bound and its application in finding necessary conditions on the number of measurements for reliable ℓ_2-norm support recovery. What remains is to find conditions under which the HCR bound is achievable which consequently provides us with the sufficient number of measurements for reliable ℓ_2-norm support recovery.

4 Achievability of the HCR Bound

We now analyze the performance of the Maximum-Likelihood (ML) estimator for the ℓ_2-norm support recovery and find conditions under which it becomes unbiased and in addition, its performance moves towards that of the HCR bound. Provided that any $2k$ columns of the measurement matrix $\boldsymbol{\Phi}$ are linearly independent, the noiseless measurement vector $\boldsymbol{x} = \boldsymbol{\Phi}\boldsymbol{\theta}$ belongs to one and *only* one of the N possible subspaces. Since the noise $\epsilon \in \mathbb{R}^m$ is i.i.d. Gaussian, the ML estimator selects the subspace closest to the observed vector $\boldsymbol{y} \in \mathbb{R}^m$. More precisely,

$$\hat{\boldsymbol{s}}_{\mathrm{ML}} = \underset{\boldsymbol{s}:|\boldsymbol{s}|=k}{\operatorname{argmin}} \ \|\boldsymbol{y} - p_{\boldsymbol{s}}\boldsymbol{y}\|_2. \tag{30}$$

Now, consider another subspace $\boldsymbol{\Phi}_{\boldsymbol{s}'}$, of dimension k where $\boldsymbol{s} \neq \boldsymbol{s}'$. Clearly an error happens, when ML selects the support $\boldsymbol{s}'$ in place of the true support $\boldsymbol{s}$. Let $\Pr_{\text{ML}}(\boldsymbol{s}')$ denote the probability that ML selects the subspace $\boldsymbol{s}'$ instead of $\boldsymbol{s}$ among all the subspaces.

Lemma 4.1. *Let $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{x} = \boldsymbol{\Phi\theta} \in \boldsymbol{\Phi}_{\boldsymbol{s}}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ and $\boldsymbol{s}'$ be a support sequence different from $\boldsymbol{s}$. Then*

$$\Pr_{ML}(\boldsymbol{s}') < \Pr\left(\|\boldsymbol{\epsilon}\| \geq \frac{\|\boldsymbol{x} - p_{\boldsymbol{s}'}\boldsymbol{x}\|}{2}\right). \tag{31}$$

Proof. ML chooses $\boldsymbol{s}'$ over $\boldsymbol{s}$ if and only if

$$\min_{\boldsymbol{t}' \in \boldsymbol{\Phi}_{\boldsymbol{s}'}} \|\boldsymbol{y} - \boldsymbol{t}'\| < \min_{\boldsymbol{t} \in \boldsymbol{\Phi}_{\boldsymbol{s}}} \|\boldsymbol{y} - \boldsymbol{t}\|. \tag{32}$$

Let us assume that $\|\boldsymbol{\epsilon}\| < \|\boldsymbol{x} - p_{\boldsymbol{s}'\boldsymbol{x}}\|/2$. Then, for any $\boldsymbol{t}' \in \boldsymbol{\Phi}_{\boldsymbol{s}'}$, we have

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{t}'\|^2 &= \|\boldsymbol{x} - \boldsymbol{t}' + \boldsymbol{\epsilon}\|^2 \\
&\geq \|\boldsymbol{\epsilon}\|^2 + \|\boldsymbol{x} - \boldsymbol{t}'\|^2 - 2\|\boldsymbol{x} - \boldsymbol{t}'\|\|\boldsymbol{\epsilon}\| \\
&\overset{(a)}{>} \|\boldsymbol{\epsilon}\|^2 \\
&= \|\boldsymbol{y} - \boldsymbol{x}\|^2 \\
&\geq \min_{\boldsymbol{t} \in \boldsymbol{s}} \|\boldsymbol{y} - \boldsymbol{t}\|^2,
\end{aligned}$$

where in (a), we used the mentioned assumption. This implies that if $\|\boldsymbol{\epsilon}\| \leq \|\boldsymbol{x} - p_{\boldsymbol{s}'}\boldsymbol{x}\|/2$, the ML estimator will not choose $\boldsymbol{s}'$ over $\boldsymbol{s}$. Since the probability that the ML estimator *picks* $\boldsymbol{s}'$ instead of $\boldsymbol{s}$ is less than the probability that it *prefers* $\boldsymbol{s}'$ to $\boldsymbol{s}$, we get (31). □

Lemma 4.2. *Let the number of measurements m, be even. Then*

$$\Pr_{ML}(\boldsymbol{s}') < e^{-r/2} \sum_{t=0}^{m/2-1} \frac{(r/2)^t}{t!}, \tag{33}$$

where $r = \frac{\|\boldsymbol{x} - p_{\boldsymbol{s}'}\boldsymbol{x}\|^2}{4\sigma^2}$.

Proof. By Lemma 4.1 we have

$$\Pr_{\text{ML}}(\boldsymbol{s}') < 1 - \Pr\left(\frac{\|\boldsymbol{\epsilon}\|^2}{\sigma^2} < r\right).$$

The random variable $\frac{\|\boldsymbol{\epsilon}\|^2}{\sigma^2}$ is distributed according to the chi-square distribution with m degrees of freedom. By using the cdf of the chi-square distribution, we obtain

$$\Pr_{\text{ML}}(\boldsymbol{s}') < 1 - \frac{\gamma(m/2, r/2)}{\Gamma(m/2)}, \tag{34}$$

where $\Gamma(m)$, is the Gamma function and $\gamma(m, x)$, is the lower incomplete Gamma function. It is easy to show that for an even number m,

$$\frac{\gamma(m/2, r/2)}{\Gamma(m/2)} = e^{-r/2} \sum_{t=\frac{m}{2}}^{\infty} \frac{(r/2)^t}{t!}.$$

Since by the Taylor expansion $e^{r/2} = \sum_{t=0}^{\infty} \frac{(r/2)^t}{t!}$, we obtain

$$\frac{\gamma(m/2, r/2)}{\Gamma(m/2)} = 1 - e^{-r/2} \sum_{t=0}^{m/2-1} \frac{(r/2)^t}{t!}. \tag{35}$$

Combining (34) and (35) will lead to the lemma. □

Lemma 4.3. *Let $r = \alpha m$ for some constant $\alpha > 1$. Then we have*

$$\Pr_{ML}(\boldsymbol{s}') < \frac{r}{2\alpha}\, c(\alpha)^{-r}, \tag{36}$$

in which $c(\alpha) = e^{(\alpha-1)/2\alpha}/\alpha^{1/2\alpha} > 1$ and $c(\alpha) \longrightarrow \sqrt{e}$ as α grows.

Proof. Note that for $t < \frac{r}{2}$, the function $f(t) = \left(\frac{r}{2}\right)^t / t!$ is strictly increasing. By observing that $\frac{m}{2} - 1 < \frac{r}{2}$, and employing Lemma 4.2 we get:

$$\begin{aligned}
\Pr_{\text{ML}}(\boldsymbol{s}') &< e^{-r/2} \sum_{t=0}^{\frac{m}{2}-1} \frac{(r/2)^t}{t!} \\
&< e^{-r/2} \frac{m}{2} \frac{(r/2)^{m/2}}{(m/2)!} \\
&\overset{(a)}{<} e^{-r/2} \frac{m}{2} \frac{(r/2)^{m/2}}{(m/2e)^{m/2}} \\
&= \frac{r}{2\alpha} \left(\frac{e^{(\alpha-1)/2\alpha}}{\alpha^{1/2\alpha}} \right)^{-r},
\end{aligned}$$

where in (a) we used the inequality $m! > (m/e)^m$. It can be verified that $c(\alpha) > 1$ for $\alpha > 1$. Although we do not prove it here, it is not hard to see that the upper bound shows a linear decay for $\alpha \leq 1$. □

In the following theorem, we provide an upper bound on the performance of the ML estimator. Based on Lemma 4.1, the ML probability of error is related to the minimum distance between $\boldsymbol{x}$ and its projections onto the other subspaces. Let $d_{\min} \triangleq \min_{\boldsymbol{s}':\boldsymbol{s}'\neq\boldsymbol{s}} \|\boldsymbol{x} - p_{\boldsymbol{s}'}\boldsymbol{x}\|$, $\beta = d_{\min}^2/4m\sigma^2$ and $r_{\min} = \beta m$.

Theorem 4.4. *For $\beta > 1$, the performance of the ML estimator is upper bounded as*

$$\mathrm{tr}[\underset{\mathrm{ML}}{\mathrm{cov}}(\hat{\boldsymbol{s}})] < \frac{kmp^2}{2}\, c(\beta)^{-r_{min}}. \tag{37}$$

Proof. By Lemma 4.1, we know that if $\|\boldsymbol{\epsilon}\| < d_{\min}/2$, ML makes the correct choice. Therefore, from Lemma 4.3, we obtain

$$\underset{\mathrm{ML}}{\Pr}(\mathrm{err}) < \frac{r_{\min}}{2\beta}\, c(\beta)^{-r_{\min}}, \tag{38}$$

Since $\|\boldsymbol{s} - \boldsymbol{s}_i\|^2 < kp^2$ and $\Pr_{\mathrm{ML}}(\mathrm{err}) = \sum_{i=2}^{N} \Pr_{\mathrm{ML}}(\boldsymbol{s}_i)$, we obtain

$$\begin{aligned}
\mathrm{tr}[\underset{\mathrm{ML}}{\mathrm{cov}}(\hat{\boldsymbol{s}})] &= \sum_{i=2}^{N} \underset{\mathrm{ML}}{\Pr}(\boldsymbol{s}_i)\|\boldsymbol{s} - \boldsymbol{s}_i\|^2 \\
&< kp^2 \frac{r_{\min}}{2\beta}\, c(\beta)^{-r_{\min}} \\
&= \frac{kmp^2}{2}\, c(\beta)^{-r_{\min}}.
\end{aligned}$$

□

In general, the ML estimator can be biased and its performance cannot be compared with unbiased estimators. The following theorem provides us with the condition under which it becomes unbiased.

Theorem 4.5. *Under the conditions $m \geq (1+\varepsilon)\log(p)/\beta \log c(\beta)$, for some fixed $\varepsilon > 0$ and β bounded away from 1, the ML estimator is asymptotically unbiased as $p \to \infty$.*

Proof. Let $\hat{\boldsymbol{s}} = (\hat{n}_1, \hat{n}_2, \ldots, \hat{n}_k)$ be the ML estimate for the true support set $\boldsymbol{s} = (n_1, n_2, \ldots, n_k)$. Then

$$\begin{aligned}
\mathbb{E}(\hat{\boldsymbol{s}}) &= \sum_{i=1}^{N} \hat{\boldsymbol{s}}_i \underset{\mathrm{ML}}{\Pr}(\hat{\boldsymbol{s}}_i) \\
&= \boldsymbol{s} \underset{\mathrm{ML}}{\Pr}(\boldsymbol{s}) + \sum_{\hat{\boldsymbol{s}}_i \neq \boldsymbol{s}} \hat{\boldsymbol{s}}_i \underset{\mathrm{ML}}{\Pr}(\hat{\boldsymbol{s}}_i).
\end{aligned}$$

Since $\sum_{\hat{\boldsymbol{s}}_i \neq \boldsymbol{s}} \Pr_{\text{ML}}(\hat{\boldsymbol{s}}_i) = \Pr_{\text{ML}}(\text{err})$ and $1 \leq \hat{n}_i \leq p$ for $1 \leq i \leq k$, we have

$$\sum_{\hat{\boldsymbol{s}}_i \neq \boldsymbol{s}} \hat{\boldsymbol{s}}_i \Pr_{\text{ML}}(\hat{\boldsymbol{s}}_i) \leq (p, p, \ldots, p) \Pr_{\text{ML}}(\text{err}).$$

Since $\beta > 1$, from (38) we get

$$\lim_{p \to \infty} \sum_{\hat{\boldsymbol{s}}_i \neq \boldsymbol{s}} \hat{\boldsymbol{s}}_i \Pr_{\text{ML}}(\hat{\boldsymbol{s}}_i) \leq \lim_{p \to \infty} (p, p, \ldots, p) \frac{m}{2} c(\beta)^{-\beta m} \overset{(a)}{=} \underline{\mathbf{0}},$$

where in (a), we used $m \geq (1+\varepsilon) \log(p) / \beta \log c(\beta)$. Obviously, $\Pr_{\text{ML}}(\boldsymbol{s}) \to 1$ as $p \to \infty$. Hence $\mathbb{E}(\hat{\boldsymbol{s}}) = \boldsymbol{s}$. □

As we observe, our results do not depend on any specific measurement matrix. On the one hand, Theorem 4.4 provides us with an upper bound on the error of the ML estimator. On the other hand, since by Theorem 4.5, the ML estimator is unbiased under the mentioned conditions, its estimation error is lower bounded by the HCR bound, which shows a 9 dB gap in the denominator with the upper bound. Therefore, such asymptotic behavior of the ML estimator, shows the achievability of the HCR bound, under the mentioned conditions.

In the following, we see how Theorem 4.4 leads to find the sufficient number of measurements for reliable ℓ_2-norm support recovery where the Gaussian measurement ensemble is used.

4.1 Sufficient Conditions

Theorem 4.4 provides us with an upper bound on the performance of the ML estimator. For reliable ℓ_2-norm support recovery, the RHS of (37) should go to zero as $p \to \infty$. To that end, one should make sure that first, β is bounded away from one, which is a property of the underlying measurement matrix and second, that the number of measurements is at least of the order of $\log p$ which assures that the ML estimator is unbiased.

Theorem 4.6. *Let the measurement matrix $\boldsymbol{\Phi}$ be drawn with i.i.d. elements from a Gaussian distribution $\mathcal{N}(0,1)$. If the minimum value remains constant (meaning $\theta_{\min} = \Theta(1)$), then $m = \Theta(k \log \frac{p-k}{k})$ number of measurements suffices to ensure reliable ℓ_2-norm support recovery.*

Proof. The proof follows the same lines as [6] to show that both mentioned conditions are simultaneously satisfied. In this part, we analyze the upper bound on the performance of the ML estimator for the measurement matrix

Φ, with elements i.i.d. $\mathcal{N}(0,1)$. To ensure that $\beta > 1$, we need to find the scaling for which

$$\Pr(\min_{s'} \frac{\|x - P_{s'}x\|}{4\sigma^2} > m) \to 1,$$

where $x = \Phi\theta$. We have,

$$\begin{aligned}\|x - P_{s'}x\|^2 &= \|P_{s'}^{\perp}\Phi\theta\|^2 \\ &= \|P_{s'}^{\perp}\Phi_{s/s'}\theta_{s/s'}\|^2,\end{aligned}$$

where s/s' denotes the elements of s which does not belong to s'. It can be shown that,

$$X_{s,s'} = \frac{\|P_{s'}^{\perp}\Phi_{s/s'}\theta_{s/s'}\|^2}{\|\theta_{s/s'}\|^2} \sim \chi^2(m-k).$$

Moreover, note that

$$\begin{aligned}\Pr(\min_{s'} \frac{\|x - P_{s'}x\|}{4\sigma^2} > m) &= \Pr\left[\bigcap_{s'_j} \frac{\|x - P_{s'_j}x\|}{4\sigma^2} > m\right] \\ &= \Pr\left(\left[\bigcup_{s'_j} \frac{\|P_{s'_j}^{\perp}x\|^2}{4\sigma^2} < m\right]^c\right) \\ &\overset{(a)}{\geq} 1 - \sum_j \Pr\left[\frac{\|P_{s'_j}^{\perp}x\|^2}{4\sigma^2} < m\right],\end{aligned}$$

where in (a) we have used the union bound. For any subspace $s' \neq s$, we have

$$\Pr\left[\frac{\|P_{s'}^{\perp}x\|^2}{4\sigma^2} < m\right] = \Pr\left[X_{s,s'} < \frac{4m\sigma^2}{\|\theta_{s/s'}\|^2}\right].$$

Under the condition

$$m - k > \frac{4m\sigma^2}{\|\theta_{s/s'}\|^2}, \tag{39}$$

we use the following bound for the the $\chi^2(m-k)$ distribution,

$$\Pr(X_{s,s'} - (m-k) < -2\sqrt{(m-k)x}) \leq \exp(-x),$$

where x is given by,

$$x = \frac{(\frac{m-k}{2} - \frac{2m\sigma^2}{\|\theta_{s/s'}\|^2})^2}{m-k}.$$

Note that the condition given in (39) is satisfied, irrespective of the subspace s' if

$$m - k > \frac{4m\sigma^2}{\theta_{\min}^2}. \tag{40}$$

Thus, we can write

$$\Pr\left[X_{s,s'} < \frac{4m\sigma^2}{\|\theta_{s/s'}\|^2}\right] < \exp\left(-\left(\frac{\sqrt{m-k}}{2} - \frac{2m\sigma^2}{\|\theta_{s/s'}\|^2\sqrt{m-k}}\right)^2\right).$$

Note that we have $\|\theta_{s/s'}\|^2 \geq \ell\,\theta_{\min}^2$ where $\ell = |s/s'|$. Under condition (39), we have

$$\triangle = \frac{\sqrt{m-k}}{2} - \frac{2m\sigma^2}{\|\theta_{s/s'}\|^2\sqrt{m-k}} > 0$$

and,

$$\bigtriangledown = \frac{\sqrt{m-k}}{2} - \frac{2m\sigma^2}{\ell\theta_{\min}^2\sqrt{m-k}} > 0.$$

Since

$$\triangle \geq \bigtriangledown,$$

we get

$$\exp\left(-\left(\frac{\sqrt{m-k}}{2} - \frac{2m\sigma^2}{\|\theta_{s/s'}\|^2\sqrt{m-k}}\right)^2\right) \leq \exp\left(-\left(\frac{\sqrt{m-k}}{2} - \frac{2m\sigma^2}{\ell\theta_{\min}^2\sqrt{m-k}}\right)^2\right),$$

thus, we conclude

$$\begin{aligned}
\sum_j \Pr\left[\frac{\|P_{s_j'}^{\perp}x\|^2}{4\sigma^2} < m\right] &\leq \sum_{\ell=1}^{k}\binom{k}{\ell}\binom{p-k}{\ell}\exp\left(-\left(\frac{\sqrt{m-k}}{2} - \frac{2m\sigma^2}{\ell\theta_{\min}^2\sqrt{m-k}}\right)^2\right) \\
&\leq k\max_{1\leq\ell\leq k}\left\{\binom{k}{\ell}\binom{p-k}{\ell}\exp\left(-\left(\frac{\sqrt{m-k}}{2} - \frac{2m\sigma^2}{\ell\theta_{\min}^2\sqrt{m-k}}\right)^2\right)\right\} \\
&= k\max_{1\leq\ell\leq k}\left\{\binom{k}{\ell}\binom{p-k}{\ell}\exp\left(-\frac{(m-k)(\ell\theta_{\min}^2 - \frac{4\sigma^2}{1-k/m})^2}{4\ell^2\theta_{\min}^4}\right)\right\}.
\end{aligned}$$

Note that we require

$$\max_{1\leq\ell\leq k}\left\{\log k + \log\binom{k}{\ell} + \log\binom{p-k}{\ell} - \frac{(m-k)(\ell\frac{\theta_{\min}^2}{\sigma^2} - \frac{4}{1-k/m})^2}{4\ell^2\frac{\theta_{\min}^4}{\sigma^4}}\right\} \to -\infty.$$

From now on, w.l.o.g., we assume that $\sigma^2 = 1$. For the scalings $k = o(m)$ or $k = \theta(m)$, we have

$$\frac{4}{1 - k/m} \to c,$$

therefore, asymptotically, we should have

$$m - k > \max_{1 \le \ell \le k} \left\{ \left(\frac{2\ell\theta_{\min}^2}{\ell\theta_{\min}^2 - c} \right)^2 \left(\log k + \ell \log \frac{ke}{\ell} + \ell \log \frac{(p-k)e}{\ell} \right) \right\}.$$

We divide our derivation of the sufficient condition for the support recovery into three different regimes:

1. $\ell = \Theta(k)$

   We get,

   $$m - k > \left(\frac{2k\theta_{\min}^2}{k\theta_{\min}^2 - c} \right)^2 \left(\log k + c_1 k + k \log \frac{p-k}{k} \right),$$

   which is asymptotically equivalent to

   $$m > \left(\frac{2k\theta_{\min}^2}{k\theta_{\min}^2 - c} \right)^2 k \log \frac{p-k}{k} + c_2 k.$$

2. $\ell = \Theta(1)$

   In this regime, we have

   $$m > c_1 k + c_2 \frac{\theta_{\min}^4}{(\theta_{\min}^2 - c)^2} \log (p-k)$$

3. $\ell = o(k)$
   Note that in this regime, we have

   $$\lim_{\frac{\ell}{k} \to 0} \frac{\ell \log \frac{k}{\ell}}{k} = 0. \tag{41}$$

   Moreover,

   $$\ell \log \frac{p-k}{\ell} < k \log \frac{p-k}{k}. \tag{42}$$

To show (42), note that in the regime $k = \Theta(p)$, we have

$$\frac{\frac{\ell}{k}\log\frac{p-k}{\ell}}{\log\frac{p-k}{k}} \to 0. \tag{43}$$

On the other hand, in the regime $k = o(p)$, we have

$$\begin{aligned}\frac{\ell\log\frac{p-k}{\ell}}{k\log\frac{p-k}{k}} &= \frac{\ell}{k}\frac{\log\frac{p}{\ell}}{\log\frac{p}{k}} \\ &= \frac{\ell}{k}\frac{1-\frac{\log\ell}{\log p}}{1-\frac{\log k}{\log p}} \to 0.\end{aligned}$$

Therefore, the lower bound on m in cases 1 and 2 covers the current regime.

In total, we get the following sufficient condition on the number of measurements needed for the perfect support reconstruction

$$m > \max\left\{c_1 k\log\frac{p-k}{k}, c_2 k + c_3\frac{\theta_{\min}^4}{(\theta_{\min}^2 - c)^2}\log p - k\right\}$$

where, we assume that the condition

$$\theta_{\min}^2 > \frac{4}{1-k/m}$$

is satisfied. Thus, for the case $\theta_{\min}^2 = \Theta(1)$, we find the following sufficient conditions for perfect support recovery

$$\begin{aligned}k = \Theta(p) &\Longrightarrow m = \Theta(p) \\ k = o(p) &\Longrightarrow m = \Theta(k\log\frac{p}{k}).\end{aligned}$$

□

Sufficient conditions in different regimes are shown in Table 1.

Acknowledgment

The authors would like to thank Prof. M. J. Wainwright for his help and many useful comments.

	Necessary conditions	Sufficient conditions
$k = \Theta(p)$ $\theta^2_{\min} = \Theta\left(\frac{1}{k}\right)$	$\Theta\left(p \log p\right)$	$*$
$k = \Theta(p)$ $\theta^2_{\min} = \Theta\left(1\right)$	$\Theta(p)$	$\Theta(p)$
$k = o(p)$ $\theta^2_{\min} = \Theta\left(\frac{1}{k}\right)$	$\Theta\left(k \log(p-k)\right)$	$*$
$k = o(p)$ $\theta^2_{\min} = \Theta\left(1\right)$	$\Theta\left(k\right)$	$\Theta\left(k \log \frac{p}{k}\right)$

Table 1: Necessary and sufficient conditions on the number of measurements required for reliable ℓ_2-norm support recovery.

References

[1] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[2] E.J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[3] M. Akcakaya and V. Tarokh, "Shannon theoretic limits on noisy compressive sampling," Tech. Rep. arXiv:cs.IT:0711.0366, Harvard Univ., November 2007.

[4] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions on sparsity pattern recovery," Tech. Rep. arXiv:cs.IT:0804.1839, UC Berkeley, April 2008.

[5] S. Sarvotham, D. Baron, and R. G. Baraniuk, "Measurements versus bits: Compressed sensing meets information theory," in *Proc. Allerton Conference on Control, Communication and Computing*, September 2006.

[6] M. J. Wainwright, "Information-theoretic bounds for sparsity recovery: Dense versus sparse measurements," Tech. Rep. 726, Department of Statistics, UC Berkeley, April 2008.

[7] J. M. Hammersley, "On estimating restricted parameters," *J. Roy. Stat. Soc.*, vol. 12, no. 2, pp. 192–240, 1950.

[8] D.G. Chapman and H. Robbins, "Minimum variance estimation without regularity assumption," *Annals Math. Stat.*, vol. 22, no. 4, pp. 581–586, 1951.

[9] P. Stoica and R. Moses, *Introduction to Spectral Analysis*, Prentice-Hall, 2000.

[10] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, vol. 28, no. 5, pp. 1303–1338, 1998.