Bernoulli **20**(4), 2014, 2217–2246
DOI: 10.3150/13-BEJ555

Adaptive sensing performance lower bounds for sparse signal detection and support estimation

RUI M. CASTRO

Eindhoven University of Technology, The Netherlands. E-mail: rmcastro@tue.nl

In memory of Yuri Ingster

This paper gives a precise characterization of the fundamental limits of adaptive sensing for diverse estimation and testing problems concerning sparse signals. We consider in particular the setting introduced in (*IEEE Trans. Inform. Theory* **57** (2011) 6222–6235) and show necessary conditions on the minimum signal magnitude for both detection and estimation: if $\mathbf{x} \in \mathbb{R}^n$ is a sparse vector with s non-zero components then it can be reliably detected in noise provided the magnitude of the non-zero components exceeds $\sqrt{2/s}$. Furthermore, the signal support can be exactly identified provided the minimum magnitude exceeds $\sqrt{2 \log s}$. Notably there is no dependence on n, the extrinsic signal dimension. These results show that the adaptive sensing methodologies proposed previously in the literature are essentially optimal, and cannot be substantially improved. In addition, these results provide further insights on the limits of adaptive compressive sensing.

Keywords: adaptive sensing; minimax lower bounds; sequential experimental design; sparsity-based models

1. Introduction

This paper addresses the characterization of the fundamental limits of adaptive sensing in sparse settings, when a potentially infinite number of observations is available but there is a restriction on the sensing precision budget available. One of the key aspects of adaptive sensing is that the data collection process is sequential and adaptive. In different fields these sensing/experimenting paradigms are known by different names, such as *sequential experimental design* in statistics and economics (see Wald [35], Bessler [5], Fedorov [19], El-Gamal [18], Hall and Molchanov [21], Lai and Robbins [29], Blanchard and Geman [6]), *active learning* or *adaptive sensing/sampling* in computer science, engineering and machine learning (see Cohn, Ghahramani and Jordan [12], Freund *et al.* [20], Novak [33], Korostelev and Kim [27], Dasgupta [13], Castro, Willett and Nowak [9], Dasgupta, Kalai and Monteleoni [15], Dasgupta [14], Hanneke [22], Koltchiinskii [28], Balcan, Beygelzimer and Langford [4], Castro and Nowak [10]).

The extra flexibility of adaptive sensing can sometimes (but not always) yield significant performance gains. In this paper, we are particularly concerned with the setting introduced in Haupt, Castro and Nowak [24], where the authors propose an adaptive sparse signal recovery method that provably improves on the best possible non-adaptive sensing methods. However, in that work there is no indication on the fundamental performance limitations in such sensing sce-

1350-7265 © 2014 ISI/BS

narios. This paper addresses those breeches in our understanding, and shows that the proposed procedures are essentially asymptotically optimal for estimation problems. Furthermore, with some modifications, the procedure of Haupt, Castro and Nowak [24] is also nearly optimal when testing for the presence of a sparse signal. In addition, we also present results characterizing the fundamental limitations in several other settings, such as exact support recovery, as in Malloy and Nowak [31], Malloy and Nowak [30] or in Arias-Castro, Candès and Davenport [2].

2. Problem setting

Let $\mathbf{x} \in \mathbb{R}^n$ be an unknown vector. We assume this vector is sparse in the sense that only a reduced number of its entries are not-zero. In particular, let S be a subset of $\{1, \dots, n\}$ and assume that for all $i \in \{1, \dots, n\}$ such that $i \notin S$ we have $x_i = 0$. We refer to S as the signal support set and this is our main object of interest. In this paper, we consider two distinct classes of problems: (i) signal support estimation, where we desire to estimate S; (ii) signal detection, where we simply want to test if S belongs to some particular class.

In our model the signal $\mathbf{x}$ is unknown, but we can collect partial information through noisy observations. In particular, we observe

$$Y_k = x_{A_k} + \Gamma_k^{-1} W_k, \qquad k = 1, 2, \dots, \tag{2.1}$$

where A_k, Γ_k are taken to be measurable functions of $\{Y_i, A_i, \Gamma_i\}_{i=1}^{k-1}$, and W_k are standard normal random variables, independent of $\{Y_i\}_{i=1}^{k-1}$ and also independent of $\{A_i, \Gamma_i\}_{i=1}^{k}$. In this model, $A_k \in \{1, \dots, n\}$ corresponds to the entry of $\mathbf{x}$, that is, measured at time k, therefore A_k can be viewed as the *sensing action* taken at time k. Similarly, Γ_k^2 is the *precision* of the measurement taken at time k. Finally, there is a total sensing budget constraint that must be satisfied, namely

$$\sum_{k=1}^{\infty} \Gamma_k^2 \le m, \tag{2.2}$$

where $m > 0$. It is important to note that we can consider both deterministic sequential designs or random sequential designs. In the latter, we allow the choices A_k and Γ_k to incorporate extraneous randomness, which is not explicitly described in the model. Besides being more general this extra flexibility often facilitates the analysis. The collection of conditional distributions of A_k, Γ_k given $\{Y_i, A_i, \Gamma_i\}_{i=1}^{k-1}$ for all k is referred to as the *sensing strategy*, and denoted by $\mathcal{A}$. Note that, within the sensing model above, we can also consider non-adaptive sensing frameworks, meaning the choice of sensing actions and precision allocation must be made before collecting any data. Formally, this means that $\{A_k, \Gamma_k\}_{k \in \mathbb{N}}$ is statistically independent from $\{Y_k\}_{k \in \mathbb{N}}$. Note that a non-adaptive design can still be random.

The case $m = n$ is of particular interest and this is often considered in literature as it allows direct comparison between adaptive and non-adaptive sensing methodologies. If $m = n$ we allow, on average, one unit of precision for each one of the n signal entries. Therefore if we assume the signal $\mathbf{x}$ belongs a class for which there is no reason to give a priori preference to any particular signal entry the optimal non-adaptive sensing strategy amounts to measuring each vector entry exactly once, with precision one.[1] This is obviously the classical normal means model.

In the following sections, we consider two different scenarios: signal detection/testing and signal estimation. In both cases, the extra flexibility of adaptive sensing is shown to be extremely rewarding. We characterize the fundamental performance limits of adaptive sensing in those scenarios and show that these limits can be achieved by practical inference methodologies.

3. Signal detection

In this setting, we are interested in a binary hypothesis testing problem, where we test a simple null hypothesis against a composite alternative. In particular, the null hypothesis H_0 is simply $S = \varnothing$, and the alternative hypothesis H_1 is $S \in \mathcal{C}$, where $\mathcal{C}$ is some class of non-empty subsets of $\{1, \dots, n\}$. We are particularly interested in the case when under the alternative H_1 all the sets in $\mathcal{C}$ have cardinality s, meaning that for all $S \in \mathcal{C}$ we have $|S| = s$. We will consider only such classes as this greatly simplifies the presentation and is not, for the most part, a restrictive condition.

Define

$$x_{\min} = \min\bigl\{|x_i| : x_i \neq 0, i \in \{1, \dots, n\}\bigr\}.$$

In the following, we characterize the fundamental signal detection limits, in particular identifying conditions on $x_{\min}$ as a function of $\mathcal{C}$ and n, such that no procedure is able to reliably distinguish the two hypotheses. Furthermore, these bounds are essentially tight, in the sense that there exist practical procedures matching them. For simplicity, we consider only non-negative signals, meaning that $x_i \geq 0$ for all $i \in \{1, \dots, n\}$. This greatly simplifies the analysis, without hindering the generality of the results. More comments about this are issued in Remark 3.2. Furthermore, the hardest signals to detect or estimate are of the form

$$x_i = \begin{cases} \mu & \text{if } i \in S, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

This means that we can restrict our analysis to signals of the form above, which are entirely described by the signal support set S and signal amplitude μ. This is also the class of signals considered in Addario-Berry *et al.* [1] or in Donoho and Jin [16] in a non-adaptive sensing context.

Let

$$D = \{Y_i, A_i, \Gamma_i\}_{i \in \mathbb{N}},$$

and let $d = \{y_i, a_i, \gamma_i\}_{i=1}^{\infty}$ be a particular realization of the experimental procedure. Let $\mathcal{A}$ denote a particular sensing strategy, and $\hat{\Phi}(D) \in \{0, 1\}$ be an arbitrary testing function, taking the value 1 if the null hypothesis is to be rejected, and zero otherwise. For notational convenience we write

[1] Due to statistical sufficiency there is no gain in measuring each signal entry more than once.

simply $\hat{\Phi}$ where the hat indicates the dependency on the data D. The *risk* of this procedure is given by

$$R(\hat{\Phi}) = \mathbb{P}_\varnothing(\hat{\Phi} \neq 0) + \max_{S \in \mathcal{C}} \mathbb{P}_S(\hat{\Phi} \neq 1),$$

where $\mathbb{P}_S$ denotes the joint probability distribution of $\{Y_i, A_i, \Gamma_i\}_{i=1}^{\infty}$ for a given value of S. Likewise we use $\mathbb{E}_S$ to denote expectation under $\mathbb{P}_S$.

Now define

$$c(\mu, \mathcal{C}) = \inf_{\hat{\Phi}, \mathcal{A}} R(\hat{\Phi}) = \inf_{\hat{\Phi}, \mathcal{A}} \Big\{ \mathbb{P}_\varnothing(\hat{\Phi} \neq 0) + \max_{S \in \mathcal{C}} \mathbb{P}_S(\hat{\Phi} \neq 1) \Big\}. \tag{3.2}$$

Our formal goal is to identify the values of the signal magnitude μ for which we have necessarily $c(\mu, \mathcal{C}) \geq \varepsilon$ for $\varepsilon > 0$.

Remark 3.1. The choice of risk above is obviously not the only one possible, and in the literature other choices of risk have been considered, such as

$$\tilde{R}(\hat{\Phi}) = \max \Big\{ \mathbb{P}_\varnothing(\hat{\Phi} \neq 0), \max_{S \in \mathcal{C}} \mathbb{P}_S(\hat{\Phi} \neq 1) \Big\}, \tag{3.3}$$

or

$$\bar{R}(\hat{\Phi}) = \mathbb{P}_\varnothing(\hat{\Phi} \neq 0) + \frac{1}{N} \sum_{S \in \mathcal{C}} \mathbb{P}_S(\hat{\Phi} \neq 1). \tag{3.4}$$

As discussed in Addario-Berry *et al.* [1], the latter measure of risk corresponds to the view that, under the alternative hypothesis, a set $S \in \mathcal{C}$ is selected uniformly at random from $\mathcal{C}$. Clearly

$$\bar{R}(\hat{\Phi}) \leq R(\hat{\Phi}) \leq 2\tilde{R}(\hat{\Phi}) \leq 2R(\hat{\Phi}).$$

If there is sufficient symmetry in $\mathcal{C}$ and $\hat{\Phi}$ these three risk measures are essentially identical. Whenever possible we characterize the fundamental limits of adaptive sensing for each one of the risk measures, but focus primarily on $R(\hat{\Phi})$.

3.1. Main results – Detection

The class $\mathcal{C}$ of all subsets of $\{1, \ldots, n\}$ with cardinality s is one of particular interest. This is the class of maximal size, and obviously the one for which we expect the worst performance for detection. Perhaps surprisingly, under the adaptive sensing paradigm, the exact same performance lower bound is obtained for *any* class $\mathcal{C}$ exhibiting some very mild symmetry. This means that, in many situations, the structure of the class $\mathcal{C}$ does not really help under the adaptive sensing scenario. This is in stark contrast with non-adaptive sensing scenarios, where the structure of the set $\mathcal{C}$ can play a very prominent role, as well documented in Addario-Berry *et al.* [1], Arias-Castro *et al.* [3], Butucea and Ingster [7]. To state the main result of this section, we need the following definitions:

Definition 3.1 (Symmetric class/full range). *Let $\Xi = \bigcup_{S \in \mathcal{C}} S$ and S be drawn uniformly at random from $\mathcal{C}$. If for all $i \in \Xi$ we have $\mathbb{P}(i \in S) = s/|\Xi|$ the class $\mathcal{C}$ is said to be symmetric. Furthermore, if $|\Xi| = n$ the class is said to be full range.*

It is remarkable that many classes $\mathcal{C}$ of interest satisfy this mild symmetry, as for instance, all the classes in Addario-Berry *et al.* [1].

Theorem 3.1. *Consider the setting above and let $\mathcal{C}$ be a symmetric class. Let $\hat{\Phi}(D)$ be an arbitrary testing procedure, where $D = \{Y_i, A_i, \Gamma_i\}_{i \in \mathbb{N}}$. Finally, let $0 < \varepsilon < 1$ be arbitrary. If $R(\hat{\Phi}) \leq \varepsilon$ then necessarily*

$$x_{\min} \geq \sqrt{\frac{2|\Xi|}{sm} \log \frac{1}{2\varepsilon}}. \tag{3.5}$$

This result gives a condition on the minimal signal magnitude necessary to ensure the detection risk is not too large. Perhaps surprisingly the lower bound does not include any factor involving specific structural properties of $\mathcal{C}$, but only the range and cardinality of the corresponding sets. A possible way to understand this comes from the following observation: for detection, it suffices to identify a *single* element of S, and there is no need to identify all the elements. Therefore, cues provided by the structure are not very informative. In addition, note that the above theorem also applies to non-symmetric classes provided they contain a symmetric class. Before proving this result, it is interesting to present a simple corollary for the case of full range classes, emphasizing the asymptotic behavior.

Corollary 3.1. *Let $\mathcal{C}$ be a symmetric and full range class of sets with cardinality s, where s can be a function of n (this dependence is not explicitly stated). Let $\hat{\Phi}_n$ be an arbitrary adaptive sensing testing procedure. If*

$$\lim_{n \to \infty} R(\hat{\Phi}_n) = 0$$

then necessarily

$$x_{\min} \geq \omega_n \sqrt{\frac{n}{sm}},$$

where ω_n is a sequence for which $\lim_{n \to \infty} \omega_n = \infty$.

This corollary gives a necessary condition for detection consistency. As shown in Proposition 3.3, this bound is actually tight, meaning there are adaptive sensing procedures that can detect signals satisfying the above condition. The case $m = n$ is particularly interesting, as it allows the comparison between adaptive and non-adaptive sensing performance. For that case, adaptive sensing detection is possible if $x_{\min} = \omega_n \sqrt{1/s}$. Since ω_n can diverge arbitrarily slowly we see that the extrinsic signal dimension n plays no significant role in this bound, and only the intrinsic dimension s is relevant. Keep in mind, however, that ω_n is related to the rate of convergence of the risk $R(\hat{\Phi}_n)$ to zero. Corollary 3.1 is in stark contrast to what is known for the same problem if one restricts to the classical setting of non-adaptive sensing, as in Ingster [25], Ingster and Suslina [26], Donoho and Jin [16], Donoho [17]. For instance, for the class of all subsets with cardinality s it is necessary to have $x_{\min} \geq c\sqrt{\log n}$ if $s = \mathrm{o}(\sqrt{n})$, where the factor $c > 0$ depends on the specific relation between s and n. In Meinshausen and Rice [32], the authors considered estimation of the proportion of significant components $|S|/n$. Their setting is more general, as the distributions corresponding to significant and insignificant signal component observations can be non-normal. Their approach can be used to test the hypothesis $|S| = 0$. For the Gaussian case, they recover essentially the $\sqrt{\log n}$ scaling. Finally, in Cai, Jin and Low [8] the authors consider again the estimation of the fraction of significant signal components in the normal means case, and show results beyond consistency, including minimax rates of convergence of the risk. We now proceed with the proof of the theorem and a discussion about tightness of the bounds.

Proof of Theorem 3.1. The proof of this lower bound hinges, as usual, on the analysis of likelihood ratios. Begin by defining the joint probability density function of $\{Y_k, A_k, \Gamma_k\}_{k=1}^{\infty}$ under S, which we denote by

$$f(d; S) = f(y_1, a_1, \gamma_1, y_2, a_2, \gamma_2, \dots; S).$$

Note that this is properly defined for a certain dominating measure (mixed continuous and discrete). Taking into account the conditional dependences in our observation model we can factorize this probability density function as follows

$$\begin{aligned} f(d; S) = {} & f_{A_1,\Gamma_1}(a_1, \gamma_1) \times f_{Y_1|A_1,\Gamma_1}(y_1|a_1, \gamma_1; S) \\ & \times f_{A_2,\Gamma_2|Y_1,A_1,\Gamma_1}(a_2, \gamma_2|y_1, a_1, \gamma_1) \times f_{Y_2|A_2,\Gamma_2}(y_2|a_2, \gamma_2; S) \times \cdots. \end{aligned}$$

Note that in this factorization only some terms involve the underlying true set S, while all the other terms depend solely on the sensing strategy used. This greatly simplifies the computation of likelihood ratios, as all the terms not involving S cancel out. In particular, the likelihood ratio between two hypotheses is given simply by

$$\mathrm{LR}_{S,S'}(d) = \frac{f(d; S)}{f(d; S')} \tag{3.6}$$

$$= \prod_{k=1}^{\infty} \frac{f_{Y_k|A_k,\Gamma_k}(y_k|a_k, \gamma_k; S)}{f_{Y_k|A_k,\Gamma_k}(y_k|a_k, \gamma_k; S')}. \tag{3.7}$$

As usual, in order to effectively distinguish if the underlying true distribution is parameterized by S or S' the corresponding likelihood ratio needs to be significantly different than 1. We proceed by formally stating this. Our analysis is heavily inspired by the approach in Chernoff [11].

The first step is to relate the probabilities of type I and type II errors to the likelihood ratio, namely giving a relation between $\mathbb{P}_S(\hat{\Phi} \neq 1)$ and $\mathbb{P}_\varnothing(\hat{\Phi} \neq \varnothing)$ where S is an arbitrary element of $\mathcal{C}$. Begin by defining the total variation and the Kullback–Leibler divergence between two probability measures.

Definition 3.2. *Let $\mathbb{P}_0$ and $\mathbb{P}_1$ be two probability measures defined on a common measurable space $(\Omega, \mathcal{B})$. The total variation distance is defined as*

$$\mathrm{TV}(\mathbb{P}_0, \mathbb{P}_1) = \sup_{B \in \mathcal{B}} \big|\mathbb{P}_0(B) - \mathbb{P}_1(B)\big|.$$

The Kullback–Leibler divergence is defined as

$$\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_1) = \begin{cases} \int_\Omega \log \frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{P}_1} \, \mathrm{d}\mathbb{P}_0 & \textit{if } \mathbb{P}_0 \ll \mathbb{P}_1, \\ +\infty & \textit{otherwise}. \end{cases}$$

The total variation is a proper distance, unlike the Kullback–Leibler divergence. Both are always non-negative but the latter is not symmetric. If f_0 and f_1 are densities with respect to a measure dominating both $\mathbb{P}_0$ and $\mathbb{P}_1$ the Kullback–Leibler divergence can simply be written as

$$\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_1) = \mathbb{E}_0\left[\log \frac{f_0(X)}{f_1(X)}\right],$$

where X is a random variable with distribution given by $\mathbb{P}_0$. Therefore, this is the expected value of a log-likelihood ratio. Consider now the setting in this paper. As done in Tsybakov [34], the total variation is closely related to the infimum of the sum of type I and type II error probability, namely, for any binary (test) function $\hat{\Phi}$ we have

$$\mathbb{P}_\varnothing(\hat{\Phi} \neq 0) + \mathbb{P}_S(\hat{\Phi} \neq 1) \geq 1 - \mathrm{TV}(\mathbb{P}_\varnothing, \mathbb{P}_S).$$

Evaluating the total variation distance is generally difficult, but using Lemma 2.6 of Tsybakov [34] we can relate it to the Kullback–Leibler divergence, which is generally much easier to evaluate. Namely

$$1 - \mathrm{TV}(\mathbb{P}_\varnothing, \mathbb{P}_S) \geq \tfrac{1}{2} \exp\big(-\mathrm{KL}(\mathbb{P}_\varnothing \| \mathbb{P}_S)\big).$$

Putting these two results together we obtain a simple relation between the Kullback–Leibler divergence and the probabilities of error,

$$\mathrm{KL}(\mathbb{P}_\varnothing \| \mathbb{P}_S) \geq -\log\big(2\mathbb{P}_\varnothing(\hat{\Phi} \neq \varnothing) + 2\mathbb{P}_S(\hat{\Phi} \neq 1)\big). \tag{3.8}$$

To simplify the notation, let $\mathrm{LR}_{S,S'} \equiv \mathrm{LR}_{S,S'}(D)$. From equation (3.8) we conclude that

$$\mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}] = \mathrm{KL}(\mathbb{P}_\varnothing \| \mathbb{P}_S) \geq -\log\big(2\mathbb{P}_\varnothing(\hat{\Phi} \neq 0) + 2\mathbb{P}_S(\hat{\Phi} \neq 1)\big).$$

Since the choice of set S was completely arbitrary, we have the bound

$$\min_{S \in \mathcal{C}} \mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}] \geq \min_{S \in \mathcal{C}} \big\{-\log\big(2\mathbb{P}_\varnothing(\hat{\Phi} \neq 0) + 2\mathbb{P}_S(\hat{\Phi} \neq 1)\big)\big\}. \tag{3.9}$$

At this point it is important to note that, if we desire to have $R(\hat{\Phi}) \leq \varepsilon$ for some $0 < \varepsilon < 1$ then $\mathbb{P}_\varnothing(\hat{\Phi} \neq 0) + \mathbb{P}_S(\hat{\Phi} \neq 1) \leq \varepsilon$ (for any $S \in \mathcal{C}$), and therefore

$$\min_{S \in \mathcal{C}} \mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}] \geq \log\left(\frac{1}{2\varepsilon}\right). \tag{3.10}$$

The next step of the proof entails deriving a good upper bound on $\min_{S\in\mathcal{C}}\mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}]$ and comparing it to the lower bound just shown.

As noted before, the expected likelihood ratio is actually the Kullback–Leibler divergence between $\mathbb{P}_\varnothing$ and $\mathbb{P}_S$. This obviously depends on the sensing strategy $\mathcal{A}$ that is used. Therefore, we need to get an upper bound on

$$\sup_{\mathcal{A}}\min_{S\in\mathcal{C}}\mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}]. \tag{3.11}$$

It is instructive to compare the above expression with the one of the minimax error (3.2). Note that the roles of the max/sup and min/inf are reversed. This should not come as a surprise as larger values of $\mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}]$ correspond to lower probabilities of error. Note also that $\mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}]$ can be interpreted as the payoff matrix of a game where the sensing strategy makes the first move, and nature is the opponent that chooses a sparsity pattern in an adversarial way. Now note that

$$\begin{aligned}
\mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}] &= \sum_{k=1}^{\infty}\mathbb{E}_\varnothing\left[\log\frac{f_{Y_k|A_k,\Gamma_k}(Y_k|A_k,\Gamma_k;\varnothing)}{f_{Y_k|A_k,\Gamma_k}(Y_k|A_k,\Gamma_k;S)}\right]\\
&= \sum_{k=1}^{\infty}\mathbb{E}_\varnothing\left[\mathbb{E}_\varnothing\left[\log\frac{f_{Y_k|A_k,\Gamma_k}(Y_k|A_k,\Gamma_k;\varnothing)}{f_{Y_k|A_k,\Gamma_k}(Y_k|A_k,\Gamma_k;S)}\middle|A_k,\Gamma_k\right]\right]\\
&= \sum_{k=1}^{\infty}\mathbb{E}_\varnothing\left[\frac{\mu^2\mathbf{1}\{A_k\in S\}}{2}\Gamma_k^2\right]\\
&= \frac{\mu^2}{2}\sum_{k=1}^{\infty}\mathbb{E}_\varnothing\left[\mathbf{1}\{A_k\in S\}\Gamma_k^2\right],
\end{aligned}$$

where the final steps rely simply on the Kullback–Leibler divergence between normal random variables with the same variance and different means. At this point, we need to evaluate

$$\sup_{\mathcal{A}}\min_{S\in\mathcal{C}}\left\{\frac{\mu^2}{2}\sum_{k=1}^{\infty}\mathbb{E}_\varnothing\left[\mathbf{1}\{A_k\in S\}\Gamma_k^2\right]\right\}.$$

We need to solve the above optimization problem over the space of all possible sensing strategies. Although this might seem rather involved, this optimization can be reduced to a much simpler deterministic optimization problem. Begin by defining

$$b_i = \sum_{k=1}^{\infty}\mathbb{E}_\varnothing\left[\mathbf{1}\{A_k = i\}\Gamma_k^2\right]. \tag{3.12}$$

Note that this definition does not depend on S, as the expectation is taken under the null hypothesis. Furthermore $b_i \geq 0$, and the sensing budget equation in the observation model (2.2) can be written as $\sum_{i=1}^{n} b_i \leq m$. Therefore,

$$\begin{aligned}
&\sup_{\mathcal{A}} \min_{S\in\mathcal{C}} \left\{ \frac{\mu^2}{2} \sum_{k=1}^{\infty} \mathbb{E}_{\varnothing}\big[\mathbf{1}\{A_k \in S\}\Gamma_k^2\big] \right\} \\
&= \frac{\mu^2}{2} \sup_{\mathcal{A}} \min_{S\in\mathcal{C}} \left\{ \sum_{k=1}^{\infty} \sum_{i\in S} \mathbb{E}_{\varnothing}\big[\mathbf{1}\{A_k = i\}\Gamma_k^2\big] \right\} \\
&= \frac{\mu^2}{2} \sup_{\mathcal{A}} \min_{S\in\mathcal{C}} \left\{ \sum_{i\in S} \sum_{k=1}^{\infty} \mathbb{E}_{\varnothing}\big[\mathbf{1}\{A_k = i\}\Gamma_k^2\big] \right\} \\
&= \frac{\mu^2}{2} \sup_{\mathbf{b}\in\mathbb{R}_0^+ : \sum_{i=1}^{n} b_i \leq m} \min_{S\in\mathcal{C}} \sum_{i\in S} b_i .
\end{aligned}$$

We have now a relatively simple finite dimensional problem, where we seek to identify the vector $\mathbf{b} = (b_1, \ldots, b_n))$ maximizing a concave function. The solution of this problem obviously depends on the exact structure of $\mathcal{C}$. Remarkably, for symmetric classes, the solution is extremely simple and characterized in the first part of the following lemma, proved in the Appendix.

Lemma 3.1. *Let $\mathcal{C}$ be a symmetric class. Let $\Xi = \bigcup_{S\in\mathcal{C}} S$. Then*

1. $$\sup_{\mathbf{b}\in\mathbb{R}_0^+ : \sum_{i=1}^{n} b_i = m} \min_{S\in\mathcal{C}} \sum_{i\in S} b_i = \frac{ms}{|\Xi|},$$

2. $$\sup_{\mathbf{b}\in\mathbb{R}_0^+ : \sum_{i=1}^{n} b_i = m} \frac{1}{|\mathcal{C}|} \sum_{S\in\mathcal{C}} \sum_{i\in S} b_i = \frac{ms}{|\Xi|},$$

and in both cases the solution is attained taking $b_i = m/|\Xi|$ for $i \in \Xi$ and zero otherwise.

We are now in place to prove the theorem: by putting together the likelihood ratio lower bound (3.10) and the above upper bound we get

$$\frac{\mu^2 ms}{2|\Xi|} \geq \log \frac{1}{2\varepsilon},$$

which is equivalent to

$$\mu \geq \sqrt{\frac{2|\Xi|}{sm} \log \frac{1}{2\varepsilon}}$$

concluding the proof. □

Lower bounds for adaptive sensing in settings other than the one in this paper have been derived previously. For instance, in Castro and Nowak [10] a minimax characterization of the fundamental performance limits of active learning for a binary classification problem was provided. Such results were made possible by bringing together approximation results for smooth functional spaces and classical minimax bounding techniques (as in Tsybakov [34]), modified to incorporate the sequential experimental design aspect of the problem. In that approach the functional approximation results played the prominent role, and the stochastic part of the error had a much smaller contribution. Unfortunately this is not the case for the setting considered in the current paper and previously existing approaches were not adequate, prompting the novel approach presented here.

The proof of this theorem can be adapted for the other two risk definitions (3.3) and (3.4), and we can show that the risk behavior is qualitatively the same. These results are stated in the following proposition, proved in the Appendix.

Proposition 3.1. *Consider the setting of Theorem 3.1 and let $0 < \varepsilon < 1$. If $\tilde{R}(\hat{\Phi}) \leq \varepsilon/2$ or $\bar{R}(\hat{\Phi}) \leq \varepsilon$ then the conclusion of Theorem 3.1 is still valid and the lower bound (3.5) holds.*

3.2. Tightness of the detection lower bounds

We now proceed to show that the lower bounds derived above are indeed tight, in the sense that there are adaptive sensing testing procedures which are able to nearly attain them. As we saw, for symmetric classes $\mathcal{C}$, extra class structure does not help. Therefore, we focus exclusively on the largest class of all the subsets of $\{1, \dots, n\}$ with cardinality s. In Haupt, Castro and Nowak [24], a procedure called Distilled Sensing (DS) was introduced, and the authors proved that for the detection problem described above this procedure is able to asymptotically drive the risk to zero when $\mu > 4\sqrt{n/m}$ and $\log\log\log n < s < n^{1-\beta}$ for some $\beta \in (0,1)$. When comparing this result to the above lower bound we see that there is a huge gap, as we would expect the signal magnitude μ to scale essentially like $\sqrt{2n/(sm)}$. However, it is important to note that DS is entirely agnostic about the sparsity level and possible signal magnitude. An alternative non-agnostic methodology can be derived using DS as a black-box, which nearly achieves the lower-bounds of the previous section.

We begin by formally stating the performance results for the DS procedure. The following proposition is essentially the second part of Theorem III.1 in Haupt, Castro and Nowak [24].

Proposition 3.2 (From Haupt, Castro and Nowak [24][2]). *Assume $\log\log\log n < s \leq n^{1-\beta}$, for some $\beta \in (0,1)$. Furthermore let $\mu > 4\sqrt{n/m}$. There is a sensing strategy $\mathcal{A}_{\mathrm{DS}}$ and a test function $\hat{\Phi}_{\mathrm{DS}}$ such that*

$$R(\hat{\Phi}_{\mathrm{DS}}) \to 0,$$

as $n \to \infty$.

[2] The sparsity lower bound condition $\log\log\log n < s$ is not stated in the theorem in Haupt, Castro and Nowak [24] for presentation reasons, and the discussion on the validity of the result for $\log\log\log n < s$ appears only on the last paragraph of Section VI.

Note that this result is valid even if $s \approx \log\log\log n$, meaning s is nearly asymptotically constant. This suggests the following modification: first randomly select $\tilde{n}$ elements of $\{1, \ldots, n\}$ without replacement. Denote these by $\mathcal{E} = \{E_1, \ldots, E_{\tilde{n}}\}$. Our sensing strategy will focus exclusively on the entries $\mathcal{E}$ and ignore all the remaining ones. In other words, our observation model is now

$$Y_k = x_{E_{A_k}} + \Gamma_k^{-1} W_k \qquad \forall k \in \{1, 2, \ldots\},$$

where $A_k \in \{1, \ldots, \tilde{n}\}$. The sensing budget is, however, the same as in the original formulation

$$\sum_{k=1}^{\infty} \Gamma_k^2 \leq m.$$

In summary, we have exactly the same setting as before, but the extrinsic dimension n is now replaced by the smaller $\tilde{n}$. Now, provided we choose $\tilde{n}$ large enough so that the conditions of Proposition 3.2 are met for this new setting then an improvement in performance is possible, yielding the following result.

Proposition 3.3. *Assume $s > \log\log\log n$. Furthermore, let $\mu > \sqrt{\frac{32n \log\log\log n}{sm}}$. There is an adaptive sensing testing strategy such that*

$$R(\hat{\Phi}) \to 0,$$

as $n \to \infty$.

This result means that the statement of Corollary 3.1 is essentially tight, at least provided there are more than $\log\log\log n$ signal components under the alternative hypothesis. The constant in the bound is certainly not optimal, and the factor $\log\log\log n$ is (possibly) an artifact of the procedure. Closing the small gap between the upper and lower bounds is, however, still a direction for future research.

Remark 3.2. The results above were derived assuming the non-zero signal components are positive. Qualitatively these results remain the same even if one allows both positive and negative components. A simple way to address this setting is to write $\mathbf{x}$ as $\mathbf{x} = \mathbf{x}^+ - \mathbf{x}^-$, where $\mathbf{x}^+$ and $\mathbf{x}^-$ are sparse signal vectors with positive components (and the joint number of non-zero components is simply s). Now we can split the sensing budget into two equal parts, and make use of each one to test for the presence/absence or either signal. This approach yields the same asymptotic behavior, and will at most result in larger constants in the bounds.

Also note that, in principle, a procedure in the spirit of the one introduced in Chernoff [11] could be used to construct an adaptive sensing and testing methodology. However, the method of analysis in that paper is not adequate to deal with our setting. Nevertheless such procedure seems to work extremely well based on a short simulation study we conducted, and its analytical characterization presents an interesting direction for future work.

Proof of Proposition 3.3. The idea is simply to use the construction above, with $\tilde{n} = \frac{2n\log\log\log n}{s}$. Because of the random entry selection step (the choice of $\mathcal{E}$) the conditions of Proposition 3.2 might not always be satisfied. However, this happens with very low probability. Define $\tilde{x} \in \mathbb{R}^{\tilde{n}}$ where $\tilde{x}_i = x(E_i), i = 1, \ldots, \tilde{n}$. Suppose x has s non-zero components, and let $\tilde{s}$ be the number of non-zero components of $\tilde{x}$. Because of the sampling without replacement process, $\tilde{s}$ is an hypergeometric random variable with mean

$$\mathbb{E}[\tilde{s}] = \tilde{n}\frac{s}{n} = 2\log\log\log n,$$

and variance

$$\mathbb{V}(\tilde{s}) = \tilde{n}\frac{s}{n}\left(1 - \frac{s}{n}\right)\frac{n-\tilde{n}}{n-1} \leq \tilde{n}\frac{s}{n} = 2\log\log\log n.$$

This means that

$$\begin{aligned}\mathbb{P}(\tilde{s} < \log\log\log n) &= \mathbb{P}\bigl(\tilde{s} - \mathbb{E}[\tilde{s}] < \log\log\log n - \mathbb{E}[\tilde{s}]\bigr)\\ &= \mathbb{P}\bigl(\tilde{s} - \mathbb{E}[\tilde{s}] < -\log\log\log n\bigr)\\ &\leq \mathbb{P}\bigl(\bigl|\tilde{s} - \mathbb{E}[\tilde{s}]\bigr| > \log\log\log n\bigr)\\ &\leq \frac{\mathbb{V}(\tilde{s})}{(\log\log\log n)^2}\\ &\leq \frac{2}{\log\log\log n},\end{aligned}$$

where we used Chebyshev's inequality on the second-to-last step. This means that, with probability at least $1 - 2/\log\log\log n$ the conditions of Proposition 3.2 are fulfilled. For convenience, define the event $\Omega = \{\tilde{s} \geq \log\log\log n\}$. Since the detection risk is always bounded by 2, we have

$$R(\hat{\Phi}) \leq 2\frac{2}{\log\log\log n} + R(\hat{\Phi}|\Omega),$$

therefore it suffices to show that, conditionally on Ω, the risk of our procedure vanishes asymptotically. From Proposition 3.2, we know that if $\mu > 4\sqrt{\tilde{n}/m}$ the detection risk converges to zero, which immediately yields

$$\mu > 4\sqrt{\frac{2n\log\log\log n}{sm}},$$

concluding the proof. □

4. Signal estimation

In this section we consider the signal estimation problem, where the goal is to identify the support S of the underlying signal $\mathbf{x}$ as accurately as possible. As in the detection case, we are interested in characterizing the minimum signal amplitude $x_{\min}$ for which estimation is still possible.

Clearly estimation is statistically more "difficult" than signal detection, and therefore the requirements on $x_{\min}$ are more stringent in this case. Nevertheless we show that the dependence on the extrinsic dimension n does not play a significant role in the asymptotic performance bounds.

For the same reasons as in the previous section, we focus our attention on the signal model in (3.1). Our main goal is the estimation of the signal support set $S = \{i : x_i \neq 0\}$. In other words, our goal is to use adaptive sensing observations to construct an estimate $\hat{S}$ which is "close" to S. The metric of interest is the cardinality of the symmetric set difference

$$d(\hat{S}, S) = |\hat{S} \Delta S| = \left|(\hat{S} \cap S^c) \cup (\hat{S}^c \cap S)\right|,$$

where S^c denotes the complement of S in $\{1, \dots, n\}$. Clearly $d(\hat{S}, S)$ is just the number of errors in the estimate $\hat{S}$. In a similar spirit to that of the previous section, we want to determine how small can the signal magnitude μ be so that

$$\max_{S \in \mathcal{C}} \mathbb{E}_S\left[d(\hat{S}, S)\right] \leq \varepsilon, \tag{4.1}$$

where $\mathcal{C}$ is a class of sets, and $\varepsilon > 0$ is small. A different error metric which is also popular in the literature is $\mathbb{P}_S(\hat{S} \neq S)$, that is, the probability one does not achieve exact support estimation. Clearly

$$\mathbb{P}_S(\hat{S} \neq S) \leq \mathbb{E}_S\left[d(\hat{S}, S)\right],$$

and therefore this is a less stringent metric. The tools developed in this paper pertain $\mathbb{E}_S[d(\hat{S}, S)]$ and it is not clear if adaptive sensing lower bounds for $\mathbb{P}_S(\hat{S} \neq S)$ can be derived easily using a similar approach.

In addition, we will also consider a different support estimation risk function. Define the *False Discovery Rate* (FDR) and the *Non-Discovery Rate* (NDR) as

$$\mathrm{FDR}(\hat{S}, S) = \mathbb{E}_S\left[\frac{|\hat{S} \setminus S|}{|\hat{S}|}\right] \quad \text{and} \quad \mathrm{NDR}(\hat{S}, S) = \mathbb{E}_S\left[\frac{|S \setminus \hat{S}|}{|S|}\right].$$

In the above definitions convention $0/0 = 0$. Ideally we want both these quantities to be as small as possible, and so we can naturally define the risk

$$R_{\mathrm{FDR}+\mathrm{NDR}}(\hat{S}, S) = \max_{S \in \mathcal{C}}\left\{\mathrm{FDR}(\hat{S}, S) + \mathrm{NDR}(\hat{S}, S)\right\}.$$

Obviously $\mathbb{E}_S[d(\hat{S}, S)] \geq \mathrm{FDR}(\hat{S}, S) + \mathrm{NDR}(\hat{S}, S)$ and these two measures of error can be dramatically different, therefore controlling the risk $R_{\mathrm{FDR}+\mathrm{NDR}}(\hat{S}, S)$ is significantly easier than controlling the absolute number of errors.

Our original goal is to study lower bounds for the class $\mathcal{C}$ of all subsets of $\{1, \dots, n\}$ with cardinality s. For technical reasons this is a bit challenging, and to greatly simplify the analysis we consider a different setting that nonetheless captures the essence of the problem. Let $\mathcal{C}'$ denote the class consisting of sets of cardinality s, $s + 1$ and $s - 1$. This class is only "slightly" bigger than $\mathcal{C}$. We instead consider procedures that exhibit good performance when $S \in \mathcal{C}'$, that is, estimation procedures that are "very mildly" adaptive to unknown sparsity. Generalization of the results to other classes of sets shall be considered in future work and is out of the scope of this paper.

To aid in the presentation, we introduce some new notation. Namely let $S_i = \mathbf{1}\{i \in S\}$. Similarly, for any estimator $\hat{S}$ let $\hat{S}_i = \mathbf{1}\{i \in \hat{S}\}$. Note that the joint description of $\hat{S}_i$ for all i is equivalent to $\hat{S}$. For analysis purposes, it is convenient to consider only *symmetric* procedures, meaning that for any $S \in \mathcal{C}'$

$$\forall i, j \in S \qquad \mathbb{P}_S(\hat{S}_i \neq 1) = \mathbb{P}_S(\hat{S}_j \neq 1) \tag{4.2}$$

and

$$\forall i, j \notin S \qquad \mathbb{P}_S(\hat{S}_i \neq 0) = \mathbb{P}_S(\hat{S}_j \neq 0). \tag{4.3}$$

Although this might seem overly restrictive, it is indeed not the case. Any inference procedure can be "symmetrized" without increasing its maximal risk. In other words, given an estimator $\hat{S}$ we can construct another estimator $\hat{S}^{(\mathrm{perm})}$ satisfying (4.2) and (4.3) and such that

$$\mathbb{E}_S\bigl[d\bigl(\hat{S}^{(\mathrm{perm})}, S\bigr)\bigr] \le \max_{S' \in \mathcal{C}'} \mathbb{E}_{S'}\bigl[d\bigl(\hat{S}, S'\bigr)\bigr],$$

for all sets $S \in \mathcal{C}'$. The symmetrization is achieved by randomization. Let $\mathrm{perm} : \{1, \dots, n\} \to \{1, \dots, n\}$ be a permutation of $\{1, \dots, n\}$ chosen uniformly at random among the set of $n!$ possible permutations. Let $\hat{S}$ be a particular estimator we are going to symmetrize. Proceed by exchanging the identity of the entries of $\mathbf{x}$ using this permutation, or equivalently by taking $A_k^{(\mathrm{perm})} = A_{\mathrm{perm}^{-1}(k)}$ for all k, and use the estimator $\hat{S}$ on the collected data. Finally, reverse the permutation, namely defining $\hat{S}_i^{(\mathrm{perm})} = \hat{S}_{\mathrm{perm}(i)}$, for all $i \in \{1, \dots, n\}$. Using this construction, we get the following lemma, proved in the Appendix.

Lemma 4.1. *Let $\hat{S}$ be any adaptive sensing procedure. The random symmetrization approach described in the paragraph above yields another adaptive sensing procedure $\hat{S}^{(\mathrm{perm})}$ such that, for any $S \in \mathcal{C}'$*

$$\forall i \in S \qquad \mathbb{P}\bigl(\hat{S}_i^{(\mathrm{perm})} \neq 1\bigr) = \frac{1}{|S|\binom{n}{|S|}} \sum_{S' \in \mathcal{C}' : |S'| = |S|} \sum_{j \in S'} \mathbb{P}_{S'}(\hat{S}_j \neq 1)$$

and

$$\forall i \notin S \qquad \mathbb{P}\bigl(\hat{S}_i^{(\mathrm{perm})} \neq 0\bigr) = \frac{1}{(n - |S|)\binom{n}{|S|}} \sum_{S' \in \mathcal{C} : |S'| = |S|} \sum_{j \notin S'} \mathbb{P}_{S'}(\hat{S}_j \neq 0).$$

In addition, the following is also true:

$$\mathbb{E}_S\bigl[d\bigl(\hat{S}^{(\mathrm{perm})}, S\bigr)\bigr] \le \frac{1}{\binom{n}{|S|}} \sum_{S' \in \mathcal{C} : |S'| = |S|} \mathbb{E}_{S'}\bigl[d\bigl(\hat{S}, S'\bigr)\bigr] \le \max_{S' \in \mathcal{C}' : |S'| = |S|} \mathbb{E}_{S'}\bigl[d\bigl(\hat{S}, S'\bigr)\bigr].$$

This ensures that without loss of generality we can consider only symmetric procedures. It is important to note that this approach is valid only if the class $\mathcal{C}'$ is invariant under permutations. Finally, for symmetric procedures the lower bounds we derive are also applicable to measures of risk different than (4.1), such as the *average estimation risk* $\frac{1}{|\mathcal{C}'|}\sum_{S'\in\mathcal{C}'}\mathbb{E}_{S'}[d(\hat{S},S')]$.

4.1. Main results – Estimation

Theorem 4.1. *Let $\mathcal{C}'$ denote the class of all subsets of $\{1,\dots,n\}$ with cardinality s, $s+1$ and $s-1$. Let $\hat{S}\equiv\hat{S}(D)$ be an arbitrary adaptive sensing estimator, where $D=\{Y_i,A_i,\Gamma_i\}_{i=1}^{\infty}$. If*

$$\max_{S\in\mathcal{C}'}\mathbb{E}_S\bigl[d(\hat{S},S)\bigr]\le\varepsilon,$$

where $0<\varepsilon<1$ then necessarily

$$x_{\min}\ge\sqrt{\frac{2n}{m}\left(\log s+\log\frac{n-s}{n+1}+\log\frac{1}{2\varepsilon}\right)}.$$

The proof of the theorem is presented at the end of this section. As before it is useful to look at the asymptotic behavior, and the case $s\ll n$ is particularly interesting.

Corollary 4.1. *Consider the setting of Theorem 4.1 and assume $s=\mathrm{o}(n)$ as $n\to\infty$. Let $\hat{S}_n$ be an arbitrary estimation procedure for which*

$$\lim_{n\to\infty}\max_{S\in\mathcal{C}'}\mathbb{E}_S\bigl[d(\hat{S}_n,S)\bigr]=0.$$

Necessarily

$$x_{\min}\ge\sqrt{2\frac{n}{m}(\log s+\omega_n)},$$

where ω_n is a sequence for which $\lim_{n\to\infty}\omega_n=\infty$.

For the FDR + NDR risk, we can use the same proof approach to obtain a much less restrictive bound on the signal magnitude.

Corollary 4.2. *Consider the setting of Theorem 4.1 and assume $s=\mathrm{o}(n)$. Let $\hat{S}_n$ be an arbitrary estimation procedure such that*

$$\lim_{n\to\infty}R_{\mathrm{FDR+NDR}}(\hat{S},S)=0.$$

Necessarily

$$x_{\min}\ge\omega_n\sqrt{\frac{n}{m}},$$

where ω_n is a sequence for which $\lim_{n\to\infty}\omega_n=\infty$.

A sketch of the proof of this corollary can be found in the Appendix.

Proof of Theorem 4.1. The proof follows a similar approach to that of Theorem 3.1, and capitalizes heavily on the symmetry of the estimation procedure. In light of Lemma 4.1, it suffices to consider symmetric procedures, that is, procedures that satisfy (4.2) and (4.3). Let $S \in \mathcal{C}'$ be arbitrary and assume that

$$\mathbb{E}_S\big[d(\hat{S}, S)\big] \le \varepsilon,$$

where $0 < \varepsilon < 1$. Clearly

$$\begin{aligned}
\mathbb{E}_S\big[d(\hat{S}, S)\big] &= \mathbb{E}_S\left[\sum_{i=1}^n \mathbf{1}\{\hat{S}_i \neq S_i\}\right] \\
&= \sum_{i\in S} \mathbb{E}_S\big[\mathbf{1}\{\hat{S}_i \neq 1\}\big] + \sum_{j\notin S} \mathbb{E}_S\big[\mathbf{1}\{\hat{S}_j \neq 0\}\big] \\
&= \sum_{i\in S} \mathbb{P}_S(\hat{S}_i \neq 1) + \sum_{j\notin S} \mathbb{P}_S(\hat{S}_j \neq 0).
\end{aligned}$$

As we consider symmetric procedures, we conclude that

$$\forall i \in S \qquad \mathbb{P}_S(\hat{S}_i \neq 1) \le \frac{\varepsilon}{|S|}$$

and

$$\forall i \notin S \qquad \mathbb{P}_S(\hat{S}_i \neq 0) \le \frac{\varepsilon}{n - |S|}.$$

For our purposes, it is convenient to rewrite the likelihood ratio (3.6) as

$$\begin{aligned}
\mathrm{LR}_{S,S'}(d) &= \frac{f(d; S)}{f(d; S')} \\
&= \prod_{i=1}^n \prod_{k:a_k=i} \frac{f_{Y_k|A_k,\Gamma_k}(y_k|a_k, \gamma_k; S)}{f_{Y_k|A_k,\Gamma_k}(y_k|a_k, \gamma_k; S')}.
\end{aligned}$$

Now let $S \in \mathcal{C}$ be an arbitrary set of cardinality s, and define $S^{(i)} \in \mathcal{C}'$ to be

$$S^{(i)} = \begin{cases} S \setminus \{i\} & \text{if } i \in S, \\ S \cup \{i\} & \text{if } i \notin S, \end{cases}$$

in words, we either remove element i if $i \in S$, or add it otherwise, meaning that $S \Delta S^{(i)} = \{i\}$. We proceed in a similar way as we did in the signal detection scenario. Let $i \in \{1, \dots, n\}$ be arbitrary. We conclude that

$$\forall i \in S \qquad \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] \geq -\log\bigl(2\mathbb{P}_S(\hat{S}_i \neq 1) + 2\mathbb{P}_{S^{(i)}}(\hat{S}_i \neq 0)\bigr)$$

and

$$\forall i \notin S \qquad \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] \geq -\log\bigl(2\mathbb{P}_S(\hat{S}_i \neq 0) + 2\mathbb{P}_{S^{(i)}}(\hat{S}_i \neq 1)\bigr).$$

We now take advantage of the symmetry of the estimator, to conclude that

$$\forall i \in S \qquad \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] \geq -\log\left(\frac{2\varepsilon}{s} + \frac{2\varepsilon}{n-s+1}\right) \tag{4.4}$$

and

$$\forall i \notin S \qquad \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] \geq -\log\left(\frac{2\varepsilon}{n-s} + \frac{2\varepsilon}{s+1}\right). \tag{4.5}$$

Now that we have lower bounds for $\mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}]$ we need to evaluate this quantity in terms of μ. This is easily done by noting that for $i \in \{1, \dots, n\}$

$$\begin{aligned}\mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] &= \mathbb{E}_S\left[\sum_{k:A_k=i} \log \frac{f_{Y_k|A_k,\Gamma_k}(Y_k|A_k,\Gamma_k;S)}{f_{Y_k|A_k,\Gamma_k}(Y_k|A_k,\Gamma_k;S^{(i)})}\right] \\ &= \mathbb{E}_S\left[\sum_{k:A_k=i} \frac{\mu^2}{2}\Gamma_k^2\right].\end{aligned}$$

Note that we cannot yet evaluate the above expression, as one cannot invoke the sensing budget constraint (2.2). This can be addressed by summing each of the above terms over $i \in \{1, \dots, n\}$. On one hand

$$\sum_{i=1}^n \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] = \mathbb{E}_S\left[\sum_{i=1}^n \sum_{k:A_k=i} \frac{\mu^2}{2}\Gamma_k^2\right] = \mathbb{E}_S\left[\sum_{k=1}^{\infty} \frac{\mu^2}{2}\Gamma_k^2\right] \leq \frac{m\mu^2}{2}. \tag{4.6}$$

On the other hand

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] &= \sum_{i\in S} \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] + \sum_{i\notin S} \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] \\ &\geq -s\log\left(\frac{2\varepsilon}{s} + \frac{2\varepsilon}{n-s+1}\right) - (n-s)\log\left(\frac{2\varepsilon}{n-s} + \frac{2\varepsilon}{s+1}\right).\end{aligned}$$

We can get a more insightful bound by reorganizing the various terms

$$\begin{aligned}&\sum_{i=1}^n \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] \\ &\quad \geq n\log\frac{1}{2\varepsilon} + s\log\frac{s(n-s+1)}{n+1} + (n-s)\log\frac{(n-s)(s+1)}{n+1}\end{aligned}$$

$$
\begin{aligned}
&= n\log\frac{1}{2\varepsilon} + s\log s + (n-s)\log(s+1) + s\log\frac{n-s+1}{n+1} + (n-s)\log\frac{n-s}{n+1}\\
&\geq n\left(\log s + \log\frac{n-s}{n+1} + \log\frac{1}{2\varepsilon}\right),
\end{aligned}
$$

where the last inequality follows by noting that $\log(s+1) > \log s$ and $\log(n-s+1) > \log(n-s)$. Using this together with (4.6) concludes the proof. □

4.2. Tightness of the estimation lower bounds

Similarly to what happened in the detection setting the lower bounds derived for estimation are also tight, in the sense that there are inference procedures able to achieve them. In Malloy and Nowak [30,31], a slightly different problem was considered, where each measurement had the same accuracy/precision and one desired to control the total number of errors in $\hat{S}$. Their results were stated in term of conditions on the signal magnitude μ that were necessary to ensure the risk converged to zero. In their setting, there is no strict sensing budget, but instead only control over the expect precision budget used. In other words, the procedures in Malloy and Nowak [30,31] do not always satisfy the sensing budget in equation (2.2), but instead satisfy an *expected* sensing budget constraint

$$\mathbb{E}\left[\sum_{k=1}^{\infty}\Gamma_k^2\right] \leq m.$$

Such methods can be modified to ensure that the sensing budget (2.2) is fulfilled with increasingly high probability (as n grows) without altering their asymptotic performance behavior, and we can state the following result, proved in the Appendix.

Proposition 4.1. *Assume* $s \leq \frac{n}{(\log_2^2 n)-3}$. *Let*

$$\mu \geq \sqrt{\frac{4n}{m}(2\log s + 5\log\log_2 n)}.$$

There is a sensing and estimation strategy yielding an estimator $\hat{S}$ *such that*

$$\max_{S\in\mathcal{C}'} \mathbb{E}_S\big[d(\hat{S}, S)\big] \to 0,$$

as $n \to \infty$.

This means that provided $x_{\min}$ is of the order $\sqrt{(n/m)(\log s + \log\log n)}$ we can ensure exact recovery of a sufficiently sparse signal support with probability approaching 1. The proposition is proved in the Appendix. The constants in the above result are rather loose, and can be made much tighter (see Malloy and Nowak [31]). The $\log\log n$ term is an artifact of this method (which is parameter adaptive and agnostic about s). This term can be entirely avoided by considering another procedure, namely by executing in parallel n properly calibrated sequential likelihood ratio tests, which requires the knowledge of the sparsity level s. Such a procedure achieves precisely the bound in Corollary 4.1. Lower bounds for estimation have been derived under a different set of assumptions for the class of entry-wise sequential tests in Malloy and Nowak [30]. In contrast, the results in the current paper pertain any adaptive sensing procedure (and not only entry-wise testing procedures).

Control of the FDR + NDR risk was considered in Haupt, Castro and Nowak [24] in the exact setting described in this paper, and the distilled sensing procedure proposed there is able to achieve the bound in Corollary 4.2 provided $\log\log\log n < s \leq n^{1-\beta}$ for some $0 < \beta < 1$. Therefore the lower bounds on the FDR + NDR risk are also tight for a wide range of sparsity levels.

4.3. Relation to compressed sensing

The proof technique used in Theorem 4.1 provides some important insights for the problem of adaptive compressive sensing. This setting is different than the one considered so far and the observation model is now of the form

$$\mathbf{Y} = \mathbf{A}\mathbf{x} + \mathbf{W},$$

where $\mathbf{Y} \in \mathbb{R}^l$ denotes the observations, $\mathbf{A} \in \mathbb{R}^{l\times n}$ is the design/sensing matrix, $\mathbf{x} \in \mathbb{R}^n$ is the unknown signal, and $\mathbf{W} \in \mathbb{R}^l$ is Gaussian with zero mean an identity covariance matrix. The rows of $\mathbf{A}$ can be designed sequentially, and the kth row (denoted by $\mathbf{A}_{k\cdot}$) can depend explicitly on $\{Y_j, \mathbf{A}_{j\cdot}\}_{j=1}^{k-1}$. Note that W_k is a normal random variable independent of $\{Y_j, \mathbf{A}_{j\cdot}, W_j\}_{j=1}^{k-1}$ and also independent of $\mathbf{A}_{k\cdot}$. This setting is particularly interesting when we impose some constraints on $\mathbf{A}$, namely

$$\mathbb{E}\big[\|\mathbf{A}\|_F^2\big] \leq m,$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. Like (2.2), this sensing budget condition is very natural and the issue of noise is irrelevant without it. Each row $\mathbf{A}_{k\cdot}$ plays the role of the sensing action A_k in our original scenario, and $\|\mathbf{A}_{k\cdot}\|_2^2$ plays the role of the precision parameter Γ_k^2 in (2.2). As before, we do not impose any restrictions on the total number of measurements l, which can be potentially infinite. We can show the following result using an approach similar to that of Theorem 4.1.

Proposition 4.2. *Consider the adaptive compressed sensing setting as described above, with observations* $\mathbf{Y} = \mathbf{A}\mathbf{x} + \mathbf{W}$, *where* $\mathbf{W}$ *is Gaussian zero mean with identity covariance matrix and* $\mathbb{E}[\|\mathbf{A}\|_F^2] \leq m$. *Let* $\mathcal{H}(\mu) \subset \mathbb{R}^n$ *be the class of all vectors* $\mathbf{x}$ *with support in* $\mathcal{C}'$ (*i.e. the support*[3]

[3] Define $\operatorname{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$.

has cardinality s, $s+1$ or $s-1$) and the magnitude of the minimum non-zero entries greater or equal than μ. That is

$$\mathcal{H}(\mu) = \Bigl\{\mathbf{x} \in \mathbb{R}^n : \operatorname{supp}(x) \in \mathcal{C}' \textit{ and } \min_i \bigl\{|x_i| : x_i \neq 0\bigr\} \geq \mu\Bigr\}.$$

Let $D = \{\mathbf{Y}, \mathbf{A}\}$ and $\hat{S}(D)$ be an arbitrary estimator. If

$$\max_{\mathbf{x} \in \mathcal{H}\mu} \mathbb{E}_{\mathbf{x}}\bigl[d(\hat{S}, S)\bigr] \leq \varepsilon, \tag{4.7}$$

where $0 < \varepsilon < 1$ then necessarily

$$\mu \geq \sqrt{\frac{2n}{m}\left(\log s + \log\frac{n-s}{n+1} + \log\frac{1}{2\varepsilon}\right)}.$$

The proof of the proposition can be found in the Appendix. In Arias-Castro, Candès and Davenport [2], the authors derived lower bounds for both support recovery and mean square error risk for adaptive compressive sensing. In their setting $l = m$, and each row of the matrix $\mathbf{A}$ has expected norm at most 1. These two constraints imply the Frobenius norm constraint in Proposition 4.2. Theorem 2 in that paper states that the minimum signal amplitude $x_{\min}$ must be greater than $\sqrt{n/m}$ to ensure that support recovery is possible within the class of all possible s-sparse signals. In contrast, our result shows that the lower bound is not entirely tight. Formally, if $s = \mathrm{o}(n)$ and

$$\lim_{n \to \infty} \max_{S \in \mathcal{C}'} \mathbb{E}_S\bigl[d(\hat{S}_n, S)\bigr] = 0$$

we have necessarily

$$x_{\min} = \sqrt{2\frac{n}{m}(\log s + \omega_n)},$$

as $n \to \infty$. So, the above result improves the bound in Arias-Castro, Candès and Davenport [2] by a $\log s$ factor. In light of the recent results in Haupt *et al.* [23], it seems plausible that this is a necessary and sufficient term. However, a precise characterization of these limits remains an open problem.

5. Conclusion

In this paper, we presented several lower bounds for detection and estimation of sparse signals using adaptive sensing. These results bridge a gap in our understanding of adaptive sensing and show that methodologies recently proposed in the literature are nearly optimal. A very interesting insight is that, for signal detection, the sparsity structure is essentially irrelevant. The intuition being that for detection it suffices to identify one non-zero component, and cues provided by the structure are not too useful under adaptive sensing scenarios. However, for signal estimation it is not clear if structure helps, which raises many interesting directions for future research.

Appendix

Proof of Lemma 3.1. We begin by proving the first result. Let

$$b'_i = \begin{cases} m/|\Xi| & \text{if } i \in \Xi, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \dots, n.$$

Begin by noticing that

$$\sup_{\mathbf{b} \in \mathbb{R}_0^+ : \sum_{i=1}^n b_i = m} \min_{S \in \mathcal{C}} \sum_{i \in S} b_i \geq \min_{S \in \mathcal{C}} \sum_{i \in S} b'_i = \frac{ms}{|\Xi|}.$$

The proof proceeds by contradiction, and makes use of a probabilistic argument. Suppose there is a vector $\mathbf{b}^* \in \mathbb{R}_0^+$ such that $\sum_{i=1}^n b_i^* \leq m$ and

$$\min_{S \in \mathcal{C}} \sum_{i \in S} b_i^* > \frac{ms}{|\Xi|}. \tag{5.1}$$

We show next that this in contradiction with the symmetry assumption.

Let J be a uniform random variable with range Ξ. Then

$$\mathbb{E}\left[b_J^*\right] = \frac{1}{|\Xi|} \sum_{j \in \Xi} b_j^* \leq \frac{1}{|\Xi|} \sum_{j=1}^n b_j^* \leq \frac{m}{|\Xi|}. \tag{5.2}$$

Now construct another random variable K in a hierarchical fashion: first take S drawn uniformly over $\mathcal{C}$, and given S take K drawn uniformly over S. Then clearly

$$\begin{aligned} \mathbb{E}\left[b_K^*\right] &= \mathbb{E}\left[\mathbb{E}\left[b_K^* | S\right]\right] \\ &= \mathbb{E}\left[\frac{1}{s} \sum_{k \in S} b_k^*\right] \\ &\geq \mathbb{E}\left[\min_{S \in \mathcal{C}} \frac{1}{s} \sum_{k \in S} b_k^*\right] \\ &= \frac{1}{s} \mathbb{E}\left[\min_{S \in \mathcal{C}} \sum_{k \in S} b_k^*\right] \\ &> \frac{m}{|\Xi|}, \end{aligned} \tag{5.3}$$

where the strict inequality follows from (5.1). To conclude the proof, we just need to notice that J and K have exactly the same distribution if the class $\mathcal{C}$ is symmetric. Let $k \in \Xi$ be arbitrary. Then

$$\begin{aligned}\mathbb{P}(K = k) &= \mathbb{E}\big[\mathbf{1}\{K = k\}\big] \\ &= \mathbb{E}\big[\mathbb{E}\big[\mathbf{1}\{K = k\}|S\big]\big] \\ &= \mathbb{E}\left[\frac{1}{s}\mathbf{1}\{k \in S\}\right] \\ &= \frac{1}{s}\mathbb{P}(k \in S) \\ &= \frac{1}{s}\frac{s}{|\Xi|} = \frac{1}{|\Xi|}.\end{aligned}$$

Therefore, both J and K are uniformly distributed over Ξ and so $\mathbb{E}[b_J^*] = \mathbb{E}[b_K^*]$. This creates a contradiction between (5.2) and (5.3) invalidating the existence of vector $\mathbf{b}^*$, concluding the proof.

For the second result, note simply that

$$\begin{aligned}\frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}}\sum_{i\in S} b_i &= \frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}}\sum_{i=1}^{n} b_i\mathbf{1}\{i \in S\} \\ &= \sum_{i=1}^{n} b_i \frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}}\mathbf{1}\{i \in S\} \\ &= \sum_{i=1}^{n} b_i \frac{s}{|\Xi|},\end{aligned}$$

where the last step follows from the symmetry assumption. The result of the lemma is now immediate. □

Proof of Proposition 3.1. If $\tilde{R}(\hat{\Phi}) < \varepsilon/2$, the result follows immediately from the simple fact that $R(\hat{\Phi}) \le 2\tilde{R}(\hat{\Phi})$. Therefore, $\tilde{R}(\hat{\Phi}) < \varepsilon/2$ implies that $R(\hat{\Phi}) < \varepsilon$ and we just apply the result of the theorem. For the second statement, it is useful to look at S as a uniform random variable with range $\mathcal{C}$. In the proof of Theorem 3.1, we showed that, for any $S \in \mathcal{C}$

$$\mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}\,|S] \ge -\log\big(2\mathbb{P}_\varnothing(\hat{\Phi} \ne 0) + 2\mathbb{P}_1(\hat{\Phi} \ne 1|S)\big),$$

where $\mathbb{P}_1$ denotes the probability measure under the alternative hypothesis. By taking the expectation on both sides, we have

$$\mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}] \ge -\frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}}\log\big(2\mathbb{P}_\varnothing(\hat{\Phi} \ne 0) + 2\mathbb{P}_1(\hat{\Phi} \ne 1|S)\big).$$

To simplify the notation, let $p_0 \equiv \mathbb{P}_\varnothing(\hat{\Phi} \neq 0)$ and $p_S \equiv \mathbb{P}_1(\hat{\Phi} \neq 1|S)$. The statement $\bar{R}(\hat{\Phi}) \leq \varepsilon$ is equivalent to $p_0 + \frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}} p_S \leq \varepsilon$. Accordingly define the constraint set $\mathcal{P} \subseteq \mathbb{R}^{1+|\mathcal{C}|}$ as

$$\mathcal{P} = \left\{ p_0, \{p_S\}_{S\in\mathcal{C}} : p_0 + \frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}} p_S \leq \varepsilon \right\}.$$

We have that

$$\mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}] \geq \min_{\mathcal{P}} \left\{ -\frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}} \log(2p_0 + 2p_S) \right\} \tag{5.4}$$

$$= \log\frac{1}{2\varepsilon}, \tag{5.5}$$

where the last step follows from a straightforward Lagrange multiplier argument, to conclude that the minimum is attained by taking $p_0 + p_S = \varepsilon$ for all $S \in \mathcal{C}$.

The next step, similar to the proof of Theorem 3.1, is to solve $\sup_{\mathcal{A}} \mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}]$, where it is important to recall that S is random. Following the same approach as in the proof of the theorem yields

$$\sup_{\mathcal{A}} \mathbb{E}_\varnothing[\log \mathrm{LR}_{\varnothing,S}] = \frac{\mu^2}{2} \sup_{\mathbf{b}\in\mathbb{R}_0^+ : \sum_{i=1}^n b_i = n} \frac{1}{|\mathcal{C}|}\sum_{S\in\mathcal{C}}\sum_{i\in S} b_i,$$

where b_i is defined in (3.12). The second result of Lemma 3.1 characterizes the solution of this optimization problem, and therefore

$$\frac{\mu^2 m s}{2|\Xi|} \geq \log\frac{1}{2\varepsilon}.$$

Simple algebraic manipulation concludes the proof. □

Proof of Lemma 4.1. To ease the notation let $\mathcal{C}_s$ denote the class of all subsets of $\{1,\dots,n\}$ with cardinality s. Let $S \in \mathcal{C}_s$ and $i \in S$ be fixed, but arbitrary. Note that the permutation perm maps this set to another set $S^{(\mathrm{perm})} = \mathrm{perm}(S) \in \mathcal{C}_s$ with the same cardinality. Furthermore, since the permutation is chosen uniformly over the set of all permutations this set is uniformly distributed over $\mathcal{C}_s$, that is

$$S^{(\mathrm{perm})} \sim \mathrm{Unif}(\mathcal{C}_s).$$

In addition define the random variable $J = \mathrm{perm}(i)$. This is obviously uniformly distributed over $\{1,\dots,n\}$. More importantly, conditionally on $S^{(\mathrm{perm})}$, J is uniformly distributed over the set $S^{(\mathrm{perm})}$. In other words, for arbitrary $k \in \{1,\dots,n\}$

$$\begin{aligned}\mathbb{P}(J = k|S^{(\mathrm{perm})}) &= \mathbb{P}(\mathrm{perm}(i) = k|S^{(\mathrm{perm})}) \\ &= \mathbb{P}(\mathrm{perm}^{-1}(k) = i|S^{(\mathrm{perm})}) \\ &= \begin{cases} 1/s & \text{if } k \in S^{(\mathrm{perm})}, \\ 0 & \text{otherwise.}\end{cases}\end{aligned}$$

Therefore

$$\begin{aligned}\mathbb{P}\big(\hat{S}_i^{(\text{perm})} \neq 1\big) &= \mathbb{E}\big[\mathbf{1}\{\hat{S}_{\text{perm}(i)} \neq 1\}\big]\\ &= \mathbb{E}\big[\mathbb{E}\big[\mathbf{1}\{\hat{S}_{\text{perm}(i)} \neq 1\} | S^{(\text{perm})}\big]\big]\\ &= \mathbb{E}\Bigg[\frac{1}{s}\sum_{j\in S^{(\text{perm})}} \mathbb{P}_{S^{(\text{perm})}}(\hat{S}_j \neq 1)\Bigg]\\ &= \frac{1}{|\mathcal{C}_s|}\sum_{S'\in\mathcal{C}_s}\frac{1}{s}\sum_{j\in S'}\mathbb{P}_{S'}(\hat{S}_j \neq 1),\end{aligned}$$

where the two last steps follow from the distribution of $S^{(\text{perm})}$ and perm(i). The case $i \notin S$ is entirely analogous. Using these two results we obtain the first two statements of the lemma for the class $\mathcal{C}' = \mathcal{C}_{s-1} \cup \mathcal{C}_s \cup \mathcal{C}_{s+1}$. Finally, the last result in the lemma follows trivially from the other two statements. □

Sketch proof of Corollary 4.2. The result in the corollary follows in the same manner as the result in Theorem 4.1, but noticing that for symmetric estimation procedures the requirements on the estimator $\hat{S}_i$ for each $i \in \{1, \dots, n\}$ are much less stringent. In particular let $S \in \mathcal{C}'$ be arbitrary and assume that

$$R_{\text{FDR}+\text{NDR}}(\hat{S}, S) \leq \varepsilon,$$

where $\varepsilon > 0$, which implies that both FDR and NDR are less than ε. Now consider symmetric procedures and let $\alpha = \mathbb{P}_S(\hat{S}_i \neq 0)$ for $i \notin S$ and $\beta = \mathbb{P}_S(\hat{S}_i \neq 1)$ for $i \in S$. Clearly, the constraint in NDR implies that

$$\varepsilon \geq \text{NDR}(\hat{S}, S) = \mathbb{E}\left[\frac{|S \setminus \hat{S}|}{|S|}\right] = \frac{|S|\beta}{|S|} = \beta.$$

The constraint on FDR is a bit more difficult to analyze, due to the random denominator its definition. However, a very sloppy bound suffices, namely

$$\varepsilon \geq \text{FDR}(\hat{S}, S) = \mathbb{E}\left[\frac{|\hat{S} \setminus S|}{|\hat{S}|}\right] \geq \mathbb{E}\left[\frac{|\hat{S} \cap S^c|}{n}\right] = \frac{(n - |S|)\alpha}{n}.$$

Therefore, we conclude that $\alpha \leq \varepsilon$ suffices. Note that this is a very loose but nevertheless sufficient bound. The rest of the proof proceeds now in the same fashion as Theorem 4.1 and Corollary 4.1. □

Proof of Proposition 4.2. The proof of this result mimics closely the proof of Theorem 4.1, with the necessary changes to account for the different sensing model. The first step is to reduce the class of signals under consideration. Clearly signals of the form (3.1) are also in the class $\mathcal{H}(\mu)$. Therefore

$$\max_{\mathbf{x}\in\mathcal{H}\mu} \mathbb{E}_{\mathbf{x}}\big[d(\hat{S}, S)\big] \geq \max_{S\in\mathcal{C}'} \mathbb{E}_S\big[d(\hat{S}, S)\big],$$

where the expectation on the right-hand-side is taken assuming $\mathbf{x}$ is of the form (3.1) with support S. Condition (4.7) therefore implies that

$$\max_{S\in\mathcal{C}'} \mathbb{E}_S\big[d(\hat{S}, S)\big] \le \varepsilon,$$

so, for the purpose of computing a lower bound it suffices to consider on the signals where all the non-zero components are valued μ. It is important to note that this subclass of signals might not correspond to the "hardest" signals to estimate, and no claim is made about this. However, this subclass seems to capture the essential aspects of the problem in light of the bounds derived. As the class of signals under consideration is the same as in Theorem 4.1 the only change in that proof stems from the different observation model, which in turn results in a different log-likelihood ratio. Notice that, as before, we can consider only symmetric procedures in the sense of Lemma 4.1.

To aid in the presentation, let A_{ij} denote the entry in the ith row and jth column of the matrix $\mathbf{A}$, and let $\mathbf{A}_{i\cdot}$ and $\mathbf{A}_{\cdot j}$ denote respectively the ith row of and the jth column of $\mathbf{A}$. The log-likelihood ratio is therefore given by

$$\begin{aligned}\log \mathrm{LR}_{S,S'}(\mathbf{Y},\mathbf{A}) &= \log \frac{f(\mathbf{Y},\mathbf{A};S)}{f(\mathbf{Y},\mathbf{A};S')}\\ &= \sum_{k=1}^{\ell} \log \frac{f_{Y_k|\mathbf{A}_{k\cdot}}(Y_k|\mathbf{A}_{k\cdot};S)}{f_{Y_k|\mathbf{A}_{k\cdot}}(Y_k|\mathbf{A}_{k\cdot};S')}\\ &= \frac{1}{2}\sum_{k=1}^{\ell}\left[\left(Y_k-\mu\sum_{j\in S'}A_{kj}\right)^2-\left(Y_k-\mu\sum_{j\in S}A_{kj}\right)^2\right].\end{aligned}$$

Given this, the expected log-likelihood ratio can be computed quite easily as before, and we get

$$\mathbb{E}_S\big[\log \mathrm{LR}_{S,S'}(\mathbf{Y},\mathbf{A})\big] = \frac{\mu^2}{2}\sum_{k=1}^{\ell}\mathbb{E}_S\left[\left(\left(\sum_{j\in S}A_{kj}\right)-\left(\sum_{j\in S'}A_{kj}\right)\right)^2\right]. \tag{5.6}$$

Now consider the sets $S^{(i)}$ as in the proof of Theorem 4.1. Since we have $S\Delta S^{(i)}=\{i\}$, we get from equation (5.6)

$$\begin{aligned}\mathbb{E}_S\big[\log \mathrm{LR}_{S,S^{(i)}}(\mathbf{Y},\mathbf{A})\big] &= \frac{\mu^2}{2}\mathbb{E}\left[\sum_{k=1}^{\ell}A_{ki}^2\right]\\ &= \frac{\mu^2}{2}\mathbb{E}\big[\|\mathbf{A}_{\cdot i}\|_F^2\big].\end{aligned} \tag{5.7}$$

From this point on, the proof proceeds in exactly the same fashion as that of Theorem 4.1. Begin by summing the terms (5.7) over $i\in\{1,\dots,n\}$ to get an upper bound on the expected likelihood ratio

$$\sum_{i=1}^{n} \mathbb{E}_S[\log \mathrm{LR}_{S,S^{(i)}}] = \frac{\mu^2}{2}\mathbb{E}\left[\sum_{i=1}^{n} \|\mathbf{A}_{\cdot i}\|_F^2\right] = \frac{\mu^2}{2}\mathbb{E}\big[\|\mathbf{A}\|_F^2\big] \leq \frac{m\mu^2}{2}. \tag{5.8}$$

Finally, the lower bounds on the log-likelihood ratio in (4.4) and (4.5) are not dependent on the nature of the likelihood ratio itself, but rather on the desired risk performance. So these bounds are valid in the compressed sensing setting as well. As in the proof of Theorem 4.1, using these lower bounds together with (5.8) concludes the proof. □

Proof of Proposition 4.1. We begin by introducing an algorithm that achieves the desired performance bound. Algorithm 1 is described here for convenience of presentation and explained in detail in the next paragraphs. It is essentially the algorithm presented in Malloy and Nowak [30] for the case of Gaussian observation noise.

Algorithm 1: Simple distilled sensing.

```
Parameters: Number of steps $l$ and per-measurement precision $p$
Initialization:
    $k \leftarrow 0$, $i \leftarrow 1$, $\hat{S} \leftarrow \varnothing$
    $c_i \leftarrow 0$ for $i = 1, \ldots, n$
    $\Gamma_j^2 \leftarrow p$ for $j = 1, 2, \ldots$
for $i \leftarrow 1$ to $n$ do
    repeat
        $k \leftarrow k + 1$
        $c_i \leftarrow c_i + 1$
        Measure $Y_k \equiv Y_i^{(c_i)} = x_i + \Gamma_k^{-1} W_k$
        if $p(k+1) > m$ then
            Terminate: Output $\hat{S}$
        end
    until $c_i = l$ or $Y_k < 0$;
    if $c_i = l$ and $Y_k \geq 0$ then
        $\hat{S} \leftarrow \hat{S} \cup \{i\}$
    end
end
Terminate: Output $\hat{S}$
```

Sensing is performed coordinate-wise in a sequential way, until all the signal entries have been explored or the total sensing budget is exhausted. Note that all the measurements are made with the same precision p. For each signal entry i the algorithm performs at most l measurements. If any of these measurements is negative then entry i is deemed not to belong to the support estimate $\hat{S}$. If all the l measurements are non-negative, then entry i is deemed to belong to the support estimate. For convenience, we identify the measurements of entry i by $Y_i^{(j)}$, where $j \in \{1, \dots, l\}$.

In a sense, the algorithm is a very crude version of a sequential likelihood ratio test. Given that we are interested in the general rates of error decay we do not optimize the algorithm parameters for performance and instead make crude choices that are sufficient to prove the result. In particular we take $p = m/(4n)$ and $l = \log_2^2 n$.

The proof goes by showing first that, with high probability, the algorithm terminates before reaching the total sensing budget. Therefore, for the analysis we consider a modification of the algorithm were termination upon the event $p(k+1) > m$ is removed. Note that the number of measurements collected for entry i is simply c_i. These are independent random variables. The total number of measurements collected is $\sum_{i=1}^n c_i$. Note that for all i we have $0 \le c_i \le l$. Furthermore, note that for $i \notin S$ the corresponding measurements Y_i^j are zero mean normal random variables, which means that $\mathbb{P}_S(Y_i^{(j)} < 0) = 1/2$. Therefore, c_i corresponds to a truncated geometric random variable:

$$i \notin S, \qquad \mathbb{P}_S(c_i = x) = \begin{cases} (1/2)^x & \text{if } x = 1, \dots, l-1, \\ (1/2)^{l-1} & \text{if } x = l, \\ 0 & \text{otherwise.} \end{cases}$$

Since these are truncated geometric random variables it is clear that $\mathbb{E}_S(c_i) \le 2$ and $\mathbb{V}_S(c_i) \le 2$. Now, Bernstein's inequality (as stated in Wasserman [36], page 9) tells us immediately that

$$\mathbb{P}_S\left(\sum_{i \notin S} c_i - 2(n-s) \ge t\right) \le \exp\left(-\frac{1}{2}\frac{t^2}{2(n-s) + lt/3}\right).$$

Taking $t = n - s$, and noting that $\sum_{i=1}^n c_i \le sl + \sum_{i \notin S} c_i$ we conclude that

$$\mathbb{P}_S\left(\sum_{i=1}^n c_i < 3(n-s) + sl\right) \ge 1 - \exp\left(-\frac{1}{2}\frac{n-s}{2 + l/3}\right). \tag{5.9}$$

Now, provided $s \le n/(l-3)$, we conclude that the total number of measurements of the algorithm is smaller than $4n$ with probability approaching 1 as n grows, that is

$$\mathbb{P}_S\left(\sum_{i=1}^n c_i < 4n\right) \ge 1 - \exp\left(-\frac{1}{2}\frac{n-s}{2 + l/3}\right). \tag{5.10}$$

Therefore the total amount of precision used is under $4np$ with high probability. For the choice $p = m/(4n)$, the total amount of precision used is less than m with high probability. In other words,

$$\mathbb{P}_S\big(p(k+1) > m\big) \le \exp\left(-\frac{1}{2}\frac{n-s}{2 + l/3}\right). \tag{5.11}$$

This result ensures the modified algorithm is essentially the same as the original one, as in the latter we will rarely encounter the event $p(k+1) > m$ (this statement will be made precise later).

Therefore, we can proceed by analyzing the performance of the modified algorithm. This can be done in a entry-wise fashion and we must consider the cases $i \in S$ and $i \notin S$. For $i \notin S$, note that

$$\mathbb{P}_S(i \in \hat{S}) = \mathbb{P}_S\left(\bigcap_{i=1}^{l} \{Y_i^{(j)} \geq 0\}\right) = \frac{1}{2^l}.$$

For $i \in S$, we have

$$\mathbb{P}_S(i \notin \hat{S}) \leq \mathbb{P}_S\left(\bigcup_{i=1}^{l} \{Y_i^{(j)} < 0\}\right) = \frac{l}{2}\exp\left(-\frac{p\mu^2}{2}\right),$$

where the result follows from a Gaussian tail and the union (of events) bounds. These two results together give

$$\begin{aligned}\mathbb{E}_S\big[d(\hat{S}, S)\big] &= \sum_{i \notin S} \mathbb{P}_S(i \in \hat{S}) + \sum_{i \in S} \mathbb{P}_S(i \notin \hat{S}) \\ &\leq \frac{n-s}{2^l} + \frac{sl}{2}\exp\left(-\frac{p\mu^2}{2}\right) \\ &\leq \frac{n-s}{2^l} + \frac{1}{2}\exp\left(-\frac{p\mu^2 - 2\log s - 2\log l}{2}\right).\end{aligned}$$

Now, given the choice $l = \log_2^2 n$ we conclude that the first term in the above summation converges to 0 as $n \to \infty$, and the second term also converges to zero provided

$$-p\mu^2 - 2\log s - 2\log l \to \infty$$

as $n \to \infty$. Clearly if $\mu \geq \sqrt{\frac{4n}{m}(2\log s + 5\log\log_2 n)}$ this condition is satisfied for n large enough. To conclude the proof all that remains to be done is to take equation (5.11) into account to conclude that, for the original algorithm

$$\begin{aligned}\mathbb{E}_S\big[d(\hat{S}, S)\big] &\leq \mathbb{E}_S\big[d(\hat{S}, S) | p(k+1) \leq m\big] + \mathbb{E}_S\big[d(\hat{S}, S) | p(k+1) > m\big]\mathbb{P}_S\big(p(k+1) > m\big) \\ &\leq \mathbb{E}_S\big[d(\hat{S}, S) | p(k+1) \leq m\big] + n\mathbb{P}_S\big(p(k+1) > m\big) \\ &\leq \frac{n-s}{2^l} + \frac{1}{2}\exp\left(-\frac{p\mu^2 - 2\log s - 2\log l}{2}\right) + n\exp\left(-\frac{1}{2}\frac{n-s}{2 + \log_2^2 n/3}\right).\end{aligned}$$

Clearly, under the condition $s \leq n/(l-3)$ all the terms above converge to zero as $n \to \infty$, concluding the proof. □

Acknowledgements

I want to thank Nikhil Bansal for suggesting the elegant proof of Lemma 3.1. The modification of DS proposed in Section 3.2 came into being after discussions with Jarvis Haupt. Finally, the author wants to thank the two anonymous referees and the associate editor for their valuable comments and suggestions.

References

[1] Addario-Berry, L., Broutin, N., Devroye, L. and Lugosi, G. (2010). On combinatorial testing problems. *Ann. Statist.* **38** 3063–3092. MR2722464

[2] Arias-Castro, E., Candès, E.J. and Davenport, M.A. (2013). On the fundamental limits of adaptive sensing. *IEEE Trans. Inform. Theory* **59** 472–481. MR3008159

[3] Arias-Castro, E., Candès, E.J., Helgason, H. and Zeitouni, O. (2008). Searching for a trail of evidence in a maze. *Ann. Statist.* **36** 1726–1757. MR2435454

[4] Balcan, N., Beygelzimer, A. and Langford, J. (2006). Agostic active learning. In 23*rd International Conference on Machine Learning* 65–72.

[5] Bessler, S.A. (1960). Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments: Part I – Theory. Technical Report 55, Stanford Univ., Applied Mathematics and Statistics Laboratories.

[6] Blanchard, G. and Geman, D. (2005). Hierarchical testing designs for pattern recognition. *Ann. Statist.* **33** 1155–1202. MR2195632

[7] Butucea, C. and Ingster, Y. (2013). Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli* **19** 2652–2688.

[8] Cai, T.T., Jin, J. and Low, M.G. (2007). Estimation and confidence sets for sparse normal mixtures. *Ann. Statist.* **35** 2421–2449. MR2382653

[9] Castro, R., Willett, R. and Nowak, R. (2005). Faster rates in regression via active learning. In *Advances in Neural Information Processing Systems* **18** 179–186.

[10] Castro, R.M. and Nowak, R.D. (2008). Minimax bounds for active learning. *IEEE Trans. Inform. Theory* **54** 2339–2353. MR2450865

[11] Chernoff, H. (1959). Sequential design of experiments. *Ann. Math. Statist.* **30** 755–770. MR0108874

[12] Cohn, D., Ghahramani, Z. and Jordan, M. (1996). Active learning with statistical models. *J. Artificial Intelligence Res.* **4** 129–145.

[13] Dasgupta, S. (2004). Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems* **17** 337–344.

[14] Dasgupta, S. (2005). Coarse sample complexity bounds for active learning. In *Advances in Neural Information Processing Systems* **18** 235–242.

[15] Dasgupta, S., Kalai, A. and Monteleoni, C. (2005). Analysis of perceptron-based active learning. In *Eighteenth Annual Conference on Learning Theory* (*COLT*) 249–263.

[16] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.* **32** 962–994. MR2065195

[17] Donoho, D.L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52** 1289–1306. MR2241189

[18] El-Gamal, M.A. (1991). The role of priors in active Bayesian learning in the sequential statistical decision framework. In *Maximum Entropy and Bayesian Methods* (*Laramie*, *WY*, 1990). *Fund. Theories Phys.* **43** (W.T. Grandy and L.H. Schich, eds.) 33–38. Dordrecht: Kluwer Academic. MR1173460

[19] Fedorov, V.V. (1972). *Theory of Optimal Experiments*. New York: Academic Press. MR0403103

[20] Freund, Y., Seung, H.S., Shamir, E. and Tishby, N. (1997). Selective sampling using the query by committee algorithm. *Machine Learning* **28** 133–168.

[21] Hall, P. and Molchanov, I. (2003). Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces. *Ann. Statist.* **31** 921–941. MR1994735

[22] Hanneke, S. (2011). Rates of convergence in active learning. *Ann. Statist.* **39** 333–361. MR2797849

[23] Haupt, J., Baraniuk, R., Castro, R. and Nowak, R. (2012). Sequentially designed compressed sensing. In *IEEE Statistical Signal Processing Workshop* (*IEEE SSP*) *Proceedings* 401–404. Available at http://www.win.tue.nl/~rmcastro/publications/SCS.pdf.

[24] Haupt, J., Castro, R.M. and Nowak, R. (2011). Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Trans*. *Inform*. *Theory* **57** 6222–6235. MR2857969

[25] Ingster, Y.I. (1997). Some problems of hypothesis testing leading to infinitely divisible distributions. *Math*. *Methods Statist*. **6** 47–69. MR1456646

[26] Ingster, Y.I. and Suslina, I.A. (2003). *Nonparametric Goodness-of-fit Testing Under Gaussian Models*. *Lecture Notes in Statistics* **169**. New York: Springer. MR1991446

[27] Kim, J.-C. and Korostelev, A. (2000). Rates of convergence for the sup-norm risk in image models under sequential designs. *Statist*. *Probab*. *Lett*. **46** 391–399. MR1743998

[28] Koltchinskii, V. (2010). Rademacher complexities and bounding the excess risk in active learning. *J*. *Mach*. *Learn*. *Res*. **11** 2457–2485. MR2727771

[29] Lai, T.L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Adv*. *in Appl*. *Math*. **6** 4–22. MR0776826

[30] Malloy, M. and Nowak, R. (2011). On the limits of sequential testing in high dimensions. In *Asilomar Conference on Signals*, *Systems and Computers* 1245–1249. Available at http://arxiv.org/abs/1105.4540.

[31] Malloy, M. and Nowak, R. (2011). Sequential analysis in high-dimensional multiple testing and sparse recovery. In *The IEEE International Symposium on Information Theory* 2661–2665. Available at http://arxiv.org/abs/1103.5991v1.

[32] Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann*. *Statist*. **34** 373–393. MR2275246

[33] Novak, E. (1996). On the power of adaption. *J*. *Complexity* **12** 199–237. MR1408328

[34] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. New York: Springer. MR2724359

[35] Wald, A. (1947). *Sequential Analysis*. New York: Wiley. MR0020764

[36] Wasserman, L. (2006). *All of Nonparametric Statistics*. *Springer Texts in Statistics*. New York: Springer. MR2172729

Received June 2012 and revised March 2013