Improved Bounds for Sparse Recovery from Adaptive Measurements

Jarvis Haupt, Rui Castro, and Robert Nowak
Rice University, Columbia University, and the University of Wisconsin–Madison

***Abstract*—It is shown here that adaptivity in sampling results in dramatic improvements in the recovery of sparse signals in white Gaussian noise. An adaptive sampling-and-refinement procedure called *distilled sensing* is discussed and analyzed, resulting in fundamental new asymptotic scaling relationships in terms of the minimum feature strength required for reliable signal detection or localization (support recovery). In particular, reliable detection and localization using non-adaptive samples is possible only if the feature strength grows logarithmically in the problem dimension. Here it is shown that using adaptive sampling, reliable detection is possible provided the feature strength exceeds a constant, and localization is possible when the feature strength exceeds any (arbitrarily slowly) growing function of the problem dimension.**

I. Introduction

In high dimensional multiple hypothesis testing problems the aim is to identify the subset of the hypotheses that differ from the null distribution or simply to decide if one or more of the hypotheses do not follow the null. There is now a well developed theory and methodology for this problem, and the fundamental limitations in the high dimensional setting are well understood. However, most existing treatments of the problem assume a non-adaptive measurement process. The question of how the limitations might differ under a more flexible, sequential adaptive measurement process had not been addressed, prior to our own initial work in [1], [2]. This paper builds upon those initial results, establishing improved bounds for sparse recovery from adaptive measurements.

For concreteness let $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ be an unknown sparse vector, such that most (or all) of its components x_i are equal to zero. The locations of the non-zero components are arbitrary. This vector is observed in additive white Gaussian noise, and we consider two problems: *localization*–inferring the locations of non-zero components, and *detection*–deciding whether x is the all-zero vector. Given a single, non-adaptive noisy measurement of each entry of x, a common approach entails coordinate-wise thresholding of the observed data at a given level, identifying the number and locations of entries for which the corresponding observation magnitude exceeds a certain value. In such settings there are sharp asymptotic thresholds that the magnitude of the non-zero components must exceed in order for the signal to be localizable and/or detectable. Such characterizations have been given in [3], [4] for the localization problem and [5], [6] for the detection problem. A more thorough review of these sort of characterizations is given in Section II-A.

In this paper we investigate these problems under a more flexible measurement process. Suppose we are able to sequentially collect multiple noisy measurements of each component of x, and that the data so obtained can be modeled as

$$y_{i,j} \;=\; x_i \;+\; \gamma_{i,j}^{-1/2}\, w_{i,j},\;\; i = 1, \ldots, p, \quad j = 1, \ldots, k \;. \quad (1)$$

In the above a total of k measurement steps is taken, j indexes the measurement step, $w_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and $\gamma_{i,j} \geq 0$ quantifies the precision of the jth measurement of entry i. When $\gamma_{i,j} = 0$ we adopt the convention that component x_i was not observed at step j. The crucial feature of this model is that it does not preclude sequentially adaptive measurements, in which the $\gamma_{i,j}$ can depend on past observations $\{y_{i,\ell}\}_{\ell<j}$[1].

In order to make fair comparisons to non-adaptive measurement processes, the total precision budget is limited in the following way. Let $R(p)$ be an increasing function of p, the dimension of the problem (that is, the number of hypotheses under scrutiny). The precision parameters $\{\gamma_{i,j}\}$ are required to satisfy $\sum_{j=1}^{k}\sum_{i=1}^{p}\gamma_{i,j} \;\leq\; R(p)$. For example, the usual non-adaptive, single measurement model corresponds to taking $R(p) = p$, $k = 1$, and $\gamma_{i,1} = 1$, $i = 1, \ldots, p$. This baseline can be compared with adaptive procedures by keeping $R(p) = p$, but allowing $k > 1$ and variables $\{\gamma_{i,j}\}$ satisfying the precision budget constraint.

The multiple measurement process (1) is applicable in many interesting and relevant scenarios. For example in gene association and expression studies, two-stage sampling approaches are quite popular (see [7], [8], [9] and references therein): in the first stage a large number of genes is initially tested to identify a promising subset of them, and in the second-stage these promising genes are subject to further testing. Such ideas have been extended to multiple-stage approaches; see, for example [10]. Similar two-stage approaches have also been examined in the signal processing literature–see [11]. More broadly, sequential experimental design has been popular in other fields as well, such as in computer vision where it is known as *active vision* [12], or in machine learning, where it is known as *active learning* [13], [14]. These types of procedures can potentially impact other areas such as microarray-based studies and astronomical surveying. The main contribution of

[1]The precision for a measurement at location i at step j may be controlled in practice by collecting multiple independent samples and averaging to reduce the effective observation noise, the result of which would be an observation described by the model (1). In this case, the parameters $\{\gamma_{i,j}\}$ are proportional to the number of samples collected at each such step. For exposure-based sampling modalities common in many imaging scenarios, the precision parameters $\{\gamma_{i,j}\}$ can be interpreted as proportional to the length of time for which the component at location i is observed at step j.

this paper is a theoretical analysis that reveals the dramatic gains that can be attained using such sequential procedures.

Our focus here is on a sequential adaptive sampling procedure called *distilled sensing* (DS). The idea behind DS is simple: use a portion of the total precision budget to crudely measure all components; using those measurements, eliminate a fraction of the components that appear least promising from future consideration; and iterate this process several times. When the vector x is sparse the DS algorithm, whose pseudocode is given in Algorithm 1, is shown to gradually focus the measurement process preferentially on non-zero components of the signal[2]. As mentioned above, similar procedures have been proposed in experimental science, however to the best of our knowledge the quantification of performance gains had not been established prior to our own previous work in [1], [2] and the results shown in this paper. In this manuscript we significantly extend our previous work by providing stronger results for the localization problem, and an entirely novel characterization of the detection problem.

This paper is organized as follows. Following a brief discussion of the fundamental limits of non-adaptive sampling for detection and localization, our main results, that DS can reliably solve the localization and detection problems for dramatically weaker signals than what is possible using non-adaptive measurements, are stated in Sect. II. A sketch of the proof of the main result is given in Sect. III. Simulation results demonstrating the theory are provided in Sect. IV, and conclusions and extensions are discussed in Sect. V.

II. Main Results

The main results of our theoretical analysis of DS are stated later in this section, but first we begin by reviewing the asymptotic thresholds for localization and detection from non-adaptive measurements. As mentioned above, these thresholds are now well known [3], [4], [5], [6], but here we provide a concise summary of the main ideas, in terms that will facilitate our comparison with DS. We then highlight some of the surprising gains achievable using DS.

A. Non-adaptive Localization and Detection of Sparse Signals

The non-adaptive measurement model we will consider as the baseline for comparison is as follows. We have a single observation of x in noise:

$$y_i \;=\; x_i \;+\; w_i,\; i = 1, \ldots, p\,, \tag{2}$$

where $w_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. As noted above, this is a special case of our general setup (1) where $k = 1$ and $\gamma_{i,1} = 1$, $i = 1, \ldots, p$, implying a precision budget $R(p) = \sum_{i=1}^{p} \gamma_{i,1} = p$.

To describe the asymptotic (large p) thresholds for localization we need to introduce some notation. Define the *false-discovery proportion* (FDP) and *non-discovery proportion* (NDP) as follows.

[2]We assume that the non-zero components are positive for simplicity, though it is trivial to extend the algorithm and its analysis to handle both positive and negative components by simply repeating the entire process twice; once as described and again with $y_{i,j}$ replaced with $-y_{i,j}$ in the refinement step of Algorithm 1.

Definition II.1. *Let $\mathcal{S} := \{i : x_i \neq 0\}$ be the signal support set, and let $\widehat{\mathcal{S}} = \widehat{\mathcal{S}}(y)$ denote an estimator of $\mathcal{S}$. The false-discovery proportion is given by* $\text{FDP}(\widehat{\mathcal{S}}) := |\widehat{\mathcal{S}} \backslash \mathcal{S}|/|\widehat{\mathcal{S}}|$. *In words, the* FDP *of $\widehat{\mathcal{S}}$ is the ratio of the number of components falsely declared as non-zero to the total number of components declared non-zero. The non-discovery proportion is given by* $\text{NDP}(\widehat{\mathcal{S}}) := |\mathcal{S} \backslash \widehat{\mathcal{S}}|/|\mathcal{S}|$. *In words, the* NDP *of $\widehat{\mathcal{S}}$ is the ratio of the number of non-zero components missed to the number of actual non-zero components.*

We focus on a specific class of estimators of $\mathcal{S}$ obtained by *coordinate-wise thresholding*

$$\widehat{\mathcal{S}}_\tau(y) \;:=\; \{i \in \{1, \ldots, p\} : y_i \geq \tau > 0\}\;, \tag{3}$$

where the threshold τ may depend implicitly on x, or on y itself. The following result establishes the limits of localization using non-adaptive sampling, and is similar in spirit to [15], where related results were obtained under a random signal model. Due to lack of space the result is stated without proof.

Theorem II.2. *Assume $x \geq 0$ with $p^{1-\beta}$, $\beta \in (0,1)$, non-zero components of amplitude $\sqrt{2r \log p}$, $r > 0$, and measurement model (2). There exists a coordinate-wise thresholding procedure that yields an estimator $\widehat{\mathcal{S}}(y)$ such that if $r > \beta$, then*

$$\text{FDP}(\widehat{\mathcal{S}}) \;\overset{P}{\to}\; 0\,, \quad \text{NDP}(\widehat{\mathcal{S}}) \;\overset{P}{\to}\; 0\;,$$

as $p \to \infty$, where $\overset{P}{\to}$ denotes convergence in probability. Moreover, if $r \leq \beta$, then there does not exist a coordinate-wise thresholding procedure that can guarantee that both quantities above tend to 0 as $p \to \infty$.

The detection problem, which amounts to a hypothesis test between the null distribution $x = 0$ and a sparse alternative, has also been addressed in the literature under a random signal model [5], [6]. Consider the hypothesis testing problem:

$$\begin{aligned} \text{H}_0 \;&:\; y_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1),\; i = 1, \ldots, p \\ \text{H}_1 \;&:\; y_i \overset{\text{iid}}{\sim} (1 - \theta(p))\,\mathcal{N}(0,1) + \theta(p)\,\mathcal{N}(\mu(p), 1), \\ &\qquad\qquad i = 1, \ldots, p \end{aligned} \tag{4}$$

where $\theta(p) = p^{-\beta}$ and $\mu(p) = \sqrt{2r \log p}$. These hypotheses model measurements of either the zero vector, or of a randomly generated signal x (with each entry having amplitude $\sqrt{2r \log p}$ independently with probability $p^{-\beta}$, and amplitude zero with probability $1 - p^{-\beta}$) according to the measurement model (2). Note that under the alternative, the signal has $p^{1-\beta}$ non-zero components in expectation. We recall the following.

Theorem II.3. *Consider the hypotheses in (4). Define*

$$\rho(\beta) \;:=\; \begin{cases} 0, & 0 < \beta < 1/2 \\ \beta - 1/2, & 1/2 < \beta \leq 3/4 \\ (1 - \sqrt{1-\beta})^2, & 3/4 < \beta < 1 \end{cases}$$

If $r > \rho(\beta)$, then there exists a test for which the sum of the false alarm and miss probabilities tends to 0 as $p \to \infty$. Conversely, if $r < \rho(\beta)$, then for any test the sum of the false alarm and miss probabilities tends to 1 as $p \to \infty$.

Theorem II.3 was proved in [6] relying heavily on the ideas presented in [5]. Although it is stated for the random sparsity model (4) it is possible to relate the results to the deterministic sparsity model that we consider in the paper, namely using the ideas presented in Chapter 8 of [16].

B. Distilled Sensing

Algorithm 1 describes the DS measurement process. The algorithm proceeds in steps, each of these using a portion R_j of the total precision budget $R(p)$. At each step we retain only the components with non-negative observations, meaning that roughly half of the components are eliminated from further consideration when the number of non-zero components is very small. The key is to identify conditions under which the crude thresholding at 0 at each step does not remove a significant number of the non-zero components.

The following theorem summarizes the main result for DS. In contrast to the results provided above, which require that the signal amplitude be $\Omega(\sqrt{\log p})$ for non-adaptive localization and detection, DS is capable of reliably localizing and detecting much weaker sparse signals.

Theorem II.4. *Assume $x \geq 0$ with $p^{1-\beta}$, $\beta \in (0,1)$, non-zero components of amplitude $\mu(p)$, and sequential measurement model using Distilled Sensing with $k = k(p) = \max\{\lceil \log_2 \log p \rceil, 0\} + 2$ observation steps, and precision budget distributed over the measurement steps so that $\sum_{j=1}^{k} R_j \leq p$, $R_{j+1}/R_j \geq \delta > 1/2$, and $R_1 = c_1 p$ and $R_k = c_k\, p$ for some $c_1, c_k \in (0,1)$. Then the estimator formed from the final set of observations of the DS procedure,*

$$\widehat{\mathcal{S}}_{\rm DS} \;\; := \;\; \{i \in I_k : y_{i,k} > \sqrt{2/c_k}\}$$

has the following properties:

(i) if $\mu(p) \to \infty$ as a function of p then as $p \to \infty$

$$\mathrm{FDP}(\widehat{\mathcal{S}}_{\rm DS}) \;\; \stackrel{P}{\to} \;\; 0\,, \quad \mathrm{NDP}(\widehat{\mathcal{S}}_{\rm DS}) \;\; \stackrel{P}{\to} \;\; 0\,.$$

(ii) if $\mu(p) \geq \max\left\{\sqrt{4/c_1}, 2\sqrt{2/c_k}\right\}$ (a constant) then

$$\lim_{p\to\infty} \Pr(\widehat{\mathcal{S}}_{\rm DS} = \emptyset) = \left\{ \begin{array}{ll} 1\,, & \text{if } x = 0 \\ 0\,, & \text{if } x \neq 0 \end{array} \right. ,$$

where $\emptyset$ is the empty set.

The result (ii) is entirely novel, and (i) improves on the result stated in [2] which required $\mu(p)$ to grow faster than an arbitrary iteration of the logarithm (i.e., $\mu(p) \sim \log\log\cdots\log p$).

III. Analysis of Distilled Sensing

In this section we prove the main result characterizing the performance of DS, Theorem II.4. We begin with three lemmas that quantify the finite sample behavior of DS.

Lemma III.1. *If $\{y_i\}_{i=1}^m \stackrel{\rm iid}{\sim} \mathcal{N}(0,\sigma^2)$, $\sigma > 0$, then for any $0 < \varepsilon < 1/2$,*

$$\left(\frac{1}{2} - \varepsilon\right) m \leq \Big|\{i \in \{1,\ldots,m\} : y_i > 0\}\Big| \leq \left(\frac{1}{2} + \varepsilon\right) m,$$

Algorithm 1: Distilled Sensing.

Input:
Number of observation steps: k;
Resource budget: $R(p)$;
Resource allocation sequence satisfying $\sum_{j=1}^{k} R_j \leq R(p)$;

Initialize:
Initial index set: $I_1 \longleftarrow \{1, 2, \ldots, p\}$;

Distillation:
for $j = 1$ **to** k **do**
Allocate: $\gamma_{i,j} = \left\{ \begin{array}{cc} R_j/|I_j| & i \in I_j \\ 0 & i \notin I_j \end{array} \right\}$;
Observe: $y_{i,j} = x_i + \gamma_{i,j}^{-1/2} w_{i,j},\ i \in I_j$;
Refine: $I_{j+1} \longleftarrow \{i \in I_j : y_{i,j} > 0\}$;
end

Output:
Final index set: I_k;
Distilled observations: $y_k = \{y_{i,k} : i \in I_k\}$;

with probability at least $1 - 2\exp(-2m\varepsilon^2)$.

Lemma III.2. *Let $\{y_i\}_{i=1}^m \stackrel{\rm iid}{\sim} \mathcal{N}(\mu,\sigma^2)$, with $\sigma > 0$ and $\mu \geq 2\sigma$. Define $\epsilon = \frac{\sigma}{\mu\sqrt{2\pi}} < 1$. Then*

$$(1-\epsilon)m \leq \Big|\{i \in \{1,2,\ldots,m\} : y_i > 0\}\Big| \leq m,$$

with probability at least $1 - \exp\left(-\frac{\mu m}{4\sigma\sqrt{2\pi}}\right)$.

The results follow from Hoeffding's inequality, and from a standard Gaussian tail inequality together with a characterization of the Binomial distribution from Chernoff, respectively. See [2] for details.

Now, refer to Algorithm 1 and define $s_j := |\mathcal{S} \bigcap I_j|$ and $z_j := |\mathcal{S}^c \bigcap I_j|$, the number of non-zero and zero components, respectively, present at the beginning of step $j = 1, \ldots, k$. Let $\varepsilon > 0$, and for $j = 1, \ldots, k-1$ define

$$\epsilon_j^2 \;\; := \;\; \frac{s_1 + (1/2+\varepsilon)^{j-1} z_1}{2\pi\mu^2 R_j}\,, \tag{5}$$

The output of the DS procedure is quantified in the following.

Lemma III.3. *Let $0 < \varepsilon < 1/2$ and assume that $R_j > \frac{4}{\mu^2}\left(s_1 + (1/2+\varepsilon)^{j-1} z_1\right)$, $j = 1, \ldots, k-1$. If $|\mathcal{S}| > 0$, then with probability at least*

$$1 - \sum_{j=1}^{k-1} \exp\left(\frac{-s_1 \prod_{\ell=1}^{j-1}(1-\epsilon_\ell)}{\sqrt{8\pi}}\right) - 2\sum_{j=1}^{k-1} \exp\left(-2 z_1 (1/2-\varepsilon)^{j-1}\varepsilon^2\right)$$

$\prod_{\ell=1}^{j-1}(1-\epsilon_\ell) s_1 \leq s_j \leq s_1$ *and* $\left(\frac{1}{2}-\varepsilon\right)^{j-1} z_1 \leq z_j \leq \left(\frac{1}{2}+\varepsilon\right)^{j-1} z_1$ *for $j = 2, \ldots, k$. If $|\mathcal{S}| = 0$, then with*

probability at least

$$1 - 2\sum_{j=1}^{k-1} \exp\left(-2z_1(1/2-\varepsilon)^{j-1}\varepsilon^2\right)$$

$\left(\frac{1}{2}-\varepsilon\right)^{j-1} z_1 \leq z_j \leq \left(\frac{1}{2}+\varepsilon\right)^{j-1} z_1$ *for* $j = 2, \ldots, k$.

Proof: The results follow from Lemmas III.1 and III.2 and the union bound. First assume that $s_1 = |\mathcal{S}| > 0$. Let $\sigma_j^2 := |I_j|/R_j = (s_j + z_j)/R_j$ and $\widetilde{\epsilon}_j := \frac{\sigma_j}{\mu\sqrt{2\pi}}$, $j = 1, \ldots, k$. Now, we proceed by conditioning on the outcome of all prior refinement steps. In particular, assume that $(1-\widetilde{\epsilon}_{\ell-1})s_{\ell-1} \leq s_\ell \leq s_{\ell-1}$ and $\left(\frac{1}{2}-\varepsilon\right) z_{\ell-1} \leq z_\ell \leq \left(\frac{1}{2}+\varepsilon\right) z_{\ell-1}$for $\ell = 1, \ldots, j$. Then apply Lemma III.1 with $m = z_j$, Lemma III.2 with $m = s_j$ and $\sigma^2 = \sigma_j^2$, and the union bound to obtain that with probability at least

$$1 - \exp\left(-\frac{\mu s_j}{4\sigma_j\sqrt{2\pi}}\right) - 2\exp\left(-2z_j\varepsilon^2\right) \tag{6}$$

$(1-\widetilde{\epsilon}_j)s_j \leq s_{j+1} \leq s_j$, and $\left(\frac{1}{2}-\varepsilon\right) z_j \leq z_{j+1} \leq \left(\frac{1}{2}+\varepsilon\right) z_j$. Note that the condition $R_j > \frac{4}{\mu^2}\left(s_1 + (1/2+\varepsilon)^{j-1} z_1\right)$ along with the assumptions on prior refinement steps ensure that $\mu > 2\,\sigma_j$, which is required for Lemma III.2. The condition $\mu > 2\,\sigma_j$ also allows us to simplify probability bound (6), so that the event above occurs with probability at least

$$1 - \exp\left(-\frac{s_j}{2\sqrt{2\pi}}\right) - 2\exp\left(-2z_j\varepsilon^2\right).$$

Next, we can recursively apply the union bound and the bounds on s_j and z_j above to obtain for $j = 1, \ldots, k-1$,

$$\epsilon_j \;=\; \sqrt{\frac{s_1 + (1/2+\varepsilon)^{j-1} z_1}{2\pi\mu^2 R_j}} \;\geq\; \widetilde{\epsilon}_j \;=\; \frac{\sigma_j}{\mu\sqrt{2\pi}}$$

with probability at least

$$1 - \sum_{j=1}^{k-1} \exp\left(\frac{-s_1\prod_{\ell=1}^{j-1}(1-\epsilon_\ell)}{\sqrt{8\pi}}\right) - \sum_{j=1}^{k-1} 2\exp\left(-2z_1(1/2-\varepsilon)^{j-1}\varepsilon^2\right).$$

Note that the condition $R_j > \frac{4}{\mu^2}\left(s_1 + (1/2+\varepsilon)^{j-1} z_1\right)$ implies that $\epsilon_j < 1$. The first result follows directly. If $s_1 = |\mathcal{S}| = 0$, then consider only z_j, $j = 1, \ldots, k$. The result follows again by the union bound. Note that for this statement the condition on R_j is not required. ∎

Remark: It is noteworthy to examine the condition $R_j > \frac{4}{\mu^2}\left(s_1 + (1/2+\varepsilon)^{j-1} z_1\right)$, $j = 1, \ldots, k$ more closely. Define $c := s_1/[(1/2+\varepsilon)^{k-1} z_1]$. Then the conditions on R_j are satisfied if

$$R_j > \frac{4z_1(1/2+\varepsilon)^{j-1}}{\mu^2}(c(1/2+\varepsilon)^{k-j}+1).$$

Since $z_1 \leq p$, the following condition is sufficient

$$R_j > \frac{4p(1/2+\varepsilon)^{j-1}}{\mu^2}(c(1/2+\varepsilon)^{k-j}+1).$$

This condition condenses several problem specific parameters (s_1, z_1, and k) into the scalar parameter c, and in particular the more stringent condition $R_j > \frac{4(c+1)p(1/2+\varepsilon)^{j-1}}{\mu^2}$ will suffice. It is now easy to see that if $s_1 \ll z_1$ (e.g., so that $c \leq 1$), then the sufficient conditions become $R_j > \frac{8p}{\mu^2}(1/2+\varepsilon)^{j-1}$, $j = 1, \ldots, k$. Thus, for the sparse situations we consider, the precision allocated to each step must be just slightly greater than $1/2$ of the precision allocated in the previous step. This is the key to guarantee the results of Theorem II.4.

A. Sketch of Proof of Theorem II.4

The proof of the main result follows from a careful application of Lemma III.3. We provide only a sketch of the complete proof here due to page limitations; for complete details, see [17]. The main idea of the proof is to show that, with probability tending to one as $p \to \infty$, the DS procedure retains most of the signal components (part (i) of the theorem) or at least a significant fraction of those (part (ii) of the theorem), while discarding a large number of the zero components, thereby increasing the precision of the final set of measurements dramatically. The proof proceeds by analyzing the event

$$\Gamma \;=\; \left\{ z_1\left(\frac{1}{2}-\varepsilon\right)^{k-1} \;\leq\; z_k \;\leq\; z_1\left(\frac{1}{2}+\varepsilon\right)^{k-1} \right\} \bigcap \left\{ s_1\prod_{j=1}^{k-1}(1-\epsilon_j) \;\leq\; s_k \;\leq\; s_1 \right\}.$$

By using the choice of k in the Theorem and taking $\varepsilon = p^{-1/3}$ we conclude that $\Pr(\Gamma) \to 1$ as $p \to \infty$. We now proceed by conditioning on Γ, and we note that the output of the DS procedure consists of a total of $s_k + z_k$ independent Gaussian random variables with variance $(s_k + z_k)/R_k$, where s_k of them have mean μ and z_k have mean zero. Provided μ is large enough, in particular $\mu(p) \geq \max\left\{\sqrt{4/c_1}, 2\sqrt{2/c_k}\right\}$, we can show that the threshold in the procedure for the computation of $\widehat{\mathcal{S}}_{\mathrm{DS}}$ is such that, conditionally on Γ we will retain all the s_k signal components and discard all the z_k non-signal components with high probability (this ensues from Gaussian tail bounds and a union of events bound). The proof of part (ii) of the Theorem follows by noting that s_k/s_1 is bounded away from zero, given the condition on μ above, and so if $\mathcal{S} \neq \emptyset$ we guarantee that $\widehat{\mathcal{S}}_{\mathrm{DS}} \neq \emptyset$ with increasingly high probability. Similarly if $\mathcal{S} = \emptyset$ then clearly $\widehat{\mathcal{S}}_{\mathrm{DS}} = \emptyset$ with increasingly high probability. Furthermore if μ is a diverging sequence in p we can also show that $s_k/s_1 \to 1$. This ensures that, with increasingly high probability the $\mathrm{FDP}(\widehat{\mathcal{S}}) = 0$ (as the thresholding procedure is guaranteed to retain only signal components), and that the $\mathrm{NDP}(\widehat{\mathcal{S}}) = (s_1 - s_k)/s_1 \to 0$.

IV. Numerical Experiments

This section presents numerical experiments with DS. We consider three cases, corresponding to $p = 2^{14}, 2^{17}$, and 2^{20}, and in each case the number of non-zero entries is given by $\lfloor p^{1/2} \rfloor$. We choose $k = \max\{\lceil \log_2 \log p \rceil, 0\} + 2$ as in

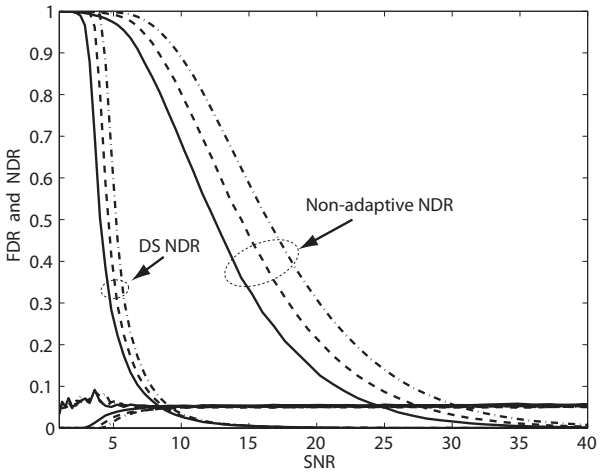

Fig. 1. FDR and NDR vs. SNR comparison. The FDR and NDR (false- and non-discovery rates) are the average FDP and NDP over 500 independent trials at each SNR (SNR = μ^2). Thresholds were chosen to achieve FDR $= 0.05$. The solid, dashed, and dash-dot lines correspond to $p = 2^{14}, 2^{17}$, and 2^{20}, respectively, and in each case the number of non-zero entries is $\lfloor p^{1/2} \rfloor$. The nearly-flat curves at the bottom of the plot correspond to the FDRs in each case, which do not differ dramatically from each other.

Theorem II.4, which corresponds to $k = 6$ for each of the three cases. The precision allocation used throughout the simulations is given by $R_j = (0.75)^{j-1} R_1$ for $j = 2, \ldots, k-1$, with $R_k = R_1$, and R_1 chosen so that $\sum_{j=1}^{k} R_j = p$.

Figure 1 compares the performance of non-adaptive sensing and DS for the cases $p = 2^{14}, 2^{17}$, and 2^{20}, which correspond to the solid, dashed, and dash-dot lines, respectively. The plot depicts the false- and non-discovery rates (average FDP and NDP) as a function of SNR for each case, averaged over 500 independent trials. Thresholds were chosen so that the FDRs were approximately 0.05 in each case. Not only does DS achieve significantly lower NDRs than non-adaptive sampling over the entire SNR range, its performance also exhibits much less dependence on the signal dimension p.

V. Concluding Remarks

There has been a tremendous interest in high-dimensional testing and detection problems in recent years. A well-developed theory exists for such problems when using a single, non-adaptive observation model [3], [4], [5], [6]. However, in practice and theory, multistage adaptive designs have shown promise [7], [8], [9], [10], [11]. This paper quantifies the improvements such methods can achieve. We analyzed a specific multistage design called Distilled Sensing (DS), and established that DS is capable of detecting and localizing much weaker sparse signals than non-adaptive methods. The main result shows that adaptivity allows reliable detection and localization at a signal-to-noise ratio (SNR) that is $\log p$ lower than the minimum required by non-adaptive methods, where p is the problem dimension. The results presented here can be extended also to very sparse signals; in particular, the analysis presented here also shows that DS enables recovery of signals having as few as $\Omega(\log \log \log p)$ nonzero entries.

Note that the DS procedure as described requires about $2n$ total measurements: n for the first step, about $n/2$ for the second, about $n/4$ for the third, and so on. This requirement can be reduced by considering alternate measurement models. For example, rather than direct measurements, each measurement could be a linear combination of the entries of x. If the linear combinations are non-adaptive, this leads to a regression model commonly studied in the Lasso and Compressed Sensing literature [18], [19]. However, sequentially tuning the linear combinations leads to an adaptive version of the regression model which can be shown to provide significant improvements, as well [20].

Acknowledgment

This work was partially supported by AFOSR grant FA9550-09-1-0140.

References

[1] J. Haupt, R. Castro, and R. Nowak, "Adaptive discovery of sparse signals in noise," in *Proc Asilomar Conf on Signals, Systems, and Computers*, October 2008.

[2] ——, "Distilled sensing: Selective sampling for sparse signal recovery," in *Proc Conf on Artificial Intelligence and Statistics*, April 2009.

[3] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone, "Adapting to unknown sparsity by controlling the false discovery rate," *Ann Stat*, vol. 34, pp. 584–653, 2006.

[4] D. Donoho and J. Jin, "Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data," *Ann Stat*, vol. 34, no. 6, pp. 2980–3018, 2006.

[5] Y. Ingster, "Some problem of hypothesis testing leading to infinitely divisible distributions," *Math Methods Statist*, vol. 6, no. 1, pp. 47–69, 1997.

[6] D. Donoho and J. Jin, "Higher criticsm for detecting sparse heterogenous mixtures," *Ann Stat*, vol. 32, no. 3, pp. 962–994, 2004.

[7] H.-H. Muller, R. Pahl, and H. Schafer, "Including sampling and phenotyping costs into the optimization of two stage designs for genomewide association studies," *Genet Epidemiol*, vol. 31, pp. 844–852, 2007.

[8] S. Zehetmayer, P. Bauer, and M. Posch, "Two-stage designs for experiments with large number of hypotheses," *Bioinformatics*, vol. 21, pp. 3771–3777, 2005.

[9] J. Satagopan and R. Elston, "Optimal two-stage genotyping in population-based association studies," *Genet Epidemiol*, vol. 25, no. 149–157, 2003.

[10] S. Zehetmayer, P. Bauer, and M. Posch, "Optimized multi-stage designs controlling the false discovery or the family-wise error rate," *Stat Med*, vol. 27, pp. 4145–4160, 2008.

[11] E. Bashan, R. Raich, and A. Hero, "Optimal two-stage search for sparse targets using convex criteria," *IEEE T Signal Proces*, vol. 56, no. 11, pp. 5389–5402, Nov. 2008.

[12] Various, "Promising directions in active vision," in *Int J Comput Vision*, M. J. Swain and M. A. Stricker, Eds., vol. 11, no. 2, 1991, pp. 109–126.

[13] D. Cohn, "Neural network exploration using optimal experiment design," *Neural Networks*, vol. 6, pp. 679–686, 1994.

[14] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *J Artif Intell Res*, pp. 129–145, 1996.

[15] C. Genovese, J. Jin, and L. Wasserman, "Revisiting marginal regression," *Submitted*, 2009.

[16] Y. Ingster and I. Suslina, *Nonparametric Goodness-of-Fit Testing under Gaussian Models*, ser. Lect Notes Stat. Springer, 2003, vol. 169.

[17] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *submitted*, Jan. 2010, online: `www.ece.rice.edu/~jdh6/publications/sub10_DS.pdf`.

[18] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.

[19] E. J. Candès, "Compressive sampling," in *Proc. Int. Congress of Mathematics*, vol. 3, Madrid, Spain, 2006, pp. 1433–1452.

[20] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak, "Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements," in *Proc Asilomar Conf on Signals, Systems, and Computers*, November 2009.