Distilled Sensing: Adaptive Sampling for Sparse Detection and Estimation

Jarvis Haupt, Rui Castro, and Robert Nowak

Abstract

Adaptive sampling results in significant improvements in the recovery of sparse signals in white Gaussian noise. A sequential adaptive sampling-and-refinement procedure called *Distilled Sensing* (DS) is proposed and analyzed. DS is a form of multi-stage experimental design and testing. Because of the adaptive nature of the data collection, DS can detect and localize far weaker signals than possible from non-adaptive measurements. In particular, reliable detection and localization (support estimation) using non-adaptive samples is possible only if the signal amplitudes grow logarithmically with the problem dimension. Here it is shown that using adaptive sampling, reliable detection is possible provided the amplitude exceeds a constant, and localization is possible when the amplitude exceeds any arbitrarily slowly growing function of the 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 quite clear. 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 has not been addressed. This paper shows that this additional flexibility can yield surprising and significant performance gains.

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:

J.H. is with the Department of Electrical and Computer Engineering, University of Minnesota; R.C. is with the Department of Mathematics and Computer Science of the Eindhoven University of Technology (TU/e) (part of the work was conducted when the author was affiliated with Columbia University); R.N. is with the Department of Electrical and Computer Engineering, University of Wisconsin—Madison. E-mail: jdhaupt@umn.edu, rmcastro@tue.nl, nowak@ece.wisc.edu.

This work was supported in part by NSF Grant CCF-0353079 and AFOSR Grant FA9550-09-1-0140, and is dedicated to the memory of Dr. Dennis Healy, who inspired and supported this direction of research through the DARPA Analog-to-Information Program. Dennis' guidance, vision, and inspiration will be missed.

Manuscript completed January 28, 2010; revised May 27, 2010 and April 10, 2011.

A preliminary version of this report appeared at the International Symposium on Information Theory (ISIT) in Austin, TX in June 2010.

Localization: Infer the locations of the few non-zero components.

Detection: Decide whether x is the all-zero vector.

Given a single, non-adaptive noisy measurement 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 exceeds the threshold. 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 various contexts in [1]–[3] for the localization problem and [4]–[6] for the detection problem. A more thorough review of these sorts of characterizations is given in Section II.

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

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

In the above a total of k measurement steps is taken, j indexes the measurement step, $w_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ are zero-mean Gaussian random variables with unit variance, and $\gamma_{i,j} \geq 0$ quantifies the precision of each measurement. 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, where the $\gamma_{i,j}$ can depend on past observations $\{y_{i,\ell}\}_{i \in \{1,\ldots,p\}, \ell < j}$.

In practice, the precision for a measurement at location i at step j may be controlled, for example, 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 $\{\gamma_{i,j}\}$ can be thought of as proportional to the number of samples collected at location i at step j. For exposure-based sampling modalities common in many imaging scenarios, the precision parameters $\{\gamma_{i,j}\}$ can be interpreted as being proportional to the length of time for which the component at location i is observed at step j.

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

$$\sum_{j=1}^{k} \sum_{i=1}^{p} \gamma_{i,j} \;\leq\; R(p)\,. \tag{2}$$

For example, the usual non-adaptive, single measurement model corresponds to taking $R(p) = p$, $k = 1$, and $\gamma_{i,1} = 1$ for $i = 1, \ldots, p$. This baseline can be compared with adaptive procedures by keeping $R(p) = p$, but allowing $k > 1$ and variables $\{\gamma_{i,j}\}$ satisfying (2).

The multiple measurement process (1) is applicable in many interesting and relevant scenarios. For example in gene association and expression studies, two-stage approaches are gaining popularity (see [7]–[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 this paper is a theoretical analysis that reveals the significant gains that can be attained using such sequential procedures. Our focus here is on a particular sequential, adaptive sampling procedure called *Distilled Sensing* (DS). The idea behind DS is simple: use a portion of the precision budget to crudely measure all components; eliminate a fraction of the components that appear least promising from further consideration after this measurement; and iterate this procedure several times, at each step measuring only components retained after the previous step. As mentioned above, similar procedures have been proposed in the context of experimental design. However, to the best of our knowledge the quantification of performance gains had not been established prior to our own initial work in [15], [16] and the results established in this paper. In this manuscript we significantly extend our previous results by providing stronger results for the localization problem, and an entirely novel characterization of the detection problem.

The sequential refinement that is inherent to the DS procedure is somewhat reminiscent of recently proposed hierarchical and "conquer and divide" strategies for multiple hypothesis testing [17]–[19]. However, there are a few important differences between the settings examined in those works and the problem being examined here. Namely, the works [17]–[19] examine particular types of coarse-to-fine testing procedures which implicitly assume some form of hierarchical structure among the signal components. On the other hand, as mentioned above, here we consider signals for which the locations of the nonzero components are arbitrary. In addition, the works [17]–[19] consider sequential testing from a single collection of *non-adaptively* collected measurements, while here we examine a more flexible sequential *adaptive* measurement model.

This paper is organized as follows. Following a brief review of the fundamental limits of non-adaptive sampling for detection and localization in Section II, our main result—that DS can reliably solve the localization and detection problems for significantly weaker signals than what is possible using non-adaptive measurements—is stated in Section III. A proof of the main result is given in Section IV. Simulation results demonstrating the theory are provided in Section V, and conclusions and extensions are discussed in Section VI. A proof of the threshold for localization from non-adaptive measurements and several auxiliary lemmas are provided in the appendices.

II. Review of Non-adaptive Localization and Detection of Sparse Signals

In this section we review the known thresholds for localization and detection from non-adaptive measurements. As mentioned above, such thresholds have been established in a variety of problem settings [1]–[6]. Here we provide a concise summary of the main ideas along with supporting proofs as needed, to facilitate comparison with our main results concerning recovery from adaptive measurements which appear in the next section.

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:

$$y_i \;=\; x_i \;+\; w_i,\; i = 1, \ldots, p\,, \tag{3}$$

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) in which $k = 1$ and $\gamma_{i,1} = 1$ for $i = 1, \ldots, p$. This implies 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 terminology and notation. The performance could be quantified in terms of error probabilities (e.g., probabilities of one or more false-positive or false-negative events). However, aiming to control the error probabilities in the high-dimensional setting (e.g,. using the Bonferoni threshold level $\sqrt{2 \log p}$) is often too conservative and this has motivated alternative criteria that are less restrictive, but capture the important requirements in many scientific applications. The so-called *False-discovery Rate* [1] is perhaps the most commonly employed criterion, and our analysis is based on this sort of approach. We define the *false-discovery proportion* (FDP) and *non-discovery proportion* (NDP) as follows.

Definition II.1. *Let $\mathcal{S} := \{i \,:\, x_i \neq 0\}$ denote the signal support set and let $\widehat{\mathcal{S}} = \widehat{\mathcal{S}}(y)$ denote an estimator of $\mathcal{S}$. The false-discovery proportion is*

$$\text{FDP}(\widehat{\mathcal{S}}) \;:=\; \frac{|\widehat{\mathcal{S}} \backslash \mathcal{S}|}{|\widehat{\mathcal{S}}|}\;.$$

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*

$$\text{NDP}(\widehat{\mathcal{S}}) \;:=\; \frac{|\mathcal{S} \backslash \widehat{\mathcal{S}}|}{|S|}\;.$$

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.*

Note that the expectation of the FDP is the well-known False-discovery Rate (FDR). As an alternative to the NDP one could also consider the *False Non-discovery Proportion* (FNP), as suggested in [20], [21], defined as $|\mathcal{S} \backslash \widehat{\mathcal{S}}|/(p - |\widehat{\mathcal{S}}|)$. However, for very sparse signals (such that $|S| \ll p$) the FNP is not very sensitive, being always close to zero for any sparse support estimate.

In this paper we focus in particular on the scenario where $x_i \geq 0$ for all $i \in \{1, \ldots, p\}$. We elaborate on possible extensions in Section III. Under this assumption it is quite natural to focus on a specific class of estimators of $\mathcal{S}$.

Definition II.2. *A* coordinate-wise thresholding procedure *is an estimator of the following form:*

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

where the threshold $\tau > 0$ may depend implicitly on x, or on y itself.

The following result establishes the limits of localization using non-adaptive sampling. A proof is provided in Appendix A.

Theorem II.3. *Assume x has $p^{1-\beta}$, $\beta \in (0,1)$, non-zero components of amplitude $\sqrt{2r\log p}$, $r > 0$, and measurement model (3). There exists a coordinate-wise thresholding procedure that yields an estimator $\widehat{\mathcal{S}} = \widehat{\mathcal{S}}(y)$ such that if $r > \beta$, then as $p \to \infty$,*

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

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

We also refer the reader to recent related work in [3], which considered localization under similar error metrics as those utilized here. There it was shown, using a random signal model and assuming observations in the form of noisy independent random (Gaussian) linear combinations of the entries of x, that similar sharp asymptotics hold for any recovery procedure [3, Thm. 5].

A related problem of signal detection was considered in [4]–[6]. In this case the goal is not to estimate the signal support set, but simply test if a sparse signal is present or absent. More formally, testing the null hypothesis $x = 0$ versus a composite hypothesis, where x has $p^{1-\beta}$ non-zero components of amplitude $\mu(p)$. In this case it is possible to show that for $\beta > 1/2$, it is necessary to have $\mu(p) > c(\beta)\sqrt{\log p}$ with $c(\beta) > 0$, in order to simultaneously drive to zero the probabilities of Type I and Type II errors as $p \to \infty$; see [4]–[6].

In summary, under the above non-adaptive sampling scheme, it is necessary that the signal magnitude scales like $\sqrt{\log p}$ for localization, and for detection (when $\beta > 1/2$).

III. Main Results: Adaptive Localization and Detection of Sparse Signals

In this section we present the main results of our theoretical analysis of Distilled Sensing (DS). Algorithm 1 describes the DS measurement process. At each step of the process, we retain only the components with non-negative observations. This means that when the number of non-zero components is very small, roughly half of the components are eliminated from further consideration at each step. Consequently, if the precision budget allocated at each step is slightly larger than $1/2$ of that used in the preceding step, then the effective precision of the measurements made at each step is increasing. In particular, if the budget for each step is $1/2 + c$ of the budget at the previous step, for some small constant $c > 0$, then the precision of the measured components is increasing exponentially. Therefore, the key is to show that the very crude thresholding at 0 at each step does not remove a significant number of the non-zero components. One final observation is that because the number of components measured decreases by a factor of roughly $1/2$ at each step, the total number of measurements made by DS is roughly $2p$, a modest increase relative to the p measurements made in the non-adaptive setting.

Recall from above that for non-adaptive sampling, reliable detection and localization is only possible provided the signal amplitude is $\Omega(\sqrt{\log p})$. In other words, the signal amplitude must exceed a constant (that depends on the sparsity level) times $\sqrt{\log p}$. The following theorem establishes that DS is capable of detecting and localizing significantly weaker sparse signals. For the purposes of our investigation we assume that the non-zero components are positive. It is trivial to extend the algorithm and its analysis to handle both positive and negative components

Algorithm 1: Distilled Sensing.

Input:

Number of observation steps: k;

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 resources: $\gamma_{i,j} = \left\{ \begin{array}{cc} R_j/|I_j| & i \in I_j \\ 0 & i \notin I_j \end{array} \right\}$;

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: $y_k = \{y_{i,k} : i \in I_k\}$;

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.

Theorem III.1. *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 $k = k(p) = \max\{\lceil \log_2 \log p \rceil, 0\} + 2$, 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 support set estimator constructed using the output of the DS algorithm*

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

has the following properties:

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

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

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

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

where $\emptyset$ *is the empty set.*

The above theorem is stated in full generality, for a wide range of choices for the algorithm parameters c_1, c_k and $\{R_j\}_{j=1}^k$. In particular the result holds with the following simple choices:

$$R_j \quad = \quad (0.75)^{j-1} R_1\,,\ j = 2, \ldots, k-1\,, \tag{4}$$

$$R_k \quad = \quad R_1\,, \tag{5}$$

and R_1 chosen so that $\sum_{j=1}^k R_j = p$.

In words, this result states that DS successfully identifies the sparse signal support provided only that the signal amplitude grows (arbitrarily slowly) as a function of the problem dimension p, while reliable signal detection requires only that the signal amplitude exceed a constant. The result (ii) is entirely novel, and (i) improves on our initial result in [16] which required $\mu(p)$ to grow faster than an arbitrary iteration of the logarithm (i.e., $\mu(p) \sim \log\log\ldots\log p$). Comparison with the $\Omega(\sqrt{\log p})$ amplitude required for both tasks using non-adaptive sampling illustrates the significant gains that are achieved through adaptivity. In particular note that with adaptive sampling we effectively remove the dependence on the problem dimension p, that was unavoidable with non-adaptive sampling (embodied by the factor $\sqrt{\log p}$). Therefore DS is in essence overcoming the curse of dimensionality through the use of adaptive sampling.

Although the main results in Theorem III.1 are stated in terms of asymptotic (high-dimensional) limits, the key ingredients of its proof are bounds for finite-dimensional problems. Lemma IV.3, given in the next section, characterizes the finite dimensional performance of DS. The finite-dimensional performance will be thoroughly examined through numerical experiments in Section V. To illustrate how the results apply in finite dimensional situations and to provide a bit of intuition about the algorithm, we consider the following example.

Example III.2. *Let* $p = 2^{14}$ *and sparsity level* $s := |\mathcal{S}| = 2^7$*. These choices of dimension and sparsity level are not unreasonable in, for example, genomics testing problems. For this problem size, under the non-adaptive observation model (3) the magnitude of the largest 'noise-only' component is about* $\sqrt{2\log p} = 4.4$*. The implies that the signal strength* μ *needs to exceed this level in order to reliably distinguish the signal components from noise. Let us examine the performance of the DS algorithm at this critical level. Assume the precision budget allocation proposed above in (5) and let the number of steps* $k = \lceil \log_2 \log p \rceil + 2 = 5$*, as prescribed in Theorem III.1. Each step of the DS algorithm thresholds the observations at level* 0*, thereby removing* $1/2$ *of the null (noise-only) components (because the noise is symmetric about* 0*). After* 4 *steps about* $(p-s)/16$ *null components will remain (roughly* 6% *of the original total). This progressive thinning of the null components means that a increasing fraction of the remaining precision budget is allocated to signal (non-zero) components. For a signal strength of* $\mu = 4.4$ *it is easy calculate*

the probability that a signal component survives each step. At the end of the process the probability that a signal component is not removed about 0.84*; in other words about* 84% *of the* s *signal components are retained. In other words, the DS algorithm reduces the problem size by a factor of about* $1/16$ *while retaining most all of the signal components.*

IV. Analysis of Distilled Sensing

In this section we prove the main result characterizing the performance of Distilled Sensing (DS), Theorem III.1. We begin with three lemmas that quantify the finite sample behavior of DS.

A. Preliminary Lemmas

Lemma IV.1. *Let* $\{y_i\}_{i=1}^m$ *be independent random variables such that* $\Pr(y_i > 0) = 1/2$*. For any* $0 < \varepsilon < 1/2$*,*

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

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

Proof: For any event A, let $\mathbf{1}_A$ be the indicator taking the value 1 if A is true and 0 otherwise. By Hoeffding's inequality, for any $\varepsilon > 0$

$$\Pr\left(\left|\sum_{i=1}^m \mathbf{1}_{\{y_i>0\}} - \frac{m}{2}\right| > m\varepsilon\right) \leq 2\exp\left(-2m\varepsilon^2\right).$$

Imposing the restriction $\varepsilon < 1/2$ guarantees that the corresponding fractions are bounded away from zero and one. ∎

Lemma IV.2. *Let* $\{y_i\}_{i=1}^m \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$*, 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 $1 - \exp\left(-\frac{\mu m}{4\sigma\sqrt{2\pi}}\right)$*.*

Proof: We will utilize the following standard bound on the Gaussian tail: for $Z \sim \mathcal{N}(0,1)$ and $\gamma > 0$,

$$\begin{aligned}\frac{1}{\sqrt{2\pi\gamma^2}}\left(1 - \frac{1}{\gamma^2}\right)\exp(-\gamma^2/2) &\leq \Pr(Z > \gamma) \\ &\leq \frac{1}{\sqrt{2\pi\gamma^2}}\exp(-\gamma^2/2).\end{aligned}$$

Let $q = \Pr(y_i > 0)$, then it follows that

$$1 - q \;\leq\; \frac{\sigma}{\mu\sqrt{2\pi}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right).$$

Next we use the Binomial tail bound from [22]: for any $0 < b < \mathbb{E}[\sum_{i=1}^m \mathbf{1}_{\{y_i>0\}}] = mq$,

$$\Pr\left(\sum_{i=1}^m \mathbf{1}_{\{y_i>0\}} \leq b\right) \leq \left(\frac{m - mq}{m - b}\right)^{m-b}\left(\frac{mq}{b}\right)^b.$$

Note that $\epsilon > 1-q$ (or equivalently, $1-\epsilon < q$), so we can apply this result to $\sum_{i=1}^m \mathbf{1}_{\{y_i>0\}}$ with $b=(1-\epsilon)m$ to obtain

$$\begin{aligned}\Pr\left(\sum_{i=1}^m \mathbf{1}_{\{y_i>0\}} \le (1-\epsilon)m\right) & \\ &\le \left(\frac{1-q}{\epsilon}\right)^{\epsilon m}\left(\frac{q}{1-\epsilon}\right)^{(1-\epsilon)m} \\ &\le \exp\left(-\frac{\mu^2\epsilon m}{2\sigma^2}\right)\left(\frac{1}{1-\epsilon}\right)^{(1-\epsilon)m}.\end{aligned}$$

Now, to establish the stated result, it suffices to show

$$\begin{aligned}-\frac{\mu^2}{2\sigma^2}+\left(\frac{1-\epsilon}{\epsilon}\right)\log\left(\frac{1}{1-\epsilon}\right) &\le -\frac{\mu}{4\epsilon\sigma\sqrt{2\pi}} \\ &= -\frac{\mu^2}{4\sigma^2},\end{aligned}$$

which holds provided $\mu \ge 2\sigma$, since $0<\epsilon<1$ and $\left(\frac{1-\epsilon}{\epsilon}\right)\log\left(\frac{1}{1-\epsilon}\right) \le 1$ for $\epsilon \in (0,1)$. ■

These two lemmas allow us to quantify the output of the DS procedure. 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, for $j=1,\ldots,k$. Let $\varepsilon>0$, and for $j=1,\ldots,k-1$ define

$$\epsilon_j^2 := \frac{s_1+(1/2+\varepsilon)^{j-1}z_1}{2\pi\mu^2 R_j}\ , \tag{6}$$

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

Lemma IV.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*

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

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

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

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

Proof: The results follow from Lemmas IV.1 and IV.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 $\widetilde{\epsilon}_j := \frac{\sigma_j}{\mu\sqrt{2\pi}}$, $j=1,\ldots,k$.

The argument proceeds by conditioning on the output of all prior refinement steps; in particular, suppose that $(1-\widetilde{\epsilon}_{\ell-1})s_{\ell-1} \le s_\ell \le s_{\ell-1}$ and $\left(\frac{1}{2}-\varepsilon\right)z_{\ell-1} \le z_\ell \le \left(\frac{1}{2}+\varepsilon\right)z_{\ell-1}$for $\ell=1,\ldots,j$. Then apply Lemma IV.1 with

$m = z_j$, Lemma IV.2 with $m = s_j$ and $\sigma^2 = \sigma_j^2$, and the union bound to obtain that with probability at least

$$1 - \exp\left(-\frac{\mu s_j}{4\sigma_j\sqrt{2\pi}}\right) - 2\exp\left(-2z_j\varepsilon^2\right) , \tag{7}$$

$(1-\widetilde{\epsilon}_j)s_j \leq s_{j+1} \leq s_j$, 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 $R_j > \frac{4}{\mu^2}\left(s_1 + (1/2+\varepsilon)^{j-1}z_1\right)$ and the assumptions on prior refinement steps ensure that $\mu > 2\,\sigma_j$, which is required for Lemma IV.2. The condition $\mu > 2\,\sigma_j$ also allows us to simplify probability bound (7), 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

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

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. ■

Now we examine the conditions $R_j > \frac{4}{\mu^2}\left(s_1 + (1/2+\varepsilon)^{j-1}z_1\right)$, $j = 1, \ldots, k$ more closely. Define $c := s_1/[(1/2+\varepsilon)^{k-1}z_1]$, in effect condensing several problem-specific parameters (s_1, z_1, and k) into a single scalar parameter. 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

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

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. We are now in position to prove the main theorem.

B. Proof of Theorem III.1

Throughout the proof, whenever asymptotic notation or limits are used it is always under the assumption that $p \to \infty$, and we use the standard notation $f(p) = o(g(p))$ to indicate that $\lim_{p\to\infty} f(p)/g(p) = 0$, for $f(p) \geq 0$ and $g(p) > 0$. Also the quantities $k := k(p)$, $\varepsilon := \varepsilon(p)$ and $\mu := \mu(p)$ are functions of p, but we do not denote this explicitly for ease of notation. We let $\varepsilon := p^{-1/3}$ throughout the proof.

We begin by proving part (ii) of the theorem, which is concerned with detecting the presence or absence of a sparse signal. Part (i), which pertains to identifying the locations of the non-zero components, then follows with a slight modification.

Case 1 – Signal absent ($\mathcal{S} = \emptyset$): This is the simplest scenario, but through its analysis we will develop tools that will be useful when analyzing the case where the signal is present. Here, we have $s_1 = 0$ and $z_1 = p$, and the number of indices retained at the end of the DS procedure $|I_k|$ is equal to z_k. Define the event

$$\Gamma \;=\; \left\{ \left(\frac{1}{2}-\varepsilon\right)^{k-1} p \;\le\; |I_k| \;\le\; \left(\frac{1}{2}+\varepsilon\right)^{k-1} p \right\} \;.$$

The second part of Lemma IV.3 characterizes the probability of this event; in particular

$$\Pr(\Gamma) \;\ge\; 1 - 2\sum_{j=1}^{k-1} \exp\left(-2p\left(\frac{1}{2}-\varepsilon\right)^{j-1}\varepsilon^2\right).$$

Since $k \le \log_2 \log p + 3$, for large enough p we get that

$$\begin{aligned}
\Pr(\Gamma) &\ge 1 - 2(k-1)\exp\left(-2p\left(\frac{1}{2}-\varepsilon\right)^{k-2}\varepsilon^2\right) \\
&= 1 - 2(k-1)\exp\left(-p\left(\frac{1}{2}\right)^{k-3}(1-2\varepsilon)^{k-2}\varepsilon^2\right) \\
&\ge 1 - 2(\log_2\log p + 2)\exp\left(-\frac{p^{1/3}}{\log p}(1-o(1))\right)
\end{aligned}$$

where we used Lemma B.1 to conclude that $(1-2\varepsilon)^{k-2} = 1 - o(1)$. It is clear that $\Pr(\Gamma) \to 1$. The rest of the proof proceeds by conditioning on the event Γ. To avoid encumbering the notation we will often not explicitly the conditioning, making sure this is clear from the context.

In this case we assume that $\mathcal{S} = \emptyset$, therefore the output of the DS procedure consists of $|I_k|$ i.i.d. Gaussian random variables with zero mean and variance $|I_k|/R_k = |I_k|/(c_k p)$. Conditionally on Γ holding,

$$\begin{aligned}
|I_k| \;\le p\left(\frac{1}{2}+\varepsilon\right)^{k-1} &= p\frac{1}{2}\left(\frac{1}{2}\right)^{k-2}(1+2\varepsilon)^{k-1} \\
&\le \frac{1}{2}\frac{p}{\log p}(1+o(1)) \;,
\end{aligned}$$

which follows from the fact that $k \ge \log_2 \log p + 2$, and using Lemma B.1. With this we conclude (with a slight abuse of notation, as all the probabilities are conditional on Γ) that

$$\begin{aligned}
&\Pr(\widehat{\mathcal{S}}_{\mathrm{DS}} \neq \emptyset \mid \Gamma) \\
&\quad= \Pr\left(\exists_{i\in I_k} : y_{i,k} > \sqrt{2/c_k}\right) \\
&\quad\le |I_k| \Pr\left(\mathcal{N}(0, |I_k|/c_k p) > \sqrt{2/c_k}\right) \\
&\quad= |I_k| \Pr\left(\mathcal{N}(0,1) > \sqrt{2p/|I_k|}\right) \\
&\quad\le p \Pr\left(\mathcal{N}(0,1) > \sqrt{4\log p(1-o(1))}\right) \\
&\quad\le p\exp\left(-2\log p(1-o(1))\right) \\
&\quad= p^{-1+o(1)} \to 0 \;,
\end{aligned}$$

where the last inequality follows from the standard Gaussian tail bound. This together with $\Pr(\Gamma) \to 1$ immediately shows that when $\mathcal{S} = \emptyset$ we have $\Pr(\widehat{\mathcal{S}}_{\rm DS} \neq \emptyset) \to 0$.

Case 2 – Signal present ($\mathcal{S} \neq \emptyset$): The proof follows the same idea as in the previous case, although the argument is a little more involved. Begin by applying Lemma IV.3 and constructing an event that occurs with probability tending to one. Let Γ be the event

$$\begin{aligned}\Gamma &= \left\{ z_1 \left(\frac{1}{2} - \varepsilon\right)^{k-1} \;\leq\; z_k \;\leq\; z_1 \left(\frac{1}{2} + \varepsilon\right)^{k-1} \right\} \\ &\qquad \bigcap \quad \left\{ s_1 \prod_{j=1}^{k-1} (1 - \epsilon_j) \;\leq\; s_k \;\leq\; s_1 \right\} ,\end{aligned}$$

where ϵ_j is given by equation (6). Lemma IV.3 characterizes the probability of this event under a condition on R_j that we will now verify. Note that this condition is equivalent to $\epsilon_j^2 < 1/(8\pi)$ for all $j = 1, \ldots, k-1$. Instead of showing exactly this we will show a stronger result that will be quite useful in a later stage of the proof. Recall that $R_{j+1}/R_j \geq \delta > 1/2$, $j = 1, \ldots, k-2$, and $R_1 = c_1 p$ by the assumptions of the theorem. Thus for $j = 1, \ldots, k-1$

$$\begin{aligned}\epsilon_j^2 &\leq \frac{s_1 + \left(\frac{1}{2} + \varepsilon\right)^{j-1} z_1}{2\pi\mu^2 \delta^{j-1} R_1} \\ &\leq \frac{1}{2\pi\mu^2 c_1} \left(\frac{s_1}{p} \delta^{-(j-1)} + \frac{z_1}{p} \left(\frac{\delta}{\frac{1}{2} + \varepsilon} \right)^{-(j-1)} \right) .\end{aligned}$$

Clearly we have that $\epsilon_1^2 \leq \frac{1}{2\pi\mu^2 c_1} < 1/(8\pi)$ since by assumption $\mu > \sqrt{4/c_1}$. Now consider the case $j > 1$. Recall that $k \leq \log_2 \log p + 3$. Therefore if $\delta \geq 1$, then the term $\delta^{-(j-1)}$ can be upper bounded by 1, otherwise

$$\begin{aligned}\delta^{-(j-1)} &\leq \delta^{-(k-2)} \\ &\leq \delta^{-(\log_2 \log p + 1)} \\ &= \delta^{-1} (\log p)^{-\log_2 \delta} \\ &\leq 2 \log p \; , \end{aligned} \tag{8}$$

where the last step follows from $\delta > 1/2$.

Now recall that $s_1 = p^{1-\beta}$, therefore

$$\begin{aligned}\epsilon_j^2 &\leq \frac{1}{2\pi\mu^2 c_1} \left(p^{-\beta} \delta^{-(j-1)} + \left(\frac{\delta}{\frac{1}{2} + \varepsilon} \right)^{-(j-1)} \right) \\ &\leq \frac{1}{2\pi\mu^2 c_1} \left(2p^{-\beta} \log p + \left(\frac{\delta}{\frac{1}{2} + \varepsilon} \right)^{-(j-1)} \right) . \end{aligned} \tag{9}$$

Note that, since $\varepsilon \to 0$ as $p \to \infty$ we have that, for p large enough, $\delta/(1/2 + \varepsilon) > (\delta + 1/2 + \varepsilon)$. Assume p is large enough so that this is true, then

$$\epsilon_j^2 \leq \frac{1}{2\pi\mu^2 c_1} \left(2p^{-\beta} \log p + \left(\delta + \frac{1}{2} + \varepsilon \right)^{-(j-1)} \right) .$$

Clearly since $j \leq k-1 \leq \log_2 \log p + 2$ we have that $\left(\delta + \frac{1}{2} + \varepsilon\right)^{-(j-1)} = \Omega\left(1/(\log p)^{\log_2(\delta+1/2+\epsilon)}\right)$ and so the first of the additive terms in (9) is negligible for large p. Therefore for p sufficiently large, we have, for all $j = 1, \ldots, k-1$

$$\epsilon_j^2 \;\leq\; \frac{1}{2\pi\mu^2 c_1}\left(\delta + \frac{1}{2}\right)^{-(j-1)} \quad . \tag{10}$$

Since by assumption $\mu > \sqrt{4/c_1}$, we conclude that, for all p sufficiently large, $\epsilon_j^2 < 1/(8\pi)$ for all $j = 1, \ldots, k-1$, and so $R_j > \frac{4}{\mu^2}\left(s_1 + (1/2+\varepsilon)^{j-1} z_1\right)$ for $j = 1, \ldots, k-1$. Thus, applying Lemma IV.3 we have

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

By a similar argument to that used in Case 1, it is straightforward to show that

$$2\sum_{j=1}^{k-1} \exp\left(-2z_1(1/2-\varepsilon)^{j-1}\varepsilon^2\right) \to 0 \ .$$

In addition,

$$\begin{aligned}
&\sum_{j=1}^{k-1} \exp\left(\frac{-s_1 \prod_{\ell=1}^{j-1}(1-\epsilon_\ell)}{\sqrt{8\pi}}\right) \\
&\quad\leq\quad (k-1)\exp\left(\frac{-s_1 \prod_{\ell=1}^{k-2}(1-\epsilon_\ell)}{\sqrt{8\pi}}\right) \\
&\quad\leq\quad (k-1)\exp\left(\frac{-s_1 \prod_{\ell=1}^{k-2}\left(1-\frac{(\delta+1/2)^{-(\ell-1)/2}}{\mu\sqrt{2\pi c_1}}\right)}{\sqrt{8\pi}}\right) \\
&\quad\leq\quad (k-1)\exp\left(\frac{-s_1 \prod_{\ell=1}^{k-2}\left(1-\frac{(\delta+1/2)^{-(\ell-1)/2}}{\sqrt{8\pi}}\right)}{\sqrt{8\pi}}\right) \ ,
\end{aligned}$$

where in the last step we used the fact that $\mu > \sqrt{4/c_1}$. Finally note that from Lemma B.2 we know that

$$\prod_{\ell=1}^{k-2}\left(1 - \frac{1}{\sqrt{8\pi}}\left(\delta + \frac{1}{2}\right)^{-(\ell-1)/2}\right) \to L(\delta) \ ,$$

where $L(\delta) > 0$ hence

$$\begin{aligned}
&\sum_{j=1}^{k-1} \exp\left(\frac{-s_1 \prod_{\ell=1}^{j-1}(1-\epsilon_\ell)}{\sqrt{8\pi}}\right) \\
&\quad\leq\quad (\log_2 \log p + 2)\exp\left(\frac{-p^{1-\beta}(L(\delta)+o(1))}{\sqrt{8\pi}}\right) \\
&\quad\to\quad 0 \ .
\end{aligned} \tag{11}$$

Therefore we conclude that the event Γ happens with probability converging to one.

We now proceed as before, by conditioning on event Γ. The output of the DS procedure consists of a total of $|I_k| = s_k + z_k$ independent Gaussian measurements with variance $|I_k|/R_k$, where s_k of them have mean μ and the

remaining z_k have mean zero. We will show that the proposed thresholding procedure identifies only true non-zero components (i.e., correctly rejects all the zero-valued components). In other words, with probability tending to one, $\widehat{\mathcal{S}}_{\rm DS} = \mathcal{S} \cap I_k$. For ease of notation, and without loss of generality, assume the $y_{i,k} \sim \mathcal{N}(\mu, |I_k|/R_k)$ for $i \in \{1, \ldots, s_k\}$ and $y_{i,k} \sim \mathcal{N}(0, |I_k|/R_k)$ for $i \in \{s_k + 1, \ldots, |I_k|\}$. Then

$$\begin{aligned}
&\Pr\left(\widehat{\mathcal{S}}_{\rm DS} \neq \mathcal{S} \cap I_k \;\middle|\; \Gamma\right) \\
&\quad = \Pr\left(\bigcup_{i=1}^{s_k}\left\{y_{i,k} < \sqrt{2/c_k}\right\}\right. \\
&\qquad\qquad \left. \text{or} \bigcup_{i=s_k+1}^{|I_k|}\left\{y_{i,k} > \sqrt{2/c_k}\right\} \;\middle|\; \Gamma\right) \\
&\quad \leq s_k \Pr\left(\mathcal{N}(\mu, |I_k|/R_k) < \sqrt{2/c_k}\right) \\
&\qquad\qquad + z_k \Pr\left(\mathcal{N}(0, |I_k|/R_k) > \sqrt{2/c_k}\right).
\end{aligned}$$

Note that conditioned on the event Γ (using arguments similar to those in Case 1)

$$\begin{aligned}
|I_k| &= s_k + z_k \\
&\leq s_1 + z_1\left(\frac{1}{2} + \varepsilon\right)^{k-1} \\
&\leq p^{1-\beta} + \frac{p}{2\log p}(1 + o(1)) \\
&\leq \frac{p}{2\log p}(1 + o(1)) \ . \qquad (12)
\end{aligned}$$

Finally, taking into account that $\mu > 2\sqrt{2/c_k}$ we conclude that

$$\begin{aligned}
&\Pr\left(\widehat{\mathcal{S}}_{\rm DS} \neq \mathcal{S} \cap I_k \middle| \Gamma\right) \\
&\quad \leq s_k \Pr\left(\mathcal{N}(0, |I_k|/R_k) < -\sqrt{2/c_k}\right) \\
&\qquad\qquad + z_k \Pr\left(\mathcal{N}(0, |I_k|/R_k) > \sqrt{2/c_k}\right) \\
&\quad \leq s_k \Pr\left(\mathcal{N}(0,1) > \sqrt{\frac{2p}{|I_k|}}\right) \\
&\qquad\qquad + z_k \Pr\left(\mathcal{N}(0,1) > \sqrt{\frac{2p}{|I_k|}}\right) \\
&\quad = |I_k| \Pr\left(\mathcal{N}(0,1) > \sqrt{\frac{2p}{|I_k|}}\right) \\
&\quad \leq p \Pr\left(\mathcal{N}(0,1) > \sqrt{4\log p(1 - o(1))}\right) \\
&\quad \leq p \exp\left(-2\log p(1 - o(1))\right) \\
&\quad = p^{-1+o(1)} \to 0 \ ,
\end{aligned}$$

where the last inequality follows from the standard Gaussian tail bound. Note again the slight abuse of notation, as all the probabilities are conditioned on Γ. The above together with $\Pr(\Gamma) \to 1$, and the fact that $|\mathcal{S} \cap I_k| =$

$s_k = L(\delta)(1 - o(1))s_1$ is bounded away from zero for large enough p immediately shows that $\Pr(\widehat{\mathcal{S}}_{\rm DS} = \emptyset) \to 0$, concluding the proof of part (ii) of the theorem.

Part (i) of the theorem follows from the result proved above, since if μ is any positive diverging sequence in p then a stronger version of Lemma B.2 applies. In particular, recall (10), and note that Lemma B.2 implies

$$\prod_{\ell=1}^{k-1}(1-\epsilon_\ell) \;\geq\; \prod_{\ell=1}^{k-1}\left(1 - \frac{1}{\mu\sqrt{2\pi c_1}}\left(\delta + \frac{1}{2}\right)^{-\frac{\ell-1}{2}}\right) \to 1 \ .$$

We have already established that the events Γ and $\{\widehat{\mathcal{S}}_{\rm DS} \neq \mathcal{S} \cap I_k\}$ both hold (simultaneously) with probability tending to one. Conditionally on these events we have

$$\text{FDP}(\widehat{\mathcal{S}}_{\rm DS}) = \frac{0}{s_k} = 0 \ ,$$

and

$$\text{NDP}(\widehat{\mathcal{S}}_{\rm DS}) = \frac{s_1 - s_k}{s_1} = 1 - \frac{s_k}{s_1} \to 0 \ ,$$

since from the definition of Γ we have

$$s_1 \;\geq\; s_k \;\geq\; s_1 \prod_{\ell=1}^{k-1}(1-\epsilon_\ell) \to s_1 \ .$$

Therefore we conclude that both FDP($\widehat{\mathcal{S}}_{\rm DS}$) and NDP($\widehat{\mathcal{S}}_{\rm DS}$) converge in probability to zero as $p \to \infty$, concluding the proof of the theorem.

V. Numerical Experiments

This section presents numerical experiments with Distilled Sensing (DS). The results demonstrate that the asymptotic analysis predicts the performance in finite dimensional cases quite well. Furthermore, the experiments suggest useful rules of thumb for implementing DS in practice.

There are two input parameters to the DS procedure; the number of distillation steps, k, and the distribution of precision across the steps, $\{R_j\}_{j=1}^k$. Throughout our simulations we choose $k = \max\{\lceil \log_2 \log p \rceil, 0\} + 2$, as prescribed in Theorem III.1. For the precision distribution, first recall the discussion following the proof of Lemma IV.3. There it is argued that if the sparsity model is valid, a sufficient condition for the precision distribution is $R_j > R_1(1/2 + \varepsilon)^{j-1}$, $j = 1, \ldots, k$, with $0 < \varepsilon < 1/2$. In words, the precision allocated to each step must be greater than $1/2$ the precision allocated in the previous step. In practice, we find that choosing $R_{j+1}/R_j = 0.75$ for $j = 1, \ldots, k-2$ provides good performance over the full SNR range of interest. Also, from the proof of the main result (Theorem III.1) we see that the threshold for detection is inversely proportional to the square root of the precision allocation in the first and last steps. Thus, we have found that allocating equal precision in the first and last steps is beneficial. The intuition is that the first step is the most crucial in controlling the NDP and the final step is most crucial in controlling the FDP. With this in mind, the precision allocation used throughout the simulations follows the simple formulas (4) (5).

Figure 1 compares the FDP vs. NDP performance of the DS procedure to non-adaptive (single observation) measurement at several signal-to-noise ratios (SNR = μ^2). We consider signals of length $p = 2^{14}$ having $\sqrt{p} = 128$

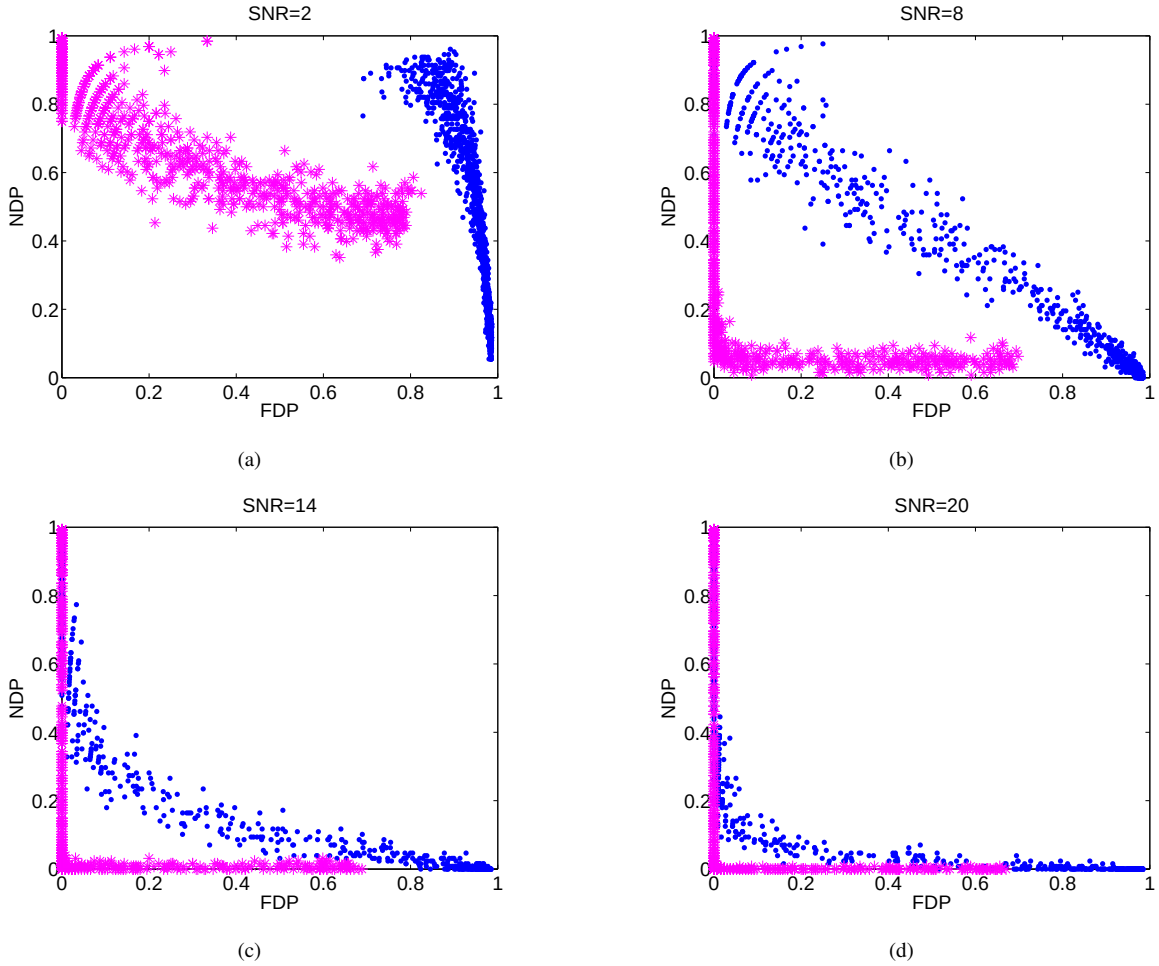

Fig. 1. FDP and NDP performance for DS (indicated with $*$) and non-adaptive sensing (indicated with $\bullet$) at different SNRs. Smaller values of FDP and NDP correspond to more accurate recovery (ie, exact support recovery occurs when NDP $=$ FDP $= 0$). The results clearly show that DS outperforms non-adaptive sensing for each SNR examined.

non-zero components with uniform amplitude with locations chosen uniformly at random. This choice of signal dimension corresponds to $k = 6$ observation steps in the DS procedure. The choice of sparsity level $\beta = 1/2$ is somewhat arbitrary, and extensive experimentation was done for other sparsity levels. The range of FDP-NDP operating points is surveyed by varying the threshold applied to the non-adaptive measurements and the output of the DS procedure for each of 1000 trials, corresponding to different realizations of randomly-generated signal and additive noise. Recall that largest squared magnitude in a realization of p i.i.d. $\mathcal{N}(0,1)$ variables grows like $2\log p$, and in our experiment, $2\log p \approx 20$. Consequently, when the SNR $= 20$ we see that both DS and non-adaptive measurements are highly successful, as expected. Another SNR level of interest is 8, since in this case this happens to approximately satisfy the condition $\mu = \sqrt{2/c_1} = \sqrt{2p/R_1}$, which according to the Theorem III.1 is a critical level for detection using DS. The simulations show that DS remains highly successful at this level while the non-adaptive results are poor. Finally, when the SNR $= 2$, we see that DS still yields useful results. For example, at

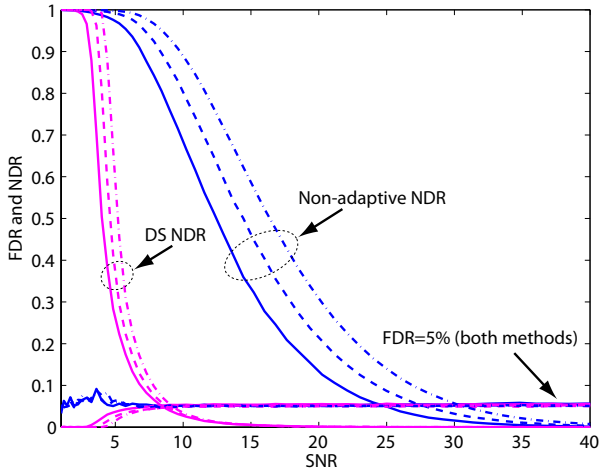

Fig. 2. FDR and NDR vs. SNR comparison. The solid, dashed, and dash-dot lines correspond to signals of length $p = 2^{14}, 2^{17}$, and 2^{20}, respectively, having $\lfloor p^{1/2} \rfloor$ non-zero entries. At each value of SNR and for each method (DS and non-adaptive sampling), thresholds were selected to achieve FDR $= 0.05$. Lower values of NDR correspond to more accurate recovery; DS clearly outperforms non-adaptive sensing over the entire SNR range and shows much less dependence on the signal dimension p.

FDP $= 0.05$, the DS procedure has an average NDP of roughly 80% (i.e., 20% of the true components are still detected, on average). This demonstrates the approximate $\log p$ extension of the SNR range provided by DS. Note the gap in the FDP values of the DS results (roughly from 0.75 to 1). The gap arises because the the output of DS has a higher SNR and is much less sparse than the original signal, and so arbitrarily large FDP values cannot be achieved by any choice of threshold. Large FDP values are, of course, of little interest in practice. We also remark on the structured patterns observed in cases of high NDP and low FDP (in upper left of figures for SNR $= 2$ and SNR $= 8$). The visually structured 'curves' of NDP-FDP pairs arise when the total number of discoveries is small, and hence the FDP values are restricted to certain rational numbers. For example, if just 3 components are discovered, then the number of false-discoveries can only take the values 0, $1/3$, $2/3$, and 1.

Figure 2 compares the performance of non-adaptive sensing and the DS procedure in terms of the false-discovery rate (FDR) and the non-discovery rate (NDR), which are the average FDP and NDP, respectively. We consider three different cases, corresponding to signals of length $p = 2^{14}, 2^{17}$, and 2^{20}, (the solid, dashed, and dash-dot lines, respectively) where for each case the number of non-zero signal components is $\lfloor p^{1/2} \rfloor$. The precision allocation and number of observation steps are chosen as described above (here, $k = 6$ for each of the three cases). For each value of SNR, 500 independent experiments were performed for DS and non-adaptive sampling, and in each, thresholds were selected so that the FDRs were fixed at approximately 0.05. The resulting average FDRs and NDRs for each SNR level are shown. The results show that 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.

Figure 3 provides another comparison of the performance of DS and the non-adaptive procedure, across several values of the ambient signal dimension p. We consider four different settings, corresponding to signals of length

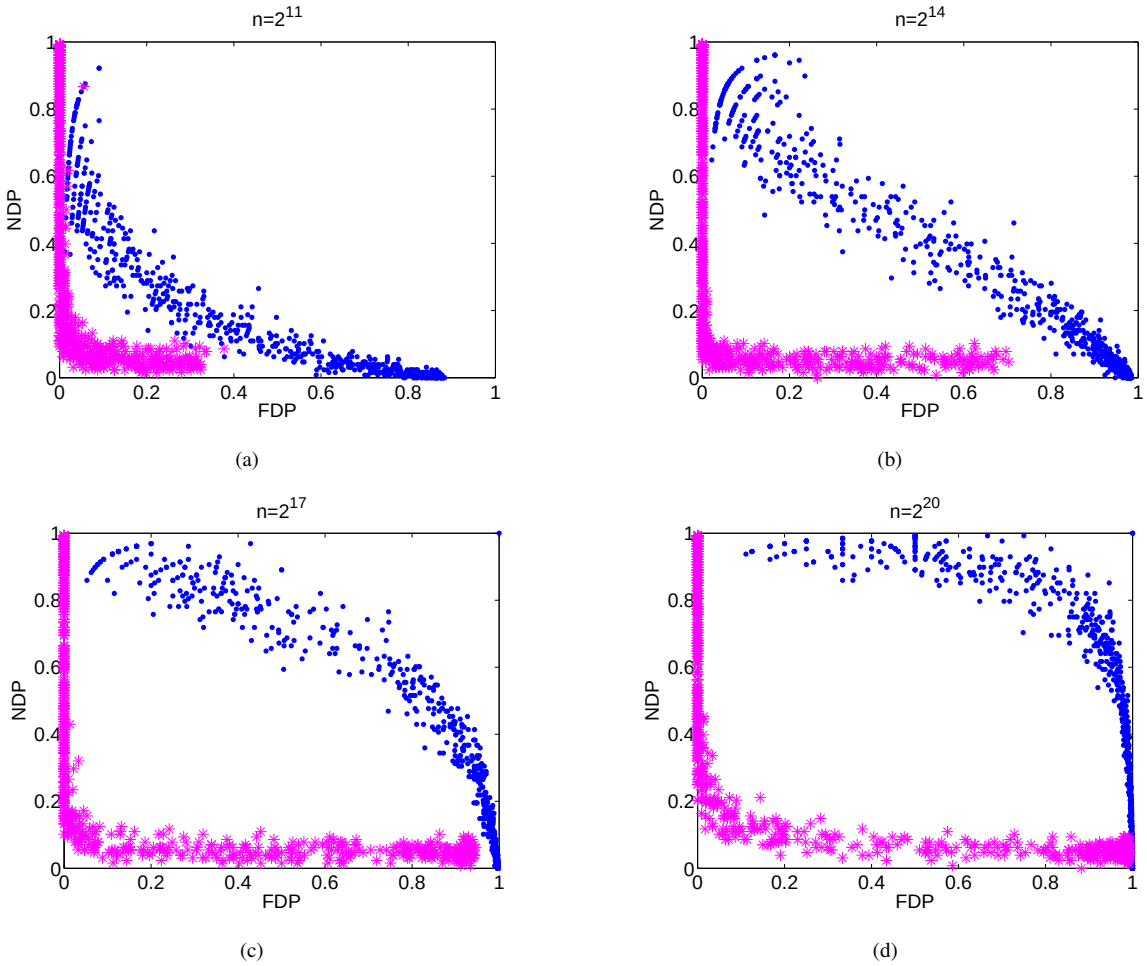

Fig. 3. FDP and NDP performance for DS (indicated with $*$) and non-adaptive sensing (indicated with $\bullet$) for different values of the signal dimension. DS shows little dependence on the increasing ambient signal dimension across 9 orders of magnitude, while the non-adaptive sensing procedure degrades markedly. This demonstrates the effectiveness of DS at mitigating the "curse of dimensionality".

$p = 2^{11}, 2^{14}, 2^{17}$ and 2^{20}. In each case the number of non-zero signal components is fixed to be $m = 128$. The precision allocation and number of observation steps are chosen as in the previous to simulation settings. Here, the number of refinement steps $k = 5$ for $p = 2^{11}$ and $k = 6$ for $p = 2^{14}, 2^{17}$, and 2^{20}. The SNR in each case is fixed by the choice $\mu^2 = 8$. As above, we examine the range of FDP-NDP operating points by varying the threshold applied to the non-adaptive measurements and the output of the DS procedure, for each of 1000 trials. The results clearly demonstrate that DS shows much less dependence on the ambient signal dimension, as its performance is relatively unchanged across 9 orders of magnitude. The non-adaptive procedure, on the other hand, degrades markedly as the dimension of the signal increases. This demonstrates the ability of adaptive sampling to mitigate the "curse of dimensionality" in these sparse testing problems.

VI. 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 [1], [2], [4]–[6]. However, in practice and theory, multistage adaptive designs have shown promise [7]–[10]. This paper quantifies the improvements such methods can achieve. We proposed and 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 roughly $\log p$ lower than the minimum required by non-adaptive methods, where p is the problem dimension. To put this in context, suppose one is interested in screening $p = 20,000$ genes, then $\log p \approx 10$. Thus, the gains can be quite significant in problem sizes of practical interest, which is why experimentalists often do employ similar methods.

An additional point worthy of future investigation is the development of lower bounds, characterizing the minimum amplitude $\mu(p)$ below which signal detection and localization are impossible for *any* sensing procedure (including adaptive sensing). In general, lower bounds are difficult to devise for sequential experimental design settings, with a few notable exceptions [23], [24]. Here, our results establish that significant improvements are achievable using adaptivity, although we relegate any general claims of optimality for adaptive sensing procedures to future work.

There are several possible extensions to DS. One is to consider even sparser signal models, where the number of nonzero entries is significantly smaller than $p^{1-\beta}$ for $\beta \in (0,1)$, as considered here. In particular, the same asymptotic results stated here follow also for signals whose sparsity levels are as small as a constant times $\log\log\log p$. Indeed, making this choice of s_1 in (9) leads to the same bound on the ϵ_j^2 given in (10), and this choice is also sufficient to ensure that (11) holds as well. In addition, for this choice of s_1 the same bound is obtained in (12), and the rest of the proof goes through as stated. Another extension is to use DS with alternate measurement models. For example, each measurement could be a linear combination of the entries of x, rather than direct measurements of individual components. If the linear combinations are non-adaptive, this leads to a regression model commonly studied in the Lasso and Compressed Sensing literature—see, for example, [25], [26]. 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 [27].

Appendix A
Thresholds for Non-Adaptive Recovery

In this section we give a proof of Theorem II.3. We will proceed by considering two cases separately: (i) $r > \beta$ and (ii) $r < \beta$. The analysis of the phase transition point $r = \beta$ is interesting, but it is beyond the scope of this paper. Begin by noticing that in the setting of the theorem the minimax optimal support estimation procedure to control the false and non-discovery proportions is a simple coordinate-wise thresholding procedure of the form

$$\widehat{\mathcal{S}} = \{i : y_i > \tau\} \ ,$$

where $\tau \geq 0$ can be chosen appropriately. A formal proof of this optimality can be done by noting that the class of hypothesis is invariant under permutations (see [4], [5] for details).

Case (i) $r > \beta$**:** In this case the signal support can be accurately identified from the observations, in the sense that $\mathrm{FDP}(\widehat{\mathcal{S}})$ and $\mathrm{NDP}(\widehat{\mathcal{S}})$ both converge in probability to zero. For this case we will take $\tau = \tau(p) = \sqrt{2\alpha \log p}$, where $\beta < \alpha < r$.

Begin by defining D_z and M_s to be the number of retained non-signal components and the number of missed signal components, respectively. Formally

$$D_z = \sum_{i=1}^{p} \mathbf{1}_{\{y_i > \tau,\ x_i = 0\}} \ ,$$

and

$$M_s = \sum_{i=1}^{p} \mathbf{1}_{\{y_i \leq \tau,\ x_i \neq 0\}} \ .$$

Note that D_z is binomially distributed, that is $D_z \sim \mathrm{Bin}(p(1-p^{-\beta}), q_z)$, where $q_z = \Pr(y_i > \tau)$ when i is such that $x_i = 0$. By noticing that $\tau > 0$ and using a standard Gaussian tail bound we have that

$$q_z \leq \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{\tau^2}{2}\right) = \frac{1}{\sqrt{4\pi\alpha \log p}} p^{-\alpha}.$$

In a similar fashion note that $M_s \sim \mathrm{Bin}(p^{1-\beta}, q_s)$, where $q_s = \Pr(y_i \leq \tau)$ when i is such that $x_i = \sqrt{2r \log p}$. Let $Z \sim \mathcal{N}(0,1)$ be an auxiliary random variable. Then

$$\begin{aligned} q_s &= \Pr(Z + \sqrt{2r \log p} \leq \tau) \\ &= \Pr(Z \leq \tau - \sqrt{2r \log p}) \\ &= \Pr(Z > \sqrt{2 \log p}(\sqrt{r} - \sqrt{\alpha})) \ , \end{aligned}$$

And so, using the Gaussian tail bound we have

$$q_s \leq \frac{1}{\sqrt{4\pi \log p}(\sqrt{r} - \sqrt{\alpha})} p^{-(\sqrt{r} - \sqrt{\alpha})^2} \ .$$

We are ready to show that both $\mathrm{FDP}(\widehat{\mathcal{S}})$ and $\mathrm{NDP}(\widehat{\mathcal{S}})$ converge in probability to zero. Begin by noticing that $\mathrm{NDP}(\widehat{\mathcal{S}}) = M_s / p^{1-\beta}$. By definition $\mathrm{NDP}(\widehat{\mathcal{S}}) \xrightarrow{P} 0$ means that for all fixed $\epsilon > 0$,

$$\Pr(|\mathrm{NDP}(\widehat{\mathcal{S}})| > \epsilon) \to 0 \ ,$$

as $p \to \infty$. Noting that $\mathrm{NDP}(\widehat{\mathcal{S}})$ is non-negative, this can be easily established using Markov's inequality.

$$\begin{aligned} \Pr(\mathrm{NDP}(\widehat{\mathcal{S}}) > \epsilon) &= \Pr\left(\frac{M_s}{p^{1-\beta}} \geq \epsilon\right) \\ &= \Pr(M_s > \epsilon p^{1-\beta}) \\ &\leq \frac{\mathbb{E}[M_s]}{\epsilon p^{1-\beta}} \\ &= \frac{p^{1-\beta} q_s}{\epsilon p^{1-\beta}} = \frac{q_s}{\epsilon} \to 0 \ , \end{aligned}$$

as $p \to \infty$ as clearly q_s converges to zero (since $r > \alpha$). For the false discovery proportion the reasoning is similar. Note that the number of correct discoveries is $p^{1-\beta} - M_s$. Taking this into account we have

$$\mathrm{FDP}(\widehat{\mathcal{S}}) = \frac{D_z}{p^{1-\beta} - M_s + D_z} \ .$$

Let $\epsilon > 0$. Then

$$\begin{aligned}
&\Pr(\mathrm{FDP}(\widehat{\mathcal{S}}) > \epsilon) \\
&\quad = \Pr\left(\frac{D_z}{p^{1-\beta} - M_s + D_z} > \epsilon\right) \\
&\quad = \Pr\left((1-\epsilon)\frac{D_z}{p^{1-\beta}} + \epsilon\frac{M_s}{p^{1-\beta}} > \epsilon\right) \\
&\quad \leq \frac{\mathbb{E}\left[(1-\epsilon)\frac{D_z}{p^{1-\beta}} + \epsilon\frac{M_s}{p^{1-\beta}}\right]}{\epsilon} \\
&\quad = \frac{1-\epsilon}{\epsilon}\frac{p(1-p^{-\beta})q_z}{p^{1-\beta}} + q_s \\
&\quad \leq \frac{1-\epsilon}{\epsilon}p^{\beta}(1-p^{-\beta})\frac{1}{\sqrt{4\pi\alpha\log p}}p^{-\alpha} \\
&\qquad\qquad + \frac{1}{\sqrt{4\pi\log p}(\sqrt{r}-\sqrt{\alpha})}p^{-(\sqrt{r}-\sqrt{\alpha})^2} \ ,
\end{aligned}$$

where the last line clearly converges to zero as $p \to \infty$, since $\beta < \alpha < r$. Therefore we conclude that $\mathrm{FDP}(\widehat{\mathcal{S}})$ converges to zero in probability.

Case (ii) $r < \beta$**:** In this case we will show that no thresholding procedure can simultaneously control the false and non-discovery proportions. Begin by noting that the smaller τ is, the easier it is to control the non-discovery proportion. In what follows we will identify an upper-bound on τ necessary for the control of the non-discovery rate. Note that if $\tau = \tau(p) = \sqrt{2r\log p}$ then $q_s = 1/2$, and therefore

$$\mathrm{NDP}(\widehat{\mathcal{S}}) = \frac{M_s}{p^{1-\beta}} \overset{a.s.}{\to} 1/2 \ ,$$

as $p \to \infty$, by the law of large numbers. Therefore a necessary, albeit insufficient, condition for $\mathrm{NDP}(\widehat{\mathcal{S}}) \overset{P}{\to} 0$ is that for all but finitely many p

$$\tau < \sqrt{2r\log p} \ . \tag{13}$$

Similarly, note that the larger τ is, the easier it is to control the false discovery rate. In the same spirit of the above derivation we will identify a lower-bound for τ that must necessarily hold in order to control the false-discovery rate. Recall the previous derivation, where we showed that, for any $\epsilon > 0$

$$\begin{aligned}
\Pr(\mathrm{FDP}(\widehat{\mathcal{S}}) > \epsilon) &= \Pr\left((1-\epsilon)\frac{D_z}{p^{1-\beta}} + \epsilon\frac{M_s}{p^{1-\beta}} \geq \epsilon\right) \\
&\geq \Pr\left((1-\epsilon)\frac{D_z}{p^{1-\beta}} \geq \epsilon\right) \\
&= \Pr\left(\frac{D_z}{p^{1-\beta}} \geq \frac{\epsilon}{1-\epsilon}\right) \ ,
\end{aligned}$$

where the last inequality follows trivially given that $M_s \geq 0$ and, without loss of generality, we assume that $\epsilon < 1$. This means that $\mathrm{FDP}(\widehat{\mathcal{S}})$ converges in probability to zero only if $\frac{D_z}{p^{1-\beta}}$ also converges in probability to zero. Namely, for any $\epsilon > 0$ we must have $\lim_{p\to\infty} \Pr(D_z/p^{1-\beta} > \epsilon) = 0$. In what follows take $\tau = \sqrt{2r\log p}$. Let $\epsilon > 0$ and note that

$$\begin{aligned}\Pr\left(\frac{D_z}{p^{1-\beta}} > \epsilon\right) &= \Pr(D_z > \epsilon p^{1-\beta}) \\ &= \Pr(D_z - \mathbb{E}[D_z] > \epsilon p^{1-\beta} - \mathbb{E}[D_z]) \\ &= \Pr(D_z - \mathbb{E}[D_z] > \epsilon p^{1-\beta} - p(1-p^{-\beta})q_z) \ .\end{aligned}$$

Define $a = \epsilon p^{1-\beta} - p(1-p^{-\beta})q_z$. Note that by the Gaussian tail bound, we have

$$\frac{1}{\sqrt{4\pi r\log p}}\left(1 - \frac{1}{2r\log p}\right)p^{-r} \leq q_z \leq \frac{1}{\sqrt{4\pi r\log p}}p^{-r} \ ,$$

or equivalently,

$$q_z = \frac{1-o(1)}{\sqrt{4\pi r\log p}}p^{-r}.$$

Given this it is straightforward to see that

$$\begin{aligned} a &= \epsilon p^{1-\beta} - (1-o(1))\frac{p(1-p^{-\beta})}{\sqrt{4\pi r\log p}}p^{-r} \\ &= \epsilon p^{1-\beta} - (1-o(1))\frac{p^{1-r}}{\sqrt{4\pi r\log p}} \\ &= p^{1-r}\left(\epsilon p^{r-\beta} - \frac{1-o(1)}{\sqrt{4\pi r\log p}}\right) \\ &= -\frac{p^{1-r}}{\sqrt{4\pi r\log p}}(1-o(1)) \ ,\end{aligned}$$

where in the last step we use the assumption that $\beta > r$. Therefore $a \to -\infty$ as p goes to infinity. Let $p_0(\epsilon) \in \mathbb{N}$ be such that $a < 0$ for all $p \geq p_0(\epsilon)$. Then

$$\begin{aligned}\Pr(D_z/p^{1-\beta} > \epsilon) &= \Pr(D_z - \mathbb{E}[D_z] > a) \\ &= 1 - \Pr(D_z - \mathbb{E}[D_z] \leq a) \\ &\geq 1 - \Pr(|D_z - \mathbb{E}[D_z]| \geq -a) \\ &\geq 1 - \frac{\mathrm{Var}(D_z)}{(-a)^2} \ ,\end{aligned}$$

where $\mathrm{Var}(D_z) = p(1-p^{-\beta})q_z(1-q_z)$ is the variance of D_z and the last step uses Chebyshev's inequality. Recalling that $p \geq p_0(\epsilon)$ we can examine the last term in the above expression easily.

$$\begin{aligned}1 - \frac{\mathrm{Var}(D_z)}{(-a)^2} &= 1 - (1-q_z)\frac{p(1-p^{-\beta})q_z}{a^2} \\ &= 1 - (1-o(1))\frac{pq_z}{a^2} \\ &= 1 - (1-o(1))\frac{p^{1-r}}{\sqrt{4\pi r\log p}}\frac{4\pi r\log p}{p^{2-2r}} \\ &= 1 - (1-o(1))\frac{\sqrt{4\pi r\log p}}{p^{1-r}} \to 1 \ ,\end{aligned}$$

as $p \to \infty$. Therefore we conclude that, for $\tau = \sqrt{2r \log p}$, $D_z/p^{1-\beta}$ does not converge in probability to zero, and therefore $\mathrm{FDP}(\widehat{\mathcal{S}})$ also does not converge to zero.

The above result means that a necessary condition for the convergence of $\mathrm{FDP}(\widehat{\mathcal{S}})$ to zero is that for all but finitely many p

$$\tau > \sqrt{2r \log p} \ .$$

This, together with (13) shows that there is no thresholding procedure capable of controlling both the false-discovery and non-discovery proportions when $r < \beta$ as we wanted to show, concluding the proof.

APPENDIX B
AUXILIARY MATERIAL

Lemma B.1. *Let $0 \leq f(p) \leq 1/2$ and $g(p) \geq 0$ be any sequences in p such that $\lim_{p\to\infty} f(p)g(p) = 0$. Then*

$$\lim_{p\to\infty} (1 + f(p))^{g(p)} = \lim_{p\to\infty} (1 - f(p))^{g(p)} = 1 \ .$$

Proof: To establish that $\lim_{p\to\infty}(1 + f(p))^{g(p)} = 1$ note that

$$\begin{aligned} 1 &\leq (1 + f(p))^{g(p)} \\ &= \exp\left(g(p) \log(1 + f(p))\right) \\ &\leq \exp\left(g(p)f(p)\right) \ , \end{aligned}$$

where the last inequality follows from $\log(1 + x) \leq x$ for all $x \geq 0$. As $g(p)f(p) \to 0$ we conclude that $\lim_{p\to\infty}(1 + f(p))^{g(p)} = 1$.

The second part of the result is established in a similar fashion. Note that

$$\begin{aligned} \log\left(1 - f(p)\right) &= -\log\left(\frac{1}{1 - f(p)}\right) \\ &= \log\left(1 + \frac{f(p)}{1 - f(p)}\right) \\ &\geq -\frac{f(p)}{1 - f(p)} \geq -2f(p) \end{aligned}$$

where the last inequality relies on the fact that $f(p) \leq 1/2$. Using this fact we have that

$$\begin{aligned} 1 &\geq (1 - f(p))^{g(p)} \\ &= \exp\left(g(p) \log(1 - f(p))\right) \\ &\geq \exp\left(-2f(p)g(p)\right) \ . \end{aligned}$$

Taking into account that $g(p)f(p) \to 0$ establishes the desired result. ■

Lemma B.2. *Let $k = k(p)$ be a positive integer sequence in p, and let $g = g(p)$ be a positive nondecreasing sequence in p. For some fixed $a > 1$ let $\epsilon_j = \epsilon_j(p) \leq a^{-j}/g(p)$. If $g(p) > a^{-1}(1 + \eta)$, for some fixed $\eta > 0$, then*

$$\lim_{p\to\infty} \prod_{j=1}^{k(p)} (1 - \epsilon_j(p)) > 0.$$

If, in addition, $g(p)$ is any positive monotone diverging sequence in p, then

$$\lim_{p\to\infty} \prod_{j=1}^{k(p)} (1-\epsilon_j(p)) = 1.$$

Proof: Note that

$$\begin{aligned}
\log\left(\prod_{j=1}^{k(p)} (1-\epsilon_j(p))\right) &\geq \sum_{j=1}^{k(p)} \log\left(1-\frac{a^{-j}}{g(p)}\right) \\
&= -\sum_{j=1}^{k(p)} \log\left(1+\frac{a^{-j}/g(p)}{1-a^{-j}/g(p)}\right) \\
&\geq -\sum_{j=1}^{k(p)} \frac{a^{-j}/g(p)}{1-a^{-j}/g(p)} \\
&\geq \frac{-1}{1-a^{-1}/g(p)} \sum_{j=1}^{k(p)} a^{-j}/g(p) \\
&= \frac{-1}{g(p)-a^{-1}} \sum_{j=1}^{k(p)} a^{-j}.
\end{aligned}$$

Now, using the formula for the sum of a geometric series, we have

$$\log\left(\prod_{j=1}^{k(p)} (1-\epsilon_j(p))\right) \geq \frac{-1}{g(p)-a^{-1}} \left[\frac{a^{-1}(1-a^{-k(p)})}{1-a^{-1}}\right],$$

from which it follows that

$$\prod_{j=1}^{k(p)} (1-\epsilon_j(p)) \geq \exp\left(\frac{-1}{g(p)-a^{-1}} \left[\frac{a^{-1}(1-a^{-k(p)})}{1-a^{-1}}\right]\right).$$

Now, assuming only that $g(p) > a^{-1}(1+\eta)$, for some fixed $\eta > 0$ it is easy to see that

$$\lim_{p\to\infty} \prod_{j=1}^{k(p)} (1-\epsilon_j(p)) > 0,$$

and if $g(p) \to \infty$ as $p \to \infty$ we have

$$\lim_{p\to\infty} \prod_{j=1}^{k(p)} (1-\epsilon_j(p)) = 1,$$

as claimed. ■

REFERENCES

[1] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone, "Adapting to unknown sparsity by controlling the false discovery rate," *Annals of Statistics*, vol. 34, no. 2, pp. 584–653, 2006.

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

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

[4] Y. Ingster, "Some problem of hypothesis testing leading to infinitely divisible distributions," *Mathematical Methods of Statistics*, vol. 6, pp. 47–69, 1997.

[5] ——, "Minimax detection of a signal for ℓ_n^p-balls," *Mathematical Methods of Statistics*, vol. 7, pp. 401–428, 1998.

[6] D. Donoho and J. Jin, "Higher criticsm for detecting sparse heterogenous mixtures," *Annals of Statistics*, 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," *Genetic Epidemiology*, vol. 31, no. 8, 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, no. 19, pp. 3771–3777, 2005.

[9] J. Satagopan and R. Elston, "Optimal two-stage genotyping in population-based association studies," *Genetic Epidemiology*, vol. 25, no. 2, pp. 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," *Statistics in Medicine*, vol. 27, no. 21, pp. 4145–4160, 2008.

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

[12] Various, "Promising directions in active vision," *International Journal of Computer Vision*, vol. 11, no. 2, pp. 109–126, 1991.

[13] D. Cohn, "Neural network exploration using optimal experiment design," in *Advances in Neural Information Processing Systems (NIPS)*, 1994, pp. 679–686.

[14] D. Cohn, Z. Ghahramani, and M. Jordan, "Active learning with statistical models," *Journal of Artificial Intelligence Research*, vol. 4, pp. 129–145, 1996.

[15] J. Haupt, R. Castro, and R. Nowak, "Adaptive discovery of sparse signals in noise," in *Proc. 42nd Asilomar Conf. on Signals, Systems, and Computers*, Monterey, CA, October 2008.

[16] ——, "Distilled sensing: selective sampling for sparse signal recovery," in *Proc. 12th Conference on Artificial Intelligence and Statistics*, Clearwater Beach, FL, 2009.

[17] N. Meinshausen, "Hierarchical testing of variable importance," *Biometrika*, vol. 95, no. 2, pp. 265–278, June 2008.

[18] N. Meinshausen, P. Bickel, and J. Rice, "Efficient blind search: Optimal power of detection under computational cost constraints," *The Annals of Applied Statistics*, vol. 3, no. 1, pp. 38–60, March 2009.

[19] W. Ehm, J. Kornmeier, and S. Heinrich, "Multiple testing along a tree," *Electronic Journal of Statistics*, vol. 4, pp. 461–471, 2010.

[20] C. R. Genovese and L. Wasserman, "Operating characteristics and extensions of the FDR procedure," *J. R. Stat. Soc. Ser. B*, vol. 64, pp. 499–518, 2002.

[21] S. K. Sarkar, "False discovery and false nondiscovery rates in single-step multiple testing procedures," *Annals of Statistics*, vol. 34, no. 1, pp. 294–415, 2006.

[22] H. Chernoff, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Annals of Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

[23] A. P. Korostelev, "On minimax rates of convergence in image models under sequential design," *Statistics & Probability Letters*, vol. 43, pp. 369–375, 1999.

[24] R. Castro and R. Nowak, "Minimax bounds for active learning," *IEEE Trans. Info. Theory*, vol. 54, no. 5, pp. 2339–2353, July 2008.

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

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

[27] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak, "Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements," in *Proc. 43rd Asilomar Conf. on Signals, Systems, and Computers*, Monterey, CA, 2009.