Distilled Sensing: Selective Sampling for Sparse Signal Recovery

Jarvis Haupt
Dept. of Elec. and Comp. Engr.
University of Wisconsin—Madison
Madison, WI 53706

Rui Castro
Dept. of Electrical Engineering
Columbia University
New York, NY 10027

Robert Nowak
Dept. of Elec. and Comp. Engr.
University of Wisconsin—Madison
Madison, WI 53706

Abstract

A selective sampling procedure called *distilled sensing* (DS) is proposed, and shown to be an effective method for recovering sparse signals in noise. Based on the notion that it is often easier to rule out locations that do not contain signal than it is to directly identify non-zero signal components, DS is a sequential method that systematically focuses sensing resources towards the signal subspace. This adaptivity in sensing results in rather surprising gains in sparse signal recovery—dramatically weaker sparse signals can be recovered using DS compared with conventional non-adaptive sensing procedures.

1 INTRODUCTION

Consider the following canonical signal model.

$$X_i \sim \mathcal{N}(\mu_i, 1)\ ,\ i = 1, \ldots, n, \tag{1}$$

where $\mathcal{N}(\mu_i, 1)$ denotes the normal distribution with mean μ_i and unit variance. The signal $\mu = (\mu_1, \ldots, \mu_n)$ is sparse if most of the components μ_i are zero. Identifying the locations of the non-zero components based on the data $X = (X_1, \ldots, X_n)$ when n is very large is a fundamental problem arising in many applications, including fMRI (Genovese et al., 2002), microarray analysis (Pawitan et al., 2005), and astronomical surveying (Hopkins et al., 2002). A common approach in these problems entails coordinate-wise thresholding of the observed data X at a given level, identifying the locations whose corresponding observation exceeds the threshold as signal components. Among such methods, false-discovery rate (FDR) analysis (Benjamini and Hochberg, 1995) tends to be the procedure of choice because it is less conservative than Bonferroni correction, making it more useful in practice, and because it enjoys asymptotically optimal performance characteristics (Abramovich et al., 2006; Benjamini and Hochberg, 1995; Donoho and Jin, 2006; Donoho and Jin, 2008; Jin, 2003).

Suppose that the number of non-zero components of μ grows sublinearly in n according to $n^{1-\beta}$ for $\beta \in (0, 1)$, and that each non-zero component takes the same (positive) value $\sqrt{2r \log n}$ for $r > 0$. For a given recovery procedure, define the *false-discovery proportion* (FDP) to be the number of falsely discovered components relative to the total number of discoveries, and the *non-discovery proportion* (NDP) to be the number of non-zero components missed relative to the total number of non-zero components. The asymptotic limits of sparse recovery for data collected according to (1) are sharply delineated in the (β, r) parameter space. Specifically, if $r < \beta$, no recovery procedure based on coordinate-wise thresholding of the observed data can drive the FDP and NDP to zero as $n \to \infty$. But, when $r > \beta$, there exists a recovery procedure based on coordinate-wise thresholding that drives both the NDP and FDP to zero as $n \to \infty$. Thus, the relation $r = \beta$ defines a sharp asymptotic boundary in the parameter space, identifying when sparse signals observed under the model (1) can be reliably recovered. Under similar sparse signal and noise models, several related works established sharp asymptotics in signal estimation and classification settings (Abramovich et al., 2006; Donoho and Jin, 2006; Donoho and Jin, 2008; Jin 2003).

Suppose that instead of a single observation of a sparse signal in noise, one were able to take multiple 'looks,' possibly adjusting the focus in a sequential fashion. Similar adaptive methods have been proposed in the signal processing literature (Rangarajan, 2007), and they certainly are conceivable in applications such as

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

microarray analysis and astronomical surveying. Here we consider an approach based on the idea that after the first look, one might be able to rule out a large number of locations where the signal is probably not present, and then focus sensing resources into the remaining locations in question. Formalizing this idea, we show that for the same budget of sensing resources, sequential adaptive sensing procedures dramatically outperform non-adaptive procedures, resulting in different scaling laws in terms of estimability. Rather than requiring that the non-zero components obey $\mu_i > \sqrt{2\beta \log n}$, we show that a novel adaptive sensing procedure called *distilled sensing* (DS) guarantees that sparse signals at level $\mu_i = \alpha(n)$, where $\alpha(n)$ is any positive monotone diverging sequence in n that exceeds the m-fold iterated logarithm function,

$$\log^{[m]} n = \underbrace{\log\log\ldots\log n}_{\text{repeated } m \text{ times}},$$

for an arbitrary finite integer m, can be recovered in the sense that there exists a coordinate-wise thresholding procedure that sends the FDP and NDP to zero as $n \to \infty$. In other words, DS can detect dramatically weaker signals than non-adaptive methods.

The paper is organized as follows. In Section 2 we review the conventional non-adaptive approach to sparse recovery. We introduce our adaptive sensing technique (DS) in Section 3, and in Section 4 we state our main results, that DS enables the recoverability of significantly weaker signals than standard, non-adaptive methods. Section 5 provides numerical simulations of DS, and a short discussion appears in Section 6. Proofs of the main results are given in the Appendix.

2 SPARSE RECOVERY BY NON-ADAPTIVE SENSING

Suppose we observe a $n \times 1$ signal μ in noise according to the model (1). The signal μ is assumed to be sparse—that is, most of the components of the signal are equal to zero. Define $\mathcal{S} = \{i : \mu_i \neq 0,\ i = 1, \ldots, n\}$. The elements of $\mathcal{S}$ are called the signal components, and the elements in the complement set, $\mathcal{S}^c = \{1, \ldots, n\} \setminus \mathcal{S}$, are called null components. The goal of a signal recovery procedure is to identify the signal components (in other words, estimate $\mathcal{S}$) using the observed data X. Let $\widehat{\mathcal{S}}(X)$ be the outcome of a given signal recovery procedure. Define the *false-discovery proportion* (FDP) to be the ratio between the number of falsely-discovered signal components and the total number of discovered components, $\text{FDP} = \left|\widehat{\mathcal{S}}(X) \setminus \mathcal{S}\right| / \left|\widehat{\mathcal{S}}(X)\right|$, and define the *non-discovery proportion* to be the ratio between the number of undiscovered signal components and the total number of signal components, $\text{NDP} = \left|\mathcal{S} \setminus \widehat{\mathcal{S}}(X)\right| / \left|\mathcal{S}\right|$. An effective signal recovery procedure must be able to control both the FDP and NDP.

Consider sparse signals having $n^{1-\beta}$ signal components each of amplitude $\sqrt{2r \log n}$, for some $\beta \in (0, 1)$ and $r > 0$, under the model (1). We consider a coordinate-wise thresholding procedure,

$$\widehat{\mathcal{S}}(X) = \{i \ : \ X_i > \tau\},\ \tau > 0, \tag{2}$$

to estimate the locations of the signal components. It follows from techniques used in (Abramovich et al., 2006; Benjamini and Hochberg, 1995; Donoho and Jin, 2006; Donoho and Jin, 2008; Jin, 2003) that if $r > \beta$, the procedure (2), with a threshold τ that may depend on r, β, and n, drives both the FDP and NDP to zero with probability one as $n \to \infty$. Conversely, if $r < \beta$, then no such coordinate-wise thresholding procedure can drive the FDP and NDP to zero simultaneously with probability tending to one as $n \to \infty$. In other words, for the specified signal parametrization and observation model, the (β, r) parameter plane is partitioned into two disjoint regions. In the region $r > \beta$, sparse signal components can be reliably located using a coordinate-wise thresholding procedure. In the complementary region where $r < \beta$, no coordinate-wise thresholding procedure is reliable in the sense of controlling both the FDP and NDP. This establishes a sharp boundary in the parameter space, $r = \beta$, for large-sample consistent recovery of sparse signals.

3 DISTILLED SENSING

We generalize the observation model (1) to allow multiple observations, indexed by j, of the form

$$X_i^{(j)} = \sqrt{\phi_i^{(j)}}\mu_i + Z_i^{(j)}, \tag{3}$$

for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, k$, where each $\phi_i^{(j)}$ is non-negative, and $Z_i^{(j)} \overset{iid}{\sim} \mathcal{N}(0, 1)$. In addition, we impose the restriction $\sum_{i,j} \phi_i^{(j)} \leq n$, limiting the total amount of sensing energy. Note that the standard observation model (1) takes the form (3) with $k = 1$ and $\phi_i^{(1)} = 1$ for $i = 1, \ldots, n$. Another possibility is to make multiple iid observations, but each with only a fraction of the total sensing energy budget. For example, set $\phi_i^{(j)} = 1/\sqrt{k}$, $i = 1, \ldots, n$, for $j = 1, \ldots, k$. Because of the independence of the $Z_i^{(j)}$, $\sum_{j=1}^k X^{(j)}$ is equivalent to X in the standard model in this case as well. There are obviously many other non-adaptive choices of $\{\phi_i^{(j)}\}_{i,j}$ that yield the same result. Furthermore, no non-adaptive sensing scheme exists that can produce better results than those obtained using observations from the standard model (1).

Algorithm 1: Distilled sensing.

Input:
Number of observation steps k;
Energy allocation strategy: $\mathcal{E}^{(j)}$, $\sum_{j=1}^{k} \mathcal{E}^{(j)} \leq n$;

Initialize:
Initial index set $I^{(1)} \longleftarrow \{1, 2, \ldots, n\}$;

Distillation:
for $j = 1$ **to** k **do**

$$X_i^{(j)} = \left\{ \begin{array}{ll} \sqrt{\frac{\mathcal{E}^{(j)}}{|I^{(j)}|}}\mu_i + Z_i^{(j)}, & i \in I^{(j)} \\ Z_i^{(j)}, & i \in I^{(1)} \setminus I^{(j)} \end{array} \right\};$$

$$I^{(j+1)} \longleftarrow \{i \in I^{(j)} : X_i^{(j)} > 0\};$$

end

Output:
Final index set $I^{(k)}$;
Distilled observations $X_{\mathrm{DS}}^{(k)} := \{X_i^{(k)} : i \in I^{(k)}\}$;

Therefore, we are interested here in adaptive, sequential designs of $\{\phi_i^{(j)}\}_{i,j}$ that tend to focus on the signal components of μ. In other words, we allow $\phi_i^{(j)}$ to depend explicitly on the past $\{\phi_i^{(\ell)}, X_i^{(\ell)}\}_{i,\ell<j}$. The principle upon which our procedure is based is simple—given a collection of noisy observations of the components of a sparse vector, it is far easier to identify a *large set* of null components (where the signal is *absent*) than it is to identify a *small set* of signal components. When multiple observations of each component are allowed, this principle suggests a process for refining observations—iteratively allocate more sensing resources to locations that are most promising while ignoring locations that are unlikely to contain signal components. This is reminiscent of the purification that occurs in the process of distillation; hence, we refer to our procedure as *distilled sensing* (DS).

Let k denote the number of observation steps in the DS process, and divide the total budget of sensing energy among the steps. Each observation takes the form (3), where the sensing energy allocated to that observation step is distributed equally among the set of locations of interest at that step. Following each of the first $k-1$ observation steps, a refinement or *distillation* is performed, identifying the subset of locations where the corresponding observation is positive. The rationale is that it is highly improbable that the signal (which is assumed to be positive) is present at locations where the observation is negative. The algorithm terminates after the final observation, and the output consists of the final observations and the set of locations that were measured in the last step. A pseudocode description of DS appears as Algorithm 1.

To quantify the performance of DS, we will show that each distillation step retains almost all of the locations corresponding to signal components, but only about *half* of the locations corresponding to null components. When the signal μ is sparse, this implies that the effective dimension is roughly halved at each step. A judicious allocation of sensing energy over observation steps provides increasing sensing energy per location in each subsequent step, resulting in a net exponential boost in the effective amplitude of each measured signal component. As a result, applying a coordinate-wise thresholding procedure to the output observations of DS results in significant improvements in recovery compared to procedures that utilize nonadaptive sensing, as shown in the next section.

4 MAIN RESULTS

We use an energy allocation scheme designed to balance the probabilities of successful retention of signal components at each step, by allocating a larger portion of the sensing energy to the first steps and decreasing the energy used in later steps when there are fewer locations to observe. The exponential decrease in the number of observed locations at each step suggests that the sensing energy allocated to each step can also decrease exponentially. To accomplish this we allocate energy for the first $k-1$ steps according to the entries of a geometric progression, and put all remaining energy on the last step. For a fixed parameter $0 < \Delta < 1$, the energy allocation scheme is

$$\mathcal{E}^{(j)} = \left\{ \begin{array}{ll} \frac{\Delta n}{2}\left(1 - \frac{\Delta}{2}\right)^{j-1}, & j = 1, \ldots, k-1 \\ n\left(1 - \frac{\Delta}{2}\right)^{k-1}, & j = k \end{array} \right\}, \quad (4)$$

and the total energy expended satisfies $\sum_{j=1}^{k} \mathcal{E}^{(j)} = n$.

Our first main result quantifies the performance gain provided by distilled sensing (DS) when the number of observation steps is fixed. The result, stated below as a theorem, establishes an expanded region of large-sample consistent recovery in the (β, r) parameter space. The proof is given in the Appendix.

Theorem 4.1. *Let k be a positive integer, and consider applying the k-step DS procedure using the sensing energy allocation strategy described in* (4) *with fixed parameter $0 < \Delta < 1$, to sparse signals $\mu \in \mathbb{R}^n$ having $n^{1-\beta}$ signal components each of amplitude $\sqrt{2r \log n}$, for some $\beta \in (0,1)$ and $r > 0$. If $r > \beta/(2-\Delta)^{k-1}$, there exists a coordinate-wise thresholding procedure of $X_{DS}^{(k)}$ of the form* (2) *that drives both the FDP and NDP to zero with probability tending to one as $n \to \infty$.*

In other words, this result shows that a small number of observation steps leads to a significant improvement in terms of recoverability—the minimum signal amplitude required for consistent recovery decreases exponentially as a function of the number of observation steps in the DS procedure. A natural question arises as to whether the number of steps can be large enough, so that signals whose amplitudes grow (with the dimension n) slower than $\sqrt{\log n}$ can be recovered. We address this question here by letting the number of observation steps tend to infinity slowly as a function of n. The result is the following theorem, for which the proof appears in the Appendix.

Theorem 4.2. *Using the sensing energy allocation strategy described in* (4) *with fixed parameter* $0 < \Delta < 1$*, let*

$$k = 1 + \left\lceil \frac{\log \log n}{\log (2 - \Delta)} \right\rceil,$$

and apply the k*-step DS procedure to sparse signals* $\mu \in \mathbb{R}^n$ *having* $n^{1-\beta}$ *signal components each of amplitude* $\alpha = \alpha(n)$*. If* $\alpha(n)$ *is any positive monotone diverging sequence in* n *that exceeds the* m*-fold iterated logarithm* $\log^{[m]} n$ *for some finite integer* m*, then there exists a coordinate-wise thresholding procedure of* $X_{DS}^{(k)}$ *of the form* (2) *that drives both the FDP and NDP to zero with probability tending to one as* $n \to \infty$*.*

In other words, DS can result in dramatic improvements in recoverability, as it succeeds at recovering signals whose amplitudes are vanishingly small relative to those of signals that can be recovered using the best non-adaptive methods.

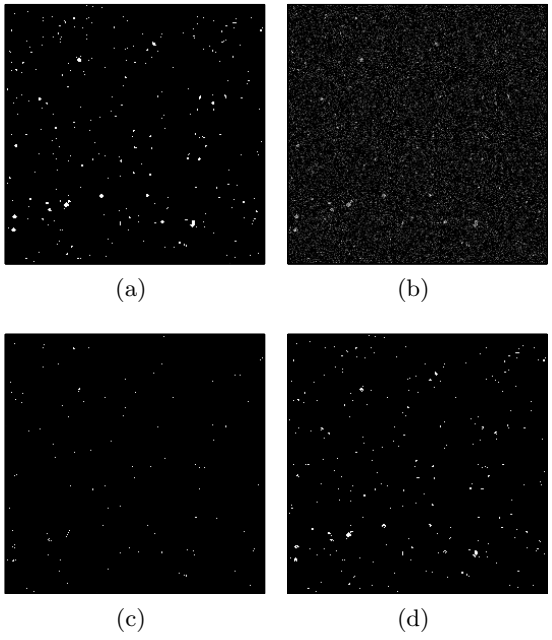

Figure 1: Experimental validation of DS procedure. Panels (a) and (b) show a noiseless radio telescope image and a simulated noisy version of the image, respectively. Panel (c) shows the result of applying BH thresholding (at FDR level 0.05) to the data in (b), while applying BH thresholding at the same level to the output of the DS procedure with $\Delta = 0.9$ and $k = 5$ gives the result depicted in panel (d).

5 EXPERIMENTAL EVALUATION

Our theoretical analysis establishes that DS is considerably more powerful than conventional non-adaptive sensing in the large-sample regime. In this section we examine the finite-sample performance of DS with a simulation experiment motivated by astronomical surveying (Hopkins et al., 2002). For recovery, we apply (2), using the data-dependent threshold identified by the Benjamini and Hochberg procedure at a specified FDR level, where FDR is defined as the expected value of the FDP. We call this recovery BH thresholding.

Fig. 1(a) depicts a portion of a real radio telescope image collected by the Phoenix Deep Survey (`www.physics.usyd.edu.au/~ahopkins/phoenix/`). The image size is 256×256 pixels and 533 pixels have nonzero amplitudes of 2.98 (implying $\beta = 0.43$ and $r = 0.4$). Fig. 1(b) depicts a simulated noisy version of the image, equivalent to a collection of non-adaptive observations from the model (1), where to improve visualization, locations whose corresponding observations were negative are mapped to black (amplitudes of positive observations are unaltered). Fig. 1(c) shows the result of applying BH thresholding at FDR level 0.05 to the non-adaptive observations, and Fig. 1(d) shows the output of DS with $\Delta = 0.9$ and $k = 5$ after BH thresholding at the same level (we chose Δ here to be conservative—in practice, larger values of Δ allocate more sensing energy to earlier steps, resulting in fewer total non-discoveries). Note that considerably more "stars" are recovered by DS.

For the same experiment, we also examine the benefit of DS in terms of false and non-discovery proportions. For each method we computed the empirical NDP for a range of FDR levels. The curves in Fig. 2 show empirical NDP vs. empirical FDP for 10 trials of each procedure (solid lines for non-adaptive sampling, dashed lines for DS with $\Delta = 0.9$ and $k = 5$). For the same FDP, DS yields lower NDPs than non-adaptive sampling, except sometimes at very high FDP levels, quantifying the improvement that is observed when visually comparing Fig. 1(c-d).

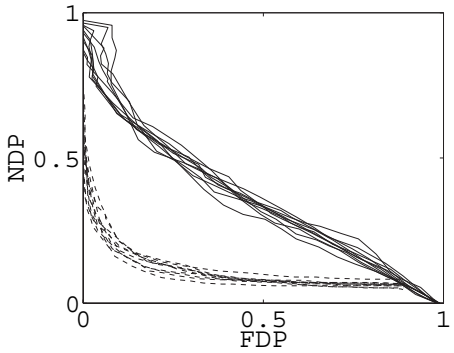

Figure 2: NDP vs. FDP for 10 independent trials of each procedure applied to the noisy star recovery task whose one-trial results are depicted graphically in Fig. 1. Curves for non-adaptive sampling followed by BH thresholding are solid lines, while curves for DS ($\Delta = 0.9$, $k = 5$) followed by BH thresholding are dashed lines. The separation of curve clusters illustrates the improvement of DS per trial and on average.

6 DISCUSSION

A fundamental difference between non-adaptive sensing and DS can be understood by comparing false and non-discovery criteria. Recovery procedures based on non-adaptive sampling methods must control the FDP and NDP simultaneously, while each step of DS only controls the the number of non-discoveries (keeping it near zero), and allows the number of false discoveries to remain large (nearly all discoveries are false when the signal is sparse). Simultaneous FDP and NDP control for DS is performed only after the last observation step, when sensing resources have been efficiently focused into the signal subspace.

An alternate way to evaluate DS is to compare the minimal sensing energy budget required to achieve the same large-sample performance as non-adaptive sensing methods (recoverability only when $r > \beta$). The results obtained here imply that procedures based on DS will be able to recover the same signals as non-adaptive sensing methods using an energy budget that grows only *sublinearly* with n, implying that DS can recover signals using less sensing energy (or in less time) than what is required by non-adaptive sensing methods.

While the theoretical results presented here are asymptotic in nature, the performance of DS can also be quantified in finite-dimensional settings using the intermediate result (Theorem 7.3) in the Appendix. Rather than exhibiting sharp asymptotics, the efficacy of DS in finite-dimensional problems is quantified by probabilities of success that vary depending on r, β, Δ, and n. In addition, in finite-dimensional applications, DS could be modified to improve the retention of true signal components, at the expense of rejecting fewer null components, by selecting a less aggressive (slightly negative) threshold at each step of the DS procedure.

The proposed DS procedure can also be applied to more general classes of signals, such as those for which μ has both positive and negative values. In this case, one approach would be to split the budget of sensing energy in half, and execute the DS procedure once assuming the signal components are positive as described above, and again assuming the signal components are negative (retaining locations at each step for which the corresponding observation is negative). The final (composite) set of observations could then be subjected to standard FDR controlling procedures.

Finally, we note that the results presented here (namely Theorem 7.3) also imply that DS followed by a recovery procedure of the form (2), but where selection of the threshold does not require prior knowledge of the signal amplitude or sparsity parameters, will recover signals that are potentially much more sparse than those described above. Specifically, signals exhibiting general sublinear sparsity having $s(n)$ signal components each with amplitude at least $\alpha(n)$, where $s(n)$ and $\alpha(n)$ are each positive monotone diverging sequences in n and $\alpha(n)$ exceeds some finite iteration of the logarithm function, such that $\alpha(n) \cdot s(n) > c \log\log\log n$ for some c depending on the energy allocation parameter Δ, are recoverable using the adaptive DS procedure. In the interest of space, we relegate a thorough exposition of this to future work.

7 APPENDIX

Establishing the main results of the paper amounts to counting how many locations corresponding to signal and null components are retained by thresholding the observations at level zero in each distillation step. We begin by considering the signal components.

Lemma 7.1. *Consider a vector μ with s components, each of amplitude $\alpha > 0$, observed according to the model $X_i = \sqrt{\phi}\mu_i + Z_i$, where Z_i is a collection of independent $\mathcal{N}(0,1)$ noises and $\phi > 0$ denotes the amount of sensing energy allocated to each location. When the amplitude satisfies $\alpha \geq 2/\sqrt{\phi}$, the number of components retained by thresholding the observations at the level zero, denoted $\widetilde{s}$, satisfies $(1-\epsilon)\, s \leq \widetilde{s} \leq s$ with probability at least*

$$1 - \exp\left(-\frac{\alpha \cdot s}{4}\sqrt{\frac{\phi}{2\pi}}\right),$$

where

$$\epsilon = \frac{1}{\alpha}\sqrt{\frac{1}{2\pi\phi}} \leq \frac{1}{2\sqrt{2\pi}}.$$

The upper bound on ϵ follows from the condition on α, and ensures that the fraction of vector components retained is bounded away from zero.

Proof. The proof amounts to counting the number of components retained by thresholding. For that, we utilize a standard bound on tail probabilities of Gaussian random variables, which states that if $Z \sim \mathcal{N}(\gamma, 1)$ for $\gamma > 0$, then $\Pr(Z < 0) \leq (\gamma\sqrt{2\pi})^{-1} \exp\left(-\gamma^2/2\right)$.

To each component observation X_i, assign a Bernoulli random variable $T_i = \mathbf{1}_{\{X_i > 0\}}$, which is equal to one whenever the observation exceeds 0, and zero otherwise. Let $p = \Pr(T_i = 1)$. The number of vector components whose observations exceed the threshold, $\widetilde{s} = \sum_i T_i$, is a Binomial random variable, with

$$1 - p \leq \frac{1}{\alpha}\sqrt{\frac{1}{2\pi\phi}} \exp\left(-\frac{\alpha^2\phi}{2}\right).$$

Establishing the lemma amounts to quantifying the probability that $\widetilde{s} \geq (1-\epsilon)s$ for an appropriately chosen ϵ. To that end, we use a bound on the tail probability of the Binomial distribution (Chernoff, 1952). Namely, for a Binomial(n, p) random variable B, whenever $b < \mathbb{E}[B] = np$,

$$\Pr(B \leq b) \leq \left(\frac{n - np}{n - b}\right)^{n-b} \left(\frac{np}{b}\right)^{b}.$$

In our context, this result implies

$$\begin{aligned} &\Pr\left(\widetilde{s} \leq (1-\epsilon)s\right) \\ &\quad \leq \quad \left(\frac{1-p}{\epsilon}\right)^{\epsilon s} \left(\frac{p}{1-\epsilon}\right)^{(1-\epsilon)s}, \end{aligned} \tag{5}$$

provided $\epsilon > 1 - p$, which is satisfied by the choice

$$\epsilon = \frac{1}{\alpha}\sqrt{\frac{1}{2\pi\phi}}.$$

Now, notice that when $\alpha > 2/\sqrt{\phi}$, the condition

$$-\frac{\alpha^2\phi\epsilon s}{2} + (1-\epsilon)s \log\left(\frac{1}{1-\epsilon}\right) \leq -\frac{\alpha s}{4}\sqrt{\frac{\phi}{2\pi}},$$

obtained by upper-bounding the logarithm of the right-hand side of (5), holds for any $\epsilon \in (0, 1)$. The result follows from exponentiating this last bound. □

Next, we quantify how many of the null components are retained by each thresholding step.

Lemma 7.2. *Consider a vector μ with z components, each of amplitude 0, observed according to the model $X_i = \sqrt{\phi}\mu_i + Z_i$, where Z_i is a collection of independent $\mathcal{N}(0, 1)$ noises and $\phi > 0$ denotes the amount of sensing energy allocated to each measured location. For any $\epsilon_0 < 1/2$, the number of components retained by thresholding the observations at the level zero, denoted $\widetilde{z}$, satisfies $(1/2 - \epsilon_0)\, z \leq \widetilde{z} \leq (1/2 + \epsilon_0)\, z$, with probability at least $1 - 2\exp\left(-2z\epsilon_0^2\right)$.*

Proof. To each observation X_i, assign a Bernoulli random variable $T_i = \mathbf{1}_{\{X_i > 0\}}$ which takes the value one when the corresponding observation exceeds 0 and zero otherwise. Since each observation is of noise only, the number of vector components whose corresponding observation exceeds the threshold, $\widetilde{z} = \sum_i T_i$, is a Binomial random variable with probability $1/2$. Applying Hoeffding's inequality we obtain $\Pr\left(|\widetilde{z} - z/2| > \epsilon_0 z\right) \leq 2\exp\left(-2z\epsilon_0^2\right)$, which holds for any $\epsilon_0 > 0$. Imposing the restriction $\epsilon_0 < 1/2$ guarantees that the fraction of components retained is within $(0, 1)$. □

Taken together, the lemmata above establish that by thresholding at level zero, almost all of the signal components and about half of the zero components are retained with high probability. Incorporating the geometric allocation of sensing energy per step specified in (4), we obtain the following.

Theorem 7.3. *Let $\mu \in \mathbb{R}^n$ be a sparse signal having $s > 0$ signal components each of amplitude α, where $\alpha > 4/\sqrt{\Delta}$, and $z = n - s > s$ null components. In the DS procedure of Algorithm 1, let $1 < k \leq 1 + \log_2(z/s)$, let Δ be a fixed parameter satisfying $0 < \Delta < 1 - 2\epsilon_0$ for some $\epsilon_0 < 1/2$, and let the energy allocation $\mathcal{E}^{(j)}$ be as described in (4). For $j = 1, \ldots, k-1$, define*

$$\epsilon^{(j)} = \frac{1}{\alpha}\sqrt{\frac{1}{2\pi\xi^{(j)}}},$$

where

$$\xi^{(j)} = \frac{\Delta}{4}\left(\frac{2 - \Delta}{1 + 2\epsilon_0}\right)^{j-1}.$$

Then, with probability at least

$$\begin{aligned} 1 - (k-1)\exp\left(-\frac{\alpha \cdot s}{8} \cdot \sqrt{\frac{\Delta}{2\pi}} \cdot \prod_{j=1}^{k-2}\left(1 - \epsilon^{(j)}\right)\right) \\ -2(k-1)\exp\left(-2 \cdot z \cdot \epsilon_0^2 \cdot \left(\frac{1}{2} - \epsilon_0\right)^{k-2}\right), \end{aligned}$$

the output of the DS procedure, $X_{DS}^{(k)}$, is equivalent in distribution to a single collection of noisy observations of a vector $\mu_{\text{eff}} \in \mathbb{R}^{n_{\text{eff}}}$ according to the observation model (1). The number of signal components in μ_{eff} is denoted by s_{eff} and satisfies

$$s\prod_{j=1}^{k-1}\left(1 - \epsilon^{(j)}\right) \leq s_{\text{eff}} \leq s,$$

the effective signal length n_{eff} *satisfies*

$$s\prod_{j=1}^{k-1}\left(1-\epsilon^{(j)}\right)+z\left(\frac{1}{2}-\epsilon_0\right)^{k-1}\leq n_{\text{eff}}\leq s+z\left(\frac{1}{2}+\epsilon_0\right)^{k-1},$$

and the effective observed amplitude α_{eff} *satisfies*

$$\alpha_{\text{eff}}\geq\alpha\sqrt{\frac{n\left(1-\Delta/2\right)^{k-1}}{s+z\left(1/2+\epsilon_0\right)^{k-1}}}.$$

Proof. We begin by applying the union bound to the result of Lemma 7.2 to enforce the condition for each of the first $k-1$ distillation steps. Using superscripts on z and s to index the observation step, such that $z^{(1)}=z$ and $s^{(1)}=s$, for $\epsilon_0<1/2$, the bounds

$$z^{(1)}\left(\frac{1}{2}-\epsilon_0\right)^{j-1}\leq z^{(j+1)}\leq z^{(1)}\left(\frac{1}{2}+\epsilon_0\right)^{j-1}$$

hold simultaneously for all $j=1,2,\ldots,k-1$ with probability exceeding

$$1-2(k-1)\exp\left(-2z^{(1)}\epsilon_0^2\left(\frac{1}{2}-\epsilon_0\right)^{k-2}\right).$$

As a result, with the same probability, the total number of locations in each set $I^{(j)}$ satisfies $|I^{(j)}|\leq s^{(1)}+z^{(1)}\left(\frac{1}{2}+\epsilon_0\right)^{j-1}$, for $j=1,2,\ldots,k$. Using these upper bounds and the energy allocation rule (4), we can lower bound the sensing energy per location at each step, $\phi_i^{(j)}=\mathcal{E}^{(j)}/|I^{(j)}|$, for $i\in I^{(j)}$—specifically,

$$\phi_i^{(j)}\geq\left\{\begin{array}{ll}\frac{\Delta n(1-\Delta/2)^{j-1}}{2\left(s^{(1)}+z^{(1)}(1/2+\epsilon_0)^{j-1}\right)}, & j=1,\ldots,k-1\\ \\ \frac{n(1-\Delta/2)^{k-1}}{s^{(1)}+z^{(1)}(1/2+\epsilon_0)^{k-1}}, & j=k\end{array}\right\},$$

for $i\in I^{(j)}$ (and $\phi_i^{(j)}=0$ for $i\notin I^{(j)}$). Notice that when $k\leq 1+\log_2\left(z^{(1)}/s^{(1)}\right)$, for each $i\in I^{(j)}$,

$$\phi_i^{(j)}\geq\left\{\begin{array}{ll}\frac{\Delta}{4}\left(\frac{2-\Delta}{1+2\epsilon_0}\right)^{j-1}, & j=1,\ldots,k-1\\ \\ \frac{1}{2}\left(\frac{2-\Delta}{1+2\epsilon_0}\right)^{k-1}, & j=k\end{array}\right\}.$$

Since $\Delta<1-2\epsilon_0$, this shows that the amount of sensing energy allocated to each retained location *increases exponentially* with the number of observation steps.

Now, conditioned on the above event, we can invoke Lemma 7.1 and apply the union bound again so that with probability at least

$$1-(k-1)\exp\left(-\frac{\alpha\cdot s^{(1)}}{8}\cdot\sqrt{\frac{\Delta}{2\pi}}\cdot\prod_{j=1}^{k-2}\left(1-\epsilon^{(j)}\right)\right),$$

when $\alpha\geq 4/\sqrt{\Delta}$, the bounds

$$\left(1-\epsilon^{(j)}\right)s^{(j)}\leq s^{(j+1)}\leq s^{(j)}$$

hold simultaneously for all $j=1,2,\ldots,k-1$.

Applying the union bound to both composite events, we obtain that with the specified probability, the number of signal and null components present in the kth observation step are given by $s^{(k)}$ and $z^{(k)}$, respectively, as defined above. Thus, the final effective signal dimension is $s^{(k)}+z^{(k)}$, and the effective observed amplitude of each signal component is obtained using the lower bound on $\phi_i^{(j)}$, establishing the claim. □

Before we can prove the main results, we need one final lemma quantifying the (limiting) fraction of signal components retained throughout the DS procedure.

Lemma 7.4. *Let* $g=g(n)$ *and* $k=k(n)$ *be positive monotone diverging sequences in* n*, where* $g(n)$ *exceeds the* m*-fold iteration* $\log^{[m]}n$ *for some finite integer* m *and* $k(n)\leq n$*. Define* $\epsilon^{(j)}(n)=c^{-j}/g(n)$ *for some constant* $c>1$*, and assume* $\epsilon^{(k)}(n)<1$*. Then*

$$\lim_{n\to\infty}\prod_{j=1}^{k}\left(1-\epsilon^{(j)}(n)\right)=1.$$

Proof. Since $c>1$, $(1-\epsilon^{(j)}(n))<(1-\epsilon^{(j+1)}(n))$ and thus the expression of interest satisfies

$$\left(1-\epsilon^{(1)}(n)\right)^k=\left(1-\frac{1}{g(n)c}\right)^k\leq\prod_{j=1}^{k}\left(1-\epsilon^{(j)}(n)\right).$$

If $k(n)/g(n)\to 0$ as $n\to\infty$, then the limit of the left hand side is easily seen to be 1 by a Taylor Series expansion about $\epsilon^{(1)}(n)=0$, and the lemma is established. Suppose this is not the case. Then, notice that for any $1\leq k'<k$,

$$\begin{aligned}&\prod_{j=1}^{k}\left(1-\epsilon^{(j)}(n)\right)\\ &\quad=\prod_{j=1}^{k'}\left(1-\epsilon^{(j)}(n)\right)\cdot\prod_{j=k'+1}^{k}\left(1-\epsilon^{(j)}(n)\right)\\ &\quad\overset{(a)}{\geq}\prod_{j=1}^{k'}\left(1-\epsilon^{(j)}(n)\right)\cdot\prod_{j=k'+1}^{k}\left(1-\epsilon^{(k'+1)}(n)\right)\\ &\quad\overset{(b)}{\geq}\prod_{j=1}^{k'}\left(1-\epsilon^{(j)}(n)\right)\cdot\left(1-\epsilon^{(k'+1)}(n)\right)^k,\end{aligned}$$

where (a) follows because $c>1$, and (b) is the result of multiplying by additional terms that are positive and less than one. Now, choosing k' so that $c^{k'+1}>k^2$, say

$k' = \max\{\lfloor 2\log_c k \rfloor, 0\} + 1$, is sufficient to ensure that $k\epsilon^{k'+1}(n) \to 0$, and thus $\lim_{n\to\infty}(1-\epsilon^{(k'+1)}(n))^k = 1$ by a Taylor Series argument. In this case we have

$$\lim_{n\to\infty} \prod_{j=1}^{k} \left(1-\epsilon^{(j)}(n)\right) \geq \lim_{n\to\infty} \prod_{j=1}^{k'} \left(1-\epsilon^{(j)}(n)\right).$$

Now, either $k'(n)/g(n) \to 0$ as $n \to \infty$ and the limit of the lower bound is 1 (by a Taylor Series argument), or we repeat the process above by introducing some $k'' < k'$ such that $c^{k''+1} > (k')^2$, say $k'' = \max\{\lfloor 2\log_c k' \rfloor, 0\} + 1 \sim \log_c \log_c k$. Since $g(n)$ exceeds a finite iteration of the log function and $k(n) \leq n$, this reduction process will eventually terminate in a finite number of steps, and the limit in this terminating case will be 1 by a Taylor Series argument, establishing the claim. □

7.1 PROOF OF THEOREM 4.1

Let $s = n^{1-\beta}$ for some $\beta \in (0,1)$, let $\alpha = \sqrt{2r\log n}$ for some $r > 0$, and let $k \in \mathbb{N}$ be a fixed integer. Choose $\epsilon_0 = n^{-1/3}$ and $\Delta < 1 - 2\epsilon_0$, and note that all of the conditions of Theorem 7.3 are satisfied when $n > \exp(8/\Delta r)$. By Lemma 7.4 we have that

$$\lim_{n\to\infty} \prod_{j=1}^{k-1} \left(1-\epsilon^{(j)}\right) = 1,$$

which is easy to see by making the substitutions

$$g(n) = \sqrt{\frac{\pi r\Delta(1+2\epsilon_0)\log n}{2-\Delta}},\ c = \sqrt{\frac{2-\Delta}{1+2\epsilon_0}}.$$

Thus, we obtain that with probability tending to one as $n \to \infty$, $s_{\text{eff}} \to s$, $n_{\text{eff}} \to s + z\cdot 2^{-(k-1)}$, and $\alpha_{\text{eff}} \geq \sqrt{2r(2-\Delta)^{k-1}\log n} > \sqrt{2r(2-\Delta)^{k-1}\log n_{\text{eff}}}$. Leveraging the results in the non-adaptive setting, there exists a thresholding procedure of $X_{DS}^{(k)}$ of the form (2) that will drive the FDP and NDP to zero with probability tending to one as $n \to \infty$ whenever $r(2-\Delta)^{k-1} > \beta$, as claimed.

7.2 PROOF OF THEOREM 4.2

Let $s = n^{1-\beta}$ for some $\beta \in (0,1)$, and let $\alpha = \alpha(n)$ be any positive monotone diverging sequence in n exceeding the m-fold iteration $\log^{[m]} n$ for an arbitrary finite integer m. Let the number of observation steps be

$$k = k(n) = 1 + \left\lceil \frac{\log\log n}{\log(2-\Delta)} \right\rceil.$$

Choose $\epsilon_0 = n^{-1/3}$ and $\Delta < 1 - 2\epsilon_0$. Applying Theorem 7.3 and Lemma 7.4, we obtain that with probability tending to one as $n \to \infty$, $s_{\text{eff}} \to s$,

$$n_{\text{eff}} \to s + \frac{z}{(\log n)^{\log(2)/\log(2-\Delta)}}$$

and $\alpha_{\text{eff}} \geq \alpha\sqrt{\log n} > \alpha\sqrt{\log n_{\text{eff}}}$. Since $\alpha = \alpha(n)$ diverges, for large enough n the effective observed amplitude will exceed $\sqrt{2\beta\log n_{\text{eff}}}$ for any fixed β. Now, applying the non-adaptive results, there exists a thresholding procedure of $X_{DS}^{(k)}$ of the form (2) that will drive the FDP and NDP to zero with probability tending to one as $n \to \infty$.

Acknowledgement

The authors wish to thank Jiashun Jin for discussing the details of his work with them.

References

F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone (2006), Adapting to unknown sparsity by controlling the false discovery rate, *Ann. Statist.*, **34**(2), 584-653.

Y. Benjamini and Y. Hochberg (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statist. Soc. B*, **57**, 289-300.

H. Chernoff (1952), A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Ann. Statist.*, **23**, 493-507.

D. Donoho and J. Jin (2006), Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data, *Ann. Statist.*, **34**(6), 2980-3018.

D. Donoho and J. Jin (2008), Feature selection by higher criticism thresholding: Optimal phase diagram, *submitted manuscript.*

C. Genovese, N. Lazar, and T. Nichols (2002), Thresholding of statistical maps in functional neuroimaging using the false discovery rate, *NeuroImage*, **15**, 870-878.

A. M. Hopkins, C. J. Miller, A. J. Connolly, C. Genovese, R. C. Nichol, and L. Wasserman (2002), A new source detection algorithm using the false-discovery rate, *Astron. J.*, **123**, 1086-1094.

J. Jin (2003), Detecting and estimating sparse mixtures, Ph.D. Dissertation, Dept. of Statistics, Stanford University.

Y. Pawitan, S. Michiels, S. Koscielny, A. Gusnanto, and A. Ploner (2005), False discovery rate, sensitivity and sample size for microarray studies, *Bioinformatics*, **21**, 3017-3024.

R. Rangarajan, R. Raich, and A. Hero (2007), Optimal sequential energy allocation for inverse problems, *IEEE J. Sel. Top. Sign. Proces.*, **1**(1), 67-78.