Compressive Distilled Sensing: Sparse Recovery Using Adaptivity in Compressive Measurements

Jarvis D. Haupt[1], Richard G. Baraniuk[1], Rui M. Castro[2], and Robert D. Nowak[3]

[1]*Dept. of Electrical and Computer Engineering, Rice University, Houston TX 77005*
[2]*Dept. of Electrical Engineering, Columbia University, New York NY 10027*
[3]*Dept. of Electrical and Computer Engineering, University of Wisconsin, Madison WI 53706*

***Abstract*—The recently-proposed theory of *distilled sensing* establishes that adaptivity in sampling can dramatically improve the performance of sparse recovery in noisy settings. In particular, it is now known that adaptive point sampling enables the detection and/or support recovery of sparse signals that are otherwise too weak to be recovered using any method based on non-adaptive point sampling. In this paper the theory of distilled sensing is extended to highly-undersampled regimes, as in compressive sensing. A simple adaptive sampling-and-refinement procedure called *compressive distilled sensing* is proposed, where each step of the procedure utilizes information from previous observations to focus subsequent measurements into the proper signal subspace, resulting in a significant improvement in effective measurement SNR on the signal subspace. As a result, for the same budget of sensing resources, *compressive distilled sensing* can result in significantly improved error bounds compared to those for traditional compressive sensing.**

I. Introduction

Let $x \in \mathbb{R}^n$ be a sparse vector supported on the set $\mathcal{S} = \{i : x_i \neq 0\}$, where $|\mathcal{S}| = s \ll n$, and consider observing x according to the linear observation model

$$y = Ax + w, \tag{1}$$

where A is an $m \times n$ real-valued matrix (possibly random) that satisfies $\mathbb{E}\left[\|A\|_F^2\right] \leq n$, and where $w_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ for some $\sigma \geq 0$. This model is central to the emerging field of *compressive sensing* (CS), which deals primarily with recovery of x in highly-underdetermined settings (that is, where the number of measurements $m \ll n$).

Initial results in CS establish a rather surprising result—using certain observation matrices A for which the number of rows is a constant multiple of $s \log n$, it is possible to recover x *exactly* from $\{y, A\}$, and in addition, the recovery can be accomplished by solving a tractable convex optimization [1]–[3]. Matrices A for which this exact recovery is possible are easy to construct in practice. For example, matrices whose entries are i.i.d. realizations of certain zero-mean distributions (Gaussian, symmetric Bernoulli, etc.) are sufficient to allow this recovery with high probability [2]–[4].

In practice, however, it is rarely the case that observations are perfectly noise-free. In these settings, rather than attempt to recover x exactly the goal becomes to estimate x to high accuracy in some metric (such as ℓ_2 norm) [5], [6]. One such estimation procedure is the *Dantzig selector*, proposed in [6], which establishes that CS recovery remains stable in the presence of noise. We state the result here as a lemma.

Lemma 1 (Dantzig selector)**.** *For $m = \Omega(s \log n)$, generate a random $m \times n$ matrix A whose entries are i.i.d. $\mathcal{N}(0, 1/m)$, and collect observations y according to* (1)*. The estimate*

$$\widehat{x} = \arg\min_{z \in \mathbb{R}^n} \|z\|_{\ell_1} \text{ subject to } \|A^T(y - Az)\|_{\ell_\infty} < \lambda,$$

where $\lambda = \Theta(\sigma\sqrt{\log n})$, satisfies $\|\widehat{x} - x\|_{\ell_2}^2 = O(s\sigma^2 \log n)$, with probability $1 - O(n^{-C_0})$ for some constant $C_0 > 0$.

Remark 1. *The constants in the above can be specified explicitly (or bounded appropriately), but we choose to present the results here and where appropriate in the sequel in terms of scaling relationships*[1] *in the interest of simplicity.*

On the other hand, suppose that an oracle were to identify the locations of the nonzero signal components (or equivalently, the support $\mathcal{S}$) prior to recovery. Then one could construct the least-squares estimate $\widehat{x}_{LS} = (A_{\mathcal{S}}^T A_{\mathcal{S}})^{-1} A_{\mathcal{S}}^T y$, where $A_{\mathcal{S}}$ denotes the submatrix of A formed from the columns indexed by the elements of $\mathcal{S}$. The error of this estimate is $\|\widehat{x}_{LS} - x\|_{\ell_2}^2 = O(s\sigma^2)$ with probability $1 - O(n^{-C_1})$ for some $C_1 > 0$, as shown in [6]. Comparing this oracle-assisted bound with the result of Lemma 1, we see that the primary difference is the presence of the logarithmic term in the error bound of the latter, which can be interpreted as the "searching penalty" associated with having to learn the correct signal subspace.

Of course, the signal subspace will rarely (if ever) be known a priori. But suppose that it were possible to *learn* the signal subspace from the data, in a sequential, adaptive fashion, *as the data are collected*. In this case, sensing energy could be focused only into the true signal subspace, gradually improving the effective measurement SNR. Intuitively, one might expect that this type of procedure could ultimately yield an estimate whose accuracy would be closer to that of

This work was partially supported by the ARO, grant no. W911NF-09-1-0383, the NSF, grant no. CCF-0353079, and the AFOSR, grant no. FA9550-09-1-0140.

[1]Recall that for functions $f = f(n)$ and $g = g(n)$, $f = O(g)$ means $f \leq cg$ for some constant c for all n sufficiently large, $f = \Omega(g)$ means $f \geq c'g$ for a constant c' for all n sufficiently large, and $f = \Theta(g)$ means that $f = O(g)$ *and* $f = \Omega(g)$. In addition, we will use the notation $f = o(g)$ to indicate that $\lim_{n\to\infty} f/g = 0$.

the oracle-assisted estimator, since the effective observation matrix would begin to assume the structure of A_S. Such adaptive compressive sampling methods have been proposed and examined empirically [7]–[9], but to date the performance benefits of adaptivity in compressive sampling have not been established theoretically.

In this paper we take a step in that direction by analyzing the performance of a multi-step adaptive sampling-and-refinement procedure called *compressive distilled sensing* (CDS), extending our own prior work in *distilled sensing*, where the theoretical advantages of adaptive sampling in "uncompressed" settings were quantified [10], [11]. Our main results here guarantee that, for signals having not too many nonzero entries, and for which the dynamic range is not too large, a total of $O(s \log n)$ adaptively-collected measurements yield an estimator that, with high probability, achieves the $O(s\sigma^2)$ error bound of the oracle-assisted estimator.

The remainder of the paper is organized as follows. The CDS procedure is described in Sec. II, and its performance is quantified as a theorem in Sec. III. Extensions and conclusions are briefly described in Sec. IV, and a sketch of the proof of the main result and associated lemmata appear in the Appendix.

Algorithm 1: Compressive distilled sensing (CDS).

Input:
Number of observation steps k;
$R^{(j)}$, $j = 1, \ldots, k$, such that $\sum_{j=1}^{k} R^{(j)} \leq n$;
$m^{(j)}$, $j = 1, \ldots, k$, such that $\sum_{j=1}^{k} m^{(j)} \leq m$;

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

Distillation:
for $j = 1$ **to** k **do**
Compute $\tau^{(j)} = R^{(j)}/|\mathcal{I}^{(j)}|$;
Construct $A^{(j)}$, where $A^{(j)}_{u,v} \overset{\text{iid}}{\sim}$
$$\begin{cases} \mathcal{N}\left(0, \frac{\tau^{(j)}}{m^{(j)}}\right), & u \in \{1, \ldots, m^{(j)}\},\ v \in \mathcal{I}^{(j)} \\ 0, & u \in \{1, \ldots, m^{(j)}\},\ v \notin \mathcal{I}^{(j)} \end{cases};$$
Collect $y^{(j)} = A^{(j)} x + w^{(j)}$;
Compute $\widehat{x}^{(j)} = \left(A^{(j)}\right)^T y^{(j)}$;
Refine $\mathcal{I}^{(j+1)} = \{i \in \mathcal{I}^{(j)} : \widehat{x}^{(j)}_i > 0\}$;
end

Output:
Distilled observations $\left\{y^{(j)}, A^{(j)}\right\}_{j=1}^{k}$;

II. Compressive Distilled Sensing

In this section we describe the *compressive distilled sensing* (CDS) procedure, which is a natural generalization of the *distilled sensing* (DS) procedure [10], [11]. The CDS procedure, given in Algorithm 1, is an adaptive procedure comprised of an alternating sequence of sampling (or observation) steps and refinement (or distillation) steps, and for which the observations are subject to a global *budget* of sensing resources (or "sensing energy") that effectively quantifies the average measurement precision. The key point is that the adaptive nature of the procedure allows for sensing resources to be allocated non-uniformly; in particular, proportionally more of the resources can be devoted to subspaces of interest as they are identified.

In the jth sampling step (for $j = 1, \ldots, k$), we collect measurements only at locations of x corresponding to indices in a set $\mathcal{I}^{(j)}$ (where $\mathcal{I}^{(1)} = \{1, \ldots, n\}$ initially). The jth refinement step (for $j = 1, \ldots, k-1$) identifies the set of locations $\mathcal{I}^{(j+1)} \subset \mathcal{I}^{(j)}$ for which the corresponding signal components are to be measured in step $j+1$. It is clear that in order to leverage the benefit of adaptivity, the distillation step should have the property that $\mathcal{I}^{(j+1)}$ contains most (or ideally, all) of the indices in $\mathcal{I}^{(j)}$ that correspond to true signal components. In addition, and perhaps more importantly, we also want the set $\mathcal{I}^{(j+1)}$ to be significantly smaller than $\mathcal{I}^{(j)}$, since in that case we can realize an SNR improvement from focusing our sensing resources into the appropriate subspace.

In the DS procedure examined in [10], [11], observations were in the form of noisy samples of x at any location $i \in \{1, \ldots, n\}$ at each step j. In that case it was shown a simple refinement operation—identifying all locations for which the corresponding observation exceeded a threshold—was sufficient to ensure that (with high probability) $\mathcal{I}^{(j+1)}$ would contain most of the indices in $\mathcal{I}^{(j)}$ corresponding to true signal components, but only about half of the remaining indices, even when the signal is very weak. On the other hand, here we utilize a compressive sensing observation model where at each step the observations are in the form of a low-dimensional vector $y \in \mathbb{R}^m$, with $m \ll n$. In an attempt to mimic the uncompressed case, here we propose a similar refinement step applied to the "back-projection" estimate $\left(A^{(j)}\right)^T y^{(j)} = \widehat{x}^{(j)} \in \mathbb{R}^n$, which can essentially be thought of as one of many possible estimates or reconstructions of x that can be obtained from $y^{(j)}$ and $A^{(j)}$. The results in the next section quantify the improvements that can be achieved using this approach.

III. Main Results

To state our main results, we set the input parameters of Algorithm 1 as follows. Choose $\alpha \in (0, 1/3)$, let $b = (1-\alpha)/(1-2\alpha)$, and let $k = 1 + \lceil \log_b \log n \rceil$. Allocate sensing resources according to

$$R^{(j)} = \left\{ \begin{array}{cc} \alpha n \left(\frac{1-2\alpha}{1-\alpha}\right)^{j-1}, & j = 1, \ldots, k-1 \\ \alpha n, & j = k \end{array} \right\},$$

and note that this allocation guarantees that $R^{(j+1)}/R^{(j)} > 1/2$ and $\sum_{j=1}^{k} R^{(j)} \leq n$. The latter inequality ensures that the total sensing energy does not exceed the total sensing energy used in conventional CS. The number of measurements acquired in each step is

$$m^{(j)} = \left\{ \begin{array}{cc} \rho_0 s \log n/(k-1), & j = 1, \ldots, k-1 \\ \rho_1 s \log n, & j = k \end{array} \right\},$$

for some constants ρ_0 (which depends on the dynamic range) and ρ_1 (sufficiently large so that the results of Lemma 1 hold). Note that $m = O(s \log n)$, the same order as the minimum number of measurements required by conventional CS.

Our main result of the paper, stated below and proved in the Appendix, quantifies the error performance of one particular estimate obtained from adaptive observations collected using the CDS procedure.

Theorem 1. *Assume that $x \in \mathbb{R}^n$ is sparse with $s = n^{\beta/\log\log n}$ for some constant $0 < \beta < 1$. Furthemore, assume that each non-zero component of x satisfies $\sigma\mu \leq x_i \leq \mathrm{D}\sigma\mu$, for some $\mu > 0$. Here σ is the noise standard deviation, $\mathrm{D} > 1$ is the dynamic range of the signal, and μ^2 is the SNR. Adaptively measure x according to Algorithm 1 with the input parameters as specified above, and construct the estimator $\widehat{x}_{\mathrm{CDS}}$ by applying the Dantzig selector with $\lambda = \Theta(\sigma)$ to the output of the algorithm (i.e., with $A = A^{(k)}$ and $y = y^{(k)}$).*

1) *There exists $\mu_0 = \Omega(\sqrt{\log n/\log\log n})$ such that if $\mu \geq \mu_0$, then $\|\widehat{x}_{\mathrm{CDS}} - x\|_{\ell_2}^2 = O(s\sigma^2)$, with probability $1 - O(n^{-C_0'/\log\log n})$, for some $C_0' > 0$.*
2) *There exists $\mu_1 = \Omega(\sqrt{\log\log\log n})$ such that if $\mu_1 \leq \mu < \mu_0$, then $\|\widehat{x}_{\mathrm{CDS}} - x\|_{\ell_2}^2 = O(s\sigma^2)$, with probability $1 - O(e^{-C_1'\mu^2})$, for some $C_1' > 0$.*
3) *If $\mu < \mu_1$, then $\|\widehat{x}_{\mathrm{CDS}} - x\|_{\ell_2}^2 = O(s\sigma^2 \log\log\log n)$, with probability $1 - O(n^{-C_2'})$, for some $C_2' > 0$.*

In words, when the SNR is sufficiently large, the estimate achieves the error performance of the oracle-assisted estimator, albeit with a lower (slightly sub-polynomial) convergence rate. For a class of slightly weaker signals the oracle-assisted error performance is still achieved, but with a rate of convergence that is inversely proportional to the SNR. Note that we may summarize the results of the theorem with the general claim $\|\widehat{x}_{\mathrm{CDS}} - x\|_{\ell_2}^2 = O(s\sigma^2 \log\log\log n)$ with probability $1 - o(1)$. It is worth pointing out that for many problems of practical interest the $\log\log\log n$ term can be negligible, whereas $\log n$ is not; for example, $\log\log\log(10^6) < 1$, but $\log(10^6) \approx 14$.

IV. Extensions and Conclusions

Although the CDS procedure was specified under the assumption that the nonzero signal components were positive, it can be easily extended to signals having negative entries as well. In that case, one could split the budget of sensing resources in half, executing the procedure once as written and again replacing the refinement step by $\mathcal{I}^{(j+1)} = \{i \in \mathcal{I}^{(j)} : \widehat{x}_i^{(j)} < 0\}$. In addition, the results presented here also apply if the signal is sparse another basis. To implement the procedure in that case, one would generate the $A^{(j)}$ as above, but observations of x would be obtained using $A^{(j)}T$, where $T \in \mathbb{R}^{n\times n}$ is an appropriate orthonormal transformation matrix (discrete wavelet or cosine transform, for example). In either case the qualitative behavior is the same—observations are collected by projecting x onto a superposition of basis elements from the appropriate basis.

We have shown here that the compressive distilled sensing procedure can significantly improve the theoretical performance of compressive sensing. In experiments, not shown here due to space limitations, we have found that CDS can perform significantly better than CS in practice, like similar previously proposed adaptive methods [7]–[9]. We remark that our theoretical analysis shows that CDS is sensitive to the dynamic range of the signal. This is an artifact of the method for obtaining the signal estimate $\widehat{x}^{(j)}$ at each step. As alluded at the end of Section II, $\widehat{x}^{(j)}$ could be obtained using any of a number of methods including, for example, Dantzig selector estimation (with a smaller value of λ) or other mixed-norm reconstruction techniques such as LASSO with sufficiently small regularization parameters. Such extensions will be explored in future work.

V. Appendix

A. Lemmata

We first establish several key lemmata that will be used in the sketch of the proof of the main result. In particular, the first two results presented below quantify the effects of each refinement step.

Lemma 2. *Let $x \in \mathbb{R}^n$ be supported on $\mathcal{S}$ with $|\mathcal{S}| = s$, and let $x_{\mathcal{S}}$ denote the subvector of x composed of entries of x whose indices are in $\mathcal{S}$. Let A be an $m\times n$ matrix whose entries are i.i.d. $\mathcal{N}(0, \tau/m)$ for some $0 < \tau_{\min} \leq \tau$, and let $A_{\mathcal{S}}$ and $A_{\mathcal{S}^c}$ be submatrices of A composed of the columns of A corresponding to the indices in the sets $\mathcal{S}$ and $\mathcal{S}^c$, respectively. Let $w \in \mathbb{R}^m$ be independent of A and have i.i.d. $\mathcal{N}(0, \sigma^2)$ entries. For the $z \times 1$ vector $U = A_{\mathcal{S}^c}^T A_{\mathcal{S}} x_{\mathcal{S}} + A_{\mathcal{S}^c}^T w$, where $z = |\mathcal{S}^c| = n - s$, we have $(1/2 - \epsilon)\, z \leq \sum_{j=1}^{z} \mathbf{1}_{\{U_i > 0\}} \leq (1/2 + \epsilon)\, z$ for any $\epsilon \in (0, 1/2)$ with probability at least $1 - 2\exp(-2\epsilon^2 z)$.*

Proof: Define $Y = Ax + w = A_{\mathcal{S}} x_{\mathcal{S}} + w$, and note that given Y, the entries of $U = A_{\mathcal{S}^c}^T Y$ are i.i.d. $\mathcal{N}(0, \|Y\|_2^2 \tau/m)$. Thus, when $Y \neq 0$ we have $\Pr(U_i > 0) = 1/2$ for all $i = 1, \ldots, z$. Let $T_i = \mathbf{1}_{\{U_i > 0\}}$ and apply Hoeffding's inequality to obtain that for any $\epsilon \in (0, 1/2)$,

$$\Pr\left(\left|\sum_{i=1}^{z} T_i - \frac{z}{2}\right| > \epsilon z \;\middle|\; Y : Y \neq 0\right) \leq 2\exp\left(-2\epsilon^2 z\right).$$

Now, we integrate to obtain

$$\begin{aligned}
&\Pr\left(\left|\sum_{i=1}^{z} T_i - \frac{z}{2}\right| > \epsilon z\right) \\
&\quad\leq \int_{Y:Y\neq 0} 2\exp\left(-2\epsilon^2 z\right)\, dP_Y + \int_{Y:Y=0} 1\, dP_Y \\
&\quad\leq 2\exp\left(-2\epsilon^2 z\right).
\end{aligned}$$

The last result follows from the fact that the event $Y = 0$ has probability zero since Y is Gaussian-distributed. ∎

Lemma 3. *Let x, $\mathcal{S}$, $x_{\mathcal{S}}$, A, $A_{\mathcal{S}}$, and w be as defined in the previous lemma. Assume further that the entries of x satisfy $\sigma\mu \leq x_i \leq D\sigma\mu$ for $i \in \mathcal{S}$ for some $\mu > 0$ and fixed $D > 1$. Define*

$$\Delta = \exp\left(-\frac{m}{32\left(sD^2 + m\mu^{-2}/\tau_{\min}\right)}\right) < 1,$$

then for the $s \times 1$ vector $V = A_{\mathcal{S}}^T A_{\mathcal{S}} x_{\mathcal{S}} + A_{\mathcal{S}}^T w$, either of the following bounds are valid:

$$\Pr\left(\sum_{i=1}^{s} \mathbf{1}_{\{V_i > 0\}} \neq s\right) \leq 2s\Delta^2,$$

or

$$\Pr\left(\sum_{i=1}^{s}\mathbf{1}_{\{V_i>0\}} < s(1-3\Delta)\right) \le 4\Delta.$$

Proof: Given A_i (the ith column of A) we have

$$V_i \sim \mathcal{N}\left(\|A_i\|_{\ell_2}^2 x_i, \|A_i\|_{\ell_2}^2\left[\frac{\tau}{m}\sum_{\substack{j=1\\ j\neq i}}^{s} x_j^2 + \sigma^2\right]\right),$$

and so, by a standard Gaussian tail bound

$$\begin{aligned}\Pr(V_i \le 0 \mid A_i) &= \Pr\left(\mathcal{N}(0,1) > \frac{\|A_i\|_{\ell_2} x_i}{\sqrt{\frac{\tau}{m}\sum_{\substack{j=1\\ j\neq i}}^{s} x_j^2 + \sigma^2}}\right)\\ &\le \exp\left(-\frac{\|A_i\|_{\ell_2}^2 x_i^2}{2(\tau\|x\|^2/m + \sigma^2)}\right)\end{aligned}$$

Now, we can leverage a result on the tails of a chi-squared random variable from [12] to obtain that, for any $\gamma \in (0,1)$, $\Pr\left(\|A_i\|^2 \le (1-\gamma)\tau\right) \le \exp\left(-m\gamma^2/4\right)$. Again we employ conditioning to obtain

$$\begin{aligned}\Pr(V_i \le 0) &\le \int_{A_i:\|A_i\|^2\le(1-\gamma)\tau} 1\; dP_{A_i}\\ &+ \int_{A_i:\|A_i\|^2>(1-\gamma)\tau} \Pr(V_i \le 0 \mid A_i)\; dP_{A_i}\\ &\le \exp\left(-\frac{m\gamma^2}{4}\right) + \exp\left(-\frac{\tau(1-\gamma)x_i^2}{2(\tau\|x\|^2/m+\sigma^2)}\right)\\ &\le \exp\left(-\frac{m\gamma^2}{4}\right) + \exp\left(-\frac{\tau(1-\gamma)\mu^2}{2(\tau s D^2\mu^2/m+1)}\right),\end{aligned}$$

where the last bound follows from the conditions on the x_i. Now, to simplify, we choose $\gamma = \gamma^* \in (0,1)$ to balance the two terms, obtaining

$$\gamma^* = \left(sD^2 + \frac{m}{\tau\mu^2}\right)^{-1}\left(\sqrt{1+2\left(sD^2+\frac{m}{\tau\mu^2}\right)} - 1\right).$$

Using the fact that

$$\frac{\sqrt{1+2t}-1}{t} > \frac{1}{2\sqrt{t}},$$

for $t > 1$, we can conclude

$$\gamma^* > \frac{1}{2}\left(sD^2 + \frac{m}{\tau\mu^2}\right)^{-1/2},$$

since $s > 1$ by assumption. Now, using the fact that $\tau \ge \tau_{\min}$, we have that $\Pr(V_i \le 0) \le 2\Delta^2$, where

$$\Delta = \exp\left(-\frac{m}{32\left(sD^2 + m\mu^{-2}/\tau_{\min}\right)}\right).$$

The first result follows from

$$\begin{aligned}\Pr\left(\sum_{i=1}^{s}\mathbf{1}_{\{V_i>0\}} \neq s\right) &= \Pr\left(\bigcup_{i=1}^{s}\{V_i \le 0\}\right)\\ &\le s\max_{i\in\{1,\ldots,s\}}\Pr\left(V_i \le 0\right)\\ &\le 2s\Delta^2.\end{aligned}$$

For the second result, let us simplify notation by introducing the variables $T_i = \mathbf{1}_{\{V_i>0\}}$, and $t_i = \mathbb{E}[T_i]$. By Markov's Inequality we have

$$\begin{aligned}\Pr\left(\left|\sum_{i=1}^{s}T_i - \sum_{i=1}^{s}t_i\right| > p\right) &\le p^{-1}\mathbb{E}\left[\left|\sum_{i=1}^{s}T_i - \sum_{i=1}^{s}t_i\right|\right]\\ &\le p^{-1}\sum_{i=1}^{s}\mathbb{E}\left[|T_i - t_i|\right]\\ &\le p^{-1}s\max_{i\in\{1,\ldots,s\}}\mathbb{E}\left[|T_i - t_i|\right].\end{aligned}$$

Now note that

$$|T_i - t_i| = \begin{cases} 1 - P(V_i > 0), & V_i > 0\\ P(V_i > 0), & V_i \le 0\end{cases},$$

and so $\mathbb{E}\left[|T_i - t_i|\right] \le 2P(V_i \le 0)$. Thus, we have that $\max_{i\in\{1,\ldots,s\}}\mathbb{E}\left[|T_i - t_i|\right] = 2\Delta^2$, and so

$$\Pr\left(\left|\sum_{i=1}^{s}T_i - \sum_{i=1}^{s}t_i\right| > p\right) \le 4p^{-1}s\Delta^2.$$

Now, let $p = s\Delta$ to obtain

$$\Pr\left(\sum_{i=1}^{s}T_i < \sum_{i=1}^{s}t_i - s\Delta\right) \le 4\Delta.$$

Since $t_i = 1 - \Pr\left(V_i \le 0\right)$, we have $\sum_{i=1}^{s}t_i \ge s(1-2\Delta^2)$, and thus

$$\Pr\left(\sum_{i=1}^{s}T_i < s(1-2\Delta^2-\Delta)\right) \le 4\Delta.$$

The result follows from the fact that $2\Delta^2 + \Delta < 3\Delta$. ∎

Lemma 4. *For $0 < p < 1$ and $q > 0$, we have $(1-p)^q \ge 1 - qp/(1-p)$.*

Proof: We have $\log(1-p)^q = q\log(1-p) = -q\log\left(1 + p/(1-p)\right) \ge -qp/(1-p)$, where the last bound follows from the fact that $\log(1+t) \le t$ for $t \ge 0$. Thus, $(1-p)^q \ge \exp\left(-qp/(1-p)\right) \ge 1 - qp/(1-p)$, the last bound following from the fact $e^t \ge 1 + t$ for all $t \in \mathbb{R}$. ∎

B. Sketch of Proof of Theorem 1

To establish the main results of the paper, we will first show that the final set of observations of the CDS procedure is (with high probability) equivalent in distribution to a set of observations of the form (1), but with different parameters (smaller effective dimension n_{eff} and effective noise power σ_{eff}^2), and for which some fraction of the original signal components may be absent. To that end, let $\mathcal{S}^{(j)} = \mathcal{S}\cap\mathcal{I}^{(j)}$ and $Z^{(j)} = \mathcal{S}^c\cap\mathcal{I}^{(j)}$, for $j = 1,\ldots,k$, denote the (sub)sets of indices of $\mathcal{S}$ and its complement, respectively, that remain to be measured in step j. Note that at each step of the procedure, the "back-projection" estimate $\widehat{x}^{(j)} = \left(A^{(j)}\right)^T A^{(j)}x + \left(A^{(j)}\right)^T w^{(j)}$ can be decomposed into $\widehat{x}_{S^{(j)}} = \left(A^{(j)}_{S^{(j)}}\right)^T A^{(j)}_{S^{(j)}}x_{S^{(j)}} + \left(A^{(j)}_{S^{(j)}}\right)^T w^{(j)}$ and $\widehat{x}_{Z^{(j)}} = \left(A^{(j)}_{Z^{(j)}}\right)^T A^{(j)}_{S^{(j)}}x_{S^{(j)}} + \left(A^{(j)}_{Z^{(j)}}\right)^T w^{(j)}$, and that these subvectors are precisely of the form specified in the conditions of Lemmas 2 and 3.

Let $z^{(j)} = |Z^{(j)}|$ and $s^{(j)} = |S^{(j)}|$, and in particular note that $s^{(1)} = s$ and $z^{(1)} = z = n - s$. Choose the parameters of the CDS algorithm as specified in Section III. Iteratively applying Lemma 2 we have that for any fixed $\epsilon \in (0, 1/2)$, the bounds $(1/2-\epsilon)^{j-1} z \leq z^{(j)} \leq (1/2+\epsilon)^{j-1} z$ hold simultaneously for all $j = 1, 2, \ldots, k$ with probability at least $1 - 2(k-1)\exp\left(-2z\epsilon^2 (1/2-\epsilon)^{k-2}\right)$, which is no less than $1 - O\left(\exp\left(-c_0 n/\log^{c_1} n\right)\right)$, for some constants $c_0 > 0$ and $c_1 > 0$, for n sufficiently large[2]. As a result, with the same probability, the total number of locations in the set $\mathcal{I}^{(j)}$ satisfies $|\mathcal{I}^{(j)}| \leq s^{(1)} + z^{(1)} \left(\frac{1}{2}+\epsilon\right)^{j-1}$, for all $j = 1, 2, \ldots, k$. Thus, we can lower bound $\tau^{(j)} = R^{(j)}/|\mathcal{I}^{(j)}|$ at each step by

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

Now, note that when n is sufficiently large[3], we have $s \leq z\,(1/2+\epsilon)^{j-1}$ holding for all $j = 1, \ldots, k$. Letting $\epsilon = (1-3\alpha)/(2-2\alpha)$, we can simplify the bounds on $\tau^{(j)}$ to obtain that $\tau^{(j)} \geq \alpha/2$ for $j = 1, \ldots, k-1$, and $\tau^{(k)} \geq \alpha \log(n)/2$. The salient point to note here is the value of $\tau^{(k)}$, and in particular, its dependence on the signal dimension n. This essentially follows from the fact that the set of indices to measure decreases by a fixed factor with each distillation step, and so after $O(\log\log n)$ steps the number of indices to measure is smaller than in the initial step by a factor of about $\log n$. Thus, for the same allocation of resources ($R^{(1)} = R^{(k)}$), the SNR of the final set of observations is larger than that of the first set by a factor of $\log n$.

Now, the final set of observations is $y^{(k)} = A^{(k)} x^{(k)} + w^{(k)}$, where $x^{(k)} \in \mathbb{R}^{n_{\text{eff}}}$ (for some $n_{\text{eff}} < n$) is supported on the set $\mathcal{S}^{(k)} = \mathcal{S} \cap \mathcal{I}^{(k)}$, $A^{(k)}$ is an $m^{(k)} \times n_{\text{eff}}$ matrix, and the $\widetilde{w}_i$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. We can divide throughout by $\tau^{(k)}$ to obtain the equivalent statement $\widetilde{y} = \widetilde{A}\widetilde{x} + \widetilde{w}$, where now the entries of $\widetilde{A}$ are i.i.d. $\mathcal{N}(0, 1/m)$ and the $\widetilde{w}_i$ are i.i.d. $\mathcal{N}(0, \widetilde{\sigma}^2)$, where $\widetilde{\sigma}^2 \leq 2\sigma^2/(\alpha \log n)$. To bound the overall squared error we consider the *variance* associated with estimating the components of $\widetilde{x}$ using the Dantzig selector (cf. Lemma 1), as well as the (squared) *bias* arising from the fact that some signal components may not be present in the final support set $\mathcal{S}^{(k)}$. In particular, a bound for the overall error is given by

$$\begin{aligned} \|\widehat{x} - x\|_{\ell_2}^2 &= \|\widehat{x} - \widetilde{x} + \widetilde{x} - x\|_{\ell_2}^2 \\ &\leq 2\|\widehat{x} - \widetilde{x}\|_{\ell_2}^2 + 2\|\widetilde{x} - x\|_{\ell_2}^2. \end{aligned}$$

We can bound the first term by applying the result of Lemma 1 to obtain that (for ρ_1 sufficiently large) $\|\widehat{x} - \widetilde{x}\|_{\ell_2}^2 = O(s\sigma^2)$ holds with probability $1 - O(n^{-C_0})$, for some $C_0 > 0$. Now, let $\delta = (|\mathcal{S}| - |\mathcal{S}^{(k)}|)/s$ denote the fraction of true signal components that are rejected by the CDS procedure. Then we have $\|\widetilde{x} - x\|_{\ell_2}^2 = O(s\sigma^2\delta\mu^2)$, and so overall, we have $\|\widehat{x} - x\|_{\ell_2}^2 = O(s\sigma^2 + s\sigma^2\delta\mu^2)$, with probability $1 - O(n^{-C_0})$. The method for bounding the second term in the error bound varies depending on the signal amplitude μ; we consider three cases below.

1) $\mu \geq (8D\sqrt{3/\alpha})\sqrt{\log n/\log\log n}$: Conditioned on the event that the stated lower-bounds for $\tau^{(j)}$ are valid, we can iteratively apply Lemma 3, taking $\tau_{\min} = \alpha/2$. For $\rho_0 = 96D^2/\log b$ (where b is the parameter from the expression for k), let $m^{(j)} = \rho_0 s \log n/\log_b \log n$. Then we obtain that for all n sufficiently large, $\delta = 0$ with probability at least $1 - O(n^{-C_0'/\log\log n})$, for some constant $C_0' > 0$. Since this term governs the rate, we have overall that $\|\widehat{x} - x\|_{\ell_2}^2 = O(s\sigma^2)$ holds with probability $1 - O(n^{-C_0'/\log\log n})$ as claimed.

2) $(16\sqrt{2/(\alpha\log b)})\sqrt{\log\log\log n} \leq \mu < (8D\sqrt{3/\alpha})\sqrt{\log n/\log\log n}$: For this range of signal amplitude we will need to control δ explicitly. Conditioned on the event that the lower-bounds for $\tau^{(j)}$ hold, we iteratively apply Lemma 3 where for $\rho_0 = 96D^2/\log b$, we let $m^{(j)} = \rho_0 s \log n/\log_b \log n$. Now, we invoke Lemma 4 to obtain that for n sufficiently large, $\delta = 1 - (1-3\Delta)^{k-1} = O(e^{-C_1'\mu^2})$ with probability at least $1 - O(e^{-C_1'\mu^2})$ for some $C_1' > 0$. It follows that $\delta\mu^2$ is $O(1)$, and so overall $\|\widehat{x} - x\|_{\ell_2}^2 = O(s\sigma^2)$ with probability $1 - O(e^{-C_1'\mu^2})$.

3) $\mu < (16\sqrt{2/(\alpha\log b)})\sqrt{\log\log\log n}$: Invoking the trivial bound $\delta \leq 1$, it follows from above that for n sufficiently large, the error satisfies $\|\widehat{x} - x\|_{\ell_2}^2 = O(s\sigma^2 \log\log\log n)$, with probability $1 - O(n^{-C_2'})$ for some constant $C_2' > 0$, as claimed.

[2] In particular, we require $n \geq c_0'(\log\log\log n)(\log n)^{c_1'}/(1 - n^{c_2'/\log\log n - 1})$, where c_0', c_1', and c_2' are positive functions of ϵ and β.

[3] In particular, we require $n \geq (1+\log n)^{\log\log n/(\log\log n - \beta)}$.

REFERENCES

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

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

[3] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[4] R. Baraniuk, M. Davenport, R. A. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, 2008.

[5] J. Haupt and R. Nowak, "Signal reconstruction from noisy random projections," *IEEE Trans. Inform. Theory*, vol. 52, no. 9, pp. 4036–4048, Sept. 2006.

[6] E. J. Candès and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Statist.*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.

[7] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 56, no. 6, pp. 2346–2356, June 2008.

[8] R. Castro, J. Haupt, R. Nowak, and G. Raz, "Finding needles in noisy haystacks," in *Proc. IEEE Conf. Acoustics, Speech, and Signal Proc.*, Honolulu, HI, Apr. 2008, pp. 5133–5136.

[9] J. Haupt, R. Castro, and R. Nowak, "Adaptive sensing for sparse signal recovery," in *Proc. IEEE 13th Digital Sig. Proc./5th Sig. Proc. Education Workshop*, Marco Island, FL, Jan. 2009, pp. 702–707.

[10] J. Haupt, R. Castro, and R. Nowak, "Adaptive discovery of sparse signals in noise," in *Proc. 42nd Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2008, pp. 1727–1731.

[11] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Selective sampling for sparse signal recovery," in *Proc. 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Clearwater Beach, FL, Apr. 2009, pp. 216–223.

[12] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Ann. Statist.*, vol. 28, no. 5, pp. 1302–1338, Oct. 2000.