FINDING NEEDLES IN NOISY HAYSTACKS

R. M. Castro, J. Haupt, R. Nowak

University of Wisconsin-Madison, ECE Department
1415 Engineering Drive, Madison, WI 53606 USA

G. M. Raz

GMR Research and Technology
Concord, MA 01742-3819 USA

ABSTRACT

The theory of compressed sensing shows that samples in the form of random projections are optimal for recovering sparse signals in high-dimensional spaces (i.e., finding needles in haystacks), *provided the measurements are noiseless*. However, noise is almost always present in applications, and compressed sensing suffers from it. The signal to noise ratio per dimension using random projections is very poor, since sensing energy is equally distributed over all dimensions. Consequently, the ability of compressed sensing to locate sparse components degrades significantly as noise increases. It is possible, in principle, to improve performance by "shaping" the projections to focus sensing energy in proper dimensions. The main question addressed here is, can projections be adaptively shaped to achieve this focusing effect? The answer is yes, and we demonstrate a simple, computationally efficient procedure that does so.

Index Terms— sparse approximation, compressed sensing, reconstruction, adaptive sampling

1. INTRODUCTION

Surprising mathematical findings and stunning practical results have propelled *compressed sensing* into the signal processing limelight and have had a profound effect on our understanding of signal acquisition and sampling. Consider a signal that can be represented (exactly or approximately) by a sparse representation (the superposition of a small number of basis vectors). The basic idea of compressed sensing is that if one takes samples in the form of projections of the signal and if these projections are incoherent with the basis vectors, then the sparse representation can be recovered from a small number of such samples (roughly proportional to the number of components in the sparse representation) provided the observations are noise-free [1, 2]. In addition, compressed sensing remains stable in the presence of random noise; i.e., the recovery degrades gracefully, but markedly, as the noise level is increased [3, 4]. This paper investigates the noise sensitivity phenomenon and proposes an improved approach based on adaptive sensing.

Incoherence between the projection vectors and the signal basis vectors is essential to compressed sensing, and is required for successful recovery from a small number of *non-adaptive* samples. The incoherence condition guarantees that one "spreads" the sensing energy over all the dimensions of the coordinate system of the basis. In essence, each compressive sample deposits an equal fraction of sensing energy in every dimension, making it possible to locate the sparse components without sensing directly in each and every dimension, which would require a number of samples equal to the length of the signal. When the observations are corrupted by noise, however, the signal to noise ratio (SNR) *per dimension* is necessarily much lower using this approach than if we had used all sensing energy to probe a single coordinate. Thus, noise can make the recovery of the sparse components much more difficult.

It is intuitively clear that focused samples can be tremendously helpful. Indeed, if a genie were to provide the locations of the sparse signal components a priori, then we would know that the optimal samples would be projections on to the corresponding basis vectors themselves, maximizing the SNR per sample. Without a genie, it is sensible to attempt to recover the locations directly so that subsequent samples can be focused into the correct subspace. The potential advantages of an adaptive projection scheme are demonstrated in [5], but this procedure does not scale well with problem dimension. Here we propose a different adaptive strategy for which the shaping of the projections can be computed in time linear in the length of the signal, and therefore is no more computationally demanding than standard compressed sensing. Begin with an incoherent projection sample, which should provide a crude indication of potential locations for the sparse components. Now, use this information to shape the next projection so that it is a bit less incoherent and a bit more focused on these potential locations. Repeat this procedure until the projections are mostly focused on one location, which hopefully corresponds to an actual signal component. Keep iterating this process, with the previously identified components removed, until no additional significant components are found.

The remainder of the paper is organized as follows. A brief review of traditional (non-adaptive) compressive sensing is given in Section 2. In Section 3 we describe our strategy for projection focusing that is based on a general-purpose Bayesian model for sparse components and an (approximate) entropy-maximizing projection shaping at each step. Computational experiments in Section 4 demonstrate that significant performance gains are possible through this adaptive procedure, especially when the signal is very sparse and the SNR per dimension is low. Finally, some conclusions are discussed in Section 5.

2. COMPRESSIVE SENSING REVIEW

Compressive sensing (CS) describes a collection of methods by which sparse high-dimensional signals can be accurately and efficiently recovered from a small (relative to the dimension) number of observations. CS employs a sampling model which is a natural generalization of conventional point sampling. Each observation of an m-sparse vector $\boldsymbol{x} \in \mathbb{R}^n$ is described by

$$Y(t) = \boldsymbol{\phi}(t)^T \boldsymbol{x} + W(t), \tag{1}$$

for $t = 1, 2, \ldots, k$, where the sampling vector $\boldsymbol{\phi}(t) \in \mathbb{R}^n$ is chosen by and known to the observer and satisfies $\|\boldsymbol{\phi}(t)\|_2 = 1$, and $W(t) \sim \mathcal{N}\left(0, \sigma_w^2\right)$ is independent of $\boldsymbol{\phi}(t)$.

This work was partially supported by the DARPA Analog-to-Information Program.

1-4244-1484-9/08/$25.00 ©2008 IEEE

The earliest contributions to CS considered noiseless settings where the sampling vectors $\{\boldsymbol{\phi}(t)\}_{t=1}^{k}$ were a collection of random vectors whose entries were drawn independently according to some distribution (e.g., Gaussian). In these settings, it was shown that Basis Pursuit (identifying the vector with minimum ℓ_1 norm[1] that agrees with the observations) efficiently recovers any m-sparse signal with overwhelming probability, provided the number of observations satisfies $k \geq Cm \log n$ where C is some constant that does not depend on the problem dimension [1, 2]. In practice, it has been observed that between $3m$ and $5m$ samples often suffice.

In settings where sampling noise is present, the provable performance of CS degrades markedly. The Basis Pursuit approach does not apply directly in this setting, and one possible estimation strategy is to minimize the weighted sum of a squared error term and a complexity term, given by

$$\widehat{\boldsymbol{x}}_k = \arg \min_{\boldsymbol{g} \in \mathbb{R}^n} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{\Phi} \boldsymbol{g}\|_2^2 + \tau \|\boldsymbol{g}\|_1, \qquad (2)$$

where $\boldsymbol{y}$ is a vector of the observations $\{y(t)\}_{t=1}^{k}$, $\boldsymbol{\Phi}$ is a matrix with rows given by the corresponding $\boldsymbol{\phi}(t)$, and τ is an appropriate tolerance. Other similar strategies have been proposed and analyzed, yielding estimates that satisfy

$$\mathbb{E}\left[\frac{\|\widehat{\boldsymbol{x}}_k - \boldsymbol{x}\|^2}{n}\right] \leq C \left(\frac{k}{m \log n}\right)^{-1}, \qquad (3)$$

where C is a constant that depends on the noise power, and the expectation is over the distribution of the noise and the projection vectors [3, 4]. It is interesting to note that this bound is meaningful only when the number of observations is at least $O(m \log n)$. This is similar to the number of observations required in the noise-free setting – the difference here is that the error decays relatively slowly after this point.

3. ADAPTIVE PROJECTIONS FOR SPARSE RECOVERY

In this section we present an adaptive projection algorithm targeting problems where the signal is very sparse (e.g., described by a small number of components). The proposed approach consists of a greedy procedure that attempts to recover the signal sequentially, component-by-component, and is inspired by our earlier work [6] where we considered a parametric model. In this work we use a related model for which it is easy to use a Bayesian approach to estimate the parameters. In [6] this is done using non-adaptive random projections. Here we propose a technique to adapt the projections based on previous observations, in order to significantly improve the estimation performance. We first describe our methodology when the signal has a single non-zero component, and later we generalize this approach for sparse signals with multiple non-zero components.

3.1. A Single Needle in the Haystack

Let $\boldsymbol{x} \in \mathbb{R}^n$, $n \in \mathbb{N}$ be a vector with at most one non-zero entry. The adaptive projection procedure proposed follows a Bayesian style approach, and so we have a generative model for the signal $\boldsymbol{x}$. Let t index the sequential sampling process. At step t, define the random variable $L(t) \in \{1, \ldots, n\}$, with probability mass function $p_i(t) = \Pr(L(t) = i)$. That is, $L(t)$ is a discrete random variable over the indices of the signal, modeling that entry i is nonzero with probability $p_i(t)$. Conditional on the value of $L(t)$ the amplitude of the non-zero signal component is modeled as a Gaussian random variable, $A(t)|L(t) = i \sim \mathcal{N}(\mu_i(t), \sigma_i^2(t))$. Thus, our model has the form

$$\boldsymbol{X}(t) = (0, \ldots, 0, A(t), 0 \ldots, 0) \ ,$$

where only the entry $L(t)$ of $\boldsymbol{X}(t)$ is non-zero. We assume $\boldsymbol{x}$ is a realization of random variable $\boldsymbol{X}(t)$. Notice that the distribution is parameterized by three quantities: $\boldsymbol{p}(t) \triangleq (p_1(t), \ldots, p_n(t))$, $\boldsymbol{\mu}(t) \triangleq (\mu_1(t), \ldots, \mu_n(t))$, and $\boldsymbol{\sigma^2}(t) \triangleq (\sigma_1^2(t), \ldots, \sigma_n^2(t))$. Initially, when $t = 0$ and no samples have been taken, we start with a uniform prior on the location, and zero mean distribution for the conditional amplitude, specifically $\boldsymbol{p}(0) \triangleq (1/n, \ldots, 1/n)$, $\boldsymbol{\mu}(0) = (0, \ldots, 0)$ and $\boldsymbol{\sigma^2}(0) \triangleq (\sigma_0^2, \ldots, \sigma_0^2)$, where $\sigma_0^2 > 0$. This prior distribution is updated in a Bayesian manner as samples are acquired, giving rise to the model at step t, as described above.

Recall the observation model in (1). Using Bayes rule we can update the posterior distribution, and straightforward calculations yield the following update rules

$$\begin{aligned}
\mu_i(t+1) &= \frac{\phi_i(t)\sigma_i^2(t)y(t) + \mu_i(t)\sigma_w^2}{\phi_i^2(t)\sigma_i^2(t) + \sigma_w^2} \ , \\
\sigma_i^2(t+1) &= \frac{\sigma_i^2(t)\sigma_w^2}{\phi_i^2(t)\sigma_i^2(t) + \sigma_w^2} \ , \\
p_i(t+1) &\propto \frac{p_i(t) \exp\left(-\frac{1}{2}\frac{(y(t)-\phi_i(t)\mu_i(t))^2}{\phi_i^2(t)\sigma_i^2(t)+\sigma_w^2}\right)}{\sqrt{\phi_i^2(t)\sigma_i^2(t) + \sigma_w^2}} \ ,
\end{aligned}$$

where $y(t)$ is a realization of $Y(t)$, and in the update of $\boldsymbol{p}(t+1)$ we omit the explicit expression of the normalization constant.

The choice of the projection vectors $\boldsymbol{\phi}(t)$ is critical for good performance. If we are constrained not to use adaptive projections it is known that random projections are as uniformly informative as possible. These can be, for example, Rademacher random vectors (n-vectors comprised of i.i.d. random variables taking values $\pm 1/\sqrt{n}$ with equal probability). However, if that constraint is removed and adaptivity is allowed, then one can use information gleaned from previous samples to "focus" the projection vectors, leading to better performance.

We propose the following methodology: define the "shaped" random projection

$$\boldsymbol{\phi}(t+1) = (\sqrt{p_1(t)}B_1, \sqrt{p_2(t)}B_2, \ldots, \sqrt{p_n(t)}B_n)$$

where $\{B_i\}$ are i.i.d. random variables, taking value ± 1 with equal probability. Note that since $\sum_{i=1}^{n} p_i(t) = 1$ (because $\boldsymbol{p}$ is a discrete probability distribution) we have $\|\boldsymbol{\phi}(t)\|_2 = 1$. If at time t we are very confident that i is the only non-zero entry of $\boldsymbol{x}$, that is $p_i(t)$ is close to 1, then the shaped projection vector is going to put a large amount of mass on that entry. While this may appear intuitively reasonable, there is also a principled rationale for this particular shaping procedure, namely it is an attempt to make observation $Y(t)$ as informative as possible.

A way of characterizing the information content of $Y(t)$ is to compute its differential entropy, as defined in [7]. In other words we want to find $\boldsymbol{\phi}(t+1)$ solving

$$\arg \max_{\boldsymbol{h} : \|\boldsymbol{h}\|_2 = 1} H(\boldsymbol{h}^T \boldsymbol{X}(t) + W(t+1)) \ , \qquad (4)$$

where $H(\cdot)$ is the differential entropy and $\boldsymbol{X}(t)$ is a random variable distributed according our generative model at step t. In other

[1] The ℓ_1 norm is defined by $\|\boldsymbol{x}\|_1 \triangleq \sum_{i=1}^{n} |x_i|$, where x_i is the ith component of $\boldsymbol{x}$.

Table 1. *Empirical probabilities of successful support identification for the adaptive procedure and standard random projections (using one step of OMP). For high noise levels (small S), more than 15 times as many random projections are needed for OMP to match the performance of the adaptive procedure.*

S	10	5.0	2.0	1.5	1.0	0.9	0.8	0.5	0.3	0.1
Average k'	16.46	17.09	20.23	21.84	26.56	27.79	30.01	39.94	58.46	153.9
$P_s(\text{Adaptive}, k')$	0.989	0.985	0.960	0.963	0.952	0.953	0.969	0.977	0.978	0.995
$P_s(\text{OMP}, k')$	0.018	0.020	0.016	0.015	0.030	0.021	0.022	0.025	0.030	0.028
$P_s(\text{OMP}, 5k')$	0.485	0.412	0.412	0.379	0.392	0.397	0.387	0.384	0.386	0.419
$P_s(\text{OMP}, 10k')$	0.944	0.927	0.856	0.860	0.836	0.808	0.812	0.774	0.761	0.783
$P_s(\text{OMP}, 15k')$	0.993	0.994	0.982	0.981	0.967	0.966	0.962	0.938	0.910	0.891
$P_s(\text{OMP}, 30k')$	1.000	1.000	1.000	1.000	0.998	1.000	1.000	0.998	0.994	0.993

words $\boldsymbol{X}(t)$ reflects our knowledge of $\boldsymbol{x}$ at time t. Now note that under our model $\boldsymbol{h}^T\boldsymbol{X}(t)$ is distributed as a Gaussian mixture with n components (recall that at most one entry of $\boldsymbol{X}(t)$ is non-zero). In particular the density of $\boldsymbol{h}^T\boldsymbol{X}(t)$ is

$$\sum_{i=1}^{n} \frac{p_i(t)}{\sqrt{2\pi h_i^2\,\sigma_i^2}} \exp\left(-\frac{(x-h_i\mu_i(t))^2}{2h_i^2\sigma_i^2(t)}\right) .$$

There is no closed form expression for the differential entropy of a Gaussian mixture. Instead, using the fact that the conditional differential entropy is a lower bound for the differential entropy [7], and conditioning on the selection of the mixture component, we obtain

$$H(\boldsymbol{h}^T\boldsymbol{X}(t)) \geq \frac{1}{2}\log\left(2\pi \mathrm{e}\prod_{i=1}^{n}(h_i^2\sigma_i^2(t))^{p_i(t)}\right) .$$

Replacing the entropy in (4) by the lower bound yields

$$\begin{aligned}\boldsymbol{\phi}(t+1) &= \underset{\boldsymbol{h}:\|\boldsymbol{h}\|_2=1}{\arg\max}\ \frac{1}{2}\log\left(2\pi \mathrm{e}\prod_{i=1}^{n}(h_i^2\sigma_i^2(t))^{p_i(t)}\right)\\ &= \underset{\boldsymbol{h}:\|\boldsymbol{h}\|_2=1}{\arg\max}\ \sum_{i=1}^{n} p_i(t)\log(h_i^2) .\end{aligned}$$

It is easily shown that $\phi_i(t+1) = \pm\sqrt{p_i(t)}$, which motivates our choice of projection vectors.

When a budget of k projective observations is allowed one can use the above algorithm to collect all the observations, and the final estimate can be computed from the posterior (different estimates should be used, to minimize the desired cost function). If optimizing mean squared error, then the best estimate is simply $\widehat{\boldsymbol{x}}_k = (\mu_1(k)p_1(k), \ldots, \mu_n(k)p_n(k))$.

3.2. Multiple Needles in the Haystack

Here we describe a modification of the procedure above when multiple entries of the signal are active (i.e., $\boldsymbol{x}$ might have more than a single non-zero entry). The idea is to search for the significant entries of $\boldsymbol{x}$ one at the time, using the previously developed method. Once an entry is found, no more observation energy is allocated to it. As time proceeds one gets closer to the single needle model.

The procedure starts exactly as in the single spike case, and proceeds until one entry of $\boldsymbol{p}(t)$ exceeds a threshold, say 0.9. As this point we infer there is significant signal value in the corresponding location, and proceed by measuring that entry directly using a projection vector that is just a singleton. The observed value becomes our estimate for the signal value at that location. We then restart the entire estimation procedure, but zero-out in $\boldsymbol{p}(t+1)$ the entry that we just measured. All the other entries of $\boldsymbol{p}(t+1)$ are equal (uniform prior). The procedure is iterated until the observation budget is expended. Unlike in the single needle model it is important to measure each detected entry directly because model mismatch often makes the estimates obtained directly from the algorithm inaccurate.

4. EXPERIMENTAL COMPARISON

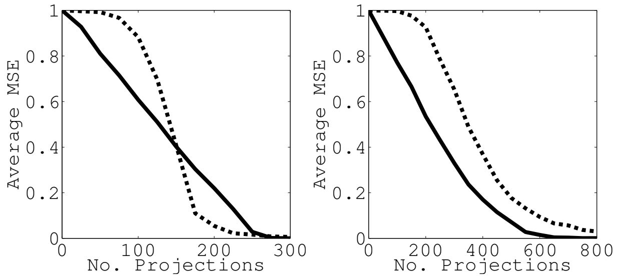
In this section we demonstrate the benefits of our proposed adaptive procedure relative to traditional random projections in several recovery tasks. First, we show that our adaptive procedure can identify true signal components much more effectively than orthogonal matching pursuit (OMP) [8] applied to standard (non-adaptive) random projection observations. To achieve comparable performance, OMP requires as many as *15-30 times* as many observations as the adaptive procedure. Second, we demonstrate that our adaptive sampling procedure often yields lower average reconstruction errors than standard random projections, and the benefit becomes more pronounced as the noise power increases. For all experiments, we considered target signals $\boldsymbol{x} \in \mathbb{R}^n$, $n = 2^{13}$, with $m = 15$ nonzero entries of the same amplitude (with random signs) at random locations, and we enforced $\|\boldsymbol{x}\|_2 = 1$. Noise power is quantified by the SNR, $S \triangleq \|\boldsymbol{x}\|^2/n\sigma_w^2$.

4.1. Support Identification

First we demonstrate the effectiveness of the adaptive procedure in support identification. For a fixed SNR, we generated a target signal as above and ran the adaptive procedure until one of the entries of the posterior probability vector exceeded 0.9. The required number of observations (k') was recorded, along with the index of the maximum of the posterior vector (the estimate of the support). For comparison we obtained support estimates using one index-selection step of OMP[2] applied to collections of non-adaptive random projection observations (using n-vectors with i.i.d. $\pm 1/\sqrt{n}$ entries). The number of non-adaptive observations for each of the OMP trials was a multiple of k'. Each experiment was termed a success if the support estimate contained the index of at least one true signal component. The average number of observations required (Average k') for one step of the adaptive procedure and the empirical probabilities of success (P_s) for each setting were determined by averaging over 1000 trials.

The results are given in Table 1. We see that adaptive sampling clearly outperforms random sampling, and in some cases up to 30 times as many random samples are required to achieve the detection

[2] The OMP index-selection step identifies the index i (or indices, in the case of a tie) for which $|r_i| = \max_i |r_i| \triangleq \|\boldsymbol{r}\|_\infty$, where $\boldsymbol{r} = \boldsymbol{\Phi}^T\boldsymbol{y}$.



(a) (b)

Fig. 1. *MSE comparisons between reconstructions obtained from adaptive samples and random projections (solid and dashed lines, respectively) for $S = 10$ and $S = 1.0$.*

performance of the adaptive method. It is also interesting to note that the adaptive procedure consistently identified true components of the signal with less than 5% error for each SNR considered. The increasing noise power essentially affected only the number of observations needed for the algorithm to converge to a true component.

4.2. Signal Reconstruction

Next we demonstrate the advantage of adaptive samples over random projections for signal reconstruction. To ascertain the effectiveness of the sampling procedure (independent of the reconstruction algorithm) we reconstruct in each case using (2) followed by debiasing. In addition, we eliminated the dependence of (2) on the regularization parameter by clairvoyantly selecting the value that gave the reconstruction with the lowest mean-square error (MSE). We used the GPSR (Gradient Projection for Sparse Reconstruction) software [9] to efficiently perform the optimization.

Fixing the number of observations k, we ran each sampling procedure to obtain the associated sampling matrices and observation vectors. Estimates $\widehat{\boldsymbol{x}} = \widehat{\boldsymbol{x}}(\alpha)$ were obtained for 41 distinct values of τ, given by $\tau = \alpha \|\boldsymbol{\Phi}^T \boldsymbol{y}\|_\infty$, where α ranged from 0 to 1 uniformly in increments of 0.025, and for each estimate the mean-square error $\|\widehat{\boldsymbol{x}}(\alpha) - \boldsymbol{x}\|_2^2$ was computed.[3] The error associated with a given sampling procedure was chosen to be the minimum error achieved over all tested values of α. This entire procedure was performed 40 times for each value of k, and the resulting minimum MSE's were averaged. The results of this experiment for two different noise levels ($S = 10$ and $S = 1.0$) are shown in Fig. 1(a) and (b), respectively.

The data in Table 1 suggest that the adaptive procedure sequentially identifies true components of the signal, and the number of observations for each discovery depends on the SNR. Thus, it is natural to predict that the reconstruction error of the adaptive procedure will qualitatively match the best approximation error of the target signal. Since all of the nonzero entries have the same amplitude, the (noise-free) approximation error will decay linearly in the number of components that are identified – retaining T components gives a squared approximation error of $1 - T(1/m)$. For the low noise setting simulated in Fig. 1(a), the data in Table 1 suggest that one true signal component is identified for every 16.5 observations, resulting in a predicted MSE of $1 - (k/16.5)(1/m)$ and full signal recovery after $(16.5)(15) \approx 250$ observations. This agrees with the observed behavior except that as the SNR decreases, the slope of the error decay changes with the instantaneous SNR, explaining the “flattening” of the curve. The same behavior is exhibited in the higher-noise setting.

The reconstruction errors using random projections exhibit a different behavior. When the SNR is high the performance is well-predicted by noiseless CS results – the reconstruction error decays to zero exponentially in the number of observations, provided enough observations are collected to ensure that certain submatrices of the observation matrix are well-conditioned. This explains the transitional error behavior for traditional compressed sensing that is apparent in Fig. 1(a). As the noise level increases, the rate of error decay becomes only polynomial in the number of observations (see (3)). It is also interesting to note that when the number of observations is less than about 50 in Fig. 1(a) and 100 in Fig. 1(b), the adaptive procedure succeeds at identifying some of the true signal components while the best reconstructions using random projections have MSE comparable to the all-zero solution.

5. CONCLUSIONS AND OPEN PROBLEMS

This paper presented a novel adaptive scheme for compressive sensing and demonstrated that it improves performance in many situations compared to non-adaptive random projection methods, providing evidence that while non-adaptive random projections are effective in noiseless situations, adaptivity can be very helpful in real-world problems. We compared our approach with the adaptive projection method of [5], and although the performance of the latter is competitive, it is only computationally feasible for relatively small problem sizes, making it intractable for the settings considered in this paper. Currently, we are investigating methodologies with provable performance, in the spirit of [6], which also provides evidence that adaptive sampling can outperform compressed sensing in noisy conditions.

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

[4] E. Candès and T. Tao, “The Dantzig selector: statistical estimation when p is much larger than n,” *Annals of Statistics*, vol. 35, no. 6, pp. 2313–2351, Dec. 2007.

[5] S. Ji, Y. Xue, and L. Carin, “Bayesian compressive sensing,” *IEEE Trans. Signal Processing*, 2007, accepted.

[6] R. Castro, J. Haupt, and R. Nowak, “Compressed sensing vs active learning,” in *Proc. IEEE Intl. Conf. Acoustics, Speech and Sig. Proc.*, Toulouse, FR., May 2006, vol. 3, pp. 820–823.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York, 1991.

[8] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in *Proceedings of the 27th Asilomar Conference on Signals, Systems, and Computers*, Nov. 1993.

[9] M. T. Figueiredo, R. D. Nowak, and S. J. Wright, “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems,” *IEEE J. Select. Topics Signal Processing*, vol. 1, no. 4, pp. 586–597, Dec. 2007.

[3] As noted in [9], choosing $\tau = \|\boldsymbol{\Phi}^T \boldsymbol{y}\|_\infty$ guarantees an all-zero solution while $\tau = 0$ gives the least-squares solution, so this parametrization covers the entire usable range of parameter values.