COMPRESSED SENSING VS. ACTIVE LEARNING

Rui Castro, Jarvis Haupt, Robert Nowak

Department of Electrical and Computer Engineering, Rice University, Houston, TX
Department of Electrical and Computer Engineering, University of Wisconsin, Madison, WI

ABSTRACT

Compressive sampling (CS), or *Compressed Sensing*, has generated a tremendous amount of excitement in the signal processing community. Compressive sampling, which involves non-traditional samples in the form of randomized projections, can capture most of the salient information in a signal with a relatively small number of samples, often far fewer samples than required using traditional sampling schemes. Adaptive sampling (AS), also called *Active Learning*, uses information gleaned from previous observations (*e.g.*, feedback) to focus the sampling process. Theoretical and experimental results have shown that adaptive sampling can dramatically outperform conventional (non-adaptive) sampling schemes. This paper compares the theoretical performance of compressive and adaptive sampling in noisy conditions, and it is shown that for certain classes of piecewise constant signals and high SNR regimes both CS and AS are near-optimal. This result is remarkable since it is the first evidence that shows that compressive sampling, which is non-adaptive, cannot be significantly outperformed by any other method (including adaptive sampling procedures), even in presence of noise.

1. INTRODUCTION

Compressive sampling (CS), also called *Compressed Sensing*, has generated a tremendous amount of excitement in the signal processing community. CS involves taking non-traditional samples in the form of randomized projections, such as random binary, Gaussian, or Fourier projection vectors. Specifically, the samples of a signal vector $\boldsymbol{f} \in \mathbb{R}^n$ are inner products of the form

$$y_j = \boldsymbol{\phi}^T(j)\boldsymbol{f}, \quad j = 1, \ldots, k,$$

where $\{\boldsymbol{\phi}(j)\}$ are random vectors (*e.g.*, normalized n-vectors comprised of i.i.d. binary or Gaussian random variables). Recent theoretical results indicate that extremely accurate signal reconstructions are possible from a relatively small number of noiseless random projections [1, 2]. We extended these results to show that many signals can be very accurately recovered from random projections contaminated with noise [3], in many cases much more accurately than possible using conventional sampling methods. More recently, similar results were confirmed using alternative analysis techniques [4]. Despite these encouraging results, there seems to be a significant gap between the performance bounds for the noiseless and noisy scenarios. This yields pessimistic bounds in regimes where the SNR is high. Also, it is not known whether or not CS, which is non-adaptive, performs optimally in noisy situations.

Adaptive sampling, also known as *Active Learning*, involves sequential sampling schemes that use information gleaned from previous observations to guide the sampling process. Several empirical and theoretical studies have shown that adaptively selecting samples in order to learn a target function can outperform conventional sampling schemes, for example see [5, 6]. In particular, it was shown that adaptive sampling can recover certain classes of one-dimensional piecewise constant functions in noise with an error that decays exponentially fast in the number of samples taken [7]. This is significantly faster than conventional (uniform) sampling schemes whose errors converge at a much slower polynomial rate, with or without noise present. Similarly encouraging results have been obtained for the recovery of multidimensional piecewise constant functions [8, 9], in which case AS achieves the optimal minimax-rate among all possible sampling schemes [9].

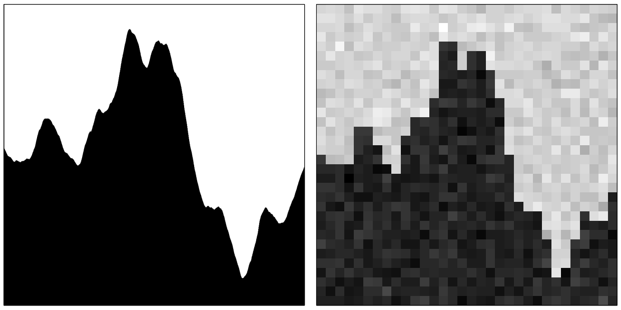
The optimality of adaptive sampling for recovering piecewise constant functions from noisy samples suggests an intriguing question. Can non-adaptive CS perform comparatively as well as adaptive sampling in such situations? This paper provides an affirmative answer to this question. This result is remarkable since it is the first theoretical evidence that shows that compressive sampling, which is non-adaptive, cannot be significantly outperformed by any other method (including every possible adaptive sampling procedure), at least in high SNR regimes. Our results hold only for certain classes of piecewise constant functions, but this is a quite rich family of signals that has many interesting potential applications, particularly in image processing. These results provide some understanding about the gap between existing error bounds for CS in the noiseless [1, 2] and noisy scenarios [3, 4]. Our results may also serve as a starting point for investigations of the optimality of CS in more general signal spaces.

2. COMPRESSIVE AND ADAPTIVE SAMPLING

We focus our attention on classes of piecewise constant functions in one or more dimensions. For illustration, consider the piecewise constant image depicted in Fig. 1 below. This image belongs to the so-called “boundary fragment” class [10], also called the “horizon” image class [11]. It consists of two constant regions (valued $+1$ and -1 for our purposes) separated by a one-dimensional curve with *functional* form $y = g(x)$; *i.e.*, the vertical coordinate of the boundary, y, is determined by a smooth function, g, of the horizontal coordinate, x.

Our primary concern is how well one can recover the original image in Fig. 1(a) from noisy samples, such as the noisy pixel samples depicted in Fig. 1(b). We assume that the boundary function g is Lipschitz smooth; *i.e.*, $|g(x_1) - g(x_2)| \le \beta|x_1 - x_2|$, with a Lipschitz constant $\beta > 0$. In this case it is known that standard wavelet denoising methods can reconstruct the image from k uniformly spaced and noisy pixel samples with a mean square error of $O(k^{-1/2})$, and that no estimation procedure based on these samples can perform significantly better ($k^{-1/2}$ is the minimax rate) [10]. However, if one allows the possibility of taking pixel samples in

This research was supported in part by the NSF under grants CCR-0113385 and CCR-0310889.



(a) (b)

Fig. 1. *Example of an* 1024×1024 *pixel image from the boundary fragment class (image (a)) [10]; and a* 32×32 *noisy pixel samples with* $\sigma = 0.15$ *(image (b)).*

an adaptive fashion, sequentially monitoring the sample values and carefully "focusing-in" on the boundary region which dominates the overall error, then it is possible to achieve a rate of $O(k^{-1})$, which is the best possible error rate among all adaptive and non-adaptive sampling schemes and estimation procedures [8, 9]. In other words, adaptive sampling can produce vastly superior image reconstructions with far fewer samples than conventional uniform sampling schemes, simply because most of the samples are *wasteful*; in conventional schemes most samples are far away from the boundary.

The main result of this paper, stated formally in the theorem below, shows that non-adaptive CS can have a performance that is similar to the one of adaptive sampling. Before stating the theorem we define the following piecewise constant function classes. First consider a space of one-dimensional n-point signals $\boldsymbol{f} = (f_1, \ldots, f_n)$,

$$\mathcal{F}_1 \quad = \quad \left\{ \boldsymbol{f} : f_i = -\mathbf{1}_{\{i \leq \theta\}} + \mathbf{1}_{\{i > \theta\}},\ \theta \in \{0, \ldots, n\} \right\},$$

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. The vectors in $\mathcal{F}_1$ correspond to step functions. For two-dimensional images, consider the following class of $n \times n$ arrays $\boldsymbol{f}$

$$\mathcal{F}_2 \quad = \quad \left\{ \boldsymbol{f} : f_{i,j} = 2\mathbf{1}_{\{j/n \leq g(i/n)\}} - 1,\ g \in \mathrm{Lip}(\beta) \right\}$$

where $\mathrm{Lip}(\beta)$ denotes the space of one-dimensional Lipschitz (boundary) functions on $[0, 1]$ (i.e., functions satisfying $|g(x) - g(y)| \leq \beta|x - y|$). Higher dimensional analogs of $\mathcal{F}_2$ are constructed in an analogous fashion from $d - 1$ dimensional Lipschitz boundary functions (*e.g.*, see [10]).

Theorem 1. *Suppose that* $f \in \mathcal{F}_d$ *for some integer* $d \geq 1$*, and assume that we take* $k \leq n^d$ *samples of the form*

$$y_j = \boldsymbol{\phi}^T(j)\boldsymbol{f} + w_j, \quad j = 1, \ldots, k,$$

where $\{\boldsymbol{\phi}(j)\}$ *are Rademacher random vectors (*n*-vectors comprised of i.i.d. random variables taking values* $\pm 1/\sqrt{n}$ *with equal probability), and* $\{w_j\}$ *are i.i.d. Gaussian random variables with zero mean and variance* σ^2*, and independent of* $\{\boldsymbol{\phi}(j)\}$*. A function estimate* $\widehat{\boldsymbol{f}}_k$ *can be derived from* $\{y_j, \boldsymbol{\phi}(j)\}$ *satisfying the following mean square error bounds.*

$$\mathbb{E}\left[n^{-d}\|\boldsymbol{f} - \widehat{\boldsymbol{f}}_k\|^2\right] \leq \begin{cases} 4n\left[\alpha(n, \sigma^2)\right]^k & , \quad d = 1 \\ C(n, \beta, \sigma^2)\, k^{-1/(d-1)} & , \quad d > 1 \end{cases},$$

where $0 < \alpha(n, \sigma^2) < 1$ *(see Section 4 for precise value), and* $0 < C(n, \beta, \sigma^2) \leq 6(\beta + 1)(1 + n\sigma^2)\log n$*, for* $n \geq 4$.

Note that in the one-dimensional setting, the error bound decays exponentially fast with the number of samples, far faster than the k^{-1} rate one usually encounters in parametric estimation (which typically considers only non-adaptive sampling). As far as the optimality of the error decay rates is concerned, first consider the one-dimensional case. We know from [7] that the exponential decay of the expected error in the number of samples k is the best one can hope for. This claim comes from results in information theory: estimation of θ, the step location, can be viewed as a communication problem where we want transmit θ through an additive white Gaussian noise channel. Due to the noise in the channel the probability of error can only decay exponentially with the number of channel transmission (equivalent to the number of samples or random projections). However, it may be possible to improve the value of α governing the exponential decay in our bound, even in the non-adaptive scenario. For the multi-dimensional setting the bound of the theorem is dramatically different than the bounds obtained using estimation-theoretic techniques, like in [3], where the mean squared error is bounded by a constant (independent of k and n) times $k^{-1/d} \log n$. The new bound is equal to a constant times $k^{-1/(d-1)} \log n$ for small values of σ^2 (i.e., $\sigma^2 \sim 1/n$). The above theorem, together with the results in [7, 8], indicates that in the high SNR regime the performance of CS is comparable with the performance of the best adaptive sampling technique. Moreover, it is known that the $k^{-1/(d-1)}$ decay rate is the minimax optimal rate [8, 9], implying that no other adaptive or non-adaptive sampling scheme and estimation procedure can significantly improve on AS or CS under these conditions. Furthermore, in the next section we describe a relatively simple Bayesian procedure for constructing $\widehat{\boldsymbol{f}}_k$ from the noisy compressive samples.

3. SIGNAL RECONSTRUCTION ALGORITHM

First we consider the reconstruction problem for the one-dimensional class $\mathcal{F}_1$. Each element $\boldsymbol{f} \in \mathcal{F}_1$ is parameterized by $\theta \in \{0, \ldots, n\}$, that is $\boldsymbol{f} \equiv \boldsymbol{f}(\theta)$. The basic reconstruction algorithm used is the maximum likelihood estimator of θ. For analysis purposes it is convenient to formulate the algorithm in a Bayesian way: Let $\boldsymbol{p}(j) \equiv \{p_0(j), \ldots, p_n(j)\}$ parameterize the posterior after j measurements, that is

$$\Pr(\theta = l | y_1, \ldots, y_j, \boldsymbol{\phi}(1), \ldots \boldsymbol{\phi}(j)) \equiv p_l(j).$$

We start with a uniform prior on θ, that is, $p_l(0) = 1/(n + 1)$ for all $l \in \{0, \ldots, n\}$. Whenever we get a new measurement we update the posterior using Bayes rule. This amounts simply to multiplication by the likelihood of the measurement (because $\{w_i\}_{i=1}^j$ are all independent) followed by a normalization, therefore

$$p_l(j + 1) = \frac{p_l(j) \exp\left(-\frac{1}{2\sigma_u^2}\left(y_{j+1} - \boldsymbol{\phi}^T(j + 1)\boldsymbol{f}(l)\right)\right)^2}{\sum_{m=0}^n p_m(j) \exp\left(-\frac{1}{2\sigma_u^2}\left(y_{j+1} - \boldsymbol{\phi}^T(j + 1)\boldsymbol{f}(m)\right)\right)^2},$$

where $\sigma_u^2 = 2\sigma^2$ for reasons stated in the next section. We consider the maximum *a posteriori* (MAP) estimator

$$\widehat{\theta}_k \equiv \arg\max_l p_l(k).$$

Note that the outcome of the estimator does not depend on σ_u^2 as long as $\sigma_u^2 > 0$. Finally our estimate of $\boldsymbol{f}$ is simply $\widehat{\boldsymbol{f}}_k \equiv \boldsymbol{f}(\widehat{\theta}_k)$.

For the multidimensional classes $\mathcal{F}_d$, $d > 1$, it suffices to note that the multidimensional signals of interest can be interpreted as a collection of one-dimensional step function signals from the class $\mathcal{F}_1$. For example, in the two-dimensional case, such as that depicted in Fig. 1, each column of the image matrix is a one-dimensional signal in $\mathcal{F}_1$. Thus, we can apply the one-dimensional CS and reconstruction process on image columns separately (although this procedure might not be entirely adequate if the total number of samples is very small, *i.e.,* $k \ll n$). Details of the method are provided in the proof below. Conversion of the multi-dimensional problem into a series of one-dimensional problems is a standard technique in the analysis of signal models in this class [10, 11]. Note that the samples are still completely non-adaptive, however this CS scheme differs slightly from other CS proposals in multiple dimensions in which the random projections are taken over the entire array [1, 2, 3], rather than column by column.

4. PROOF OF THEOREM 1

To begin, we consider the one-dimensional class $\mathcal{F}_1$. The proof of Theorem 1 employs an analysis technique similar in spirit to one used in the study of adaptive sampling in $\mathcal{F}_1$ [7]. First define

$$M_\theta(j) = \frac{1-p_\theta(j)}{p_\theta(j)}, \text{ and } N_\theta(j+1) = \frac{M_\theta(j+1)}{M_\theta(j)}.$$

Noticing that $\sum_{l=0}^n p_l(j) = 1$ we have

$$\begin{aligned} \Pr(\hat{\theta}(k) \neq \theta) &\leq \Pr\left(p_\theta(k) < \frac{1}{2}\right) = \Pr(M_\theta(k) > 1) \\ &\leq \mathbb{E}[M_\theta(k)], \end{aligned}$$

where the last inequality follows from Markov inequality. The definition of $M_\theta(j)$ is chosen to get more leverage out of Markov's inequality (akin to Chernoff bounding techniques). Now we proceed by conditioning

$$\begin{aligned} &\mathbb{E}[M_\theta(k)] = \mathbb{E}[M_\theta(k-1)N_\theta(k)] \\ &\quad = \mathbb{E}\left[M_\theta(k-1)\mathbb{E}[N_\theta(k)|\boldsymbol{p}(k-1)]\right] \\ &\quad\ \vdots \\ &\quad = M_\theta(0)\mathbb{E}\big[\mathbb{E}[N_\theta(1)|\boldsymbol{p}(0)] \times \cdots \\ &\qquad \cdots \times \mathbb{E}[N_\theta(k)|\boldsymbol{p}(k-1)]\big] \\ &\quad \leq M_\theta(0)\left\{\max_{j\in\{0,\ldots,k-1\}} \max_{\boldsymbol{p}(j)} \mathbb{E}[N_\theta(j+1)|\boldsymbol{p}(j)]\right\}^k . \end{aligned}$$

The remainder of the proof entails upper bounding $\mathbb{E}[N_\theta(j+1)|\boldsymbol{p}(j)]$. Plugging in the definitions we get

$$\begin{aligned} &\mathbb{E}[N_\theta(j+1)|\boldsymbol{p}(j)] \\ &\quad = \frac{1}{1-p_\theta(j)} \sum_{m\neq\theta} p_m(j)\mathbb{E}\left[\frac{\mathrm{e}^{-\frac{1}{2\sigma_u^2}\left(y_{j+1}-\boldsymbol{\phi}^T(j+1)\boldsymbol{f}(m)\right)^2}}{\mathrm{e}^{-\frac{1}{2\sigma_u^2}\left(y_{j+1}-\phi^T(j+1)\boldsymbol{f}(\theta)\right)^2}}\right]. \end{aligned}$$

To evaluate the above summation we consider two separate cases: (i) $m < \theta$; (ii) $m > \theta$. After some tedious but straightforward algebra we conclude that

$$\begin{aligned} &\mathbb{E}\left[\frac{\mathrm{e}^{-\frac{1}{2\sigma_u^2}\left(y_{j+1}-\boldsymbol{\phi}^T(j+1)\boldsymbol{f}(m)\right)^2}}{\mathrm{e}^{-\frac{1}{2\sigma_u^2}\left(y_{j+1}-\boldsymbol{\phi}^T(j+1)\boldsymbol{f}(\theta)\right)^2}}\right] = \\ &\qquad \mathbb{E}\left[\exp\left(-2\Big(\frac{1}{\sigma_u^2}-\frac{\sigma^2}{\sigma_u^4}\Big)\Big(\sum_{\substack{t:\ m<t\leq\theta, \text{ or} \\ \theta<t\leq m}} \phi_t(j+1)\Big)^2\right)\right]. \end{aligned}$$

The above expression is minimized when $\sigma_u^2 = 2\sigma^2$, justifying our choice for σ_u^2. Although it is not easy to compute the above expectations for general values of m and θ, it is relatively easy to conclude that those are largest when $|m-\theta| = 1$ or $|m-\theta| = 2$, therefore

$$\begin{aligned} &\mathbb{E}\left[\frac{\mathrm{e}^{-\frac{1}{2\sigma_u^2}\left(y_{j+1}-\boldsymbol{\phi}^T(j+1)\boldsymbol{f}(m)\right)^2}}{\mathrm{e}^{-\frac{1}{2\sigma_u^2}\left(y_{j+1}-\boldsymbol{\phi}^T(j+1)\boldsymbol{f}(\theta)\right)^2}}\right] \\ &\qquad \leq \max\left\{\mathrm{e}^{-\frac{1}{2n\sigma^2}}, \frac{1}{2}+\frac{1}{2}\mathrm{e}^{-\frac{2}{n\sigma^2}}\right\} \equiv \alpha(n,\sigma^2). \end{aligned}$$

Consequently $\mathbb{E}[N_\theta(j+1)|\boldsymbol{p}(j)] \leq \alpha(n,\sigma^2)$ and therefore

$$\Pr(\hat{\theta}(k) \neq \theta) \leq n\left[\alpha(n,\sigma^2)\right]^k.$$

A bound on the expected error then follows trivially, by considering a worst case scenario when $\widehat{\theta} \neq \theta$,

$$\mathbb{E}\left[n^{-1}\|\hat{f}_k - f\|^2\right] \leq 4n\left[\alpha(n,\sigma^2)\right]^k.$$

If instead of compressive samples we used carefully chosen adaptive point samples (using ideas similar to the ones in [7]) then we would get bounds with the same structure, but instead of $\alpha(n,\sigma^2)$ the exponent would be $1/2 + (1/2)\exp(-1/(2\sigma^2))$, independent of n.

For the multidimensional classes $\mathcal{F}_d$, $d > 1$, again note that the multidimensional signals of interest can be interpreted as a collection of one-dimensional step function signals from the class $\mathcal{F}_1$. Furthermore, we know from standard approximation theory that any Lipschitz function can be reasonably approximated by a piecewise constant function. These two observations along with the results for the one-dimensional case suffice to prove the general results for $d > 1$.

Let us first consider the two-dimensional case. Let h_L be the best piecewise constant fit to g on L equal-width intervals. Then $|g - h_L| \leq \beta L^{-1}$ by the Lipschitz assumption. We can estimate the levels of h_L using the one-dimensional CS method described previously, considering projections over image columns. We will consider L columns of the image, therefore using k/L samples per column. Putting all these fact together yields the bound

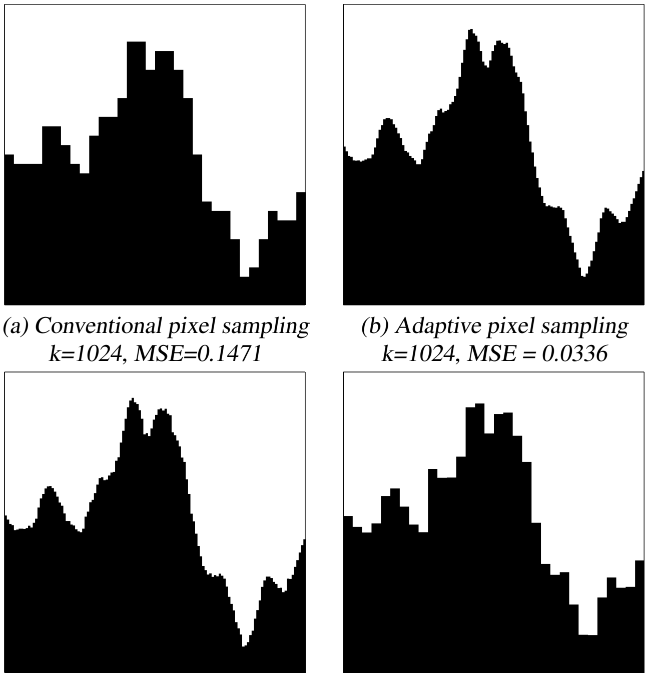
$$\mathbb{E}\left[n^{-2}\|\boldsymbol{f} - \widehat{\boldsymbol{f}}_k\|^2\right] \leq \beta\frac{1}{L} + 4n[\alpha(n,\sigma^2)]^{k/L}. \tag{1}$$

To minimize this bound we simply have to choose

$$\begin{aligned} L &= k\log(\alpha(n,\sigma^2))/\log(\beta/(-4nk\log(\alpha(n,\sigma^2)))) \\ &\sim k\log(\alpha^{-1}(n,\sigma^2))/\log(nk), \end{aligned}$$

and therefore $\mathbb{E}\left[n^{-2}\|\boldsymbol{f} - \widehat{\boldsymbol{f}}_k\|^2\right] \leq C(n,\beta,\sigma^2)k^{-1}$, where $C(n,\beta,\sigma^2)$ can be computed using the value of L given above into the bound (1). The analysis and reasoning in the higher dimensional cases is analogous, and one can easily verify that taking k samples leads to a bound on the reconstruction error of $C(n,\sigma^2,\beta)k^{-1/(d-1)}$, where

$$\begin{aligned} C(n,\sigma^2,\beta) &= \beta\frac{\log(4nk^{1/(d-1)})}{-\log\alpha(n,\sigma^2)} + 1 \\ &\leq 6(\beta+1)(1+n\sigma^2)\log n\ . \qquad \square \end{aligned}$$



(a) Conventional pixel sampling k=1024, MSE=0.1471

(b) Adaptive pixel sampling k=1024, MSE = 0.0336

(c) Compressive sampling k=1024, MSE = 0.0359

(d) Compressive sampling k=256, MSE = 0.1431

Fig. 2. *Reconstructions of image in Fig. 1(a) based on k noisy samples with $\sigma = 0.15$.*

5. EXPERIMENTS

To illustrate the theory and method developed in this paper, we consider the problem of reconstructing the 1024×1024 boundary fragment image f depicted in Fig. 1(a) from a limited number of noisy samples. We compare conventional (non-adaptive) pixel sampling, adaptive pixel sampling, and compressive sampling, with a Gaussian noise of standard deviation $\sigma = 0.15$ added to the samples in each case (equivalent to the noise level depicted in Fig. 1(b)). In the experimental results depicted in Fig. 2(a–c) we compare the methods using $k = 1024$ samples in each case. For the conventional pixel sampling case we subsample the original image on a 32×32 pixel lattice and add noise, resulting in the data depicted in Fig. 1(b). To reconstruct the image from the noisy pixel samples, we simply compute the maximum likelihood estimate in each column, using the fact that it is know that the noiseless column is one of 32 possible step functions. The resulting reconstruction is shown in Fig. 2(a). In the adaptive sampling case, 128 uniformly spaced columns are selected and 8 adaptive pixel samples are taken in each column based on the method in [7]. The resulting reconstruction is shown in Fig. 2(b). Similarly, the compressive sampling is carried out by selecting 128 columns and taking 8 random projection samples in each column. The resulting reconstruction is shown in Fig. 2(c). As expected from our theory, compressive sampling and adaptive pixel sampling perform similarly, and both significantly outperform conventional pixel sampling. In Fig. 2(d) we present the results of compressed sensing using even fewer samples, namely only $k = 256$ random projections, split among 32 uniformly spaced columns and 8 random projections per column. The result is quite impressive since we get about the same performance as conventional pixel sampling, but with four times fewer samples.

The results depicted in Fig. 2 are representative of the performance comparison of the three methods at different sampling rates k and similar noise levels. Notice that the noise level in the simulation is relatively high, much larger than one might expect based on the upper bounds given by our theoretical analysis. This shows that our bounds (for CS in particular) are somewhat loose, and that even better performance than they predict can be expected in practice. The reason for this is may be that our error bounds derive from bounds on the probability of deciding on the wrong changepoint in each column, not the expected squared error directly. In practice mistakes in the decision process often identify changepoints in the near vicinity of the true changepoint, leading to relatively small square errors.

6. CONCLUSIONS

The theory and method in this paper demonstrate that for certain classes of piecewise constant signals, compressive sampling is as effective as adaptive sampling, provided the SNR is sufficiently high. This is a significant step forward in our understanding of compressive sampling, since previous results only demonstrated the optimality of compressive sampling in noiseless conditions. The method of reconstruction employed in our work differs markedly from the usual reconstruction strategies employed in compressive sampling (based on l_1 minimization techniques). We do not know whether or not those strategies, in particular the methods that handle noisy samples proposed in [3, 4], provide the same near-optimal convergence rates as the Bayesian reconstruction proposed here. Our future work is aimed at extending the theory and methods developed in this paper to more general classes of signals.

7. REFERENCES

[1] E. Candes and T. Tao, “Near optimal signal recovery from random projections: Universal encoding strategies?,” Tech. Rep., Caltech, 2004.

[2] D. Donoho, “Compressed sensing,” Tech. Rep., Stanford, 2004.

[3] J. Haupt and R. Nowak, “Signal reconstruction from noisy random projections,” in *Proceedings of the IEEE Statistical Signal Processing Workshop*, 2005, long version submitted to *IEEE* Trans. Info. Th..

[4] E. J. Candes and T. Tao, “The dantzig selector: statistical estimation when p is much larger than n.,” Tech. Rep., Caltech, 2005.

[5] Y. Freund, H. S. Seung, E. Shamir, and N. Tishby, “Information, prediction, and query by committee,” *Proc. Advances in Neural Information Processing Systems*, 1993.

[6] K. Sung and P. Niyogi, “Active learning for function approximation,” *Proc. Advances in Neural Information Processing Systems*, vol. 7, 1995.

[7] M. V. Burnashev and K. Sh. Zigangirov, “An interval estimation problem for controlled observations,” *Problems in Information Transmission*, vol. 10, pp. 223–231, 1974.

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

[9] R. Castro, R. Willett, and R. Nowak, “Faster rates in regression via active learning,” in *Proceedings of Neural Information Processing Systems (NIPS)*, 2005.

[10] A.P. Korostelev and A.B. Tsybakov, *Minimax Theory of Image Reconstruction*, Springer Lecture Notes in Statistics, 1993.

[11] D. Donoho, “Wedgelets: Nearly minimax estimation of edges,” *Ann. Stat.*, vol. 27, pp. 859–897, 1999.