Wireless Sensing, Active Learning, and Compressive Sampling

Rui Castro, Jarvis Haupt, and Robert Nowak
Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706
Email: rcastro@cae.wisc.edu, jdhaupt@wisc.edu, nowak@ece.wisc.edu

***Abstract*— Wireless sensor networks promise a fundamentally new approach for gathering information about the physical world via a distributed network of sensors that can communicate with each other and/or with a (usually distant) fusion center through radio-frequency wireless links. Limited energy resources make power conservation essential in these envisioned sensing systems. Thus, it becomes crucial to strategically decide when, where and how to collect samples and communicate information. Active learning methods adaptively select sample locations based on previous observations in order to "learn" a target function using as few samples as possible, which could clearly be advantageous in sensor network operations. Compressive sampling refers to taking non-traditional samples in the form of randomized projections of data. Recent results show that compressive sampling can allow one to reconstruct signals from very few samples, again suggesting promising opportunities for wireless sensing. This paper compares the theoretical performance of adaptive and compressive sampling to conventional Shannon-Nyquist sampling, and it is shown that for certain classes of piecewise constant (spatial) signals, both compressive and adaptive schemes can dramatically outperform conventional sampling. Furthermore, we show that in high SNR regimes the performance of compressive sampling approaches that of adaptive sampling, achieving a near-optimal rate of convergence.**

I. Introduction

Wireless sensor networks offer a tantalizing new paradigm for measuring the world around us. These envisioned systems, however, face highly non-trivial energy and bandwidth limitations that have motivated considerable research into new and innovative schemes for sampling and communications. Perhaps the most basic approach to energy conservation is to simply limit the number of samples acquired and communicated to a bare minimum. While this may seem rather simple-minded, it turns out that sensor networks allow one considerable flexibility in when, where and how samples are taken. This paper investigates and compares two strategic sampling schemes motivated largely by problems arising in wireless sensor networks. Both schemes are promising alternatives to traditional Shannon-Nyquist sampling (uniformly-spaced point samples). A major question addressed here is whether or not adaptation is necessary to achieve optimal reconstruction error rates.

Compressive sampling (CS), also called *Compressed Sensing*, is the first alternative sampling scheme we consider. 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 unnecessarily loose bounds in regimes where the SNR is high.

Adaptive sampling (AS), also known as *Active Learning* in the machine learning literature, is the second scheme. AS 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 Shannon-Nyquist 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].

Both compressive and adaptive sampling suggest promising opportunities for wireless sensing applications. Consider a network of n nodes, each of which can sense a spatial field of interest (e.g., temperature or chemical distribution) at its location. The sensor readings may be erroneous or noisy. In addition to sensing, the nodes can wirelessly communicate

information to a distant destination or *fusion center* with the ultimate goal of obtaining an accurate estimate of the function at the fusion center. Adaptive sampling can be used in a sequential adaptive process, in which the fusion center selectively queries nodes in order to rapidly locate changepoints in the piecewise constant function being sensed. Communications between the fusion center and sensor nodes could be carried out via digital communications. Compressive sampling offers another promising alternative. A small number of non-adaptive, but randomized, projections of the sensor readings can provide sufficient information about the changepoint locations. Such projections may be communicated efficiently to the fusion center by organizing the network into a array and transmitting the projections via analog beamforming. We will not delve into the details of communications further in this paper, but additional discussion on this topic can be found in our related work [10]. The main focus here is on the reconstruction error decay as a function of the number of samples taken (and ultimately communicated).

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. The results also 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.

II. Compressive and Adaptive Sampling

We focus our attention on classes of piecewise constant functions in one dimension. Extensions to multidimensional classes (images in particular) can be analyzed in a similar manner [11]. Here we consider a space of one-dimensional n-point signals $\boldsymbol{f} = (f_1, \ldots, f_n)$,

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

where $\mathbf{1}_{\{\cdot\}}$ denotes the indicator function. The vectors in $\mathcal{F}$ correspond to step functions. To cast this in the context of wireless sensor networks, imagine that each element of a vector $\boldsymbol{f}$ corresponds to a sensor and the value of θ corresponds to a changepoint in the spatial field being sensed. Our primary concern is how well one can recover signals in this class from noisy samples. Conventional Shannon-Nyquist sampling involves taking k uniformly-spaced point samples. It is easy to see that the mean square reconstruction error will be at least as large as the bias due to undersampling, which in this case is on the order of k^{-1} (since we can only determine the location of the changepoint to within that resolution). However, if one allows the possibility of adapting the sampling locations based on previous observations, that is sequentially monitoring the sample values and carefully "focusing" samples near the perceived changepoint location, then it is possible to achieve an exponential rate of convergence [7]. A constructive approach for this AS process was proposed in [7], [12] can be roughly described as "probabilistic bisection," wherein each new sample is selected near the median of the posterior distribution of the changepoint location derived from previous samples. Using this procedure, the mean square reconstruction error after taking k samples at adaptively chosen locations is bounded as

$$\mathbb{E}\left[n^{-1}\|\boldsymbol{f} - \widehat{\boldsymbol{f}}_k\|^2\right] \;\leq\; 4n\,[\beta(\sigma^2)]^k \quad ,$$

where $0 < \beta(\sigma^2) < 1$ is given by

$$\beta(\sigma^2) \;=\; 1/2 + 1/2\ \mathrm{e}^{-1/(2\sigma^2)}$$

and σ^2 is the variance of zero-mean, additive white Gaussian noises contaminating each sample. Moreover, we know from [7], [12] 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 noise in the channel the accuracy in the transmission of θ can decay at most exponentially with the number of symbols transmitted (analogous to the number of samples in the AS procedure).

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.

Theorem 1. *Suppose that $\boldsymbol{f} \in \mathcal{F}$ and assume that we take $k \leq n$ 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 bound.

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

where $0 < \alpha(n,\sigma^2) < 1$ is given by

$$\alpha(n,\sigma^2) \;=\; \max\left\{\mathrm{e}^{-\frac{1}{2n\sigma^2}}, \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-\frac{2}{n\sigma^2}}\right\}$$

Note that form of the rate parameter $\alpha(n,\sigma^2)$ is quite similar to that in the AS case, $\beta(\sigma^2)$, with one important distinction: the rate parameter for CS depends on the signal length n. The result of this dependence is that in the high SNR regimes, CS approaches the performance of AS, but lags

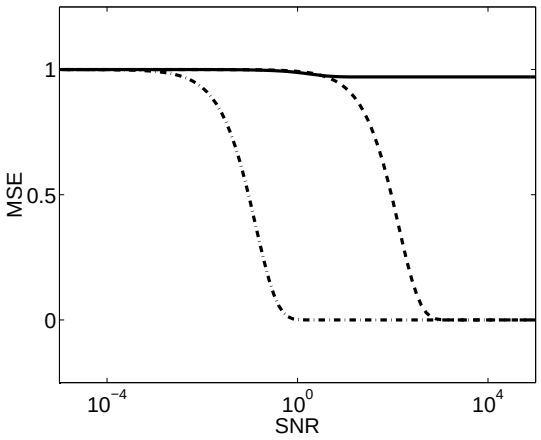

Fig. 1. *Behavior of (normalized) error bounds as a function of SNR ($1/\sigma^2$). Signal is an $n = 1000$ length vector and $k = 30$ samples are taken using Shannon-Nyquist sampling (solid), Compressive Sampling (dashed), and Adaptive Sampling (dot-dashed).*

significantly behind in lower SNR situations. This effect can be intuitively understood by noting that, due to the non-adaptive nature of the random projections, CS spreads energy evenly across the entire signal space, whereas AS focuses the energy of its sampling process in the vicinity of the signal subspace. Fig. 1 depicts the behavior of the error bounds as a function of SNR.

III. Signal Reconstruction Algorithm

Each element $\boldsymbol{f} \in \mathcal{F}$ is parameterized by $\theta \in \{0, \ldots, n\}$, that is $\boldsymbol{f} \equiv \boldsymbol{f}(\theta)$. Let $y_1, \ldots, y_j$ denote samples acquired either by the AS method of [7] or the CS method described above. 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 $\widehat{\boldsymbol{f}}_k \equiv \boldsymbol{f}(\widehat{\theta}_k)$. One final point, the AS method of [7] selects the $j+1$ sample near the median of $p_l(j)$.

IV. Proof of Theorem 1

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

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

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

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

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

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

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

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

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

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

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

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

V. Conclusions

The theory and methods discussed in this paper show that adaptive and compressive sampling can offer significant advantages over traditional Shannon-Nyquist sampling in certain situations (e.g., spatial changepoint detection). Furthermore, our new result demonstrates that compressive sampling can be nearly as effective as adaptive sampling in such situations, 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. Also, in combination with the beamforming approach to communications proposed in [10], CS may be a very attractive form of sampling for wireless sensor net applications. We also note that 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.

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] W. Bajwa, A. Sayeed, and R. Nowak, "Matched source-channel communication for field estimation in wireless sensor networks," in *Proceedings of the Fourth International Symposium on Information Processing in Sensor Networks (IPSN'05)*, 2005.

[11] R. Castro, J. Haupt, and R. Nowak, "Compressed sensing vs. active learning," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006.

[12] M. Horstein, "Sequential transmission using noiseless feedback," *IEEE Trans. Info. Th.*, vol. 9, pp. 136–143, 1963.