COMPRESSIVE SAMPLING FOR SIGNAL CLASSIFICATION

Jarvis Haupt, Rui Castro[*]*, and Robert Nowak*

University of Wisconsin - Madison
Dept. of Electrical and Computer Engineering
1415 Engineering Drive, Madison, WI 53706

Gerald Fudge and Alex Yeh

L-3 Communications, Integrated Systems
10001 Jack Finney Boulevard
Greenville, TX 75402

ABSTRACT

Compressive Sampling (CS), also called Compressed Sensing, entails making observations of an unknown signal by projecting it onto random vectors. Recent theoretical results show that if the signal is sparse (or nearly sparse) in some basis, then with high probability such observations essentially encode the salient information in the signal. Further, the signal can be reconstructed from these "random projections," even when the number of observations is far less than the ambient signal dimension. The provable success of CS for signal reconstruction motivates the study of its potential in other applications. This paper investigates the utility of CS projection observations for signal classification (more specifically, m-ary hypothesis testing). Theoretical error bounds are derived and verified with several simulations.

1. INTRODUCTION AND MOTIVATION

Compressive Sampling (CS), also called Compressed Sensing, involves sampling signals in a non-traditional way - each observation is obtained by projecting the signal onto a randomly chosen vector. Formally, we describe these "random projections" as inner products between the unknown vector being observed and a set of random vectors (for example, vectors whose entries are independent and identically distributed (i.i.d.) Gaussian or Binary random variables, or random Fourier basis vectors). Recent theoretical results show that if the signal is sparse (or nearly sparse) in some basis, then with high probability, such observations essentially encode the salient information in the signal. Further, the unknown vector can be estimated from these random projections observations to within a controllable mean-squared error, even when the observations are corrupted with additive noise [1, 2, 3, 4, 5]. These remarkable results show that the number of samples required for reconstruction can be far fewer than the ambient dimension of the observations!

One key advantage offered by the CS approach is that samples can be collected without assuming any prior information about the signal being observed (aside from sparsity in some basis). The fact that consistent estimation is possible using (a relatively small number of) these agnostic observations quantifies the *universality* of CS, and hence CS has become an attractive technology for several applications of ever-increasing importance. Specifically, CS has been proposed as a viable candidate for wideband communications monitoring systems, where the goal is to intercept communication signals over frequency range so large that conventional Nyquist sampling is technologically impossible. CS is also being examined in imaging systems, where the potential benefits include improved storage efficiency, reduced form-factor, and reduced cost [6].

While the abilities of CS for signal reconstruction have been examined, the full capability of random projection sampling is still not fully known. For example, in some practical settings the goal might be to reliably determine, from a set of noisy observations, which signal from a set of known candidates is present - a far less aggressive goal than full signal reconstruction. If the class of candidate signals is known prior to the observation process than specialized projection observations can be made, yielding optimal performance in the presence of Gaussian noise. This procedure is called matched filtering. This specialized approach offers no guarantee of universality over a broad class of signals, and in many scenarios the observations are made before knowledge of the signal class is available, precluding the construction of a matched filter. A natural question emerges, can the universality of CS be leveraged in this hypothesis testing (candidate identification) problem? The answer, of course, is yes, and this paper examines the performance such a system.

The CS classification problem we investigate here has received only a minimal treatment to date. Preliminary work by some of the authors of this paper considered a problem where the goal was to detect the presence or absence of a known signal. Bounds on the probability of error were obtained (for detecting very simple signals) by analyzing the likelihood ratio test that balances the probability of false alarm with the probability of missed detection [7]. Related to this work is an experimental evaluation of CS for signal detection based on partial signal reconstruction [8]. Some of the authors of this work also considered signal classification via CS for special classes of signals [9]; the work presented here essentially gen-

[*]Dual Affiliation: Rice University, Houston, TX

eralizes that contribution to more general classes of signals.

To clarify the exposition in following sections, we make a brief summary of notation here. We will use bold-face capital letters to denote matrices (such as $\boldsymbol{A}$). Vectors will be written using bold-face lower-case letters ($\boldsymbol{f}$) or superscripted upper-case letters, such as $\boldsymbol{A}^j$, which denotes the j^{th} row vector of the matrix $\boldsymbol{A}$. The inner product of two vectors $\boldsymbol{f}, \boldsymbol{g} \in \mathbb{R}^n$ is defined as $\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \sum_{i=1}^{n} \boldsymbol{f}(i)\boldsymbol{g}(i)$. We use the notation $\|\boldsymbol{f}\|^2 = \langle \boldsymbol{f}, \boldsymbol{f} \rangle$ to denote the standard Euclidean distance.

The remainder of the paper is organized as follows. Section 2 provides an overview of known CS reconstruction results. The main contribution of this paper is in Section 3, in which the problem is described and the new results are presented. The theory is verified in Section 4, where simulations of the main result are presented. Conclusions are given in Section 5, and a proof of the main result is given in Section 6.

2. COMPRESSIVE SAMPLING BACKGROUND

Suppose that an observer makes measurements of an unknown vector $\boldsymbol{f} \in \mathbb{R}^n$, where each measurement is the inner product between the signal vector $\boldsymbol{f}$ and a sampling vector $\boldsymbol{A}^j$ chosen by and known to the observer. The observations may be corrupted by additive noise. Formally, the observer measures

$$y(j) = \left\langle \boldsymbol{A}^j, \boldsymbol{f} \right\rangle + w(j), \;\; j = 1, \ldots, k, \tag{1}$$

where $\boldsymbol{A}^j \in \mathbb{R}^n$ are the sampling vectors, and $\{w(j)\}$ is a collection of independent and identically distributed (i.i.d.) $\mathcal{N}(0, \sigma^2)$ random variables, modeling observation noise. For convenience and ease of notation, we will sometimes use the matrix representation of the observation model, $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{f} + \boldsymbol{w}$.

The appeal of CS is that in many situations, the signal $\boldsymbol{f}$ can be accurately recovered from k CS observations, even when $k < n$. CS reconstruction algorithms essentially attempt to identify a solution which matches the observations, but also has a low-complexity (sparse) representation in some basis. Many formulations are possible, but all entail a search for a candidate reconstruction vector $\tilde{\boldsymbol{f}}$, simultaneously satisfying a goodness of fit criteria (such as minimizing $\|\boldsymbol{y} - \boldsymbol{A}\tilde{\boldsymbol{f}}\|^2$) and a complexity criteria (promoting sparsity of $\tilde{\boldsymbol{f}}$, quantified with the l_0 quasinorm or its closest convex surrogate, the l_1 norm). If the signal $\boldsymbol{f}$ is m-sparse in some representation (*i.e.*, has m or fewer non-zero coefficients), then an estimate $\widehat{\boldsymbol{f}}_k$ can be derived from $(\boldsymbol{y}, \boldsymbol{A})$ that satisfies

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

where C_S is a constant that depends on noise power σ^2 and the signal energy $\|\boldsymbol{f}\|^2$, and the expectation is over the distribution of the noise and the projection vector entries. Similar techniques can be applied to derive error bounds for nearly sparse signals [4, 5], though we omit discussion of those here.

3. MAIN RESULT

In this paper, we investigate an additional utility of the observations $(\boldsymbol{y}, \boldsymbol{A})$, defined in equation (1), in another sparse setting. Recall that the reconstruction results presented in Section 2 are possible because the signals of interest were assumed to be sparse (or nearly sparse) in some representation. In contrast, here we will be concerned with discrimination among members of a known set (a set of vectors in $\mathbb{R}^n$). The results we will present here implicitly leverage the fact that the known set is a sparse subset of the set of all possible n-dimensional vectors.

Formally, the problem we consider can be described as follows. Let $\mathcal{F} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_m\}$ denote a known set of m unique candidate signals, and assume that each element $\boldsymbol{f}_i$ in $\mathcal{F}$ satisfies $\|\boldsymbol{f}_i\|^2 = 1$. We denote by $\boldsymbol{f}_T \in \mathcal{F}$ the "true" signal being presented to the observer, who makes k random projection observations of the signal. The goal is to use the observations $(\boldsymbol{y}, \boldsymbol{A})$ to select a candidate function $\widehat{\boldsymbol{f}}_k \in \mathcal{F}$ such that $\widehat{\boldsymbol{f}}_k = \boldsymbol{f}_T$ most of the time. Success will be quantified in terms of the error probability, $\Pr\left(\widehat{\boldsymbol{f}}_k \neq \boldsymbol{f}_T\right)$.

One point is worthy of mention here - in order to exploit the "universality" of CS observations, we assume that the set $\mathcal{F}$ is not known at the time that observations are collected (if it were known, then specialized observations would be more powerful). The benefit of this approach is that the same set of observations $(\boldsymbol{y}, \boldsymbol{A})$ could be used for reconstruction (described in Section 2) *and* in the classification problem described here (along with other applications not described here, such as signal detection [7]). We make no specific claims about how the "auxiliary information" about the set $\mathcal{F}$ is obtained, though various intelligence channels could all be feasible sources.

The main result of this paper is stated here as a theorem. A sketch of the proof is given in Section 6.

Theorem 1 *Suppose that $\boldsymbol{f}_T \in \mathcal{F}$ where the class $\mathcal{F}$ is as defined above, and we make k observations of the form*

$$y(j) = \left\langle \boldsymbol{A}^j, \alpha \boldsymbol{f}_T \right\rangle + w(j), j = 1, \ldots, k,$$

where $\alpha \in \mathbb{R}$, the entries of $\boldsymbol{A}^j$ are i.i.d. $\mathcal{N}(0, 1/n)$ random variables, and w_j are i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$ and independent of the projection vector entries. The empirical risk minimizer

$$\widehat{\boldsymbol{f}}_k = \arg\min_{\boldsymbol{f} \in \mathcal{F}} \|\boldsymbol{y} - \alpha \boldsymbol{A}\boldsymbol{f}\|^2$$

satisfies

$$\mathcal{P}\left(\widehat{\boldsymbol{f}}_k \neq \boldsymbol{f}_T\right) \leq (m-1)\left(1 + \frac{\alpha^2 d_{min}}{4n\sigma^2}\right)^{-k/2}$$

where

$$d_{min} = \min_{\boldsymbol{f}_i, \boldsymbol{f}_j \in \mathcal{F}, i \neq j} \|\boldsymbol{f}_i - \boldsymbol{f}_j\|^2$$

is the minimum Euclidean distance between elements in the set $\mathcal{F}$.

In certain practical scenarios the candidate vectors in $\mathcal{F}$ may have norm different than one. The factor α can be used to account for those cases. The results stated in Theorem 1 are derived assuming that α is known exactly, though in Section 4 we also simulate this method in a situation where α is estimated from the data.

It is important to note here that the error bound decays *exponentially* in the number of observations, in stark contrast to the polynomial rates that were achievable in the reconstruction problem. This effect illustrates the benefit of the more restricted setting. This is in essence the same relationship between the performance of classical estimation problems (polynomial error rates) and hypothesis testing problems (exponential error rates).

4. SIMULATIONS

For the first simulation, we construct the class $\mathcal{F}$ by sampling a collection of linear chirp signals. The starting frequency for each chirp is 0 Hz, and the ending instantaneous frequencies range from 1 MHz to 100 MHz in 1 MHz increments ($m = 100$ candidates). The sampling rate is 200 MHz and the observation window is 2 μsec in duration, so the vector length is $n = 400$. In other words, the elements of the class $\mathcal{F}$ are vectors $\boldsymbol{f}_l$, $l = 1, \ldots, 100$, whose entries are given by

$$\boldsymbol{f}_l(i) = \gamma_l \cos\left(2\pi \frac{l \times 10^6}{2}\left[\frac{i}{200 \times 10^6}\right]^2 \left[\frac{200 \times 10^6}{400}\right]\right),$$

for $i = 1, \ldots, 400$. The constant γ_l is a scaling factor that ensures that $\|\boldsymbol{f}_l\|^2 = 1$. We consider collecting k projection observations, and we evaluate the probability of misclassification for each value of k by averaging the empirical errors of 100000 trials of the proposed algorithm, where for each trial, independent realizations of noise and projection vectors were generated, and the "true" candidate was chosen uniformly at random from the set $\mathcal{F}$. Figure 1 shows the results of this simulation, where the SNR is $\alpha^2/n\sigma^2 = 2$ (or 3dB).

For the second simulation, we examine the detector performance on a class of unit-norm sinusoids. As in the previous example, there are $m = 100$ candidates, but this time each candidate is a cosine wave. The frequencies range from 1 MHz to 100 MHz in 1 MHz increments. The observation window is 2 μsec, the sampling frequency is 200 MHz, and the SNR is $\alpha^2/n\sigma^2 = 2$ (or 3dB). In this case, the elements of the class $\mathcal{F}$ are vectors $\boldsymbol{f}_l$, $l = 1, \ldots, 100$, with entries

$$\boldsymbol{f}_l(i) = \gamma_l \cos\left(2\pi 10^6 l \left[\frac{i}{2 \times 10^8}\right]\right),$$

for $i = 1, \ldots, 400$, where again γ_l is again chosen so that $\|\boldsymbol{f}_l\|^2 = 1$.The misclassification probabilities are again computed for each value of k by averaging over 100000 trials, and

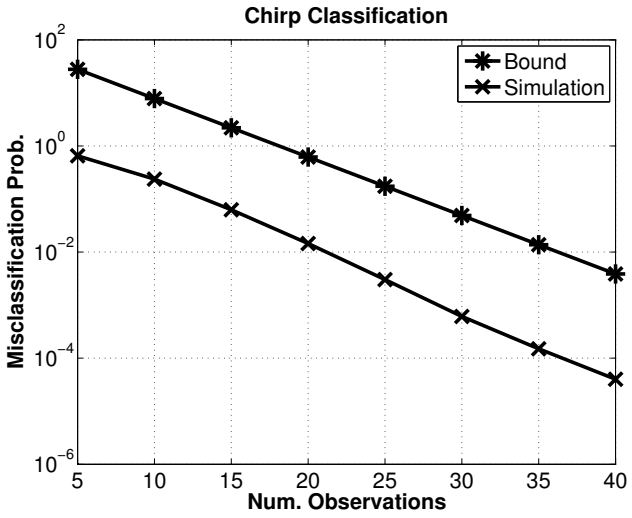

Fig. 1. Empirical and Theoretical Misclassification Probability vs. Number of Observations for Unit-Norm Chirp Signals. For this set of templates, $d_{min} = 1.325$ and the SNR is $\alpha^2/n\sigma^2 = 2$ (3dB).

the results are shown in Figure 2. Notice that the error performance in this case is similar to the performance depicted in Figure 1, although the bound is tighter in the sinusoidal case. This is because in the sinusoidal case, the worst-case behavior (quantified by d_{min}) is representative of the typical behavior, while for the chirp class the worst case and the typical cases are significantly different. As a result, the bound for the chirp case is more conservative.

The next simulation reexamines the setting of the last paragraph (depicted in Figure 2), but in this case we assume that α is not known exactly. Instead, we leverage the fact that, since the noise and projection vector entries are independent,

$$\mathbb{E}\left[\|\boldsymbol{y}\|^2\right] = \alpha^2 \frac{k}{n} + \sigma^2.$$

It is realistic to assume that the noise background is known (or could be accurately estimated a priori), so with (nearly) exact knowledge of σ^2 we can estimate α by

$$\widehat{\alpha} = \sqrt{\frac{n}{k}\left(\|\boldsymbol{y}\|^2 - \sigma^2\right)}$$

and use this estimate in the empirical risk minimization. The results presented in Figure 3 suggest that this algorithm can fairly robust to the unknown parameter α, at least for this range of SNR.

5. CONCLUSIONS

We have shown that CS observations can be effectively used in signal classification problems, where the goal is to identify a signal from a class of candidates. The misclassification

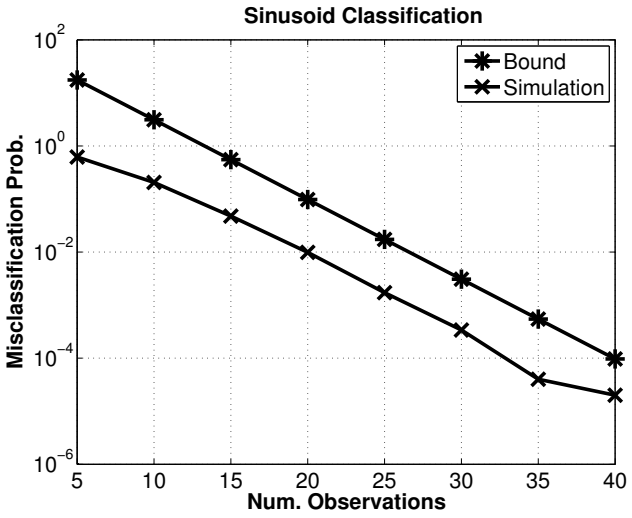

Fig. 2. Empirical and Theoretical Misclassification Probability vs. Number of Observations for Unit-Norm Cosine Signals. For this set of templates, $d_{min} = 2$ and the SNR $S = \alpha^2/n\sigma^2 = 2$ (3dB).

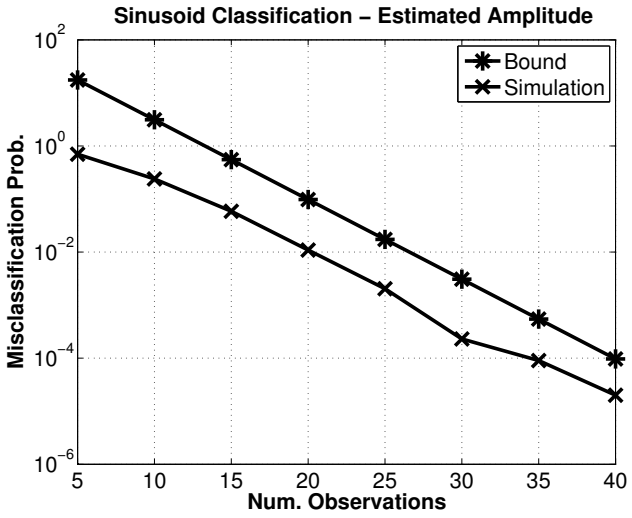

Fig. 3. Empirical and Theoretical Misclassification Probability vs. Number of Observations for Unit-Norm Cosine Signals, using an estimated signal amplitude. As in the second example, $d_{min} = 2$ and the SNR $S = \alpha^2/n\sigma^2 = 2$ (3dB).

probability was shown to decay exponentially in the number of observations, in contrast to the rates achievable when CS observations are used for full signal reconstruction. The theoretical bound presented was validated through several simulations.

6. PROOF OF THEOREM 1

The method of proof is similar to that presented in [9]. Although the proposed methodology is the maximum likelihood estimator, it is convenient to formulate the estimation strategy in a Bayesian-like setting. For each candidate $\boldsymbol{f}_l$, $l \in \{1, \ldots, m\}$, define the posterior probability given j observations to be $p_j(l) = \Pr(\boldsymbol{f}_T = \boldsymbol{f}_l | y(1), \ldots, y(j), \boldsymbol{A}^1, \ldots \boldsymbol{A}^j)$, where we let $p_0(l) = 1/m$. In other words, in the absence of any observations we assume a uniform prior over $\mathcal{F}$. Notice that under this prior assumption the posterior probabilities and likelihood values coincide. Whenever we get an observation we can update the posterior probabilities by applying Bayes' rule:

$$p_{j+1}(l) = \frac{p_j(l) \exp\left(\frac{-(y(j+1) - \alpha \langle \boldsymbol{A}^{j+1}, \boldsymbol{f}_l \rangle)^2}{2\sigma_u^2}\right)}{\sum_{i=1}^m p_j(i) \exp\left(\frac{-(y(j+1) - \alpha \langle \boldsymbol{A}^{j+1}, \boldsymbol{f}_i \rangle)^2}{2\sigma_u^2}\right)}$$

for $j \in \{0, \ldots, k-1\}$. Notice that in our update rule we implicitly assume that the variance of $w(j)$ is σ_u^2. In principle $\sigma_u^2 = \sigma^2$, but the final outcome of the estimator does not change as long as $\sigma_u^2 > 0$. Having this extra degree of freedom is going to allow tightening of the bounds. We choose our estimator $\widehat{\boldsymbol{f}}_k$ to be the candidate $\boldsymbol{f} \in \mathcal{F}$ for which the value of $p_k(l)$ is maximum, which is equivalent to

$$\widehat{\boldsymbol{f}}_k = \arg \min_{\boldsymbol{f} \in \mathcal{F}} \|\boldsymbol{y} - \alpha \boldsymbol{A} \boldsymbol{f}\|^2.$$

The proof proceeds by recasting the misclassification probability as a tractable expectation. This approach first appeared in [10], and can be viewed as Chernoff-like bounding technique. Define $M_j = (1 - p_j(T))/p_j(T)$ and

$$N_{j+1} = \frac{M_{j+1}}{M_j} = \frac{p_j(T)(1 - p_{j+1}(T))}{p_{j+1}(T)(1 - p_j(T))},$$

and let $\boldsymbol{p}_j = \{p_j(1), \ldots, p_j(m)\}$ denote the vector whose entries are the posterior values for each candidate given j observations. Note that since $\sum_{l=1}^m p_j(m) = 1$ we have

$$\begin{aligned} P\left(\widehat{\boldsymbol{f}}_k \neq \boldsymbol{f}_T\right) &\leq P\left(p_k(T) < \frac{1}{2}\right) \\ &= P(M_k > 1) \leq \mathbb{E}\left[M_k\right], \end{aligned}$$

where the last step follows from Markov's inequality. Now we condition to obtain

$$\begin{aligned} P\left(\widehat{\boldsymbol{f}}_k \neq \boldsymbol{f}_T\right) &\leq \mathbb{E}\left[M_k\right] \\ &= \mathbb{E}\left[M_{k-1} N_k\right] \\ &= \mathbb{E}\left[M_{k-1} \mathbb{E}\left[N_k | \boldsymbol{p}_{k-1}\right]\right] \\ &\vdots \\ &= M_0 \mathbb{E}\left[\prod_{j=0}^{k-1} \mathbb{E}\left[N_{j+1} | \boldsymbol{p}_j\right]\right] \\ &\leq M_0 \left\{\max_{j \in \{0,k-1\}} \max_{\boldsymbol{p}(j)} \mathbb{E}\left[N_{j+1} | \boldsymbol{p}_j\right]\right\}^k. \end{aligned}$$

The remainder of the proof consists of deriving an upper bound for the quantity

$$\mathbb{E}\left[N_{j+1}|\boldsymbol{p}_j\right]$$

that holds for $j \in \{0,\ldots,k-1\}$. Straightforward algebra gives

$$\begin{aligned}\mathbb{E}\left[N_{j+1}|\boldsymbol{p}_j\right] &= \sum_{i=1,i\neq T}^{m} \frac{p_j(i)\mathbb{E}\left[\exp\left(\frac{y(j+1)\alpha D_i}{\sigma_u^2} - \frac{\alpha^2 D_i S_i}{2\sigma_u^2}\right)\right]}{1-p_j(T)} \quad (2)\end{aligned}$$

where $D_i = \langle \boldsymbol{A}^{j+1}, \boldsymbol{f}_i - \boldsymbol{f}_T\rangle$ and $S_i = \langle \boldsymbol{A}^{j+1}, \boldsymbol{f}_i + \boldsymbol{f}_T\rangle$, and the letters "D" and "S" were chosen to denote difference and sum, respectively. Now recall that given knowledge of the projection vector $\boldsymbol{A}^{j+1}$, $y(j+1)$ is distributed as $\mathcal{N}(\alpha\langle \boldsymbol{A}^{j+1}, \boldsymbol{f}_T\rangle, \sigma^2)$. Rewriting this in terms of S_i and D_i, given $\boldsymbol{A}^{j+1}$, $y(j+1) \sim \mathcal{N}(\alpha\left(\frac{S_i - D_i}{2}\right), \sigma^2)$. Therefore

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\frac{y(j+1)\alpha D_i}{\sigma_u^2} - \frac{\alpha^2 D_i S_i}{2\sigma_u^2}\right)\right] \\ &= \mathbb{E}\left[\exp\left(\frac{-\alpha^2 D_i S_i}{2\sigma_u^2}\right)\times \right. \\ &\qquad \left. \mathbb{E}\left[\exp\left(\frac{y(j+1)\alpha D_i}{\sigma_u^2}\right)|\boldsymbol{A}^{j+1}\right]\right] \\ &= \mathbb{E}\left[\exp\left(\frac{-\alpha^2 D_i S_i}{2\sigma_u^2}\right)\times \right. \\ &\qquad \left. \exp\left(\frac{\alpha D_i}{\sigma_u^2}\frac{\alpha\left(S_i - D_i\right)}{2} + \frac{\sigma^2\alpha^2 D_i^2}{2\sigma_u^4}\right)\right] \\ &= \mathbb{E}\left[\exp\left(\left[\frac{\sigma^2\alpha^2}{2\sigma_u^4} - \frac{\alpha^2}{2\sigma_u^2}\right] D_i^2\right)\right],\end{aligned}$$

where the second-to-last step follows from the computation of the moment generating function of a Gaussian random variable. Note that the final expression no longer depends on S_i.

Since the linear combination of independent Gaussian random variable is still a Gaussian random variable, $D_i \sim N\left(0, \|\boldsymbol{f}_i - \boldsymbol{f}_T\|^2/n\right)$, and so it is straightforward to show that

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\left[\frac{\sigma^2\alpha^2}{2\sigma_u^4} - \frac{\alpha^2}{2\sigma_u^2}\right] D_i^2\right)\right] \\ &= \left(1 - 2\left[\frac{\sigma^2\alpha^2}{2\sigma_u^4} - \frac{\alpha^2}{2\sigma_u^2}\right]\frac{\|\boldsymbol{f}_i - \boldsymbol{f}_T\|^2}{n}\right)^{-1/2}. \quad (3)\end{aligned}$$

Now define

$$d_{min} = \min_{\boldsymbol{f}_i, \boldsymbol{f}_j \in \mathcal{F}, i\neq j} \|\boldsymbol{f}_i - \boldsymbol{f}_j\|^2,$$

and take $\sigma_u^2 = 2\sigma^2$ so that (3) is minimized. We can rewrite the expectation in (2) as

$$\begin{aligned}&\mathbb{E}\left[\exp\left(\frac{y(j+1)\alpha D_i}{\sigma_u^2} - \frac{\alpha^2 D_i S_i}{2\sigma_u^2}\right)\right] \\ &\qquad \leq \left(1 + \frac{\alpha^2 d_{min}}{4n\sigma^2}\right)^{-1/2},\end{aligned}$$

and thus

$$\mathbb{E}\left[N_{j+1}|\boldsymbol{p}_j\right] \leq \left(1 + \frac{\alpha^2 d_{min}}{4n\sigma^2}\right)^{-1/2}.$$

Notice that this quantity no longer depends on the posterior probabilities $\boldsymbol{p}$ or the index j. The final step is to incorporate knowledge of the prior, $M_0 = (m-1)$, from which the stated result follows.

7. REFERENCES

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

[2] E. Candès and T. Tao, "Near optimal signal recovery from random projections: Universal encoding strategies?," IEEE Trans. on Information Theory, 2004, to appear.

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

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

[5] E. Candès and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," Annals of Statistics, 2005, to appear.

[6] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, "An architecture for compressive imaging," in *Proc. International Conf. on Image Processing (ICIP)*, 2006.

[7] R. Castro, J. Haupt, and R. Nowak, "Active learning vs. compressive sampling," in *Proc. SPIE Intelligent Integrated Microsystems*, Orlando, FL, June 2006, Intl. Society for Optical Engineering (SPIE), Defense and Security Symposium.

[8] M. Duarte, M. Davenport, M. Wakin, and R. Baraniuk, "Sparse signal detection from incoherent projections," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2006.

[9] R. Castro, J. Haupt, and R. Nowak, "Compressive sensing vs. active learning," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2006.

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