Efficient Sparsity Pattern Recovery

Galen Reeves Michael Gastpar*
University of California, Berkeley
Department of Electrical Engineering and Computer Sciences
Berkeley, CA
greeves@eecs.berkeley.edu gastpar@eecs.berkeley.edu

Abstract

The theory of compressed sensing shows that sparsity pattern (or support) of a sparse signal can be recovered from a small number of appropriate linear projections (samples). Unfortunately, as soon as noise is added, the number of required samples exceeds the full signal dimension, rendering compressed sensing ineffective. In recent work, we have shown that this can be fixed if a small distortion is allowed in the signal recovery. The present paper extends our results to a simplified estimator.

1 Introduction

Estimation of sparse vectors from a limited number of noisy measurements is a problem that arises in many areas such as compressive sensing, subset selection in linear regression, graphical model selection, sparse approximation, and signal denoising. In these problems, an unknown vector $\mathbf{x} = [x_1, \cdots x_n]$ is known a priori to be sparse in the sense that it has a relatively small number of non-zero elements. However, the locations and values of the non-zero elements are unknown and must be inferred from a set of noisy linear projections (or samples) of the form

$$y_i = \langle \phi_i, \mathbf{x} \rangle + w_i, \quad \text{for} \quad i = 1, \cdots, m \tag{1}$$

where $\{\phi_i\}_{i=1}^m$ is a set of sampling vectors and w_i is an error term. Of particular interest is the setting where the number of samples m is much less than the ambient signal dimension n and hence the error-free measurements constitute an under-constrained set of linear equations. A large body of work has analyzed this problem under a variety of assumptions and recovery tasks, and results typically consist of conditions on the sparsity of $\mathbf{x}$, the sampling vectors $\{\phi_i\}$, and the signal-to-noise ratio (SNR) guaranteeing the success or failure of various algorithms.

The focus in this paper is on the task of sparsity pattern recovery (also known simply as sparsity or support recovery) which is to determine which elements of $\mathbf{x}$ are non-zero. In the noiseless setting, $m = k+1$ samples are sufficient for recovery using an NP hard combinatorial search, and a fundamental result from the field of compressed sensing by Donoho [1] and Candès, Romberg, and Tao [2,3] is that $m = O(k \log(n/k))$ measurements are sufficient to recover the entire signal $\mathbf{x}$, and hence it support, using linear programming. Here, the extra factor $\log(n/k)$ represents the additional sampling "cost" associated using efficient algorithms.

Unfortunately, in the presence of any measurement noise, $m = \Theta(k \log(n-k))$ samples are needed [4–6]. Here, the factor $\log(n-k)$ represents the fundamental sampling "cost" of measurement error. When $k = \Theta(n)$ this cost is significant because

[*] Also with Delft University of Technology, Faculty of EEMCS, Dept. of Mediamatics, Information and Communication Theory Group, Delft, The Netherlands.

	Perfect Recovery	Approximate Recovery
Necessary (Any Algorithm)	$m > k + C\dfrac{k \log k}{f(\mathsf{SNR})}$ Wei et al [6], Fletcher et al [5]	$m > C_1 k + C_2 \dfrac{k}{f(\mathsf{SNR})}$ Reeves & Gastpar [10]
Achievable for GML (Combinatorial Search)	$m > k + C\dfrac{k \log k}{f(\mathsf{SNR})}$ Wainwright [4]	$m > k + C\dfrac{k}{f(\mathsf{SNR})}$ Reeves & Gastpar [10]
Achievable for MC (Thresholding)	$m > k + Ck \log k$ Fletcher et. al [5]	$m > Ck$ This paper

Table 1: Summary of scaling results for $k = \Theta(n)$ and finite SNR. The dependence on the sparsity rate k/n and the bound on the fraction of errors is not shown. Constants are denoted by C and any increasing nonasymptotic function is denoted by f.

it means that recovery is not possible with $m < n$ samples [7]. Alternatively, if only $m = \Theta(k)$ samples are available and the noise is due to quantization error, this means that an unbounded rate (in bits) per sample is required.

These negative results have prompted our investigation of approximate support recovery for which positive results can be given [7–11]. When $k = \Theta(n)$, we showed in [10] that $m = \Theta(k)$ samples are necessary and sufficient to upper bound the fraction of errors in the estimated support. Hence, for the "cost" of a relatively small number of errors, it is possible to attain the same scalings as in the noiseless cases. This means that if some small fraction of errors (greater than zero) is allowed, and if the SNR is sufficiently large, then recovery is possible using a fixed *sampling rate* $\rho = m/n$ that is much less then one. Furthermore, if noise is due to quantization error, this means that it is possible to have both $m \ll n$ and a fixed rate per sample.

For perfect recovery from noisy samples, Wainwright [4] showed achievability of the scaling $m = \Theta(k \log(n-k))$ for a *generalized maximum likelihood* (GML) estimator that requires solving a combinatorial optimization problem. More recently Fletcher et al [5] showed that the same scaling result can also be achieved by a computationally simple thresholding algorithm termed *maximum correlation* (MC). The main difference in performance between these two algorithms is that the number of samples needed by GML estimation decreases as the SNR increases (eventually $m = k + 1$ is sufficient as $\mathsf{SNR} \to \infty$) but the number of samples needed for MC estimation asymptotes. Hence, the "cost" of the computational efficiency gained by the MC estimator is the inability to capitalize on large SNR.

The contribution of this paper is to generalize the performance of the MC algorithm to the problem of approximate sparsity patter recovery. The achievability results for approximate recovery given in [10] correspond to the same computationally expensive GML estimator investigated by Wainwright [4]. The results in this paper complete the picture by determining the complementary performance of the MC estimator. We show that MC estimation achieves the same approximate recovery scalings as the GML estimator. Also, as was the case for perfect recovery, our results show that the "cost" associated with MC estimation is that, unlike the GML estimator, the sufficient number of samples does not decrease with large SNR. This result is meaningful because it is the first one to give (bounded error) performance guarantees on sparsity recovery for a computationally simple algorithm in the setting where $m \ll n$. Table 1 shows the results of this paper in the context of the previous work for the setting $k = \Theta(n)$.

2 Problem Formulation

2.1 Notation

We use capital letters such as X and Y to denote random variables and lower case letters for their realizations x and y. Vectors such as the random vector $\mathbf{X} = (X_1, \cdots, X_n)$ or its realization $\mathbf{x} = (x_1, \cdots, x_n)$ are denoted with boldface. Any subset of the integers such as the random subset $\underline{K}$ or its realization $\underline{k}$ are denoted with an underscore. We also use a boldface capital letter to denote a random matrix such as $\mathbf{A}$ and a regular capital letter for its realization A. The transpose of a matrix A is denoted by A^T. For any vector $\mathbf{x}$ and set of integers $\underline{k}$ the notation $\mathbf{x}_{\underline{k}}$ denotes the vector of elements indexed by $\underline{k}$. Likewise for a matrix A the notation $A_{\underline{k}}$ denotes the submatrix formed by concatenating the columns of A indexed by $\underline{k}$. Assume throughout that logarithms are natural.

2.2 Sampling Rate and Distortion

We consider estimation of a signal $\mathbf{x} \in \mathbb{R}^n$ where $\mathbf{x}$ is known a priori to be exactly k-sparse but the support $\underline{k} := \{i \in \{1, \cdots, n\} : x_i \neq 0\}$ and the values of the non-zero elements indexed by $\underline{k}$ are unknown. It is assumed that $\mathbf{x}$ is sampled using m random measurement vectors $\mathbf{\Phi}_1, \cdots, \mathbf{\Phi}_m \in \mathbb{R}^n$ where $\mathbf{\Phi}_i$ are i.i.d. Gaussian $\mathcal{N}(0, \frac{1}{n} I_{n\times n})$. The resulting observations $\mathbf{Y} \in \mathbb{R}^m$ have the form

$$\mathbf{Y} = \mathbf{A}\mathbf{x} + \mathbf{W} \tag{2}$$

where $\mathbf{A}$ is the $m \times n$ matrix with rows $\mathbf{\Phi}_i$ and $\mathbf{W}$ is additive white Gaussian noise $\mathcal{N}(0, I_{m\times m})$. It is also assumed that the true support is a random variable $\underline{K}$ distributed uniformly over all $\binom{n}{k}$ supports of size k. Under these assumptions, the per-sample SNR is proportional to the average power of $\mathbf{x}$

$$\mathsf{SNR} := \frac{\mathbb{E}\langle \mathbf{\Phi}_i, \mathbf{x}\rangle^2}{\mathbb{E}W_i^2} = \frac{1}{n}\|\mathbf{x}\|^2. \tag{3}$$

Given the sampling matrix $\mathbf{A}$ and the samples $\mathbf{Y}$ the goal is to estimate the support $\underline{K}$. It is important to observe that there are two different error events: for each index $i \in \{1, \cdots, n\}$ a "missed detection" occurs if $x_i \neq 0$ but i is not included in the estimated support and a "false alarm" occurs if $x_i = 0$ but i is included in the estimated support. In this paper, the goal is to minimize the maximum number of each error type and we use the distortion

$$d(\underline{k}, \hat{\underline{k}}) := \max\Big(|\underline{k} \backslash \hat{\underline{k}}|, |\hat{\underline{k}} \backslash \underline{k}|\Big). \tag{4}$$

Since k is known, minimization of the above metric may equivalently be viewed as minimizing the total number of errors subject to the constraint that there are an equal number of each error type. For a given estimator f an error is said to occur if $d(\underline{K}, f(\mathbf{Y})) > \alpha k$ where the distortion $\alpha \in [0, 1]$ is the allowable fraction of errors ($\alpha = 0$ corresponds to perfect recovery).

It is also important to observe that the performance of any algorithm depends on the non-zero values of $\mathbf{x}$. These values are referred to as the *sparsity coefficients* and are denoted by the vector $\mathbf{s} := \mathbf{x}_{\underline{k}} \in \mathbb{R}^k$. Both deterministic and stochastic assumptions for the sparsity coefficients are considered. Under deterministic assumptions, $\mathbf{s}$ is constrained to some subset of $\mathbb{R}^k$ and the probability of error $P_e^{(n)}(\alpha) =$

$\sup_{\mathbf{s}} \mathbb{P}\left(d(\underline{K}, f(\mathbf{Y})) > \alpha k\right)$ corresponds to the worst case. Under stochastic assumptions, $\mathbf{S}$ is a random vector and the probability of error $P_e^{(n)}(\alpha) = \mathbb{E}_{\mathbf{S}} \mathbb{P}\left(d(\underline{K}, f(\mathbf{Y})) > \alpha k\right)$ corresponds to the average error with respect to the probability measure on $\mathbf{S}$.

We view the problem of support recovery from a sampling perspective where the goal is to determine how many samples m are needed as a function of the parameters k, n, α and the assumptions on the sparsity coefficients. We proceed by considering a sequence of problems indexed by the dimension n and characterize the behavior of the problem in the limit as $n \to \infty$. It is assumed throughout that $k = \Omega n$ where the *sparsity rate* $\Omega \in (0, 1)$ measures the degrees of freedom per dimension of $\mathbf{x}$ and is analogous to the "bandwidth" of a signal.

We use $\rho = m/n$ to denote the *sampling rate* and the compressed sensing setting corresponds to $\rho < 1$. A sampling rate distortion pair (ρ, α) is said to be *achievable* for an estimator f if $P_e^{(n)}(\alpha) \to 0$ as $n \to \infty$ for a sequence of problems with sampling rate ρ. Moreover, the *sampling rate function* $\rho_f(\alpha)$ is the infimum of rates ρ such that (ρ, α) is achievable using f, and $\rho(\alpha) = \inf_f \rho_f(\alpha)$ denotes the best possible sampling rate function using any algorithm.

2.3 Estimators

The design and analysis of estimators depends on several factors such as the degree of prior information about the sparsity coefficients, the loss function that is minimized, and computational constraints. Given a target distortion bound α and set of assumptions on the sparsity coefficients (either deterministic or stochastic), the optimal estimator is the one that minimizes our loss function, that is

$$f^* = \inf_f P_e(\alpha).$$

Hence the sampling rate function $\rho_{f^*}(\alpha) = \rho(\alpha)$ gives a baseline measure of the best performance (i.e. lowest sampling rate) that is achievable for any estimator. A lower bound on $\rho(\alpha)$ is given in [11] for various assumptions on the sparsity coefficients.

In contrast to optimal estimation, the estimators considered thus far in the literature for achievability results do not utilize any information about the sparsity coefficients or the distortion α. For instance, the estimator analyzed for perfect recovery in [4] and approximate recovery in [10] corresponds to the maximum likelihood estimated in the special case where there is there is uniform prior on the sparsity coefficients and $\alpha = 0$. Hence, we refer to it as the generalized maximum likelihood estimate targeting zero distortion, or simply GML-0. This estimator is computationally expensive because it requires a search through all possible sparsity patterns and can be expressed as

$$\begin{aligned} f_{\text{GML-0}}(\mathbf{y}) &= \arg\max_{\underline{k}} \Big\{ \sup_{\mathbf{s} \in \mathbb{R}^k} p_{\mathbf{Y}|\underline{K}}(\mathbf{y} | \underline{K} = \underline{k}) \Big\} \\ &= \arg\min_{\underline{k}} \|(I_{m\times m} - A_{\underline{k}}(A_{\underline{k}}^T A_{\underline{k}})^{-1} A_{\underline{k}})\mathbf{y}\|. \end{aligned}$$

An upper bound on $\rho_{\text{GML-0}}(\alpha)$ is given in [11] for various assumptions on the sparsity coefficients.

The MC estimator analyzed previously for perfect recovery in [5] and in this paper for approximate recovery provides a computationally efficient alternative to GML-0 estimation. The MC estimate corresponds to the k largest (in magnitude) elements of the n-dimensional vector $A^T\mathbf{y}$. Although such estimation amounts to sorting the elements of $A^T\mathbf{y}$, it may be equivalently represented as an optimization over supports as

$$f_{\text{MC}}(\mathbf{y}) = \arg\max_{\underline{k}} \|(A^T\mathbf{y})_{\underline{k}}\|.$$

3 Results

The results in the paper are the characterization of the sampling rate function $\rho_{MC}(\alpha)$ for various assumptions on the sparsity coefficients. Our first result applies to any setting where the empirical distribution of the sparsity coefficients converges to a non-random limit.

Theorem 1. *For each n, let the empirical distribution function of the squared value of the sparsity coefficients be given by $F_n(u) = \frac{1}{k}\sum_{i=1}^{k} \mathbf{1}(s_i^2 \le u)$. If $F_n \to F$ where F is a distribution function with $\int_{\mathbb{R}} u dF(u) = P < \infty$, then $\rho_{MC}(\alpha)$ is given by the solution to*

$$\int_{\mathbb{R}} G\left(\sqrt{\frac{\rho_{MC}(\alpha)}{1+\Omega P} u}, Q^{-1}\left(\frac{\alpha\Omega}{2(1-\Omega)}\right)\right) dF(u) = \alpha \tag{5}$$

where $G(\mu,t) = 1 - Q(t-\mu) - Q(t+\mu)$ and $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$.

In the above theorem, the quantity P corresponds to the average power of each sparsity coefficient s_i and hence $\mathsf{SNR} = \Omega P$. The following corollary addresses the setting where the limiting distribution is Gaussian.

Corollary 1 (Gaussian Sparsity Coefficients)**.** *If the sparsity coefficients are i.i.d. zero mean Gaussian with variance P then $\rho_{MC}(\alpha)$ is given by*

$$\rho_{MC}(\alpha) = \frac{1+\Omega P}{P}\left[\left(\frac{Q^{-1}\left(\frac{\alpha\Omega}{2(1-\Omega)}\right)}{Q^{-1}\left(\frac{1-\alpha}{2}\right)}\right)^2 - 1\right]. \tag{6}$$

We remark that the above result clearly illustrates the dependence on the SNR: The sampling rate function scales like $1/\mathsf{SNR}$ for small values and asymptotes for large values.

Our next result applies to any setting in which the sparsity coefficients are known to be bounded.

Theorem 2 (Bounded Sparsity Coefficients)**.** *For each n, assume that the sparsity coefficients obey the elementwise constraints $0 < B \le s_i^2 \le C < \infty$. Then $\rho_{MC}(\alpha)$ is given by the solution to*

$$\sup_F \int_{\mathbb{R}} G\left(\sqrt{\frac{\rho_{MC}(\alpha)}{1+\Omega P_F} u}, Q^{-1}\left(\frac{\alpha\Omega}{2(1-\Omega)}\right)\right) dF(u) = \alpha \tag{7}$$

where the supremum is over all distribution functions supported on the interval $[B, C]$ and $P_F = \int_{\mathbb{R}} u dF(u)$.

In the above theorem the quantity P_F corresponding to the maximizing distribution F is the average power of the sparsity coefficients (in the worst case). Although this value may be hard to compute it general, it is clear that $B \le P_F \le C$ and hence the SNR is bounded. The following corollary gives a simplified, and necessarily weaker, sufficient condition.

Corollary 2. *Under the same assumptions as Theorem 2, any sampling rate distortion pair (ρ, α) is achievable using MC estimation if $\rho > \rho_{MC}^{+}(\alpha)$ where*

$$\begin{aligned} \rho_{MC}^{+}(\alpha) &= \frac{1+\Omega C}{B}\left(Q^{-1}(\alpha) + Q^{-1}\left(\tfrac{\alpha\Omega}{2(1-\Omega)}\right)\right)^2 && (8) \\ &\ge \frac{8(1+\Omega C)}{B} \log\left(\frac{1-\Omega}{\alpha\Omega}\right) && (9) \end{aligned}$$

4 Illustration of Results

This section illustrates the sampling rate functions for MC estimation characterized in Corollaries 1 and 2. These results are compared with the corresponding results for optimal and GML-0 estimation from [11].

We first consider the setting of Corollary 1 in which the sparsity coefficients are zero mean Gaussian with variance P. Figure 1 shows the sampling rate functions for both high and low SNR, and Figure 2 shows the sampling rates as a function of the SNR for fixed distortion $\alpha = 0.2$. In the low SNR setting we see that the MC rate improves on the GML-0 upper bound. This does not necessarily mean that MC estimation performs better than GML-0 estimation but it does provide a tighter upper bound on optimal estimation. In the high SNR setting, we see that GML-0 estimation attains near optimal performance but the MC estimator has essentially the same performance as in the low SNR setting.

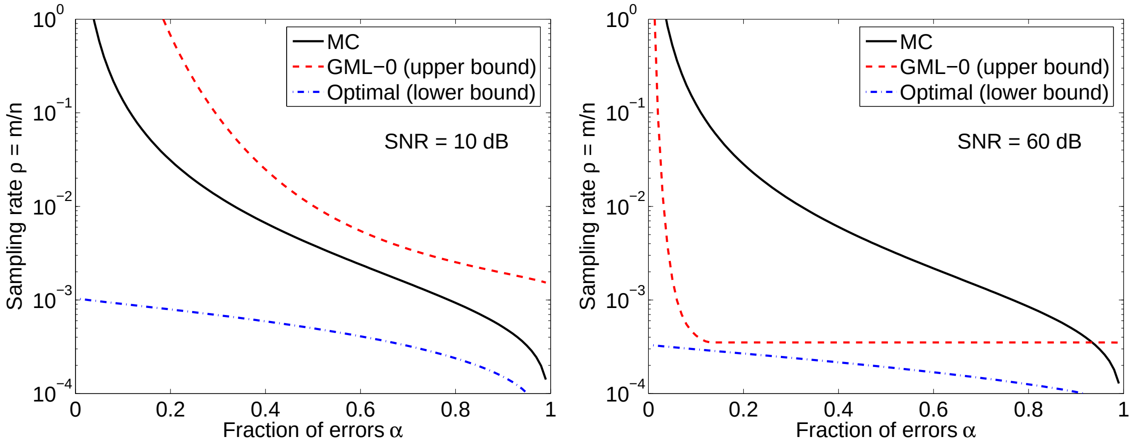

Figure 1: Characterization of sampling rate functions for high and low SNR when the sparsity coefficients are i.i.d. $\mathcal{N}(0, P)$, the sparsity rate is $\Omega = 10^{-4}$, and $\mathsf{SNR} = \Omega P$.

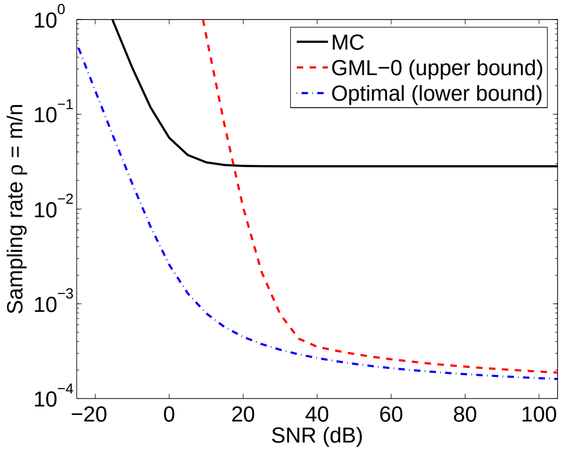

Figure 2: Characterization of sampling rate functions for $\alpha = 0.2$ as a function of the SNR when the sparsity coefficients are i.i.d. $\mathcal{N}(0, P)$ and the sparsity rate is $\Omega = 10^{-4}$.

Next, we consider the deterministic setting of Corollary 2 in which the sparsity coefficients obey the elementwise constraint $B \leq s_i^2 \leq C$. Figure 3 shows the sampling rate functions for both high and low SNR, and Figure 4 shows the sampling rates as a function of the SNR for fixed distortion $\alpha = 0.01$. All plots correspond to the bounds $B = C/2$ with $\mathsf{SNR} = \Omega B$. As in the Gaussian setting, we see that MC estimation provides a tighter bound on ML estimation in low SNR but does not improve as the SNR increases.

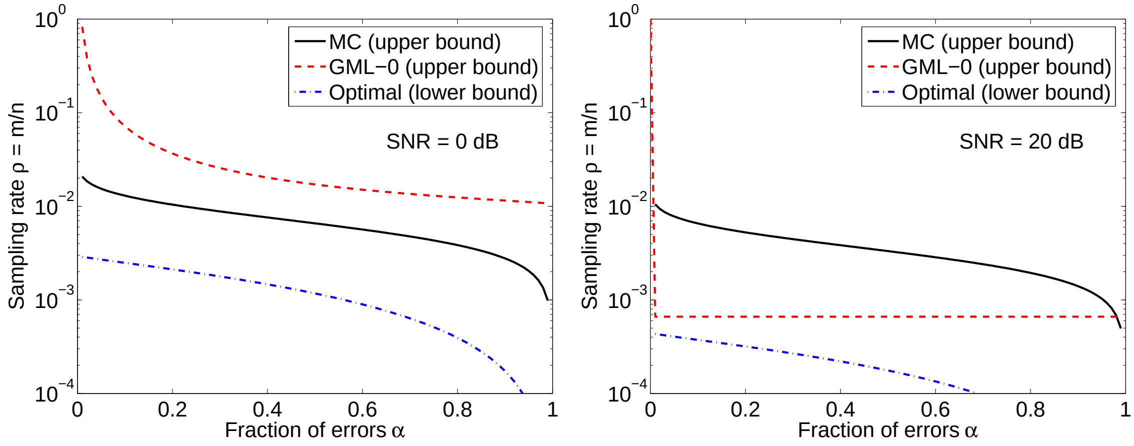

Figure 3: Characterization of sampling rate functions for high and low SNR when the sparsity coefficients are bounded $B \leq s_i^2 \leq 2B$, the sparsity rate is $\Omega = 10^{-4}$, and $\mathsf{SNR} = \Omega B$.

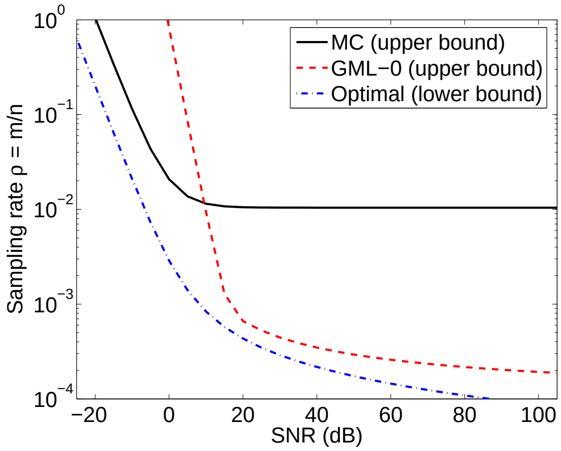

Figure 4: Characterization of sampling rate functions for $\alpha = 0.01$ as a function of the SNR when the sparsity coefficients are bounded $B \leq s_i^2 \leq 2B$, the sparsity rate is $\Omega = 10^{-4}$, and $\mathsf{SNR} = \Omega B$.

Acknowledgments

Part of this research was conducted when the authors were visiting Delft University of Technology and has been funded in part by 3TU.Cedict: Centre for Dependable ICT Systems. It has also been funded in part by ARO MURI No. W911NF-06-1-0076.

References

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

[2] 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.

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

[4] M. J. Wainwright, "Information-theoretic bounds on sparsity recovery in the high-dimensional and noisy setting," in *Proc. IEEE Int. Symp. on Inform. Theory*, Nice, France, Jun. 2007.

[5] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions on sparsity pattern recovery," May 2008, arXiv:0804.1839v1 [cs.IT].

[6] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," in *Proc. IEEE Int. Symp. on Inform. Theory*, Toronto, Canada, Jul. 2008.

[7] G. Reeves, "Sparse signal sampling using noisy linear projections," Univ. of California, Berkeley, Dept. of Elec. Eng. and Cpmp. Sci., Tech. Rep. UCB/EECS-2008-3, Jan. 2008.

[8] M. Akcakaya and V. Tarokh, "Shannon theoretic limits on noisy compressive sampling," Nov. 2007, arXiv:0711.0366v1 [cs.IT].

[9] S. Aeron, M. Zhao, and V. Saligrama, "Fundamental limits on sensing capacity for sensor networks and compressed sensing," Apr. 2008, arXiv:0804.3439v1 [cs.IT].

[10] G. Reeves and M. Gastpar, "Sampling bounds for sparse support recovery in the presence of noise," in *Proc. IEEE Int. Symp. on Inform. Theory*, Toronto, Canada, Jul. 2008.

[11] ——, "Information-theoretic sampling rates bounds for approximate sparsity recovery," preprint.