Compressed Sensing With Prior Information: Information-Theoretic Limits and Practical Decoders

Article *in* IEEE Transactions on Signal Processing · January 2013

DOI: 10.1109/TSP.2012.2225051

3 authors:

Jonathan Scarlett
National University of Singapore
95 PUBLICATIONS **686** CITATIONS

SEE PROFILE

Jamie Scott Evans
University of Melbourne
243 PUBLICATIONS **3,904** CITATIONS

SEE PROFILE

Subhrakanti Dey
University of Melbourne
161 PUBLICATIONS **1,971** CITATIONS

SEE PROFILE

Compressed Sensing With Prior Information: Information-Theoretic Limits and Practical Decoders

Jonathan Scarlett, Jamie S. Evans, *Member, IEEE*, and Subhrakanti Dey, *Senior Member, IEEE*

***Abstract*—This paper considers the problem of sparse signal recovery when the decoder has prior information on the sparsity pattern of the data. The data vector $\mathbf{x} = [x_1, \ldots, x_N]^T$ has a randomly generated sparsity pattern, where the i-th entry is non-zero with probability p_i. Given knowledge of these probabilities, the decoder attempts to recover $\mathbf{x}$ based on M random noisy projections. Information-theoretic limits on the number of measurements needed to recover the support set of $\mathbf{x}$ perfectly are given, and it is shown that significantly fewer measurements can be used if the prior distribution is sufficiently non-uniform. Furthermore, extensions of Basis Pursuit, LASSO, and Orthogonal Matching Pursuit which exploit the prior information are presented. The improved performance of these methods over their standard counterparts is demonstrated using simulations.**

***Index Terms*—Basis pursuit, compressed sensing, compressive sampling, information-theoretic bounds, Lasso, orthogonal matching pursuit, prior information, sparsity pattern recovery, support recovery.**

I. Introduction

THE problem of estimating an unknown vector from a number of noisy linear projections arises frequently in signal processing. Given a data vector $\mathbf{x} \in \mathbb{R}^N$ and a known measurement matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$, the observation vector $\mathbf{y} \in \mathbb{R}^M$ is of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} \tag{1}$$

where $\mathbf{n}$ is additive noise. In the absence of any additional information on $\mathbf{x}$, this problem is well-posed only in the case that $M \geq N$, where techniques such as least squares can be used. However, if $\mathbf{x}$ is known to be *sparse*, i.e. $K \ll N$ where K is the number of non-zero elements of $\mathbf{x}$, then an accurate estimate of $\mathbf{x}$ can be obtained even with $M \ll N$ [1]. This phenomenon is known as *compressed sensing* or *compressive sampling*. A closely related problem is that of *sparsity pattern recovery* or *support recovery*, where the aim is to recover the support set of $\mathbf{x}$, defined the be the positions of its non-zero entries. It is easy to see that if this problem is solved then the problem of sparse signal estimation is essentially solved as well, since one can then apply least squares restricted to the known support set.

Recently there has been a great deal of work on the design and analysis of both tractable and intractable methods for compressed sensing and sparsity pattern recovery. The focus of this paper is on developing and analyzing methods which exploit *prior information* (other than $\mathbf{x}$ being sparse). We are interested in information-theoretic limits on the performance of *any* decoder, as well as the design of tractable methods. These are complementary and both of great interest, with tractable methods being useful for practical systems, and information-theoretic limits being highly valuable for assessing the performance of tractable methods and determining the level of further improvement possible.

The motivation for our work is the availability of prior information in several applications of compressed sensing. In sensor networks [2] the information obtained via a particular sensor could be used as information for another sensor. In functional medical resonance imaging [3], the prior information may arise from knowing which parts of the brain are usually associated with various decision making processes. In other applications, one may be interested in performing compressed sensing at multiple time instants, where the support set is correlated between times. This occurs, for example, in multipath channel estimation [4] and real time video reconstruction [5]. Finally, in some applications compressed sensing is one of several stages of an estimation problem, and additional information can be passed from an earlier stage [6].

The above applications involve many different types of prior information, making it difficult to obtain a tractable mathematical model which is suitable for each one. However, valuable insight can still be gained by analyzing specific models. In this paper, we assume prior information on the *sparsity pattern* of $\mathbf{x}$, but not the *values* of the non-zero entries. Specifically, we assume that the sparsity pattern is generated at random, with each entry of $\mathbf{x}$ being non-zero with a given probability which is known at the decoder. A similar model is used in [7], and in [6] it is shown that such a model can provide insight into the compressed sensing problem even when prior information is not available. Specifically, [6] proposes a two-step recovery algorithm in which a hypothetical non-uniform sparsity model is generated in the first step, and a compressed sensing technique which exploits prior information is used in the second step.

A. Notation

We use the following notation throughout this paper. $\Pr(\cdot)$ denotes the probability of an event, and $E[\cdot]$ denotes statistical expectation. $\stackrel{d}{=}$ means "distributed as", $\stackrel{d}{\approx}$ means "approximately

Manuscript received August 25, 2011; revised January 27, 2012 and July 31, 2012; accepted September 22, 2012. Date of publication October 16, 2012; date of current version December 24, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Namrata Vaswani.

J. Scarlett is with the Department of Engineering, University of Cambridge, Cambridge CB2 1PZ, U.K. (e-mail: jms265@cam.ac.uk).

J. S. Evans is with the Department of Electrical and Computer Systems Engineering, Monash University, Clayton, Victoria 3800, Australia (e-mail: jamie.evans@monash.edu).

S. Dey is with the Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, Victoria 3052, Australia (e-mail: sdey@unimelb.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2012.2225051

distributed as", and $\propto$ means "proportional to" in a probabilistic sense. The Gaussian distribution is denoted by $\mathrm{N}(\cdot,\cdot)$. The support of a vector is denoted by $\mathrm{supp}(\cdot)$. All logarithms have base e, and $H_2(\cdot)$ is the binary entropy function in nats. $|\cdot|$ denotes magnitude when applied to a number, and denotes cardinality when applied to a finite set. $\|\cdot\|_p$ denotes the ℓ_p-norm of a vector if $p \geq 1$, and the number of non-zero elements in a vector (commonly referred to as the ℓ_0-*norm*) if $p = 0$.

For an index set $\mathcal{J}$ and matrix $\mathbf{M}$, $\mathbf{M}_{\mathcal{J}}$ denotes the submatrix of $\mathbf{M}$ containing the columns indexed by $\mathcal{J}$. Similarly, for a vector $\mathbf{v}$, $\mathbf{v}_{\mathcal{J}}$ denotes the subvector of $\mathbf{v}$ containing the elements indexed by $\mathcal{J}$. The transpose of a vector or matrix is denoted by $(\cdot)^T$. For two functions $g(N)$ and $h(N)$, we write $g = O(h)$ if $|g| \leq c|h|$ for some constant c when N is sufficiently large, $g = o(h)$ if $\lim_{N\to\infty} \frac{g(N)}{h(N)} = 0$, $g = \Omega(h)$ if $h = O(g)$, $g = \omega(h)$ if $h = o(g)$, and $g = \Theta(h)$ if both $g = O(h)$ and $g = \Omega(h)$ hold.

B. Problem Statement

The data vector $\mathbf{x}$ and measurement vector $\mathbf{y}$ are related by (1). We assume that $\mathbf{A}$ has entries $a_{ki} \overset{d}{=} \mathrm{N}(0,1)$ and $\mathbf{n}$ has entries $n_k \overset{d}{=} \mathrm{N}(0,\sigma^2)$ for $k \in \{1,\ldots,M\}$ and $i \in \{1,\ldots,N\}$.[1] The i-th column of $\mathbf{A}$ is denoted by $\mathbf{a}_i$. The data vector $\mathbf{x} = [x_1,\ldots,x_N]^T$ is generated as follows. The i-th entry of $\mathbf{x}$ is given by

$$x_i = g_i s_i \tag{2}$$

where $g_i \in \mathbb{R}$ is a deterministic non-zero value, and $s_i \in \{0,1\}$ is a random variable indicating whether the entry is non-zero. The probability of s_i being equal to one is denoted by p_i, and we assume that the random variables $\{s_i\}_{i=1}^N$ are independent. The *support set* $\mathcal{I} = \{i | s_i = 1\}$ of $\mathbf{x}$ is therefore distributed according to

$$\Pr(\mathcal{I} = I) = \prod_{i\in I} p_i \prod_{i\notin I}(1-p_i). \tag{3}$$

The task of the decoder is to estimate $\mathbf{x}$ from the observation vector $\mathbf{y}$ given in (1), using $\mathbf{p} = [p_1,\ldots,p_N]^T$ as prior information. We refer to p_i as the *support probability* of the i-th entry of $\mathbf{x}$.

We model the support probabilities as follows. For $n = 1,\ldots,G$ a proportion f_n of the coefficients have support probability equal to p'_n, where $\sum_{n=1}^G f_n = 1$ and $f_n > 0$ for all n. That is, the coefficients are divided into G groups, where the support probabilities of all coefficients in a given group are equal. The number of coefficients in group n is denoted by $N_n = N f_n$. We define the *average binary entropy* of the support probabilities as

$$\overline{H} = \frac{1}{N}\sum_{i=1}^N H_2(p_i) = \sum_{n=1}^G f_n H_2(p'_n) \tag{4}$$

where $H_2(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function.

[1]Some authors use different normalizations, such as each entry of $\mathbf{A}$ and $\mathbf{n}$ having variance $\frac{1}{M}$ and $\frac{\sigma^2}{M}$ respectively. This does not affect the results provided the SNR is the same.

TABLE I
SUMMARY OF DEFINITIONS

Symbol(s)	Meaning
$\mathbf{x}$, x_i	Data vector and its i-th entry
$\mathbf{A}$, $\mathbf{a}_i$, a_{ki}	Measurement matrix, its i-th column and its (k,i)-th entry
$\mathbf{y}$, y_i	Observation vector and its i-th entry
$\mathbf{n}$, n_i	Noise vector and its i-th entry
σ^2	Variance of each element of $\mathbf{n}$
N, M	Number of coefficients in $\mathbf{x}$ and number of measurements in $\mathbf{y}$
g_i, s_i	Deterministic non-zero value and indicator random variable such that $x_i = g_i s_i$
$\mu_{\min}$, $\mu_{\max}$, $\mu'_{\min}$, $\mu'_{\min}$	$\min_i \lvert x_i\rvert$, $\max_i \lvert x_i\rvert$, $\min_i \lvert g_i\rvert$ and $\max_i \lvert g_i\rvert$
$\mathcal{I}$, K, $\overline{K}$	Support set of $\mathbf{x}$, $\lvert\mathcal{I}\rvert$, and $E[\lvert\mathcal{I}\rvert]$
$\mathbf{p}$, p_i	Support probability vector and its i-th entry
G	Number of groups
f_n, N_n	Non-zero proportion and number of coefficients in group n, $n \in \{1,...,G\}$
p'_n	Support probability for coefficients in group n, $n \in \{1,...,G\}$
$\overline{H}$	Average binary entropy of the support probabilities

For a given support set $\mathcal{I}$, we denote the number of non-zero entries by $K = |\mathcal{I}|$. The average of $|\mathcal{I}|$ with respect to the distribution in (3) is denoted by $\overline{K}$, i.e.

$$\overline{K} = \sum_{i=1}^N p_i = \sum_{n=1}^G N_n p'_n. \tag{5}$$

For a given data vector $\mathbf{x}$, the smallest and largest magnitudes are denoted by $\mu_{\min} = \min_i |x_i|$ and $\mu_{\max} = \max_i |x_i|$ respectively. The smallest and largest magnitudes of $\mathbf{g} = [g_1,\ldots,g_N]^T$ are denoted by $\mu'_{\min} = \min_i |g_i|$ and $\mu'_{\max} = \max_i |g_i|$ respectively. For reference, the definitions of the main symbols used throughout the paper are summarized in Table I.

C. Contributions and Previous Work

1) Overview of Standard Techniques: Before summarizing our contributions, we outline the practical methods for compressed sensing without prior information (other than $\mathbf{x}$ being sparse) most relevant to this paper. Two commonly used ℓ_1-based methods are Basis Pursuit (BP) [8] and Least Absolute Shrinkage and Selection Operator (LASSO) [9], which are respectively described by

$$\min_{\mathbf{x}:\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_1 \tag{6}$$

$$\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 + \tau\|\mathbf{x}\|_1 \tag{7}$$

where τ is a parameter to LASSO which controls the tradeoff between sparsity and goodness of fit. These techniques are closely related, with BP being the convex relaxation of the intractable ℓ_0 minimization problem $\min_{\mathbf{x}:\mathbf{y}=\mathbf{A}\mathbf{x}} \|\mathbf{x}\|_0$, and LASSO being the convex relaxation of least squares with an ℓ_0 penalty. Both have been shown to give good performance in practical systems, with the required number of measurements generally ranging from $\Theta\left(K\log\frac{N}{K}\right)$ to $\Theta(K\log(N-K))$

depending on the performance requirements; see [8], [10] for details.

Orthogonal Matching Pursuit (OMP) uses a greedy approach rather than optimization of an objective [11]. We outline the algorithm here and refer the reader to [11] for details. The decoder repeatedly adds to the support estimate the coefficient whose column of $\mathbf{A}$ is most highly correlated with a residual vector $\mathbf{r}$. The first residual vector used is $\mathbf{y}$, and subsequent residuals are computed as $\mathbf{r} = \mathbf{y} - \mathbf{A}\widehat{\mathbf{x}}$ where $\widehat{\mathbf{x}}$ is the least squares estimate of $\mathbf{x}$ restricted to the support set obtained so far. The algorithm terminates when the number of iterations exceeds a threshold, or when the norm of the residual falls below a threshold. This technique is computationally efficient and has been shown to exhibit performance which is comparable to BP and LASSO [12].

2) Contributions: A summary of our contributions is as follows:

- We extend the techniques of [13] to obtain both sufficient and necessary conditions on the number of measurements required for exact recovery of the support set as N and K grow large, with $\mathbf{p}$ known at the decoder. Sufficient conditions are obtained via the analysis of a joint typicality decoder, and necessary conditions are obtained by an analogy to a multiple input single output (MISO) communication channel in which the decoder attempts to recover the support set. We show that the introduction of prior information can significantly reduce the number of measurements required. In particular, it is shown that $\Theta(N\overline{H})$ measurements suffice under various assumptions.
- We present three practical decoding techniques for compressed sensing with prior information, which are extensions of BP, LASSO and OMP. The idea is to introduce weights which depend on the support probabilities, so that coefficients with higher probability are favored. For the extensions of BP and LASSO, the term $\|\mathbf{x}\|_1$ in the objective is replaced by $\sum_{i=1}^{N} w_i|x_i|$, and we motivate the use of the weights $w_i = -\log p_i$. For the extension of OMP the decoder greedily makes a decision based on the sum of a correlation term and a weighting term, rather than just the correlation term. We motivate the use of weights proportional to $\log \frac{p_i}{1-p_i}$.
- Using simulations, we demonstrate the improved performance of our practical decoding techniques compared to their standard counterparts. It is seen empirically that $N\overline{H}$ has a significant effect on performance when using these techniques. Furthermore, we explore the impact on performance of noise and mismatch.

3) Previous Work on Information-Theoretic Limits: Previous work on information-theoretic limits on *exact* sparsity pattern recovery is as follows. In [14], necessary conditions for the maximum likelihood (ML) decoder are obtained, and a simple maximum correlation decoder is analyzed. In [15] an analogy is drawn between sparsity pattern recovery and the Gaussian multiple access channel in order to find both necessary and sufficient conditions on M. In [16], sufficient conditions are obtained for the ML decoder, and necessary conditions are found based on Fano's inequality. These necessary conditions are tightened in [17], and a comparison between dense and sparse ensembles is performed. In [18], sufficient conditions are derived and shown to be tight in a scaling-law sense by comparison to the necessary conditions of [17]. For information-theoretic limits with respect to other performance metrics, see [13], [19], [20]. In contrast to our work, these papers consider the case that all support sets of a given cardinality K are equiprobable.

4) Previous Work on Exploiting Prior Information: In [21] a method called Modified-CS is proposed, in which the objective function only penalizes coefficients outside a partially known support. In [7] a Grassman angle approach is used to provide a comprehensive analysis of a weighted ℓ_1 minimization problem, focusing primarily on the case that there are only two different probabilities, and hence only two weights. The authors provide a method for optimizing the weights with respect to various metrics, but due to the highly non-convex nature of the problem, significant computation is needed. The introduction of weights into LASSO and OMP is proposed in [22] for the dual problem of sparse signal approximation with prior information. We present a comparison of our techniques to those of [7], [21], [22] in Section IV-D.

In [23] the method of *model-based compressed sensing* is introduced, in which the support is known to have a particular structure. Specifically, only a subset of the $\binom{N}{K}$ supports of a given cardinality K can occur. Among others, this includes block sparsity [24] and tree sparsity [25] as special cases; see [23] for a complete set of references. While the framework of [23] does not include our setup as a special case, we draw some connections between the two in Section II-C-4.

In [26], a Kalman filter is used to exploit correlations between the *values* of $\mathbf{x}$ under the assumption that the support changes very slowly. In [27], a Bayesian learning approach is used to account for correlations in the measurement vectors, but again the focus is on correlations between the values of $\mathbf{x}$, rather than the support.

D. Paper Organization

In Section II we present a summary and discussion of our information-theoretic results, which are given in Theorems 1 and 2. In Section III we give proofs of these theorems. In Section IV we present our practical decoding techniques and compare them to existing methods, as well as analyzing their performance via simulations. Conclusions are drawn in Section V.

II. Information-Theoretic Results

Before stating the main theorems, we give a precise definition of the performance metric. For each N, let the number of groups G and the corresponding $\{f_n\}_{n=1}^{G}$ and $\{p'_n\}_{n=1}^{G}$ be given, along with the gains $\{g_i\}_{i=1}^{N}$. The support set $\mathcal{I}$ is generated according to (3), yielding the data vector $\mathbf{x}$. The measurement matrix $\mathbf{A}$ and noise vector $\mathbf{n}$ are generated at random, yielding the observation vector $\mathbf{y}$. The decoder takes $\mathbf{y}$ as input and forms an estimate $\widehat{\mathcal{I}}$ of the support set $\mathcal{I}$; the quantities $\mathbf{A}$, $\{p_i\}_{i=1}^{N}$, σ^2, $\mu_{\min}$ and $\mu_{\max}$ are assumed to be known. An error is said to have occurred if $\mathcal{I} \neq \widehat{\mathcal{I}}$.

For a *given* support set $\mathcal{I}$, we define the error probability

$$p_e(\mathcal{I}) = \Pr(\mathcal{I} \neq \widehat{\mathcal{I}}|\mathcal{I}) \tag{8}$$

where the probability is taken with respect to statistics of $\mathbf{A}$ and $\mathbf{n}$. Similarly, we define

$$p_e = \Pr(\mathcal{I} \neq \widehat{\mathcal{I}}) \tag{9}$$

where the probability is taken with respect to statistics of $\mathbf{A}$, $\mathbf{n}$ and $\mathcal{I}$. That is, the quantity in (9) is obtained by averaging (8) over the distribution of $\mathcal{I}$ in (3).

The value $\overline{H}$ in (4) represents the average uncertainty of whether the coefficients are going to be non-zero or not. It plays a major role in both the necessary and sufficient conditions of the number of measurements, as we will see in Section II-B. We remark that while $\overline{H}$ can be independent of N in the linear regime ($\overline{K} = \Theta(N)$), it tends to zero for large N in the sublinear regime ($\overline{K} = o(N)$), so the presence of the term $N\overline{H}$ in the number of measurements does not imply that the overall value is $\Theta(N)$.

Throughout this section, we will sometimes use a superscript to make the dependence of a variable on N explicit (e.g. $G^{(N)}$). The superscript will be omitted when no confusion arises from doing so.

A. Typicality

We begin by introducing the notion of typicality with respect to the support probabilities $\{p_i\}$. We define $K_n(\mathcal{J})$ to be the number of non-zero coefficients in the set $\mathcal{J}$ from group n.

Definition 1: (Typicality) A support set $\mathcal{J} \subseteq \{1, 2, \ldots, N\}$ is ϵ-typical if

$$\left| \frac{K_n(\mathcal{J})}{N_n} - p'_n \right| \leq \epsilon \min\{p'_n, 1 - p'_n\} \tag{10}$$

for all $n \in \{1, \ldots, G\}$. The set of all ϵ-typical supports is denoted by $T_1(\epsilon)$.

The following proposition gives useful properties of the typical set.

Proposition 1:

(i) For any $\mathcal{J} \in T_1(\epsilon)$, the cardinality of $\mathcal{J}$ satisfies

$$\left| |\mathcal{J}| - \overline{K} \right| \leq \epsilon \overline{K}. \tag{11}$$

(ii) The cardinality of the typical set satisfies

$$|T_1(\epsilon)| \leq \exp\left(N\overline{H}(1+\epsilon)\right). \tag{12}$$

(iii) If $G = O(1)$ and $\min\{N_n p'_n, N_n(1-p'_n)\} \to \infty$ for all $n \in \{1, \ldots, G\}$, then

$$\Pr\left(\mathcal{I} \in T_1(\epsilon)\right) \to 1. \tag{13}$$

Proof: These follow using standard proofs based on the method of types, e.g. see [28]. For property (iii), the condition $\min\{N_n p'_n, N_n(1-p'_n)\} \to \infty$ implies that the law of large numbers holds for each group, and thus (13) follows from the condition $G = O(1)$. ∎

B. Statement of Main Results

Here we present two theorems summarizing the asymptotic bounds on the number of measurements needed under various assumptions. It should be noted that the sufficient conditions are based on a joint typicality decoder which is too complex to be used in practice, whereas the necessary conditions bound the performance of *any* decoder. In contrast to most of the existing literature, we do *not* assume that the decoder has knowledge of $K = |\mathcal{I}|$.

Theorem 1: (Sufficient Conditions) Fix $\epsilon \in (0,1)$, and let $G^{(N)}$, $\{f_n^{(N)}\}_{n=1}^{G^{(N)}}$ and $\{p_n'^{(N)}\}_{n=1}^{G^{(N)}}$ be such that $\frac{\overline{K}^{(N)}}{N}$ is bounded away from 1, and $\min\{N_n^{(N)} p_n'^{(N)}, N_n^{(N)}(1 - p_n'^{(N)})\} \to \infty, \forall n \in \{1, \ldots, G^{(N)}\}$. Let $\mathcal{I}^{(N)}$ be an arbitrary sequence of support sets satisfying $\mathcal{I}^{(N)} \in T_1^{(N)}(\epsilon)$. If $\frac{N\overline{H}^{(N)}(\mu_{\min}^{(N)})^4}{\log N} \to \infty$, then there exists a decoder such that the error probability $p_e^{(N)}(\mathcal{I}^{(N)})$ tends to zero as $N \to \infty$, provided that

$$M^{(N)} > \left(\overline{K}^{(N)} + 4N\overline{H}^{(N)}\right)(1 + \overline{\epsilon}) \tag{14}$$

for some $\overline{\epsilon} > \epsilon$.

Proof: See Section III-A. ∎

Theorem 2: (Necessary Conditions) Fix $\epsilon \in (0, 0.5)$, and let $G^{(N)}$, $\{f_n^{(N)}\}_{n=1}^{G^{(N)}}$ and $\{p_n'^{(N)}\}_{n=1}^{G^{(N)}}$ be such that $\min\{N_n^{(N)} p_n'^{(N)}, N_n^{(N)}(1 - p_n'^{(N)})\} \to \infty, \forall n \in \{1, \ldots, G^{(N)}\}$. Let $\mathcal{I}^{(N)}$ be an arbitrary sequence of support sets satisfying $\mathcal{I}^{(N)} \in T_1^{(N)}(\epsilon)$, and let $\mathbf{x}^{(N)}$ be the corresponding data vector. Then the error probability $p_e^{(N)}(\mathcal{I}^{(N)})$ for any decoder is bounded away from zero as $N \to \infty$ if

$$M^{(N)} < \max\left\{ \frac{\overline{K}^{(N)}(1-\epsilon)}{G^{(N)}}, \frac{N\overline{H}^{(N)}(1-\overline{\epsilon})}{\frac{1}{2}\log\left(1 + \frac{\left\|\mathbf{x}^{(N)}\right\|_2^2}{\sigma^2}\right)} \right\} \tag{15}$$

for some $\overline{\epsilon} > H_2(\epsilon)$. Consequently, if $\mu_{\max}^{(N)} = \Theta(\mu_{\min}^{(N)}) = O(1)$ and $\frac{N\overline{H}^{(N)}(\mu_{\min}^{(N)})^4}{\log N} \to \infty$ then $\max\left\{\Omega(\overline{K}), \Omega\left(\frac{N\overline{H}}{\log \overline{K}}\right)\right\}$ measurements are necessary.

Proof: See Section III-B. ∎

Theorems 1 and 2 give conditions on M such that $p_e(\mathcal{I}) \to 0$ for a *given* sequence of typical support sets $\mathcal{I}$. We can obtain conditions on M such that $p_e \to 0$ by writing

$$p_e = \sum_I \Pr(\mathcal{I} = I) p_e(I) \tag{16}$$

$$= \sum_{I \in T_1(\epsilon)} \Pr(\mathcal{I} = I) p_e(I) + \sum_{I \notin T_1(\epsilon)} \Pr(\mathcal{I} = I) p_e(I) \tag{17}$$

$$= \sum_{I \in T_1(\epsilon)} \Pr(\mathcal{I} = I) p_e(I) + o(1) \tag{18}$$

where (18) follows from (13).

Corollary 1: Let the assumptions of Theorem 1 be satisfied. If $\frac{N\overline{H}^{(N)}(\mu_{\min}'^{(N)})^4}{\log N} \to \infty$, then there exists a decoder such that the error probability $p_e^{(N)}$ tends to zero as $N \to \infty$, provided that

$$M^{(N)} > \left(\overline{K}^{(N)} + 4N\overline{H}^{(N)}\right)(1 + \epsilon') \tag{19}$$

for some $\epsilon' > 0$.

Proof: We obtain (19) by combining (14) and (18), and noting that ϵ and $\overline{\epsilon}$ can be chosen to be arbitrarily small. The

condition $\frac{N\overline{H}^{(N)}(\mu_{\min}^{(N)})^4}{\log N} \to \infty$ in Theorem 1 is satisfied for all $\mathcal{I}$ since $\mu'_{\min} \leq \mu_{\min}$ by definition. ∎

Corollary 2: Let the assumptions of Theorem 2 be satisfied. The error probability $p_e^{(N)}$ for any decoder is bounded away from zero as $N \to \infty$ if

$$M^{(N)} < \max\left\{ \frac{\overline{K}^{(N)}(1-\epsilon')}{G^{(N)}}, \frac{N\overline{H}^{(N)}(1-\epsilon')}{\frac{1}{2}\log\left(1+\frac{\overline{K}\left(\mu'^{(N)}_{\max}\right)^2}{\sigma^2}\right)} \right\} \tag{20}$$

for some $\epsilon' > 0$.

Proof: For any $\mathcal{I} \in T_1(\epsilon)$, we have

$$\log\left(1+\frac{\|\mathbf{x}\|_2^2}{\sigma^2}\right) \leq \log\left(1+\frac{\overline{K}(1+\epsilon)\left(\mu'_{\max}\right)^2}{\sigma^2}\right) \tag{21}$$

$$\leq (1+\epsilon)\log\left(1+\frac{\overline{K}\left(\mu'_{\max}\right)^2}{\sigma^2}\right) \tag{22}$$

where we have used (11) and the definition of $\mu'_{\max}$. It follows that the necessary number of measurements in (15) can be weakened to

$$\max\left\{ \frac{\overline{K}(1-\epsilon)}{G}, \frac{N\overline{H}}{\frac{1}{2}\log\left(1+\frac{\overline{K}(\mu'_{\max})^2}{\sigma^2}\right)}\frac{1-\overline{\epsilon}}{1-\epsilon} \right\} \tag{23}$$

where $\overline{\epsilon} > H_2(\epsilon)$. The corollary follows by combining (18) and (23) and noting that ϵ, and therefore $\overline{\epsilon}$, can be chosen to be arbitrarily small. ∎

C. Discussion

Since Theorems 1 and 2 apply for any $\mathcal{I} \in T_1(\epsilon)$, it follows from (11) that the corresponding K are close to $\overline{K}$, and the resulting necessary and sufficient conditions are unchanged in a scaling-law sense when $\overline{K}$ is replaced by K. For the remainder of the section, we express the results in terms of K.

Theorem 1 implies that $\max\{\Theta(K), \Theta(N\overline{H})\}$ measurements are sufficient. Theorem 2 states that under the same conditions as Theorem 1, $\max\left\{\Omega(K), \Omega\left(\frac{N\overline{H}}{\log K}\right)\right\}$ measurements are necessary, meaning that there is a gap between the scaling of the necessary and sufficient conditions. Nevertheless, these bounds still provide valuable insight into the number of measurements needed with prior information. In particular, we show in Section II-C-3 that the sufficient number of measurements with prior information can have a lower rate of growth than the necessary number of measurements in the absence of prior information. In the case that $G = 1$, our sufficient and necessary conditions match those of [13]. This is unsurprising, since our proofs are based on techniques used in [13].

1) Discussion of Assumptions: We make the following remarks on the assumptions of Theorems 1 and 2.

- The condition that $\frac{N\overline{H}\mu_{\min}^4}{\log N} \to \infty$ in Theorem 1 is a generalization of the assumptions in [13] that $K\mu_{\min}^4 = \omega(\log K)$ in the linear regime ($K = \Theta(N)$), and $K\mu_{\min}^4 = \omega(1)$ in the sublinear regime ($K = o(N)$). The reason such restrictions on $\mu_{\min}$ are required is that due to the presence of noise, perfect recovery of the support is not possible when the coefficients can be arbitrarily small.
- The conditions that $\frac{\overline{K}}{N}$ remains bounded away from 1 and $G = O(1)$ are mainly for technical reasons. The first is very mild, since $\frac{\overline{K}}{N} \to 1$ would mean that nearly all coefficients of $\mathbf{x}$ are non-zero with high probability. The condition that $G = O(1)$ states that number of groups does not grow unbounded, hence limiting the growth rate of certain expressions in the analysis.
- The condition $\min\{N_n p'_n, N_n(1-p'_n)\} \to \infty$, $\forall n \in \{1,\dots,G\}$ states that each group has an unbounded average number of both zero and non-zero coefficients. This is for technical reasons relating to typicality (see Definition 1). Defining $V_n = \min\{N_n p'_n, N_n(1-p'_n)\}$, we see that $V_n = o(1)$ is of limited interest, since it implies that asymptotically either all or none of the coefficients of the group are zero with probability approaching one. However, $V_n = \Theta(1)$ could be of interest. For example, a group with a large number of low probability coefficients leading to a Poisson distribution would give $V_n = \Theta(1)$. In this paper, we assume there are no such groups.

2) Discussion of Linear Regime: A suitable model for the linear regime is one in which the number of groups and their associated probabilities are independent of N. In this case $\overline{H}$ is constant and hence the necessary and sufficient number of measurements is $\Theta(K)$ under the assumptions of Theorem 1, thus matching the case that the prior information is absent [13]. However, prior information can decrease the constant factor in the sufficient number of measurements. In particular, a simple application of Jensen's inequality in (4) yields $\overline{H} \leq H_2\left(\frac{\overline{K}}{N}\right)$, where the right-hand side recovers the case with no prior information.

3) Discussion of Sublinear Regime: For the sublinear regime, a comparison of our results to those without prior information is presented in Table II under various scalings of $\mu_{\min}$, where it is assumed that $\mu_{\max} = \Theta(\mu_{\min})$. We see that depending on the scaling of $N\overline{H}$, our sufficient conditions can exhibit better scaling than the necessary and sufficient conditions without prior information. For example, consider the case that $G = 2$ and $f_1 = \frac{\overline{K}}{N}$, so that $N_1 = \overline{K}$ coefficients are in group 1 and $N_2 = N - \overline{K}$ are in group 2. We let $p'_2 = \frac{g(\overline{K})}{N-\overline{K}}$ for some $g(\overline{K}) = o(\overline{K})$. Using $N_1 p'_1 + N_2 p'_2 = \overline{K}$, it follows that $p'_1 = 1 - \frac{g(\overline{K})}{\overline{K}}$, and hence

$$N\overline{H} = \overline{K}H_2\left(\frac{g(\overline{K})}{\overline{K}}\right) + (N-\overline{K})H_2\left(\frac{g(\overline{K})}{N-\overline{K}}\right) \tag{24}$$

$$= \Theta\left(g(\overline{K})\log N\right) \tag{25}$$

where (25) follows by noting that the term $(N-\overline{K})H_2\left(\frac{g(\overline{K})}{N-\overline{K}}\right)$ dominates in the sublinear regime and has the same rate of growth as $g(\overline{K})\log N$. Supposing now that $\overline{K} = \log N$, we see that $\Theta(Kg(K))$ measurements are sufficient. From Table II, the necessary number of measurements without prior information is $\Theta\left(\frac{K^2}{\log\log K}\right)$ and $\Theta\left(\frac{K^2}{\log K}\right)$ for the cases $\mu_{\min}^2 = \Theta\left(\frac{\log K}{K}\right)$ and $\mu_{\min}^2 = \Theta(1)$ respectively. We therefore achieve improved scaling in the former case for any $g(\overline{K}) = o\left(\frac{\overline{K}}{\log\log\overline{K}}\right)$, and in the latter case for any $g(\overline{K}) = o\left(\frac{\overline{K}}{\log\overline{K}}\right)$.

TABLE II
NECESSARY AND SUFFICIENT NUMBER OF MEASUREMENTS IN THE SUBLINEAR REGIME WITH AND WITHOUT PRIOR INFORMATION

Scaling	Sufficient With Prior Information	Necessary With Prior Information	Necessary and Sufficient Without Prior Information [18]
$\mu_{\min}^2 = \Theta(\frac{1}{K})$	-	$\max\{\Theta(K), \Theta(N\overline{H})\}$	$\Theta(K\log(N-K))$
$\mu_{\min}^2 = \Theta(\frac{\log K}{K})$	$\max\{\Theta(K), \Theta(N\overline{H})\}$	$\max\{\Theta(K), \Theta(\frac{N\overline{H}}{\log K})\}$	$\max\{\Theta(\frac{K\log(N-K)}{\log K}), \Theta(\frac{K\log\frac{N}{K}}{\log\log K})\}$
$\mu_{\min}^2 = \Theta(1)$	$\max\{\Theta(K), \Theta(N\overline{H})\}$	$\max\{\Theta(K), \Theta(\frac{N\overline{H}}{\log K})\}$	$\max\{\Theta(K), \Theta(\frac{K\log\frac{N}{K}}{\log K})\}$

Connections With Model-Based CS: In [23] it is assumed that the data vector's support belongs to some set $\mathcal{M}$ containing $|\mathcal{M}|$ supports of cardinality K. It is shown that with $M = c_2(K + \log|\mathcal{M}|)$ measurements, an ℓ_2 error of $\|\mathbf{x} - \widehat{\mathbf{x}}\|_2 \leq c_1\|\mathbf{n}\|^2$ is possible, where $\widehat{\mathbf{x}}$ is the decoder output and c_1 and c_2 are constants. It is interesting to note that the scaling $\Theta(K + \log|\mathcal{M}|)$ matches the scaling given in Theorem 1 when we substitute $\mathcal{M} = T_1(\epsilon)$. However, Theorem 1 does not follow from this result, since the performance metrics differ and our setup allows for $|\mathcal{I}|$ to vary with different realizations of $\mathcal{I}$.

III. PROOFS OF INFORMATION-THEORETIC RESULTS

The proof of Theorem 1, given in Section III-A, uses a joint typicality decoder which is suboptimal but simpler to analyze than the optimal decoder. The proof of Theorem 2, given in Section III-B, uses an analogy to a MISO communication channel. Before presenting the proofs, we give bounds on the term $N\overline{H}$, which appears in the upper bound on the typical set size in (12).

Proposition 2:

(i) If $\frac{\overline{K}}{N}$ is bounded away from 1 then $N\overline{H} = O\left(\overline{K}\log\frac{N}{\overline{K}}\right)$.

(ii) If $G = O(1)$ and $\min\{N_n p_n', N_n(1-p_n')\} \to \infty$ for all $n \in \{1,\ldots,G\}$, then $N\overline{H} = \omega(\log N)$.

Proof: For the first part, we apply Jensen's inequality to obtain

$$N\overline{H} \leq NH_2\left(\frac{\overline{K}}{N}\right) = \overline{K}\log\frac{N}{\overline{K}} + (N-\overline{K})\log\frac{N}{N-\overline{K}}. \quad (26)$$

It is straightforward to show that this behaves as $O\left(\overline{K}\log\frac{N}{\overline{K}}\right)$ by treating the linear and sublinear regimes separately and using the condition that $\frac{\overline{K}}{N}$ is bounded away from 1 in the linear regime. For the second part, we write $N\overline{H} = \sum_{n=1}^{G} N_n H_2(p_n')$ and focus on the group with the largest N_n. Since there are only $G = O(1)$ groups and N coefficients in total, there must be at least one group with $\Theta(N)$ coefficients. Assuming without loss of generality that group 1 has the largest number of coefficients, we write

$$N\overline{H} \geq N_1 H_2(p_1') \quad (27)$$

$$= N_1 p_1' \log\frac{N_1}{N_1 p_1'} + N_1(1-p_1')\log\frac{N_1}{N_1(1-p_1')} \quad (28)$$

where $N_1 = \Theta(N)$, and we have included N_1 in the numerators and denominators inside the logarithms for convenience. We show that this grows faster than $\log N$ by treating the cases $N_1 p_1' = o(N)$ and $N_1 p_1' = \Theta(N)$ separately. Starting with the former, the right hand side of (27) is lower bounded by $V_{\text{sublinear}} = N_1 p_1' \log\frac{N_1}{N_1 p_1'}$. If $N_1 p_1' = O(\log N)$ we obtain $V_{\text{sublinear}} = N_1 p_1' \Theta(\log N)$, which grows faster than $\log N$ since $N_1 p_1' \to \infty$. If $N_1 p_1' = \omega(\log N)$ we obtain $V_{\text{sublinear}} = \omega(\log N)\omega\left(\log\frac{N}{o(N)}\right) = \omega(\log N)$. Hence $N\overline{H} = \omega(\log N)$ in the case that $Np_1' = o(N)$. When $N_1 p_1' = \Theta(N)$, a similar argument holds using $V_{\text{linear}} = N_1(1-p_1')\log\frac{N_1}{N_1(1-p_1')}$ and $N_n^{(N)}(1-p_n'^{(N)}) \to \infty$, and treating the cases $N_1(1-p_1') = O(\log N)$ and $N_1(1-p_1') = \omega(\log N)$ separately. ■

A. Proof of Theorem 1

Along with Definition 1, we will make use of the following definition of joint typicality from [13].

Definition 2: [13] (Joint Typicality) The observation vector $\mathbf{y}$ and a set of indices $\mathcal{J} \subseteq \{1,2,\ldots,N\}$ are δ-jointly typical if $\text{rank}(\mathbf{A}_{\mathcal{J}}) = |\mathcal{J}|$ and

$$\left|\frac{1}{M}\left\|\Pi_{\mathbf{A}_{\mathcal{J}}}^{\perp}\mathbf{y}\right\|^2 - \frac{M-|\mathcal{J}|}{M}\sigma^2\right| < \delta \quad (29)$$

where $\Pi_{\mathbf{A}_{\mathcal{J}}}^{\perp} = \mathbf{I} - \mathbf{A}_{\mathcal{J}}(\mathbf{A}_{\mathcal{J}}^T\mathbf{A}_{\mathcal{J}})^{-1}\mathbf{A}_{\mathcal{J}}^T$ is an orthogonal projection matrix. The set of all sequences which are δ-jointly typical with $\mathbf{y}$ is denoted by $T_2(\delta,\mathbf{y})$.

Using the notions of typicality in Definitions 1 and 2, the decoder estimates $\mathcal{I}$ to be a support set in the intersection of $T_1(\epsilon)$ and $T_2(\delta,\mathbf{y})$. If no such support exists, an error is declared. If multiple such supports exist, the decoder chooses the one with the smallest cardinality, with ties broken arbitrarily.

We define the following events:

- E_1: The true support is not jointly typical with $\mathbf{y}$ ($\mathcal{I} \notin T_2(\delta,\mathbf{y})$)
- E_2: Another typical support is jointly typical with $\mathbf{y}$ and has a cardinality which does not exceed the true support ($\mathcal{J} \in T_2(\delta,\mathbf{y})$ for some $\mathcal{J} \in T_1(\epsilon)$, $|\mathcal{J}| \leq |\mathcal{I}|$)

The overall error probability is then bounded as

$$p_e(\mathcal{I}) \leq \Pr(E_1) + \Pr(E_2). \quad (30)$$

To perform the analysis, we make use of the following lemma from [13]. We note that the bounds given in the lemma hold for the *given* support set $\mathcal{I}$, with the associated probabilities denoted by $\Pr(\cdot|\mathcal{I})$.

Lemma 1: [13] If $\mathcal{I} \in T_1(\epsilon)$ is the true support set and $\mathcal{J} \in T_1(\epsilon)$ is another support set, then

$$\Pr\left(\mathcal{I} \notin T_2(\delta,\mathbf{y})|\mathcal{I}\right) \leq 2\exp\left(-\frac{\delta^2}{4\sigma^4}\frac{M^2}{M-K+\frac{2\delta}{\sigma^2}M}\right) \quad (31)$$

$$\Pr\left(\mathcal{J} \in T_2(\delta,\mathbf{y})|\mathcal{I}\right) \leq \exp\left(-\frac{M-|\mathcal{J}|}{4}\left(\frac{\sum_{k\in\mathcal{I}\setminus\mathcal{J}}|x_k|^2-\delta'}{\sum_{k\in\mathcal{I}\setminus\mathcal{J}}|x_k|^2+\sigma^2}\right)^2\right) \quad (32)$$

where $K = |\mathcal{I}|$ and $\delta' = \delta\frac{M}{M-|\mathcal{J}|}$.

In particular, the following corollary will be useful.

Corollary 3: [13] Choosing $\delta' = \zeta\mu_{\min}^2$ for some $0 < \zeta < 1$, the bounds in Lemma 1 imply

$$\Pr\left(\mathcal{I} \notin T_2(\delta, \mathbf{y})|\mathcal{I}\right) \leq 2\exp\left(-\frac{\zeta^2}{\sigma^4\left(1+\frac{2\zeta\mu_{\min}^2}{\sigma^2}\right)}\frac{(M-K)\mu_{\min}^4}{4}\right) \quad (33)$$

$$\Pr(\mathcal{J} \in T_2(\delta, \mathbf{y})|\mathcal{I}) \leq \exp\left(-\frac{M-|\mathcal{J}|}{4}\left(\frac{\mu_{\min}^2(L-\zeta)}{L\mu_{\min}^2+\sigma^2}\right)^2\right) \quad (34)$$

where $L = |\mathcal{I} \setminus \mathcal{J}|$.

Since we are focusing on the case that $\mathcal{I}$ and $\mathcal{J}$ are both elements of $T_1(\epsilon)$, we can further bound (33), (34) using the cardinality bound in (11), yielding

$$\Pr(\mathcal{I} \notin T_2(\delta, \mathbf{y})|\mathcal{I}) \leq 2\exp\left(-\frac{\zeta^2}{\sigma^4\left(1+\frac{2\zeta\mu_{\min}^2}{\sigma^2}\right)}\frac{\left(M-\overline{K}(1+\epsilon)\right)\mu_{\min}^4}{4}\right) \quad (35)$$

$$\Pr(\mathcal{J} \in T_2(\delta, \mathbf{y})|\mathcal{I}) \leq \exp\left(-\frac{M-\overline{K}(1+\epsilon)}{4}\left(\frac{\mu_{\min}^2(L-\zeta)}{L\mu_{\min}^2+\sigma^2}\right)^2\right). \quad (36)$$

We begin by bounding $\Pr(E_1)$. Using the bound on M in (14) and the assumption that $\overline{\epsilon} > \epsilon$, we have $\frac{M-\overline{K}(1+\epsilon)}{4} > N\overline{H}$. Since $N\overline{H}\mu_{\min}^4$ grows faster than $\log N$ by assumption, it follows that $\frac{(M-\overline{K}(1+\epsilon))\mu_{\min}^4}{4} = \omega(\log N)$. Substituting this into (35), it follows that $\Pr(E_1)$ decays to zero for large N.

The remainder of the proof shows that $\Pr(E_2)$ also decays to zero for large N, and thus the overall error probability $p_e(\mathcal{I})$ tends to zero. We define $c(L, \widetilde{L})$ as the number of supports $\mathcal{J} \in T_1(\epsilon)$ with $|\mathcal{I} \setminus \mathcal{J}| = L$, $|\mathcal{J} \setminus \mathcal{I}| = \widetilde{L}$ and $\mathrm{rank}(\mathbf{A}_{\mathcal{J}}) = |\mathcal{J}|$. A trivial upper bound gives $c(L, \widetilde{L}) \leq |T_1(\epsilon)|$, and hence (12) yields

$$c(L, \widetilde{L}) \leq \exp\left(N\overline{H}(1+\epsilon)\right). \quad (37)$$

Furthermore, a simple counting argument gives

$$c(L, \widetilde{L}) \leq \binom{K}{L}\binom{N-K}{\widetilde{L}} \quad (38)$$

which, using $\binom{a}{b} \leq a^b = \exp(b\log a)$, implies that

$$c(L, \widetilde{L}) \leq \exp\left(L\log K + \widetilde{L}\log(N-K)\right). \quad (39)$$

Since the event E_2 requires $|\mathcal{J}| \leq |\mathcal{I}|$, we can limit to our attention to the case that $|\mathcal{J} \setminus \mathcal{I}| \leq |\mathcal{I} \setminus \mathcal{J}|$, or equivalently $\widetilde{L} \leq L$. Hence, from (39), we obtain

$$c(L, \widetilde{L}) \leq \exp\left(L\log\left(\overline{K}(1+\epsilon)\right) + L\log\left(N-\overline{K}(1-\epsilon)\right)\right) \quad (40)$$

where we have bounded K using (11). Combining (37) and (40) with (36) and applying the union bound, we obtain

$$\Pr(E_2) \leq \sum_{L=1}^{K}\sum_{\widetilde{L}=1}^{L}\exp\left(q(L)\right) \quad (41)$$

where

$$q(L) = \min\left\{L\log\left(\overline{K}(1+\epsilon)\right) + L\log\left(N-\overline{K}(1-\epsilon)\right), N\overline{H}(1+\epsilon)\right\} - \frac{M-\overline{K}(1+\epsilon)}{4}\left(\frac{\mu_{\min}^2(L-\zeta)}{L\mu_{\min}^2+\sigma^2}\right)^2. \quad (42)$$

Again bounding K using (11), it follows that

$$\Pr(E_2) \leq \sum_{L=1}^{\overline{K}(1+\epsilon)}\sum_{\widetilde{L}=1}^{L}\exp\left(q(L)\right) \quad (43)$$

$$\leq \left(\overline{K}(1+\epsilon)\right)^2 \max_{L\in\left\{1,\ldots,\overline{K}(1+\epsilon)\right\}}\exp\left(q(L)\right). \quad (44)$$

Thus, a sufficient condition for $\Pr(E_2)$ to decay to zero is that $2\log\overline{K} + q(L)$ tends to $-\infty$ for all $L \in \{1, \ldots, \overline{K}(1+\epsilon)\}$. We show that this condition is satisfied by treating the cases $L\mu_{\min}^2 = O(1)$ and $L\mu_{\min}^2 = \omega(1)$ separately. Starting with the former, we have

$$2\log\overline{K} + q(L) \leq (L+2)\log\left(\overline{K}(1+\epsilon)\right) + L\log\left(N-\overline{K}(1-\epsilon)\right) - \frac{M-\overline{K}(1+\epsilon)}{4}\left(\frac{\mu_{\min}^2(L-\zeta)}{L\mu_{\min}^2+\sigma^2}\right)^2 \quad (45)$$

$$\leq (L+2)\log\left(\overline{K}(1+\epsilon)\right) + L\log\left(N-\overline{K}(1-\epsilon)\right) - N\overline{H}\mu_{\min}^4\left(\frac{(L-\zeta)}{L\mu_{\min}^2+\sigma^2}\right)^2 \quad (46)$$

$$\leq L\left[\Theta(\log N) - \omega(\log N)\Theta(L)\right] \quad (47)$$

$$\to -\infty \quad (48)$$

where (45) follows from $\min\{a, b\} \leq a$, (46) follows from the bound on M in (14) and since $\overline{\epsilon} \geq \epsilon$, and (47) follows from $\log(\overline{K}(1+\epsilon)) + \log(N-\overline{K}(1-\epsilon)) = \Theta(\log N)$, the assumption of the theorem that $N\overline{H}\mu_{\min}^4$ grows faster than $\log N$, and the assumption that $L\mu_{\min}^2 = O(1)$.

In the case that $L\mu_{\min}^2 = \omega(1)$, the term $\frac{\mu_{\min}^2(L-\zeta)}{L\mu_{\min}^2+\sigma^2}$ asymptotically tends to 1, and we have

$$q(L) \leq N\overline{H}(1+\epsilon) - \frac{M-\overline{K}(1+\epsilon)}{4}\left(1+o(1)\right) \quad (49)$$

where we have used $\min\{a, b\} \leq b$. By rearranging, it is straightforward to show that this expression tends to $-\infty$ for any M that satisfies (14) with $\overline{\epsilon} \geq \epsilon$.

B. Proof of Theorem 2

Throughout the proof, we will make use of the fact that the support set cardinalities $K^{(N)} = |\mathcal{I}^{(N)}|$ must satisfy (11), since $\mathcal{I} \in T_1(\epsilon)$ by assumption. We first prove the lower bound of $\frac{\overline{K}(1-\epsilon)}{G}$. With K non-zero coefficients and G groups, there must exist a group with at least $\frac{K}{G}$ non-zero coefficients. Without loss of generality, we assume group 1 is such a group. Suppose that all other groups' coefficients, as well as the realization of the additive noise $\mathbf{n}$, are revealed to the decoder. The decoder is left with the task of identifying which of the $K_1 \geq \frac{K}{G}$ coefficients of group 1 are non-zero. The problem is therefore reduced to finding a sparsity pattern of K_1 coefficients for a vector of dimension $N_1 > K_1$ in the absence of noise, and with no prior

information. This requires $M \geq K_1$ measurements [14], and hence using (11) we obtain that $M \geq \frac{K}{G} \geq \frac{\overline{K}(1-\epsilon)}{G}$ is necessary.

To prove the other term in the lower bound in (15), we consider a MISO channel which emulates support recovery with additional information at the decoder. We note that this proof is a more straightforward extension of [13] than that of the sufficient conditions, but we describe the setup for completeness. We assume that the decoder has access not only to the probability vector $\mathbf{p} = [p_1, \ldots, p_N]^T$, but also to the support set size K and the coefficients corresponding to the true support $\mathbf{x}_{\mathcal{I}}$. Clearly the error probability in this scenario is no higher than that of the original sparsity pattern recovery problem.

The encoder maps the support $\mathcal{I}$ to the codeword $\mathbf{z} = \mathbf{A}_{\mathcal{I}}^T \in \mathbb{R}^{K \times M}$ which is transmitted over a K-input MISO channel in M uses. The channel is specified by $\mathbf{H} = \mathbf{x}_{\mathcal{I}}^T \in \mathbb{R}^{1 \times K}$, so that the received signal is $\mathbf{y} = \mathbf{Hz} + \mathbf{n}$. Based on $\mathbf{y}$, the decoder constructs the support estimate $\widehat{\mathcal{I}}$. For a given $\mathbf{x}$, the sum capacity of the channel is given[2] by [13]

$$C_{\mathrm{MISO}}(\mathbf{x}) = \frac{M}{2} \log\left(1 + \frac{\|\mathbf{x}\|_2^2}{\sigma^2}\right). \tag{50}$$

Necessary conditions on M are obtained by comparing (50) to the amount of information which must be transmitted over the channel. Specifically, if R nats must be transmitted over the channel, then a necessary condition on M is

$$M > \frac{R}{\frac{1}{2}\log\left(1 + \frac{\|\mathbf{x}\|_2^2}{\sigma^2}\right)}. \tag{51}$$

The following proposition shows that the number of nats R required is essentially $N\overline{H}$.

Proposition 3: For sufficiently large N, the number of nats R required to identify the support set $\mathcal{I}$ with vanishing probability of error satisfies

$$R \geq (1 - \overline{\epsilon}) N \overline{H} \tag{52}$$

for any $\overline{\epsilon} > H_2(\epsilon)$.

Proof: We let $K_n(\mathcal{I})$ denote the number of non-zero coefficients in the set $\mathcal{I}$ from group n. We lower bound R by counting the number of supports $\mathcal{J}$ for which $K_n(\mathcal{J}) = K_n(\mathcal{I})$ for all n. Writing K_n as a shorthand for $K_n(\mathcal{I})$, we have

$$\exp(R) \geq \prod_{n=1}^{G} \binom{N_n}{K_n} \tag{53}$$

$$\geq \prod_{n=1}^{G} \frac{1}{N_n + 1} \exp\left(N_n H_2\left(\frac{K_n}{N_n}\right)\right) \tag{54}$$

where (54) follows from $\binom{N}{K} \geq \frac{1}{N+1}\exp\left(N H_2\left(\frac{K}{N}\right)\right)$. Taking the logarithm of (54) gives

$$R \geq \sum_{n=1}^{G} N_n H_2\left(\frac{K_n}{N_n}\right) - \sum_{n=1}^{G} \log(N_n + 1). \tag{55}$$

[2]Compared to [13], an extra factor of 1/2 appears in our expression because we are considering the real case.

From (10), $\frac{K_n}{N_n}$ differs from p_n' by no more than ϵ. From the assumption that $\epsilon \in (0, 0.5)$, it follows that $H_2\left(\frac{K_n}{N_n}\right)$ differs from $H_2(p_n')$ by no more than $H_2(\epsilon)$; this can be seen graphically by noting that the difference is maximized at the endpoints. Continuing, and also applying $N_n \leq N$, we obtain

$$R \geq N\overline{H}\,(1 - H_2(\epsilon)) - G\log(N+1). \tag{56}$$

Finally, applying $N\overline{H} = \omega(\log N)$ (see Proposition 2) and $G = O(1)$ yields

$$R \geq N\overline{H}\,(1 - H_2(\epsilon))\,(1 - o(1)). \tag{57}$$

■

Applying Proposition 3 to (51), the main part of the theorem is proved. To prove the second part of the theorem, we show that $\mu_{\max} = \Theta(\mu_{\min}) = O(1)$ and $\frac{N\overline{H}\mu_{\min}^4}{\log N} \to \infty$ imply $\log\left(1 + \frac{\|\mathbf{x}\|_2^2}{\sigma^2}\right) = \Theta(\log \overline{K})$. It follows from $\|\mathbf{x}\|_2^2 \leq K\mu_{\max}^2$, $K \leq \overline{K}(1+\epsilon)$ and $\mu_{\max} = O(1)$ that $\log\left(1 + \frac{\|\mathbf{x}\|_2^2}{\sigma^2}\right) = O(\log \overline{K})$, so it remains to show that $\log\left(1 + \frac{\|\mathbf{x}\|_2^2}{\sigma^2}\right) = \Omega(\log \overline{K})$. We rewrite $\frac{N\overline{H}\mu_{\min}^4}{\log N} \to \infty$ as $\mu_{\min}^2 = \omega\left(\sqrt{\frac{\log N}{N\overline{H}}}\right)$, and substitute $N\overline{H} = O(\overline{K}\log N)$ from Proposition 2 to obtain $\mu_{\min}^2 = \omega\left(\overline{K}^{-\frac{1}{2}}\right)$, or equivalently $\overline{K}\mu_{\min}^2 = \omega\left(\overline{K}^{\frac{1}{2}}\right)$. Since $\|\mathbf{x}\|_2^2 \geq K\mu_{\min}^2 \geq \overline{K}(1-\epsilon)\mu_{\min}^2$, it follows that $\log\left(1 + \frac{\|\mathbf{x}\|_2^2}{\sigma^2}\right) = \Omega(\log \overline{K})$.

IV. Practical Decoding Techniques

Motivated by practical systems, we present extensions to three common techniques for sparse signal recovery, namely Basis Pursuit, LASSO and Orthogonal Matching Pursuit. In order to exploit the prior information, we introduce weights dependent on the support probabilities into each of these techniques, so that the decoder is more inclined to declare a coefficient to be non-zero when it has a higher probability.

A. Extensions of BP and LASSO

Our extensions of BP and LASSO are respectively given by

$$\min_{\mathbf{x}:\mathbf{y}=\mathbf{A}\mathbf{x}} \sum_{i=1}^{N} (-\log p_i)|x_i| \tag{58}$$

$$\min_{\mathbf{x}} \left(\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + \tau \sum_{i=1}^{N} (-\log p_i)|x_i|\right). \tag{59}$$

We call these *Log-Weighted Basis Pursuit* (LW-BP) and *Log-Weighted LASSO* (LW-LASSO) respectively. As is the case in LASSO, τ is a parameter which controls the tradeoff between sparsity and goodness of fit.

It is well known that BP is the convex relaxation of ℓ_0 minimization in the absence of noise [8]. We similarly motivate LW-BP as the convex relaxation of a problem which maximizes a combination of sparsity and probability in the absence of noise. From (3), the most probable signal consistent with $\mathbf{y}$ is given by

$$\max_{\mathbf{y}=\mathbf{A}\mathbf{x}} \left(\sum_{i \in \mathrm{supp}(\mathbf{x})} \log p_i + \sum_{i \notin \mathrm{supp}(\mathbf{x})} \log(1 - p_i)\right) \tag{60}$$

since maximizing a probability is equivalent to maximizing its logarithm. The objective in (60) contains two summations; the first favors the *inclusion* of higher probability coefficients in the support, while the second favors the *exclusion* of lower probability coefficients. For the purposes of promoting sparsity, and also to allow a convex relaxation of the problem, we consider the problem containing only the former summation, given by

$$\max_{\mathbf{y}=\mathbf{A}\mathbf{x}} \left(\sum_{i \in \mathrm{supp}(\mathbf{x})} \log p_i \right). \tag{61}$$

For example, in the case that each p_i is equal, the objective in (61) is proportional to $\|\mathbf{x}\|_0$. Relaxing $\sum_{i\in\mathrm{supp}(\mathbf{x})} \log p_i$ to the concave function $\sum_{i=1}^{N} (\log p_i)|x_i|$, we obtain LW-BP in (58). The weights of $-\log p_i$ have the desirable properties of being continuous and decreasing, with an infinite penalty when the probability is zero and no penalty when the probability is 1.

Using a similar argument, LW-LASSO can be motivated by starting with the problem of maximizing $\Pr(\mathbf{x}|\mathbf{A},\mathbf{y}) \propto \Pr(\mathbf{y}|\mathbf{A},\mathbf{x})\Pr(\mathbf{x})$ when the noise term $\mathbf{n}$ contains independent Gaussian entries with variance σ^2. In this case, using the definition of the Gaussian probability density function, we have

$$\Pr(\mathbf{y}|\mathbf{A},\mathbf{x}) \propto \exp\left(-\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2 2\sigma^2\right) \tag{62}$$

and hence maximizing $\Pr(\mathbf{y}|\mathbf{A},\mathbf{x})\Pr(\mathbf{x})$ is equivalent to minimizing $\frac{\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2}{2\sigma^2} - \log\Pr(\mathbf{x})$. Assuming that the only knowledge of $\mathbf{x}$ is the support probabilities, a similar argument to the one above gives the equation for LW-LASSO in (59).

B. Extension of Orthogonal Matching Pursuit

The idea of OMP is to repeatedly add the index i which has the highest correlation $|\mathbf{a}_i^T\mathbf{r}|$ with a residual vector $\mathbf{r}$. We introduce an algorithm which we call *Logit-Weighted Orthogonal Matching Pursuit* (LW-OMP), in which instead the highest $\mathbf{a}_i^T\mathbf{r} + v(p_i)$ is added at each step, where $v(p)$ is proportional to the logit function $\log\frac{p}{1-p}$. More precisely, in Appendix A we show that among all additive weight functions,[3] the choice

$$v(p) = \frac{g}{2}\left(2K - 1 + 2\left(\frac{\sigma}{g}\right)^2\right)\log\frac{p}{1-p} \tag{63}$$

approximately minimizes the probability of incorrectly choosing a zero coefficient over a non-zero coefficient in the case that $\mathbf{r} \in \mathbb{R}^M$ is the random noisy projection of a data vector containing K non-zero coefficients of equal amplitude g, and each term of the additive noise $\mathbf{n}$ has variance σ^2.[4] We briefly discuss the resulting terms in the expression. The only part containing p is $\log\frac{p}{1-p}$, which has the desirable properties of being continuous and increasing, with an infinite penalty when p is zero and an infinite reward when p is 1. The constant factor $\frac{g}{2}\left(2K - 1 + 2\left(\frac{\sigma}{g}\right)^2\right)$ increases with σ^2, which can be explained by the fact that if the measurements are more noisy then they are less reliable, so the prior information is relied on more. A similar argument applies for the dependence on K. Finally, the term $\frac{g}{2}$ merely scales the weights to match the amplitude of the non-zero coefficients, making the decision unchanged if, for example, g and σ are both scaled by the same amount.

The expression in (63) motivates the use of the a weight function which varies with the iteration number, with (63) being used on the first iteration, but then substituting

$$K \leftarrow K - (k-1) \tag{64}$$

on the k-th iteration. In the case that K is not known, one could replace the right-hand side of (64) with $\min\{1, \overline{K} - (k-1)\}$ and terminate the algorithm when the norm of the residual falls below a threshold. For simplicity, we focus on the case that K is known.

In the case that the non-zero coefficients do not have equal amplitude but are instead independently drawn according to some distribution X, we can replace each occurrence of g in (63) with a suitable average $\overline{g}$, such as $E[|X|]$ or $\sqrt{E[X^2]}$. While this is not necessarily optimal, it keeps the algorithm simple while still effectively exploiting the prior information, as we will see via simulations in the following subsection. An alternative approach would be to take the constant factors as design parameters. The steps of LW-OMP are summarized in Algorithm 1.

Algorithm 1: Logit Weighted Orthogonal Matching Pursuit

Inputs: $\mathbf{A}$, $\mathbf{y}$, $\mathbf{p}$, K, $\overline{g}$

Outputs: Support set estimate $\widehat{\mathcal{I}} \subseteq \{1,2,\ldots,N\}$, data vector estimate $\widehat{\mathbf{x}} \in \mathbb{R}^{N\times 1}$

Algorithm:

1) Initialize the residual vector $\mathbf{r}_0 = \mathbf{y}$ and support estimate $\widehat{\mathcal{I}}_0 = \emptyset$, and set $k = 1$
2) While $k \le K$:
   a) Set $c_k = \frac{\overline{g}}{2}\left(2(K+1-k) - 1 + 2\left(\frac{\sigma}{\overline{g}}\right)^2\right)$
   b) Set $i_k = \arg\max |\mathbf{a}_i^T\mathbf{r}_{k-1}| + c_k \log\frac{p_i}{1-p_i}$, where the maximum is over $i \in \{1,\ldots,N\} \setminus \widehat{\mathcal{I}}_{k-1}$
   c) Set $\widehat{\mathcal{I}}_k = \widehat{\mathcal{I}}_{k-1} \cup \{i_k\}$ and $\mathbf{A}_k = [\mathbf{A}_{k-1}\,\mathbf{a}_{i_k}]$, where $\mathbf{A}_0$ is the empty matrix
   d) Set $\widehat{\mathbf{x}}_k = \arg\min_{\mathbf{x}} \|\mathbf{y}-\mathbf{A}_k\mathbf{x}\|_2^2$ and $\mathbf{r}_k = \mathbf{y} - \mathbf{A}_k\widehat{\mathbf{x}}_k$
   e) Increment k
3) Return $\widehat{\mathcal{I}} = \widehat{\mathcal{I}}_{k-1}$ and $\widehat{\mathbf{x}} = \widehat{\mathbf{x}}_{k-1}$

C. Simulations

In this section we simulate LW-BP, LW-LASSO and LW-OMP for various M and σ with $N = 240$, $\overline{K} = 16$, and 2000 runs per simulation. The non-zero gains $\{g_i\}_{i=1}^N$ are randomly drawn according to N(0,1). We set $G = 4$ and choose the p'_n such that $N_n p'_n = 4$ for all n. That is, each group has an average of four non-zero coefficients. We simulate four

[3] A justification for the use of an additive (rather than multiplicative) weight function is given in Section IV-D.

[4] The resulting equations will vary when different normalizations are used for the measurement matrix and noise vector, but these can be incorporated into the values of g and σ^2.

TABLE III
GROUP SIZES USED IN THE SIMULATIONS

	N_1	N_2	N_3	N_4	$\overline{H}$
Simulation 1	60	60	60	60	0.2449
Simulation 2	120	80	20	20	0.2226
Simulation 3	204	12	12	12	0.1775
Simulation 4	210	20	5	5	0.1451

different group arrangements each with different values of $\overline{H}$, as given in Table III.

We measure the performance of each technique by comparing the true support $\mathcal{I}$ to $\widehat{\mathcal{I}}$, defined to contain the $K = |\mathcal{I}|$ coefficients of $\widehat{\mathbf{x}}$ with the largest magnitude, where $\widehat{\mathbf{x}}$ is the estimate of $\mathbf{x}$. The performance metric we study is the average proportion of coefficients recovered, given by

$$p_{\text{recovered}} = \frac{1}{S}\sum_{i=1}^{S}\frac{|\mathcal{I}_i \cap \widehat{\mathcal{I}}_i|}{|\mathcal{I}_i|} \tag{65}$$

where S is the number of runs per simulation, $\mathcal{I}_i$ is the true support on the i-th run and $\widehat{\mathcal{I}}_i$ is the support estimate on the i-th run.

Fig. 1(a) shows the average proportion of coefficients recovered using LW-BP for both $\sigma^2 = 0.1$ and $\sigma^2 = 1$. Fig. 1(c) shows the performance of LW-LASSO with $\tau = \frac{1.5}{\overline{w}}$ where $\overline{w} = \frac{1}{N}\sum_{i=1}^{N}(-\log p_i)$ is the average weight. Fig. 1(e) shows the performance of LW-OMP using $\overline{g} = \sqrt{\frac{2}{\pi}}$, which is the average magnitude of an N(0,1) random variable. Simulation 1 represents the case where there is no prior information, since in this case all support probabilities are equal and the techniques reduce to the standard versions on which they were based; the corresponding curves are shown in bold. For all three techniques, the weighted versions result in better performance, particularly when the prior distribution is far from uniform. Furthermore, the improvement is evident using both $\sigma^2 = 0.1$ and $\sigma^2 = 1$.

Comparing the three methods, we see that LW-LASSO tends to give the best performance. In comparison, the performance of LW-BP is nearly identical with $\sigma^2 = 0.1$ but slightly worse with $\sigma^2 = 1$. The performance of LW-OMP is similar to LW-LASSO for high values of M, but worse at low values of M.

The information-theoretic results of Section II give sufficient and necessary conditions on M for perfect support recovery which are affine in $N\overline{H}$. We now investigate whether a similar dependence on $N\overline{H}$ exists for these practical decoding techniques under less strict performance requirements. If the proportion of coefficients recovered for given values of N and $\overline{K}$ depends inversely on an affine function of $N\overline{H}$, then we expect the plots of $p_{\text{recovered}}$ vs. $\frac{M}{c+N\overline{H}}$ to be the same for each of the four group arrangements, for some constant c. Figs. 1(b) and 1(d) plot this relationship for LW-BP and LW-LASSO respectively, both with $c = 5$. Fig. 1(f) shows the relationship for LW-OMP with $c = 10$. The plots align very closely for LW-BP and LW-LASSO in the case that $\sigma^2 = 0.1$, indicating that $N\overline{H}$ does have a significant effect on performance. Specifically, this suggests that for given values of N and $\overline{K}$, the minimum number of measurements required to recover a certain proportion of the coefficients is approximately proportional to $c + N\overline{H}$. In the noisier setting of $\sigma^2 = 1$ the dependence is somewhat weaker. For LW-OMP, the dependence is weaker at both values of σ^2.

Lastly, we examine the effect of mismatch on each technique. We focus on Simulation 3 with $M = 50$ and assume that the true value of $p_1' = \frac{1}{51} \approx 0.0196$ is used in each method, but $p_2' = p_3' = p_4' = \frac{1}{3}$ is replaced by $\widehat{p}$. For example, in LW-BP the weights of $-\log p_n'$ are replaced by $-\log \widehat{p}$ for $n \in \{2,3,4\}$. Fig. 2 plots the performance of each technique as a function of $\widehat{p}$. The left-most points correspond to $\widehat{p} = p_1'$ and therefore give the performance when the prior information is ignored altogether. We see that even with mismatch the performance is improved for the entire range plotted. Furthermore, the performance of each technique remains close to its peak for a wide range of $\widehat{p}$, suggesting a good robustness against mismatch. The peak in each plot is not at $\widehat{p} = \frac{1}{3}$, suggesting that weights are slightly suboptimal. However, the performance at $\widehat{p} = \frac{1}{3}$ is close to the peak for all three techniques, particularly for LW-BP and LW-LASSO.

D. Comparisons to Existing Work

The Modified-CS technique [21] performs a similar ℓ_1 minimization to BP, but with zero penalty on coefficients within a partially known support. This can be viewed as an extreme case of LW-BP, where $p_i = 1$ for coefficients in the partially known support, and all other p_i are equal to some constant in the range (0,1).

In [7], a similar ℓ_1 minimization technique to LW-BP is presented, with the main focus being on the case that there are only two different probabilities and hence two weights. A method is given for optimizing the weights, but it involves searching the space of potential weights and performing a significant amount of computation to decide which to use. This makes LW-BP more suitable when there are a wide range of different probabilities, or in the case that the probabilities are varying with time and a such a search is not feasible (e.g. in online applications).

An iterative reweighted ℓ_1 minimization technique is proposed in [29], in which the estimate of $\mathbf{x}$ on the k-th iteration is given by

$$\widehat{\mathbf{x}}^{(k)} = \min_{\mathbf{y}=\mathbf{A}\mathbf{x}} \sum_i w_i^{(k)}|x_i| \tag{66}$$

for some non-negative weights $\{w_i^{(k)}\}_{i=1}^{N}$. The weights are initialized as $w_i^{(0)} = 1$ and then reweighted as $w_i^{(k)} \leftarrow \frac{1}{\widehat{x}_i^{(k-1)}+\eta}$, where $\widehat{x}_i^{(k)}$ is the i-th entry of $\widehat{\mathbf{x}}^{(k)}$ and η is a positive constant. While (66) bears a strong resemblance to LW-BP, the reason for introducing the weights is different. Instead of exploiting information which is known *a priori*, the weights in (66) serve to reduce the dependence of the objective on the magnitudes of $\mathbf{x}$, hence making the optimization more similar to that of ℓ_0 minimization [29].

In [22], a similar technique to LW-LASSO appears under the name *Weighted Basis Pursuit Denoising* (W-BPDN). Specifically, the objective function is $\|\mathbf{y}-\mathbf{A}\mathbf{x}\|^2 + \tau\sum_{i=1}^{N}\frac{|x_i|}{w_i}$ where $w_i \in (0,1]$ is some measure of the a priori likelihood of each term being non-zero, but not necessarily a probability measure. Its use generally relies on already knowing a good choice of weights, or being able to tune them via a parameter search. On the other hand, in LW-LASSO the weights are a simple function

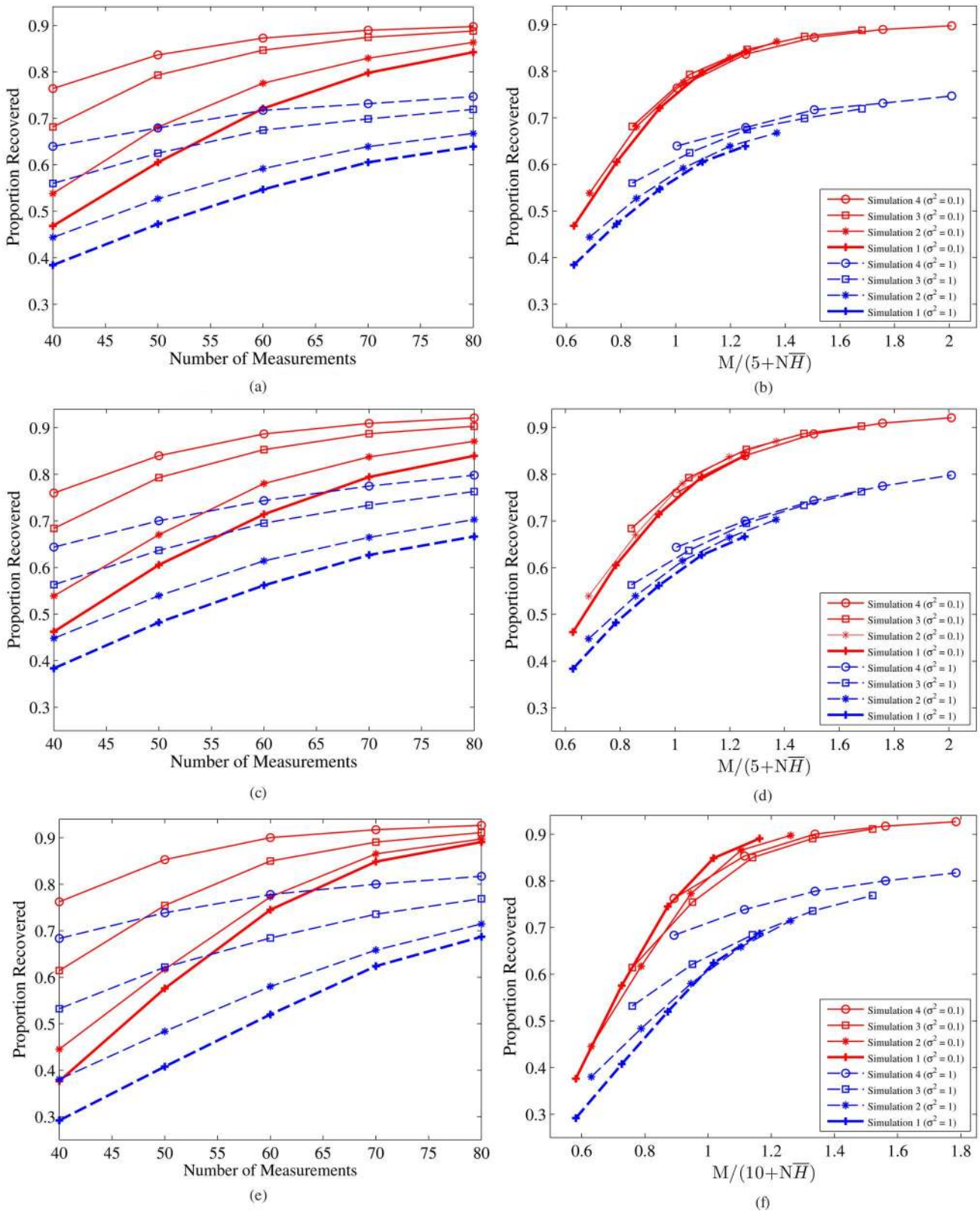

Fig. 1. Algorithm performance without prior information (simulation 1) and with prior information (simulations 2–4). (a) LW-BP; (b) LW-BP (scaled); (c) LW-LASSO; (d) LW-LASSO (scaled); (e) LW-OMP; (f) LW-OMP (scaled). The legends for (a), (c) and (e) are shown in (b), (d) and (f) respectively.

of the support probabilities, making it more suitable for applications where the prior information arises in this form.

Similarly, a technique called *Weighted Matching Pursuit* (W-MP) appears in [22]. As discussed in [22], $|\mathbf{a}_i^T\mathbf{r}|$ can be viewed as a measure of $\Pr(\mathbf{a}_i|\mathbf{r})$ when $\mathbf{r} = \mathbf{a}_i + \mathbf{n}_1$ and $\mathbf{n}_1$ is Gaussian. The authors in [22] state that since $\Pr(\mathbf{r}|\mathbf{a}_i) \propto \Pr(\mathbf{a}_i|\mathbf{r})\Pr(\mathbf{a}_i)$, prior information can be incorporated by instead adding the highest $|\mathbf{a}_i^T\mathbf{r}| \times w_i$ for some weight w_i. However, we argue that since $|\mathbf{a}_i^T\mathbf{r}|$ actually arises as a term in the *logarithm* of $\Pr(\mathbf{a}_i|\mathbf{r})$, *additive* weights are more

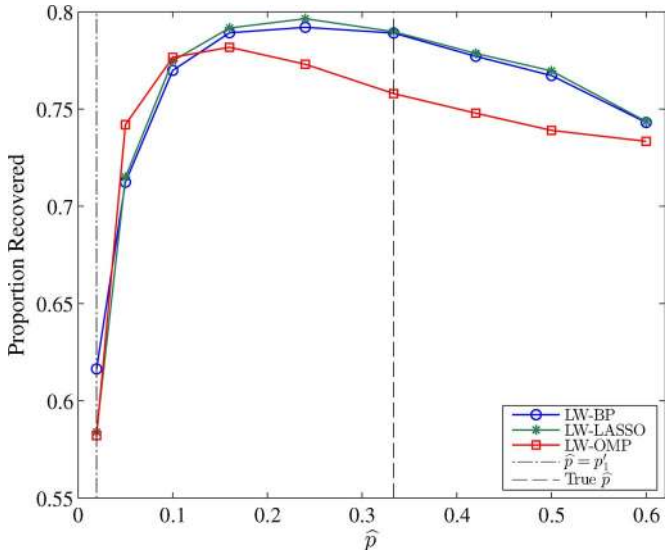

Fig. 2. Proportion of coefficients recovered using a mismatched value of $p_2 = p_3 = p_4$ in Simulation 3 with $\sigma^2 = 0.1$ and $M = 50$. The case $\widehat{p} = p_1'$ indicates performance without prior information.

suitable than multiplicative weights. Furthermore, similarly to LW-LASSO, an advantage of LW-OMP is that the weights are a simple function of the corresponding support probabilities.

V. Conclusion

We have studied the problem of sparse signal recovery with prior information on the sparsity pattern of the data. We have presented information-theoretic limits on the number of measurements required, with the main result being that $\Theta(N\overline{H})$ measurements are sufficient under various assumptions. In many cases, this is a significant improvement over the number of measurements required without prior information.

Extensions of BP, LASSO and OMP have been given, each of which uses weights in order to exploit the prior information. The proposed weights are simple functions of the corresponding support probabilities, making the techniques practical and easy to implement, while still giving good improvements in performance compared to their standard counterparts.

Appendix

A. Derivation of LW-OMP Weights

We consider the simple case in which there are K non-zero coefficients in $\mathbf{x}$, each having a fixed value g. Without loss of generality it can be assumed that $\mathcal{I} = \{1, 2, \ldots, K\}$, and thus the measurement vector is given by

$$\mathbf{y} = g \sum_{i=1}^{K} \mathbf{a}_i + \mathbf{n}. \tag{67}$$

Assuming M is sufficiently large, $\mathbf{a}_i^T \mathbf{a}_i \approx M$ by the law of large numbers. For $i \neq j$, $\mathbf{a}_i^T \mathbf{a}_j = \sum_{k=1}^{M} a_{ki} a_{kj}$ where a_{ki} and a_{kj} are independent random variables with a mean 0 and a variance of 1, meaning their product also has a mean of 0 and a variance of 1. Hence by the central limit theorem, $\mathbf{a}_i^T \mathbf{a}_j \overset{d}{\approx} \mathrm{N}(0, M)$. Similarly, for any i, $\mathbf{a}_i \mathbf{n} \overset{d}{\approx} \mathrm{N}(0, M\sigma^2)$. Therefore

$$\mathbf{a}_i^T \mathbf{y} \overset{d}{\approx} \begin{cases} \mathrm{N}\left(gM, \left(g^2(K-1)+\sigma^2\right) M\right) & i \in \mathcal{I} \\ \mathrm{N}\left(0, (g^2 K + \sigma^2) M\right) & i \notin \mathcal{I} \end{cases}. \tag{68}$$

Consider $i_1 \in \mathcal{I}$ and $i_2 \notin \mathcal{I}$, and let $v_1 = v(p_{i_1})$ and $v_2 = v(p_{i_2})$. Index i_2 is incorrectly favored over index i_1 when $\mathbf{a}_{i_1}^T \mathbf{y} + v_1 < \mathbf{a}_{i_2}^T \mathbf{y} + v_2$. We denote the probability of this event by $p_e(i_1, i_2)$. Using the expression in (68), we obtain

$$p_e(i_1, i_2) \approx Q\left(\frac{gM + v_1 - v_2}{\sqrt{M\left(g^2(2K-1) + 2\sigma^2\right)}}\right) \tag{69}$$

where $Q(x)$ is the probability that an N(0,1) random variable exceeds x. Similarly, we define $p_e(i_2, i_1)$ as the probability of i_1 being incorrectly favored over i_2 conditioned on $i_1 \notin \mathcal{I}$ and $i_2 \in \mathcal{I}$. Our aim is to minimize the probability of exactly one of the two indices being non-zero but the other one being preferred, given by

$$p_e(i_1, i_2) p_{i_1}(1 - p_{i_2}) + p_e(i_2, i_1) p_{i_2}(1 - p_{i_1}) \tag{70}$$

$$\approx Q\left(\frac{gM - \Delta v}{\sqrt{M\left(g^2(2K-1) + 2\sigma^2\right)}}\right) p_{i_1}(1 - p_{i_2}) + Q\left(\frac{gM + \Delta v}{\sqrt{M\left(g^2(2K-1) + 2\sigma^2\right)}}\right) p_{i_2}(1 - p_{i_1}) \tag{71}$$

where $\Delta v = v_2 - v_1$. Setting the partial derivative with respect to Δv to zero and simplifying, we obtain

$$\Delta v = \frac{g}{2}\left(2K - 1 + 2\left(\frac{\sigma}{g}\right)^2\right) \log \frac{p_{i_2}(1 - p_{i_1})}{p_{i_1}(1 - p_{i_2})}. \tag{72}$$

Setting $\Delta p = p_{i_2} - p_{i_1}$ and taking the limit as Δp and Δv tend to 0 yields

$$\frac{dv}{dp} = \frac{g}{2}\left(2K - 1 + 2\left(\frac{\sigma}{g}\right)^2\right)\left(\frac{1}{p} + \frac{1}{1-p}\right). \tag{73}$$

The derivation is concluded by integrating both sides of (73).

References

[1] E. J. Candès and M. B. Waki, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.
[2] W. Bajwa, J. Haupt, A. Sayeed, and R. Nowak, "Compressive wireless sensing," presented at the Int. Conf. Inf. Process. Sensor Netw., Nashville, TN, Apr. 2006.
[3] K. Lee, S. Tak, and J. C. Ye, "A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion," *IEEE Trans. Med. Imag.*, vol. 30, no. 5, pp. 1076–1089, May 2011.
[4] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, Jun. 2010.
[5] M. B. Wakin, J. N. Laska, M. F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. F. Kelly, and R. G. Baraniu, "An architecture for compressive imaging," presented at the IEEE Int. Conf. Image Process., Atlanta, GA, Oct. 2006.
[6] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi, "Improved sparse recovery thresholds with two-step reweighted ℓ_1 minimization," presented at the IEEE Int. Symp. Inf. Theory, Pasadena, CA, Jun. 2010.

[7] M. A. Khajehnejad, W. Xu, A. S. Avestimehr, and B. Hassibi, "Analyzing weighted ℓ_1 minimization for sparse recovery with nonuniform sparse models," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1985–2001, May 2011.
[8] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.
[10] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.
[11] Y. C. Pati, R. Rezaiifaarn, and P. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," presented at the Annu. Asilomar Conf., Pacific Grove, CA, Nov. 1993.
[12] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
[13] M. Akcakaya and V. Tarokh, "Shannon-theoretic limits on noisy compressive sampling," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 492–504, Jan. 2010.
[14] A. K. Fletcher, S. Rangan, and V. K. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.
[15] Y. Jin, Y. Kim, and B. D. Rao, "Support recovery of sparse signals via multiple-access communication techniques," *IEEE Trans. Inf. Theory*, vol. 57, no. 12, pp. 7877–7892, Dec. 2011.
[16] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.
[17] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2967–2979, Jun. 2010.
[18] K. R. Rad, "Nearly sharp sufficient conditions on exact sparsity pattern recovery," *IEEE Trans. Inf. Theory*, vol. 57, pp. 4672–4679, July 2011.
[19] G. Reeves and M. Gastpar, "Approximate sparsity pattern recovery: Information-theoretic lower bounds," [Online]. Available: http://arxiv.org/abs/1002.4458, arXiv:1002.4458v2 [cs.IT].
[20] S. Aeron, V. Saligrama, and M. Zhao, "Information theoretic bounds for compressed sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5111–5130, Oct. 2010.
[21] N. Vaswani and W. Lu, "Modified-CS: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4595–4607, Sep. 2010.
[22] Ò. D. Escoda, L. Granai, and P. Vandergheynst, "On the use of a priori information for sparse signal approximations," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3468–3482, Sep. 2006.
[23] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.
[24] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.
[25] R. Baraniuk, "Optimal tree approximation with wavelets," in *Proc. SPIE Wavelet App. Signal Image Process.*, Denver, CO, Jul. 1999, vol. 3813, pp. 196–207.
[26] N. Vaswani, "Kalman filtered compressed sensing," presented at the IEEE Int. Conf. Image Process., San Diego, CA, Oct. 2008.
[27] Z. Zhang and B. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning," *IEEE J. Sel. Topics in Sig. Proc.*, vol. 5, no. 5, pp. 912–926, Sep. 2011.
[28] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. : Cambridge University Press, 2011.
[29] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Analysis and Apps.*, vol. 14, no. 5, pp. 877–905, Dec. 2008.

Jonathan Scarlett was born in Melbourne, Australia, in 1988. He received the B.Eng. degree (with First Class Hons.) in electrical engineering and the B.Sci. degree in computer science from the University of Melbourne, Australia, both in 2010.

He is currently working towards the Ph.D. degree in the Signal Processing and Communications Group at the Department of Engineering, University of Cambridge, Cambridge, U.K. His research interests include mismatched decoding, multiuser information theory, and compressed sensing.

Mr. Scarlett is a holder of the Poynton Cambridge Australia International Scholarship.

Jamie S. Evans (M'98) was born in Newcastle, Australia, in 1970. He received the B.S. degree in physics and the B.E. degree in computer engineering from the University of Newcastle, U.K., in 1992 and 1993, respectively, and the M.S. and the Ph.D. degrees from the University of Melbourne, Australia, in 1996 and 1998, respectively, both in electrical engineering.

From March 1998 to June 1999, he was a Visiting Researcher in the Department of Electrical Engineering and Computer Science, University of California, Berkeley. He returned to Australia as Lecturer at the University of Sydney, Australia, where he stayed until July 2001. From July 2001 until March 2012, he was with the Department of Electrical and Electronic Engineering, University of Melbourne. He is currently a Professor in the Department of Electrical and Computer Systems Engineering at Monash University, Australia. His research interests are in communications theory, information theory, and statistical signal processing with a focus on wireless communications networks.

Dr. Evans received the University Medal upon graduation from the University of Newcastle. He was also awarded the Chancellor's Prize for Excellence for his Ph.D. thesis from the University of Melbourne.

Subhrakanti Dey (SM'06) was born in India in 1968. He received the B.Tech. and M.Tech. degrees from the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, India, in 1991 and 1993, respectively, and the Ph.D. degree from the Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, Australia, in 1996.

He has been with the Department of Electrical and Electronic Engineering, University of Melbourne, Australia, since February 2000, where he is currently a full Professor. From September 1995 to September 1997 and September 1998 to February 2000, he was a Postdoctoral Research Fellow with the Department of Systems Engineering, Australian National University. From September 1997 to September 1998, he was a post-doctoral Research Associate with the Institute for Systems Research, University of Maryland, College Park. His current research interests include networked control systems, wireless communications and networks, signal processing for sensor networks, and stochastic and adaptive estimation and control.

Prof. Dey currently serves on the Editorial Board of Elsevier *Systems and Control Letters*. He was also an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE TRANSACTIONS ON AUTOMATIC CONTROL.