Information-Theoretic Characterization of Sparse Recovery

Cem Aksoylar
Boston University

Venkatesh Saligrama
Boston University

Abstract

We formulate sparse support recovery as a salient set identification problem and use information-theoretic analyses to characterize the recovery performance and sample complexity. We consider a very general framework where we are not restricted to linear models or specific distributions. We state non-asymptotic bounds on recovery probability and a tight mutual information formula for sample complexity. We evaluate our bounds for applications such as sparse linear regression and explicitly characterize effects of correlation or noisy features on recovery performance. We show improvements upon previous work and identify gaps between the performance of recovery algorithms and fundamental information. This illustrates a trade-off between computational complexity and sample complexity, contrasting the recovery of the support as a discrete object with signal estimation approaches.

1 Introduction

We consider problems where among a set of D variables/features $X = (X_1, \ldots, X_D)$, only K variables (indexed by set S) are directly relevant to the observation/label Y. These types of problems frequently arise in a number of scenarios in high-dimensional analysis, such as compressive sensing [1], feature selection in learning [2] or other high-dimensional problems with an inherent low-dimensional structure. We formulate these problems with the following Markovian property: Given $X_S = \{X_k\}_{k \in S}$, observation Y is independent of $\{X_k\}_{k \notin S}$, i.e.,

$$P(Y|X) = P(Y|X_S). \tag{1}$$

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

Given N sample pairs $(X^N, Y^N) = \{(X^{(1)}, Y^{(1)}), \ldots, (X^{(N)}, Y^{(N)})\}$, our goal is to identify the set of relevant/salient variables S. Our analysis aims to characterize the recovery performance probabilistically and establish necessary & sufficient conditions on N in order to recover S with an arbitrarily small error probability in terms of K, D and model parameters such as the signal-to-noise ratio (SNR).

As an illustrative example, consider the sparse linear regression model given by $Y^N = X^N\beta + W^N$, where S is the support of sparse random vector β and random noise W^N independent of X^N and β. The elements in a row of the matrix X^N correspond to variables $X_1, \ldots, X_D$. Each row is a realization of X and X^N is formed from sampled rows. Markov assumption (1) is satisfied, since each $Y^{(n)}$ depends only on the linear combination of the elements $X_S^{(n)}$. The coefficients of this combination are given by β_S, viewed as a random "nuisance" parameter in the observation model. This perspective also holds for non-linear models, thus unifying many sparse recovery problems.

Information-theoretic approaches with relation to channel coding [3] have been considered in previous work for different application areas, where the salient set S is seen as a message encoded by X^N and is recovered from outputs Y^N. Specifically, the problem of group testing was formulated in a similar framework in Russian literature [4–8] and in [9]. [10] has followed a similar approach to [9] to obtain sample complexity results for general sparse signal processing models. Note that the identification problem formulated here has key differences with channel coding, namely the inability to "code" the variables X^N and different messages/sets overlapping and thus sharing codewords.

Previous work on general models, namely [10], has severe limitations related to the specific analysis tech-

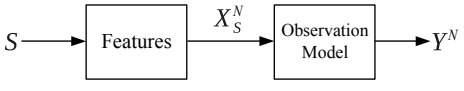


Figure 1: Channel model.

niques used. In this work we overcome several of these limitations, where we (1) consider dependent instead of independent and identically distributed (IID) variables and relate correlation to sample complexity, (2) present a non-asymptotic analysis through probability of recovery bounds instead of solely asymptotic analyses for sample complexity and (3) state more general results for the high-dimensional sparsity regime where sparsity scales with the number of variables.

The bounds we present for the sample complexity are of the form

$$N\, I(X_S; Y) > \log \binom{D}{K}, \tag{2}$$

which can be interpreted as follows: The right side of the inequality is the number of bits required to represent all sets S of size K. On the left side, the mutual information term [3] represents the uncertainty reduction on the output Y when given the input X_S, in bits per sample. This term essentially quantifies the "capacity" of the observation model $P(Y|X_S)$. (2) is then a statement that uncertainty reduction with N samples should exceed the uncertainty of set S.

Furthermore, our analysis also provides us with a sharp exponential upper bound on the probability of error in identifying the salient set. This bound can be computed easily and a closed form expression can be obtained for some applications, such as in the case of linear models. We compute these bounds for the sparse linear model described, where we explicitly characterize the effects of correlation, SNR or noisy variables.

Many models sharing the common structure of sparsity satisfy the Markovian assumption (1). These include sparse linear regression or compressive sensing (CS) [1], probit regression or 1-bit CS [11, 12], group testing [9], sparse logistic regression and multiple regression models [13] with group sparsity property. In addition, variants of these problems can be considered, e.g., with noisy or missing data where variables are not fully observed (see [14]).

1.1 Related Work and Contributions

The problem of sparse recovery and in particular information-theoretic (IT) analysis is extensive. We only describe work closely related to this paper. Much of the IT literature deals with linear models and mean-squared estimation of β in the sparse linear model with sub-Gaussian assumptions on variables X^N. Below we list the contributions of our approach and contrast it with some of the related work in the literature.

Unifying framework through Markovianity: Much of the literature on sparse recovery is specialized with tailored algorithms for different problems. For instance, Lasso for linear regression [15, 16], relaxed integer programs for group testing [17], convex programs for 1-bit quantization [12], projected gradient descent for sparse regression with noisy and missing data [14] and other general forms of penalization. While all of these problems share an underlying sparse structure, it is conceptually unclear from a purely IT perspective, how they come together from an inference point-of-view. Our Markovian viewpoint of (1) unifies these different sparse problems from an inference perspective.

Discrete objects with continuous observations: While [16, 18–23] describe IT bounds for sparsity pattern recovery to recover S, they exclusively focus on the linear sub-Gaussian setting. Furthermore, their approach is *circuitous*. Indeed, they rely on first estimating the sparse vector β, which is then thresholded to obtain S. This not only complicates the analysis and introduces unnecessary assumptions on β but also obfuscates the distinction between signal estimation vs. support discovery. It is well-known that if support is known, signal estimation is easy and least-squares estimates are reliable. At a conceptual level IT tools such as Fano's inequality and capacity theorems are powerful tools for inferring about discrete objects (messages) given continuous observations. Indeed, to exploit IT tools, [18–23] resort to one of the following strategies: (a) Use IT tools only for necessity part by assuming a special case of discrete β and derive sufficiency with some well-known algorithm (Lasso, Basis pursuit etc.); or (b) find a ϵ-cover for β in some metric space (which requires imposing extra assumptions) and reduce β to a discrete object. In contrast our approach lifts these assumptions and focuses on the natural discrete object S. Our result shows that indeed the discrete part, namely, uncertainty support pattern is the dominating factor and not β itself.

Furthermore, prior work relied heavily on the design of sampling matrices with special structures such as Gaussian ensembles and RIP matrices, which is a key difference from the setting we consider herein as for our purpose we do not always have the freedom to design the matrix X. We do not make explicit assumptions about the structure of the sensing matrix, such as the restricted isometry property [24] or incoherence properties [15], or about the distribution of the matrix elements, such as sub-Gaussianity. Also, the existing IT bounds which are largely based on Gaussian ensembles are limited to the linear CS model and hence not suitable for the non-linear models we consider herein.

Information-theoretic tight error bounds: Through our analysis of the ML decoder, we obtain a tight upper bound on the probability of error of support recovery, in addition to necessary and sufficient

conditions on the sample complexity. We compute this upper bound explicitly for popular problems such as sparse linear regression and its variants. We compare the information-theoretic bound to the performance of practical algorithms used to solve the sparse recovery problem, such as Lasso [15, 16] or orthogonal matching pursuit (OMP) variants [25] and illustrate large gaps between their performance and our bounds. The presence of these gaps show that there is still room to improve the performance of practical algorithms for solving support recovery problems.

Bounds for new sparse recovery problems: Our unifying approach also allows the study of problems that are not previously analyzed, or that are not easily analyzed using other approaches. These types of problems may include sparse recovery with novel observation models, or existing models with different distributions of variables or noise. Due to our Markovian formulation, obtaining necessary and sufficient conditions and error bounds only necessitate computation of simple mutual information and error exponent expressions.

Feature selection: The Markovian framework we consider in (1) is also a natural formulation for feature selection. However, in a learning framework it is usually not feasible to assume that the observation model or variable distributions exist and are known exactly, whereas we assume these are exactly known and use them in the computation of our bounds. Therefore, while our results are not applicable for the analysis of practical feature selection problems, they are informative when an idealized Bayesian setting with known distributions are considered. We plan to further explore the unknown distributions scenario from a robust statistics point of view in future work, to present a worst-case analysis for mismatched or estimated distributions.

As mentioned in the introduction, the identification problem was formulated in a channel coding framework in [9] and [4–8], which was extended to general sparse signal processing models with IID variables and latent variable observation model in [10] and [26]. In contrast to [10], we consider the analysis of models with correlated variables, specifically conditionally IID variables X given a latent parameter θ. We also state a non-asymptotic bound on the probability of error, which in turn allows us to identify performance gaps between practical algorithms and our information-theoretic results. In addition, we consider a general scaling regime where $K = O(D)$ for linear models and variants through this bound. We also introduce the noisy data framework and explicitly characterize recovery performance w.r.t. the noise variance.

1.2 Problem setup

Notation. We represent variables with row vectors and samples as different rows to obtain a $N \times D$ matrix, while the observation samples form a column vector. In that context, subscripts are used for column indexing and superscripts with parentheses are used for row indexing. log denotes logarithm to the base 2.

Problem setup. We observe the realizations of N variable-observation pairs (X^N, Y^N) with each sample $(X^{(n)}, Y^{(n)})$, $n = 1, 2, \ldots, N$. Observations Y are given by $P(Y|X_S, \beta_S)$ with latent model parameter $\beta_S \sim P(\beta_S)$ and satisfy the Markovian property (1), where $|S| \leq K$ with known $K \ll D$. Observation parameters β_S correspond to the coefficients on the support of the sparse vector in sparse recovery problems. For simplicity of exposition we consider the case $|S| = K$. The variables $X^{(n)}$ are IID across $n = 1, \ldots, N$. However, the observations $Y^{(n)}$ are independent for different n only when conditioned on β_S. Our goal is to identify the set S from the N samples of variables and the associated observations (X^N, Y^N), with an arbitrarily small average error probability.

We index the different sets of size K as S_ω, so that S_ω is a set of K indices corresponding to the ω-th set of variables. Since there are D variables in total, there are $\binom{D}{K}$ such sets, hence $\omega \in \left\{1, 2, \ldots \binom{D}{K}\right\}$.

Let $\hat{S}(X^N, Y^N)$ denote the estimate of the set S and let $P(E)$ denote the average probability of error, averaged over all sets S of size K, variables X^N and observations Y^N, i.e., $P(E) = \Pr[\hat{S}(X^N, Y^N) \neq S]$.

2 Recovery and Error Bounds

Central to our analysis are the following four assumptions, which we utilize in order to analyze the probability of error in recovering the salient set and to obtain bounds on sample complexity.

(A1) Equiprobable support: Any set $S_\omega \subset \{1, \ldots, D\}$ with K elements is equally likely *a priori* to be the salient set, as such we assume no prior knowledge of the salient set S among $\binom{D}{K}$ such sets.

(A2) Conditional independence: The observation Y is conditionally independent of other variables given variables with indices in S, i.e., $P(Y|X) = P(Y|X_S)$. A simple example is the sparse linear model,

$$Y = \langle X, \beta \rangle + W = \langle X_S, \beta_S \rangle + W,$$

with noise W; with non-linear extensions $Y = f(\langle X_S, \beta_S \rangle + W)$, for a function $f\colon \mathbb{R} \to \mathbb{R}$.

(A3) Conditionally IID variables: The variables

$X_1, \ldots, X_D$ are IID conditioned on a latent parameter θ. For conditionally IID variables, the joint distribution of variables can be written as

$$P(X_1, \ldots, X_D) = \int_\Theta \prod_{k=1}^{D} P(X_k|\theta)P(\theta)\,\mathrm{d}\theta,$$

where $\theta \in \Theta$ is the latent coupling parameter with density $P(\theta)$. (A3) appears restrictive and so we describe a few examples and extensions to build intuition.

Bouquet model [27] arises in sparsity-based face recognition and given by $X_k = \mu + W_k, \quad k = 1, \ldots, D$, with $W_k \sim \mathcal{N}(0, \sigma_W^2)$ IID across k and $\mu \sim \mathcal{N}(0, \sigma_\mu^2)$. It can be seen that two variables X_i and X_j are dependent and correlated with correlation coefficient $\rho = \sigma_\mu^2/(\sigma_\mu^2 + \sigma_W^2)$ but IID conditioned on μ.

Meta parameters: We can account for several possibilities by selectively introducing meta parameters. For instance, we can let $X_k = \alpha_k^\top \mu$ with $\mu \sim \mathcal{N}(0, I_D)$ and IID random vectors α_k. Here $(\{X_k\} \mid \{\alpha_k\}, \mu)$ are independent, not identically distributed with $E(X_k X_j \mid \{\alpha_k\}, \mu) = \alpha_k^\top \alpha_j$. Nevertheless, our results also extend to this setting. Note that X_k's are exchangeable. Indeed, there is a close connection between conditional IID random variables and exchangeable random variables through de Finetti's theorem [28–30].

(A4) Observation model symmetry: For any permutation mapping π, $P(Y|X_S) = P(Y|X_{\pi(S)})$, i.e., the observations are independent of the ordering of variables. This is not a very restrictive assumption since the asymmetry w.r.t. the indices can be incorporated into β_S, as the symmetry is assumed for the observation model averaged over β_S.

With only these four general assumptions, we are able to identify bounds that we state in the next section, for a general class of problems.

2.1 Recovery Conditions and Error Bounds

To derive the upper bound on recovery error and sufficiency bound for the required number of samples, we analyze the error probability of a Maximum Likelihood (ML) decoder [31]. The decoder goes through all $\binom{D}{K}$ sets of size K and chooses the set S_{ω^*} for which observation Y^N is most likely, i.e.,

$$P(Y^N|X^N_{S_{\omega^*}}) > P(Y^N|X^N_{S_\omega}), \quad \forall \omega \neq \omega^*. \tag{3}$$

An error occurs if any set other than the true set S is more likely. This ML decoder is a minimum probability of error decoder assuming uniform prior on the candidate sets of variables. Note that the ML decoder requires the knowledge of the observation model $P(Y|X_S, \beta_S)$ and the prior $P(\beta_S)$. Next, we derive an upper bound on the error probability $P(E)$ of the ML decoder, averaged over all sets, data realizations and observations.

Our methodology for the analysis is as follows: To deal with scenarios where a candidate set S_ω and true set S have overlapping elements (and thus $X^N_{S_\omega}$ and X^N_S share certain columns), we define the error event E_i as the event of mistaking the true set for a set which differs from the true set S in exactly i variables, i.e., there exists some set which differs from the true set in i variables and is more likely to the decoder. Note that $E = \bigcup_{i=1}^K E_i$. Then for each i we use an analysis based on the characterization of error exponents as in [31] to obtain an upper bound on $P(E_i)$, which leads to Theorem 2.1 and a sufficient condition on N. We derive a matching necessity bound on N with an argument based on Fano's inequality [3].

Our first main result is the following theorem, which states a non-asymptotic upper bound on the probability of error of exact support recovery.

Theorem 2.1. *Under the assumptions (A1)-(A4), the probability of error $P(E)$ that a set other than S is selected by the ML decoder is bounded from above by*

$$P(E) \leq \min_{\delta \in [0,1]} \sum_{i=1}^{K} 2^{-N\left(E_o(\delta) - \delta \frac{\log\binom{D-K}{i}\binom{K}{i}}{N}\right)}, \tag{4}$$

where

$$E_o(\delta) = -\frac{1}{N} \log E_{\theta^N}\left[\sum_{Y^N}\sum_{X^N_{\mathcal{S}^2}} P(X^N_{\mathcal{S}^2}|\theta^N) \left(\sum_{X^N_{\mathcal{S}^1}} P(X^N_{\mathcal{S}^1}|\theta^N)P(Y^N|X^N_{\mathcal{S}^1}, X^N_{\mathcal{S}^2})^{\frac{1}{1+\delta}}\right)^{1+\delta}\right],$$

for $0 \leq \delta \leq 1$. $(\mathcal{S}^1, \mathcal{S}^2)$ denotes any disjoint partition of the true set S with cardinalities i and $K-i$, $(X^N_{\mathcal{S}^1}, X^N_{\mathcal{S}^2})$ is the corresponding disjoint partition of the $N \times K$ input X^N_S of size $N \times i$ and $N \times (K-i)$, respectively. θ is the parameter in the cond. IID representation. The bound holds for any (N, K, D).

Remark 2.1. *For fixed and known β_S, observations $Y^{(n)}$ are independent and $E_o(\delta)$ simplifies to*

$$E_o(\delta) = -\log E_\theta\left[\sum_{Y}\sum_{X_{\mathcal{S}^2}} P(X_{\mathcal{S}^2}|\theta) \left(\sum_{X_{\mathcal{S}^1}} P(X_{\mathcal{S}^1}|\theta)P(Y|X_{\mathcal{S}^1}, X_{\mathcal{S}^2})^{\frac{1}{1+\delta}}\right)^{1+\delta}\right].$$

Next, we state our main result for the sample complexity of support recovery. The following theorem provides tight necessary and sufficient conditions on the number of samples N asymptotically for an arbitrarily small average error probability.

Theorem 2.2. *Let $I(X_{\mathcal{S}^1}; Y|X_{\mathcal{S}^2}, \beta_S, \theta)$ be the mutual information between $X_{\mathcal{S}^1}$ and Y conditioned on $X_{\mathcal{S}^2}$, β_S and θ. Under the assumptions (A1)-(A4), a necessary condition on the number of samples N to recover S with an arbitrarily small average probability is given by*

$$N > (1+\epsilon) \max_{i=1,\dots,K} \frac{\log \binom{D-K+i}{i}}{I(X_{\mathcal{S}^1}; Y|X_{\mathcal{S}^2}, \beta_S, \theta)}, \quad (5)$$

and $\epsilon = 0$. If $I(X_{\mathcal{S}^1}; Y|X_{\mathcal{S}^2}, \beta_S, \theta) = \omega(1/\log D)$ for all $i = 1, \dots, K$, (5) is also a sufficient condition, where $\epsilon > 0$ is an arbitrary constant. The necessary condition holds for all scalings $K = O(D)$, while the sufficiency bound holds for any fixed K as $D \to \infty$.

Note that the condition that $I(X_{\mathcal{S}^1}; Y|X_{\mathcal{S}^2}, \beta_S, \theta) = \omega(1/\log D)$ is not restrictive, since typically the mutual information per sample depends on the number of salient variables K and not on the total number of variables D and we consider the regime where K is fixed w.r.t. D for the sufficient condition.

IID variables. For IID variables, the mutual information expression in the denominator is $I(X_{\mathcal{S}^1}; Y|X_{\mathcal{S}^2}, \beta_S)$ and further reduces to $I(X_{\mathcal{S}^1}; Y|X_{\mathcal{S}^2})$ for fixed observation parameters β_S.

Interpretation. Intuitively, the condition in (5) can be explained as follows: For each i, the numerator is the number of bits required to represent all sets S_ω with $K - i$ indices known beforehand. The denominator represents the information given by the output variable Y about the remaining i indices $\mathcal{S}^1$, given the subset $\mathcal{S}^2$ of $K - i$ true indices. Hence, the ratio represents the number of samples needed to control i support errors in $\mathcal{S}^1$ and maximization accounts for all possible support errors.

Partial recovery. As we analyze the error probability separately for $i = 1, \dots, K$ support errors in order to obtain the necessity and sufficiency results, it is trivial to determine conditions for *partial* instead of *exact* support recovery. By changing the maximization from over $i = 1, \dots, K$ to $i = \lfloor \alpha K \rfloor, \dots, K$ in (5), the conditions to recover at least $(1-\alpha)K$ of the K support indices can be determined.

Support pattern recovery dominates support coefficient estimation. In the proof of Theorem 2.2, we show that β_S being unknown with prior $P(\beta_S)$ induces a penalty term in the denominator given by $I(\beta_S; X^N_{\mathcal{S}^1} | X^N_{\mathcal{S}^2}, Y^N, \theta^N)/N$, compared to the case where support coefficients β_S are fixed and known. We show that this term is always dominated by $I(X_{\mathcal{S}^1}; Y|X_{\mathcal{S}^2}, \beta_S, \theta)$ provided a mild condition on the mutual information is satisfied, therefore does not affect the sample complexity asymptotically. This shows that recovering support while knowing the support coefficients is as hard as recovering with unknown coefficients, underlying the importance of recovering the support in sparse recovery problems.

3 Applications

In this section, using the result of Theorem 2.1, we provide explicit non-asymptotic upper bounds for the error probability for sparse linear models that may include correlations or noisy variables. We also state asymptotic sample complexity results using the error bounds and Theorem 2.2. We then compare the information-theoretic error bounds we obtained with the recovery performance of practical algorithms.

3.1 Sparse linear regression

We consider the normalized model [20],

$$Y^N = X^N \beta + W^N, \quad (6)$$

where X^N is the $N \times D$ sensing matrix, β is a K-sparse vector of length D with support S and Y^N is the observation vector of length N. We assume $X^{(n)}$ are jointly Gaussian row vectors and IID across rows n. Each element $X_k^{(n)}$ is zero mean and has variance $1/N$. W^N is the IID observation noise, with $W \sim \mathcal{N}(0, \frac{1}{\mathrm{SNR}})$. The coefficients of the support, β_S, are either fixed and $|\beta_k| = \sigma$, or IID Gaussian with zero mean and variance σ^2.

We consider a generalized model, which may include correlations between sensing columns, such that $E[X_k^{(n)} X_{k'}^{(n)}] = \rho/N$. ρ is then the correlation coefficient between two columns. Note that this model is statistically equivalent to the following one: Let $X_k^{(n)} = \mu^{(n)} + U_k^{(n)}$, where μ is also a Gaussian random variable with zero mean and variance ρ/N. $U_k^{(n)}$ is IID Gaussian, with zero mean and variance $(1-\rho)/N$. We analyze the latter model, where entries are conditionally IID given μ. Correlated columns have been analyzed for Lasso in this context [15, 16]. The strongest results due to [15] require correlations to decay asymptotically to zero as $1/\log(D)$, while [16] is not strictly comparable since their results are for high-SNR limit. In contrast, we will show that fundamentally, up to constant correlation can be tolerated. The following theorem provides an upper bound to the probability of error for exact support recovery.

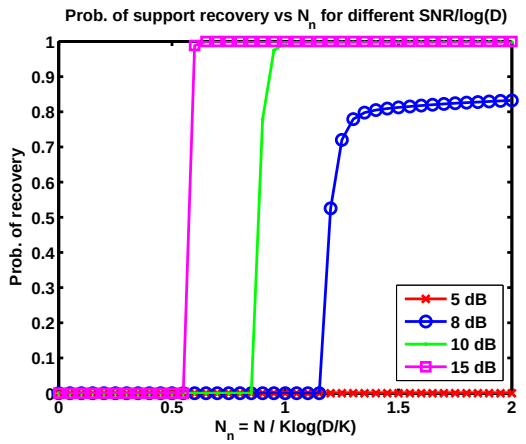


Figure 2: Illustration of SNR cutoff, $K = 32$, $D = 512$.

Theorem 3.1. *$P(E) \leq \sum_{i=1}^{K} 2^{-Nf(\rho)}$, where $f(\rho) = \frac{1}{2}\log\left(1+(1-\rho)\frac{2i\sigma^2 SNR}{N}\right) - \frac{i}{4N}\log 4 - \frac{\log\binom{D-K}{i}\binom{K}{i}}{N}$.*

The first term in $f(\rho)$ is related to the information between X and Y via SNR and ρ, while the second term is related to the uncertainty of β_S and the last term to the uncertainty of S. Note that the error bound given above precisely characterizes the achievable error for any (N, K, D), in contrast to the setting for the sufficient condition in Section 2.1 where K is fixed w.r.t. D. Using this bound, we show the probability of recovery vs. other interesting quantities (see Figs. 4, 5). Also note the relation between $f(\rho)$ and ρ, e.g., for the degenerate case where $\rho = 1$, $f(\rho)$ is negative for any N. This is expected since recovery is not possible in that case, which we prove with the necessity bound.

We now present necessary and sufficient conditions for exact support recovery. We start with a lemma describing the mutual information for this model.

Lemma 3.1.

$$I(X_{\mathcal{S}^1}; Y | X_{\mathcal{S}^2}, \beta_S, \mu) = \frac{1}{2} E\left[\log\left(1+(1-\rho)\frac{\|\beta_{\mathcal{S}^1}\|^2 SNR}{N}\right)\right],$$

where the expectation is w.r.t. $\beta_{\mathcal{S}^1}$.

The mutual information formula along with the bound given by Theorem 3.1 allow us to obtain the following necessary & sufficient condition for exact recovery.

Theorem 3.2. *Consider the correlated setup described above. First, $SNR = \Omega(\log D)$ is a necessary condition for recovery. Furthermore, for this SNR we can recover S with average error probability approaching zero if and only if $N = \Omega\left(\frac{K\log(D/K)}{\log(1+(1-\rho)\sigma^2)}\right)$.*

The necessary condition on SNR is also illustrated in Figure 2, where we plot the probability of error bound given by Theorem 3.1 for different SNR values. Indeed, we show an SNR cutoff regardless of number of measurements as well as tradeoffs beyond the cutoff point. Note that the relation between SNR and N is not explicitly described for Lasso [15, 16].

Both upper and lower bounds hold for the general case $K = O(D)$, since we use the error bound in Theorem 3.1 to obtain the upper bound instead of Theorem 2.2.

Remark 3.1. *It follows that our relatively simple analysis gives us a bound asymptotically identical to the best-known bound $N = \Omega(K\log(D/K))$ [20] with an independent Gaussian sensing matrix. Our results also incorporate correlations to explicitly characterize the effect of correlated columns on sample complexity. We have shown that the number of samples increases by $\frac{1}{\log(1+(1-\rho)C)}$ relative to $\frac{1}{\log(1+C)}$ for the independent model for some constant C.*

3.2 Noisy variables

We also analyze the additive noise model considered in [14, 25], where in the sparse linear regression model (6), a matrix Z^N is observed instead of the sensing matrix X^N, with the relation $Z^N = X^N + V^N$, where $V^{(n)} \sim \mathcal{N}(0, \frac{\nu}{N} I_D)$ IID for $n = 1, \ldots, N$. The rest of the setup is as given in Section 3.1. The model described here exhibits a non-linear relationship between the variables Z^N and the observations Y^N in contrast to Section 3.1. For this problem with noisy observations of variables, we have the following theorem for an upper bound on the probability of error of exact support recovery.

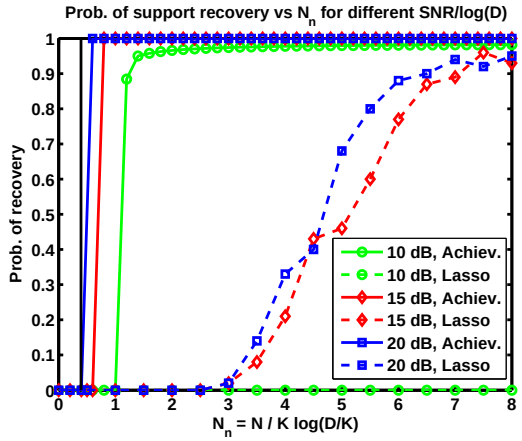
Theorem 3.3. *The error probability of exact support recovery in the noisy data model is given by $P(E) \leq \sum_{i=1}^{K} 2^{-Nf(\rho,\nu)}$, where $f(\rho,\nu) = \frac{1}{2}\log\left(1+\frac{1-\rho}{1+\nu}\frac{2i\sigma^2 SNR}{N\xi}\right) - \frac{i}{4N}\log 4 - \frac{\log\binom{D-K}{i}\binom{K}{i}}{N}$, where $\xi = 1 + \frac{(1-\rho)\nu}{1+\nu}\frac{KSNR\sigma^2}{N}$.*

The error exponent $f(\rho,\nu)$ differs from $f(\rho)$ defined in Section 3.1 mainly by an extra $1+\nu$ term in the denominator in the log term and reduces to $f(\rho)$ for $\nu = 0$. Also note that $\xi \approx 1$ for sufficiently large N.

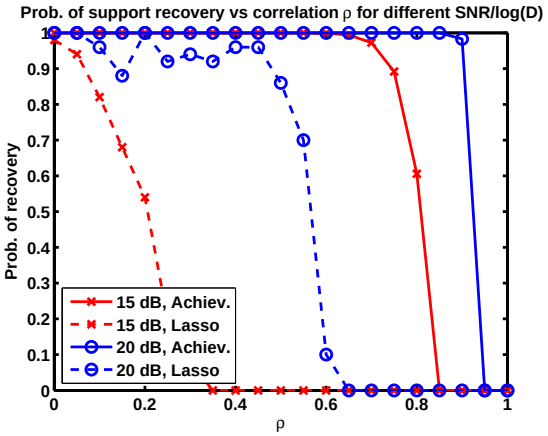
We now state a sufficient condition on the number of measurements with the theorem below, which follows from an analysis of the upper bound on recovery error provided in Theorem 3.3.

Theorem 3.4. *For $SNR = \Omega(\log D)$, a sufficient condition on the number of measurements is $N = \Omega\left(\frac{K\log(D/K)}{\log\left(1+\frac{1-\rho}{1+\nu}\sigma^2\right)}\right)$.*

Remark 3.2. *We observe that the sufficient number of measurements is affected by a factor of $\frac{1}{\log(1+C/(1+\nu))}$ in our results, which greatly improves*



(a) Measurement Cutoffs



(b) Correlation Cutoffs

Figure 3: Comparison of information-theoretic bound vs. Lasso.

upon the bound with a factor of $(1+\nu)^2$ *by [25].*

We also note that while [15, 18] analyze correlated Gaussians and others noisy or missing data [14, 25] separately, our error bounds and asymptotic sample complexity results unify these into a single expression.

3.3 Experiments and comparisons

In this section we compare the information-theoretic probability of error bounds (or equivalently, bounds on the probability of successful recovery) we derived in the above sections with the frequency of successful exact support recovery for two recovery algorithms.

For all experiments and evaluation of bounds, we set $K = 32$ and $D = 512$. The X^N and Y^N are generated according to the normalized model (6), where we choose S uniformly at random and let $\beta_S \in \{-1, 1\}^K$ with uniform probability. $N_n = N/(K \log(D/K))$ is the normalized number of measurements.

We compare our bounds for independent and correlated sensing elements with Lasso [15, 16], as defined in [15]. We set the regularization parameter as $\lambda = 2\sqrt{2 \log D}/\sqrt{\text{SNR}}$ as suggested in [15]. We also investigated different values however we have not observed any significant improvements in performance.

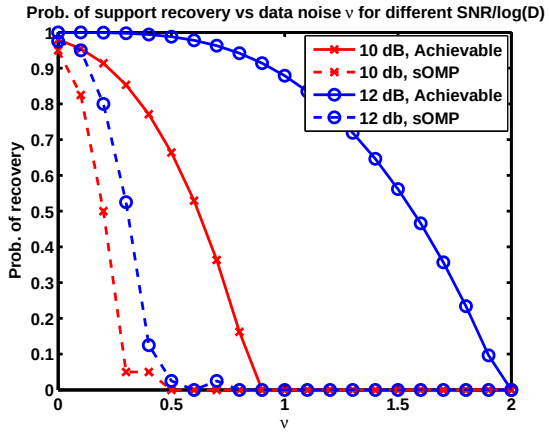
Figure 3(a) plots the recovery bound for IID variables vs. Lasso simulation performance, for different number of measurements N. The probability of recovery for Lasso is computed over 100 iterations. Our IT bound has a much sharper transition, while also being tighter, matching closely our lower bound (vertical line for $\text{SNR}/\log D = 20$ dB) obtained with Theorem 3.1. The theoretical results in [16, 18] are not strictly comparable since they require a significantly large SNR regime. Furthermore, the performance gap approaches infinity as we let K approach D. Thus Lasso works strictly in sublinear regime.

Figure 3(b) shows our probability of error bound vs. Lasso performance for different values of the correlation coefficient ρ, where $N_n = 10$. The probability of recovery for Lasso is computed over 20 iterations. This plot demonstrates clearly that while our bounds show tolerance to correlation up to a constant approaching 1 (as seen from the sample complexity bound in Theorem 3.2), Lasso can tolerate at most $\rho = 0.5$ correlation for exact recovery in this scenario, with very high SNR and N. Note that strongest results due to [15] require correlations to decay asymptotically to zero as $1/\log(D)$.

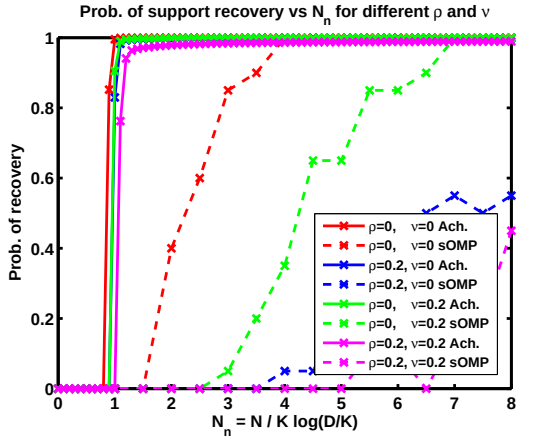
For the second set of experiments, we compare with a variant of the orthogonal matching pursuit (OMP) algorithm called support-OMP. This algorithm is proposed by [25] and shown to have good performance with theoretical guarantees for problems with noisy or missing observations of the sensing matrix as we consider in Section 3.2.

Figure 4(a) shows the performance of support-OMP vs. information-theoretic bound, for noisy variables with different noise variances ν. For support-OMP, the recovery probability is computed over 40 iterations. It can be seen that support-OMP performs reasonably well for noisy variables but fails in high variance noise, whereas our information-theoretic bound shows that recovery in much higher noise levels are achievable, especially with higher SNR.

A similar conclusion can be reached from Figure 4(b), where we plot recovery performance for both correlated and noisy variables. For support-OMP, the recovery probability is computed over 20 iterations. The gap is more pronounced for correlated variables compared to noisy variables, which shows support-OMP is highly affected by correlation and by variable noise to



(a) Noise Variance Cutoffs



(b) Effect of Both Noisy Data & Correlation

Figure 4: Comparison of information-theoretic bound vs. support-OMP.

a lesser degree.

4 Discussion

We have presented a framework for analyzing sparse recovery problems from an inference perspective by introducing a Markovian assumption. This framework unifies linear and non-linear observation models, dependent and non-Gaussian measurements matrices, and noisy data. This framework leads to a tight, exponential upper bound on the support recovery error probability and an explicit universal mutual information formula for computing the sample complexity of sparse recovery problems. The central theme here is "inference of a discrete object (sparse support pattern) in a continuous world of observations." Our approach is not algorithmic and therefore must be used in conjunction with tractable algorithms. Nevertheless, it is useful for identifying gaps between existing algorithms and fundamental information. Fundamentally, we identify a sample complexity and computational complexity trade-off: Treating the support pattern as a discrete object optimizes sample complexity, while approaches that estimate the sparse vector in a continuous space optimize computational complexity.

Although we consider sparse linear regression and its variants as applications in this paper, there are many other sparse recovery applications for which the framework we consider is applicable and the error bound or the sample complexity bounds we have described are explicitly computable through the formulas in Theorems 2.1 and 2.2. Some examples we have not included due to space considerations are group testing, quantized compressive sensing, multiple regression models or models with missing observations.

As we have shown our approach is also useful in understanding fundamental tradeoffs between different design parameters such as SNR, correlations, measurements matrices and noisy features. For instance, in the linear Gaussian setting we have shown that we could information theoretically tolerate up to constant correlation across different variables while existing results require vanishing correlation. The linear setting has also identified large sample complexity gaps between Lasso, support-OMP and information theoretic bounds. Specifically, these gaps get larger as correlation and variable noise increases.

Acknowledgements

This work was supported by NSF Grant 0932114 and NSF Grant CCF-1320547.

References

[1] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289 –1306, April 2006.

[2] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 1. Springer New York, 2006.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York: John Wiley and Sons, Inc., 1991.

[4] M. B. Malyutov and P. S. Mateev. Screening design for non-symmetric response function. *Mat. Zemetki*, 27:109–127, 1980.

[5] M. B. Malyutov. On planning of screening experiments. In *Proceedings of 1975 IEEE-USSR Workshop on Inf. Theory*, 1976.

[6] M. B. Malyutov. The separating property of random matrices. *Mat. Zametki*, 23, 1978.

[7] M. B. Malyutov. Maximal rates of screening designs. *Probability and its Applic.*, 24, 1979.

[8] A. Dyachkov. Lectures on designing screening experiments. *Moscow State Univ.*, 2003.

[9] G. Atia and V. Saligrama. Boolean compressed sensing and noisy group testing. *IEEE Trans. Inf. Theory*, 58(3), March 2012.

[10] C. Aksoylar, G. Atia, and V. Saligrama. Sparse signal processing with linear and non-linear observations: A unified shannon theoretic approach. In *Information Theory Workshop (ITW), 2013 IEEE*, Seville, Spain, 2013.

[11] P. T. Boufounos and R. G. Baraniuk. 1-bit compressive sensing. In *Proc. of Conf. on Information Sciences and Systems (CISS)*, pages 16–21, March 2008.

[12] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Inf. Theory*, 59(1):482–494, Jan 2013.

[13] S. Negahban and M. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[14] P.-L. Loh and M. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *arXiv pre-print, arXiv:1109.3714*, September 2011.

[15] E. J. Candès and Y. Plan. Near-ideal model selection by ℓ_1 minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009.

[16] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programs. In *Allerton Conf. on Communication, Control and Computing*, Monticello, IL, 2006.

[17] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri. Non-adaptive group testing: Explicit bounds and novel algorithms. In *Proc. of the Int. Symp. on Information Theory (ISIT)*, pages 1837–1841, July 2012.

[18] M. J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Inf. Theory*, 55(12):5728–5741, 2009.

[19] A. K. Fletcher, S. Rangan, and V. K. Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inf. Theory*, 55(12):5758–5772, 2009.

[20] S. Aeron, M. Zhao, and V. Saligrama. Information theoretic bounds for compressed sensing. *IEEE Trans. Inf. Theory*, 56(10):5111–5130, Oct. 2010.

[21] M. Akcakaya and V. Tarokh. Shannon-theoretic limits on noisy compressive sampling. *IEEE Trans. Inf. Theory*, 56(1):492–504, Jan. 2010.

[22] Y. Wu and S. Verdu. Optimal phase transitions in compressed sensing. *IEEE Trans. Inf. Theory*, 58(10):6241–6263, Oct.

[23] G. Reeves and M. Gastpar. The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Trans. Inf. Theory*, May 2012.

[24] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

[25] Y. Chen and C. Caramanis. Noisy and missing data regression: Distribution-oblivious support recovery. In *International Conference on Machine Learning*, 2013.

[26] C. Aksoylar, G. Atia, and V. Saligrama. Sparse Signal Processing with Linear and Non-Linear Observations: A Unified Shannon Theoretic Approach. *arXiv pre-print, arXiv:1304.0682*, April 2013.

[27] J. Wright and Y. Ma. Dense error correction via ℓ_1-minimization. *IEEE Trans. Inf. Theory*, 56(7):3540–3560, 2010.

[28] P. Diaconis and D. Freedman. Finite exchangeable sequences. *The Annals of Probability*, pages 745–764, 1980.

[29] D. Aldous. Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII1983*, pages 1–198, 1985.

[30] S. L. Lauritzen, O. E. Barndorff-Nielsen, A. P. Dawid, P. Diaconis, and S. Johansen. Extreme point models in statistics [with discussion and reply]. *Scandinavian Journal of Statistics*, pages 65–91, 1984.

[31] R. G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., New York, NY, USA, 1968.