Sparse Signal Processing with Linear and Non-Linear Observations: A Unified Shannon Theoretic Approach
Cem Aksoylar, George Atia and Venkatesh Saligrama*

Abstract
In this work we derive fundamental limits for many linear and non-linear sparse signal processing models including linear and non-linear sparse regression, group testing, multivariate regression and problems with missing features. In general, sparse signal processing problems can be characterized in terms of the following Markovian property. We are given a set of $N$ variables $X_1, X_2, \ldots, X_N$, and there is an unknown subset of variables $S \subset \{1, 2, \ldots, N\}$ that are relevant for predicting outcomes/outputs $Y$. More specifically, when $Y$ is conditioned on $\{X_n\}_{n \in S}$ it is conditionally independent of the other variables, $\{X_n\}_{n \notin S}$. Our goal is to identify the set $S$ from samples of the variables $X$ and the associated outcomes $Y$. We characterize this problem as a version of the noisy channel coding problem. Using asymptotic information theoretic analyses, we establish mutual information formulas that provide sufficient and necessary conditions on the number of samples required to successfully recover the salient variables. These mutual information expressions unify conditions for both linear and non-linear observations. We then compute sample complexity bounds for the aforementioned models, based on the mutual information expressions in order to demonstrate the applicability and flexibility of our results in general sparse signal processing models.

1 Introduction

Recent advances in sensing and storage systems have led to the proliferation of high-dimensional data such as images, video or genomic data, which cannot be processed efficiently using conventional signal processing methods due to their dimensionality. However, high-dimensional data often exhibit an inherent low-dimensional structure, so they can often be represented “sparsely” in some basis or domain. The discovery of an underlying sparse structure is important in order to compress the acquired data or to develop more robust and efficient processing algorithms.

In this paper, we are concerned with the asymptotic analysis of the sample complexity in problems where we aim to identify a set of salient variables responsible for producing an outcome. In particular, we assume that among a set of $N$ variables/features $X = (X_1, \ldots, X_N)$, only $K$ variables (indexed by set $S$) are directly relevant to the outcome $Y$. We formulate this with the assumption that given $X_S = \{X_n\}_{n \in S}$, the outcome $Y$ is independent of the other variables $\{X_n\}_{n \notin S}$, i.e.,

$$P(Y|X_S, S) = P(Y|X_S).$$

Abstractly, we consider the following generative model: $X$ is generated from distribution $Q(X)$ and the set of salient variables $S$ is generated from a distribution over sets of size $K$. Then an observation $Y$ is generated using the conditional distribution $P(Y|X_S)$ conditioned on $X$ and $S$, as in [1].

We assume we are given $T$ sample pairs $(X, Y)$ and the problem is to identify the set of salient variables, $S$, from these $T$ samples given the knowledge of the observation model $P(Y|X_S)$. Our analysis aims to establish sufficient conditions on $T$ in order to recover the set $S$ with an arbitrarily small error probability in terms of $K$, $N$, the observation model and other model parameters such as the signal-to-noise ratio.

---

*This work was partially supported by NSF Grant 0932114 and NSF Grant CCF-1320547. The material in this paper was presented in part at the 2013 IEEE Information Theory Workshop, Seville, Spain, September 2013.
this paper, we limit our analysis to the setting with independent and identically distributed (IID) variables for simplicity. It turns out that our methods can be extended to the dependent case at the cost of additional terms in our derived formulas that compensate for dependencies between the variables. Some results derived for the former setting were presented in [1] and more recently in [2].

\[ S \xrightarrow{\text{Code}} X_S^T \xrightarrow{\text{Channel}} Y^T \]

Figure 1: Channel model.

The analysis of the sample complexity is performed by posing this identification problem as an equivalent channel coding problem, as illustrated in Figure 1. The salient set \( S \) corresponds to the message transmitted through a channel. The set \( S \) is encoded by \( X^T \) of length \( T \), which is the collection of codewords \( X_n^T \) for \( n \in S \), from a codebook \( X^T \). The coded message \( X_S^T \) is transmitted through a channel \( P(Y|X_S,S) \) with output \( Y^T \). As in channel coding, our aim is to identify which message \( S \) was transmitted given the channel output \( Y^T \) and the codebook \( X^T \).

The sufficiency and necessity results we present in this paper are analogous to the channel coding theorem for memoryless channels [3]. Before we present exact statements of our results, it is useful to mention that these results are roughly of the form

\[ TI(X_S;Y|S) > \log \left( \frac{N}{K} \right), \tag{2} \]

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 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 \( P(Y|X_S,S) \). Then, the total uncertainty reduction through the \( T \) samples should exceed the uncertainty of possible salient sets \( S \), in order to reliably recover the salient set.

Sparse signal processing models analyzed in this paper have wide applicability. Below we list some examples of problems which can be formulated in the described framework.

**Sparse linear regression** [4] is the problem of reconstructing a sparse signal from underdetermined linear systems. It is assumed that the output vector \( Y^T \) can be obtained from a \( K \)-sparse vector \( \beta \) through some linear transformation with matrix \( X^T \), i.e., in the noisy case with noise \( W^T \),

\[ Y^T = X^T \beta + W^T. \tag{3} \]

Non-linear versions of the regression problem are also investigated, where the channel model also includes a quantization of the output. The sparse linear regression model with an example is illustrated in Figure 2.

Note that in our analysis the columns of the matrix \( X^T \) correspond to the variables \( X \) and the support of the sparse vector corresponds to the set \( S \). It is then easy to see that the Markovianity property [1] holds. In contrast to typical regression setup, the focus here is on the recovery of the support \( S \), and not the sparse vector \( \beta \). Hence, the non-zero coefficients \( \beta_S \) are absorbed into our observation model \( P(Y|X_S,S) \) as we elaborate on Section 3.

**Models with missing features** [5]: Our methods are also used to establish sample complexity bounds for sparse signal processing problems with missing features. The problem here is that some of the variables for some of the measurements \( Y^T \), could be missing. Specifically, we observe a \( T \times N \) matrix \( Z^T \) instead of \( X^T \), with the relation

\[ Z_i^{(t)} = \begin{cases} X_i^{(t)}, & \text{w.p. } 1 - \rho \\ m_i, & \text{w.p. } \rho \end{cases}, \quad \forall i, t \]

i.e., we observe a version of the feature matrix which may have missing entries (denoted by \( m_i \)) with probability \( \rho \), independently for each entry. Note that \( m \) can take any value as long as there is no ambiguity whether the realization is missing or not, e.g., \( m = 0 \) would be valid for continuous variables where the variable taking value 0 has zero probability. Note that if a problem satisfies assumption [1] with variables \( X \), the same
Figure 2: Sparse linear regression example and its mapping to the channel model.

Problem with missing features also satisfies the assumption with variables $Z$. Interestingly, our analysis shows that the sample complexity, $T_{\text{miss}}$ for problems with missing features is related to the sample complexity, $T$, of the fully observed case with no missing features by the simple inequality:

$$T_{\text{miss}} \geq \frac{T}{1 - \rho}.$$

**Group testing** [6] is a form of sensing with Boolean arithmetic, where the goal is to identify a set of defective items among a larger set of items. As an example, group testing has been used for medical screening to identify a set of individuals who have a certain disease from a large population while reducing the total number of tests. The idea is to pool blood samples from subsets of people and to test them simultaneously rather than conducting a separate blood test for each individual. In an ideal setting, the result of a test is positive if and only if the subset contains a positive sample. A significant part of the existing research is focused on combinatorial pool design to guarantee detection using a small number of tests. Several variants of the problem exist, such as noisy group testing with different types of errors. An interesting variant is the graph-constrained group testing problem, where the salient set is the set of defective links in a graph and each test is a random walk on the graph [7]. The group testing model can be represented graphically as in Figure 3 where $X$ is a Boolean testing matrix and $Y$ is the outcome vector. Again, the different columns of the testing matrix correspond to the variables $X$, while the defective set corresponds to set $S$. Then, a test outcome $Y$ only depends on $X_S$, which captures the presence or absence of defective items in the test.

**Sparse channel estimation** [8] is used for the estimation of multi-path channels characterized by sparse impulse responses. The output of the channel depends on the input time instances, which correspond to the non-zero coefficients of the impulse response. In an equivalent channel model, the indices of the non-zero coefficients in the impulse response correspond to the encoded set $S$ and the coefficients themselves are absorbed into the channel model.

### 1.1 Related work and contributions

A large body of research work studies the sparse recovery problem, particularly from an information-theoretic (IT) perspective. In this section, we only describe work that is closely related to this paper. The dominant stream of research in this area deals with linear models and mean-squared estimation of the sparse vector $\beta$ in [8] with sub-Gaussian assumptions on variables $X$. Below we list the contributions of our approach and contrast it to some of the related work in the literature.

**Unifying framework for linear and non-linear problems:** Much of the literature on sparse recovery has focused on particular sparse models and reconstruction algorithms were developed for a specific setting.
Figure 3: Group testing example and its mapping to the channel model. \{2,3\} is the set of defective items. The channel is a Boolean channel and the codebook \(X^N\) is the testing matrix, which determines whether a sample is included in a test.

For instance, Lasso was used for linear regression \([9,10]\), relaxed integer programs for group testing \([11]\), convex programs for 1-bit quantization \([12]\), projected gradient descent for sparse regression with missing data \([5]\) 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 standpoint. The approach presented herein unifies the different sparse models based on the conditional independence assumption in \((1)\), and a single mutual information expression \((2)\) is shown to provide an exact characterization of the sample complexity for such models.

**Direct support recovery vs. signal estimation:** Much of the existing work focuses on limits of sparse recovery in linear models based on sensing matrices drawn from the standard Gaussian ensemble \([13–18]\). While only support recovery is contemplated in \([13,14,18]\), the support is generally chosen by retaining those elements of the signal estimate \(\hat{\beta}\) that lie above a design threshold. Alternatively, the estimated support is chosen to minimize the error associated with the best estimator \([13]\). Thus, much of this related literature is focused on *estimation*, sometimes as a preliminary step towards support recovery, which we avoid in this paper. In sharp contrast to prior work, our analysis makes a clear distinction between signal estimation and support discovery. It is conceivable that if the support is known, then the signal can be reliably estimated using least-square estimates or other variants. At a conceptual level, IT tools such as Fano’s inequality and the capacity theorems are powerful tools for inference of discrete objects (messages) given continuous observations. Indeed, to exploit such tools, \([13–18]\) resort to one of the following strategies: (a) Use IT tools only to establish necessary conditions for recovery by assuming a discrete \(\beta\), and derive sufficient conditions using some of the well-known algorithms (Lasso, Basis pursuit etc.); or (b) find an \(\epsilon\)-cover for \(\beta\) in some metric space (which requires imposing some extra assumptions) and reduce \(\beta\) to a discrete object. In contrast, our approach lifts these assumptions and focuses on the discrete combinatorial component of the object \(S\). Indeed, our results in Section 3 show that the discrete part, namely, the uncertainty of the
support pattern is the dominating factor and not $\beta$ 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 [19] or incoherence properties [9], or about the distribution of the matrix elements, such as sub-Gaussianity. Also, the existing information-theoretic bounds, which are largely based on Gaussian ensembles, are limited to the linear regression model, and hence not suitable for the non-linear models we consider herein.

It is worth noting that the authors in [20] adopt a parallel approach to derive sufficiency bounds for direct support recovery, albeit their analysis is focused on a hypothesis testing framework with fixed measurement matrices. In contrast, here we consider a general Bayesian framework with random $X$ and $\beta$.

**Performance bounds for new sparse recovery problems:** Our unifying approach also allows us to study problems that were not previously analyzed, or that are not easily analyzed, using the previous approaches. This includes problems with new observation models, or existing models with different distributions of variables. Using the formulation presented herein, obtaining necessary and sufficient conditions and error bounds only requires the computation of simple mutual information expressions.

The problem of identifying relevant variables was formulated in a channel coding framework in [6] and in the Russian literature in [21–25] in the context of group testing. Both sufficient and necessary conditions on the number of tests for the group testing problem with IID test assignments were derived. One main difference between the Russian literature and [6] is that, in the former the number of defective items, $K$, is held fixed while the number of items, $N$, approaches infinity. Consequently, the earlier work suggests that the number of tests must scale poly-logarithmically in $N$ regardless of $K$ for the error probability to approach zero. In contrast, [6] considers the setting wherein both the number of defective items, as well as the number of items can approach infinity. Hence, the necessary condition was derived using Fano’s inequality [3]. This analysis was further extended to general sparse signal processing models and models with dependent variables in [1].

In this paper, we are concerned with the analysis of the problem with IID variables $X$, which encompasses many important problems, such as the classical group testing or sparse linear regression models, to name a few. While this setup and a similar approach were considered in [1,26], this paper presents a more thorough and rigorous analysis, including the analysis of problems with latent variable observation models, formally extending the analysis to continuous models, presenting results for scaling models, and bounds for many example applications.

In Section 2, we introduce our notation and provide a formal description of the problem. In Section 3, we state necessary and sufficient conditions on the number of samples required for recovery. Applications are considered in Section 4 including bounds for sparse linear regression, group testing models, and models with missing data. We summarize our results in Section 5. We defer the proofs of theorems and lemmas in Sections 3 and 4 to the Appendix.

## 2 Problem Setup

**Notation.** We use upper case letters to denote random variables, vectors and matrices, and we use lower case letters to denote realizations of scalars, vectors and matrices. Subscripts are used for column indexing and superscripts with parentheses are used for row indexing in vectors and matrices. Subscripting with a set $S$ implies the selection of columns with indices in $S$. Table 1 provides a reference and further details on the used notation. Transpose of a vector or matrix is denoted by the $^\top$ symbol. $\log$ is used to denote the natural logarithm and entropic expressions are defined using the natural logarithm, however results can be converted other logarithmic bases w.l.o.g., such as base 2 used in [6].

**Variables.** Let $X = (X_1, X_2, \ldots, X_N) \in X^N$ denote a set of IID random variables with a joint probability distribution $Q(X)$. To simplify the expressions, we do not use subscript indexing on $Q$ to denote the random
variables since the distribution is determined solely by the number of variables indexed.

Candidate sets. We index the different sets of size $K$ as $S_\omega$ with index $\omega$, so that $S_\omega$ is a set of $K$ indices corresponding to the $\omega$-th set of variables. Since there are $N$ variables in total, there are $\binom{N}{K}$ such sets, therefore $\omega \in \mathcal{I} \triangleq \{1, 2, \ldots, \binom{N}{K}\}$. For any two sets $S_i$ and $S_j$, we define $S_{i,j}$, $S_{i,j^c}$, and $S_{i,j^c}$ as the overlap set, the set of indices in $S_j$ but not in $S_i$, and the set of indices in $S_i$ but not in $S_j$, respectively. Namely, $S_{i,j} = S_i \cap S_j$, $S_{i,j} = S_i^c \cap S_j$, and $S_{i,j}^c = S_i \cap S_j^c$.

Observations. We let $Y \in \mathcal{Y}$ denote an observation or outcome, which depends only on a small subset of variables $S \subset \{1, \ldots, N\}$ of known cardinality $|S| = K$ where $K \ll N$. In particular, $Y$ is conditionally independent of the subset of variables given the index set $S$, as in $\Box$, i.e.,

$$P(Y|X,S) = P(Y|X_S,S),$$

where $X_S = \{X_k\}_{k \in S}$ is the subset of variables indexed by the set $S$. We assume true set $S = S_\omega$ for some random variable $\omega$ distributed over $\mathcal{I}$.

Latent observation parameters. We consider an observation model which is not completely deterministic and known, but depends on a latent variable $\beta_S \in \mathcal{B}^K$. Note that this is a more general model compared to $\Box$ and $\Box$. We assume $\beta_S$ IID across indices in $S$, is independent of variables $X$ and has a prior distribution $P(\beta_S|S)$. We further assume that $\beta_k$ for $k \in S$ the probability $P(\beta_k)$ is lower bounded by a constant on its support independent of $K$, $N$ and $T$. While this assumption is used to obtain the sufficiency results we present in the paper, it is not essential and may be possible to remove using a different analysis.

Observation model. The outcomes depend on both $X_S$ and $\beta_S$ and are generated according to the model $P(Y|X_S,\beta_S,S)$. As an example, this latent variable corresponds to the non-zero coefficients of the $K$-sparse vector $\beta$ in the sparse linear regression framework in Section 4.1.1, or the impulse response coefficients in the sparse channel estimation framework. Note that $\Box$ still holds in this model, where $P(Y|X_S,S)$ is $P(Y|X_S,\beta_S,S)$ averaged over $\beta_S$ conditioned on $S$.

We use the lower-case $p(\cdot|\cdot) = P(\cdot|\cdot,S)$ notation as a shorthand for the conditional distribution given the true subset of variables $S$. For instance, with this notation we have $p(Y|X_S) = P(Y|X_S,S)$, $p(Y|X_S,\beta_S) = P(Y|X_S,\beta_S,S)$, $p(\beta_S) = P(\beta_S|S)$ etc. When we would like to distinguish between the outcome distribution conditioned on different sets of variables, we use $p_\omega(\cdot|\cdot) = P(\cdot|\cdot,S_\omega)$ notation, to

| Variables | Random quantities | Realizations |
|-----------|------------------|--------------|
| 1×N random vector | $X_1, \ldots, X_N$ | $x_1, \ldots, x_N$ |
| 1×|S| random vector | $X_S$ | $x_S$ |
| $T \times N$ random matrix | $X_T$ | $x_T$ |
| $t$-th row of $X_T$ | $X^{(t)}$ | $x^{(t)}$ |
| $n$-th column of $X_T$ | $X_T^n$ | $x_T^n$ |
| $n$-th elt. of $t$-th row | $X^{(t)}_n$ | $x^{(t)}_n$ |
| $T \times |S|$ sub-matrix | $X_T^S$ | $x_T^S$ |
| Outcome | $Y$ | $y$ |
| $T \times 1$ outcome vector | $Y_T$ | $y_T$ |
| $t$-th element of $Y_T$ | $Y^{(t)}$ | $y^{(t)}$ |
emphasize that the conditional distribution is conditioned on the given variables, assuming the true set \( S \) is \( S_\omega \).

We observe the realizations \((x^T, y^T)\) of \( T \) variable-outcome pairs \((X^T, Y^T)\) with each sample realization \((x^{(t)}, y^{(t)})\) of \((X^{(t)}, Y^{(t)})\), \( t = 1, 2, \ldots, T \). The variables \( X^{(t)} \) are distributed IID across \( t = 1, \ldots, T \). However, the outcomes \( Y^{(t)} \) are independent for different \( t \) only when conditioned on \( \beta_S \). Our goal is to identify the set \( S \) from the data samples and the associated outcomes \((x^T, y^T)\), with an arbitrarily small average error probability.

**Decoder and probability of error.** We let \( \hat{S}(X^T, Y^T) \) denote an estimate of the set \( S \), which is random due to the randomness in \( S \), \( X^T \) and \( Y^T \). We further let \( P(E) \) denote the average probability of error, averaged over all sets \( S \) of size \( K \), realizations of variables \( X^T \) and outcomes \( Y^T \), i.e.,

\[
P(E) = \Pr[\hat{S}(X^T, Y^T) \neq S] = \sum_{\omega \in I} P(\omega) \Pr[\hat{S}(X^T, Y^T) \neq S_\omega|S_\omega].
\]

**Scaling variables and asymptotics.** We let \( N \in \mathbb{N} \), \( K \equiv K(N) \in \mathbb{N} \) be a function of \( N \) such that \( 1 \leq K < N/2 \) and \( T \equiv T(K, N) \in \mathbb{N} \) be a function of both \( K \) and \( N \). Note that \( K \) can be a constant function in which case it does not depend on \( N \). For asymptotic statements, we consider \( N \to \infty \) and \( K \) and \( T \) scale as defined functions of \( N \). We formally define **sufficient** and **necessary** conditions for recovery as below.

**Definition 2.1.** For a function \( g \equiv g(T, K, N) \), we say an inequality \( g \geq 1 \) (or \( g > 1 \)) is a sufficient condition for recovery if there exists a sequence of decoders \( \hat{S}_N(X^T, Y^T) \) such that \( \lim_{N \to \infty} P(E) = \lim_{N \to \infty} \Pr[\hat{S}_N(X^T, Y^T) \neq S] = 0 \) for \( g \geq 1 \) (or \( g > 1 \)) for sufficiently large \( N \), i.e., for any \( \epsilon > 0 \), there exists \( N_\epsilon \) such that for all \( N > N_\epsilon \), \( g \geq 1 \) (or \( g > 1 \)) implies \( P(E) < \epsilon \). Conversely, we say an inequality \( g \geq 1 \) (or \( g > 1 \)) is a necessary condition for recovery if \( \lim_{N \to \infty} P(E) > 0 \) for any sequence of decoders when \( g \geq 1 \) (or \( g > 1 \)).

### 3 Conditions for Recovery

In this section, we state and prove sufficient and necessary conditions for the recovery of the salient set \( S \) with an arbitrarily small average error probability.

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

(A1) **Equi-probable support:** Any set \( S_\omega \subset \{1, \ldots, N\} \) with \( K \) elements is equally likely *a priori* to be the salient set. We assume we have no prior knowledge of the salient set \( S \) among the \( \binom{N}{K} \) possible sets.

(A2) **Conditional independence:** The observation/outcome \( Y \) is conditionally independent of other variables given \( X_S \), variables with indices in \( S \), i.e., \( P(Y|X,S) = P(Y|X_S,S) \). This assumption follows directly from our formulation of sparse recovery problems. We further assume the observation model does not depend on \( S \) except through \( X_S \), i.e., \( P(Y|X_S = x, S_\omega) = P(Y|X_S = x, S_\omega) \) for any \( x \in \mathcal{X}^K \), \( \omega, \omega' \in I \).

(A3) **IID variables:** The variables \( X_1, \ldots, X_N \) are independent and identically distributed. While the independence assumption is not valid for all sparse recovery problems, many problems of interest can be analyzed within the IID framework, as in Section \( [I] \).

(A4) **Observation model symmetry:** For any permutation mapping \( \pi \), \( P(Y|X_S, S) = P(Y|X_{\pi(S)}, S) \), 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 usually incorporated into \( \beta_S \). In other words, the symmetry is assumed for the observation model when averaged over \( \beta_S \).
Remarks on support versus support coefficients: In many sparse recovery problems we are concerned with the recovery of an underlying sparse vector \( \beta \), which has a sparsity support \( S \) and coefficients \( \beta_S \) on the support. For instance, a simple example that exhibits such structure is the following linear observation model, where

\[
Y = (X, \beta) + W = (X_S, \beta_S) + W,
\]

with noise \( W \), along with extensions to non-linear models, where \( Y = f((X_S, \beta_S) + W) \), for a function \( f : \mathcal{Y} \to \mathcal{Y} \).

In this work, we are specifically concerned with the recovery of the support \( S \) and not the recovery of the support coefficients \( \beta_S \). Instead, we incorporate the effects of the support coefficients into the observation model assuming prior density \( p(\beta_S) \), such that

\[
p(Y|X) = p(Y|X_S) = \int p(Y|X_S, \beta_S)p(\beta_S) \, d\beta_S,
\]

in order to analyze errors in recovering \( S \). In contrast, other error criteria are also considered for sparse recovery problems, mostly in the compressive sensing literature, such as the \( \ell_1 \) distance between the true and the estimated \( \beta \).

3.1 Sufficiency

In this section, we prove a sufficient condition for the recovery of \( S \). The notation in this section assumes discrete variables and observations, however simply replacing the related sums with appropriate integrals generalize the notation to the continuous case. We consider models with non-scaling distributions, i.e., models where the observation model and the variable distributions/densities and number of relevant variables \( |S| = K \) do not depend on scaling variables \( N \) or \( T \). Group testing as set up in Section 4.2.2 for fixed \( K \) is an example of such a model. We defer the discussion of models with scaling distributions and \( K \) to Section 5.

To derive the sufficiency bound for the required number of samples, we analyze the error probability of a Maximum Likelihood (ML) decoder [27]. For this analysis, we assume that \( S_1 \) is the true set \( S_\omega \) among \( \omega \in \mathcal{I} \). We can assume this w.l.o.g. due to the equi-probable support, IID variables and observation model symmetry assumptions (A1)-(A4), thus we can write

\[
P(E) = \frac{1}{\binom{N}{K}} \sum_{\omega \in \mathcal{I}} \Pr[\hat{S}(X^T, Y^T) \neq S_\omega|S_\omega] = P(E|S_1).
\]

For this reason, we omit the conditioning on \( S_1 \) on the error probability expressions throughout this section.

The ML decoder goes through all \( \binom{N}{K} \) possible sets \( \omega \in \mathcal{I} \) and chooses the set \( S_\omega^{\star} \) such that

\[
p_{\omega^{\star}}(Y^T|X_{S_\omega^{\star}}^T) > p_\omega(Y^T|X_{S_\omega}^T), \quad \forall \omega \neq \omega^{\star}, \tag{4}
\]

and consequently, if any set other than the true set \( S_1 \) is more likely, an error occurs. This decoder is a minimum probability of error decoder for equi-probable sets, as we assumed in (A1). Note that the ML decoder requires the knowledge of the observation model \( p(Y|X_S, \beta_S) \) and the distribution \( p(\beta_S) \).

Remarks on typicality decoding: It is worth mentioning that a typicality decoder can also be analyzed to obtain a sufficient condition, as used in the early versions of [28]. However, typicality conditions must be defined carefully to obtain a tight bound w.r.t. \( K \), as standard typicality definitions the atypicality probability may dominate the decoding error probability in the typical set. For instance, for the group testing scenario considered in [6], where \( X_n \sim \text{Bernoulli}(1/K) \), we have \( \Pr[X_S = (1, \ldots, 1)] = (1/K)^K \), which would require the undesirable scaling of \( T \) as \( K^K \), to ensure typicality in the strong sense (as needed to apply results such as the packing lemma [29]). Redefining the typical set as in [28] is then necessary, but it is problem-specific and makes the analysis cumbersome compared to the ML decoder adopted herein and in [6]. Furthermore, the case where \( K \) scales together with \( N \) requires an even more subtle analysis,
whereas the analysis of the ML decoder analysis is more straightforward in regards to that scalability. Typically decoding has also been reported as infeasible for the analysis of similar problems, such as multiple access channels where the number of users scales with the coding block length [30].

We now derive a simple upper bound on the error probability $P(E)$ of the ML decoder, which is averaged over all sets, data realizations and observations. Define the error event $E_i$ as the event of mistaking the true set for a set which differs from the true set $S_1$ in exactly $i$ variables, thus we can write

$$P(E_i) = \Pr [\exists \omega \neq b \in \mathcal{B}^u : p_\omega(Y^T | X^T_{S_2}) \geq p_{S_1}(Y^T | X^T_{S_2}), |S_{1,\omega}^c| = |S_{1,\omega}^c| = i, |S_1| = |S_\omega| = K].$$

Using the union bound, the probability of error $P(E)$ can then be upper bounded by

$$P(E) \leq \sum_{i=1}^{K} P(E_i) = \sum_{i=1}^{K} \sum_{X^T_{S_1}} Q(X^T_{S_1}) p_1(Y^T | X^T_{S_1}) P(E_i | X^T_{S_1}, Y^T, \omega = 1),$$

where $P(E_i | X^T_{S_1}, Y^T, \omega = 1)$ is the probability of decoding error in exactly $i$ variables, conditioned on the true index $\omega = 1$, the realization $X^T_{S_1}$ for the set $S_1$, and on the sequence $Y^T$. While we use notation for discrete variables and observations throughout this section, continuous case follows by replacing sums with appropriate integrals.

Next we state our main result, which complements and generalizes the results in [6]. The following theorem provides a sufficient condition on the number of samples $T$ for an arbitrarily small average error probability.

**Theorem 3.1.** (Sufficiency). Let $(S^1, S^2)$ be any partition of true set $S$ to $i$ and $K-i$ indices respectively, $\epsilon > 0$ be an arbitrary constant, let $I(X_S^1; Y | X_S^2, \beta_S = b, S)$ be the conditional mutual information conditioned on fixed $\beta_S = b$ and $I(X_S^1; Y | X_S^2, \beta_{\min}, S)$ be the worst-case (w.r.t. $\beta_S$) conditional mutual information conditioned on fixed $\beta_S = \beta\min$, where

$$\beta\min \in \arg \min_{b \in \mathcal{B}^n} I(X_S^1; Y | X_S^2, \beta_S = b, S) = \arg \min_{b \in \mathcal{B}^n} E_{S,Y,X_S^1,X_S^2} \left[ \log \frac{P(Y | X_S^1, X_S^2, \beta_S = b, S)}{P(Y | X_S^2, \beta_S = b, S)} \right].$$

Then, if assumptions (A1)-(A4) are satisfied,

$$T > (1 + \epsilon) \cdot \max_{i=1, \ldots, K} \frac{\log \left( \binom{N-K}{i} \right)}{I(X_S^1; Y | X_S^2, \beta_{\min}, S)}$$

is a sufficient condition for the average error probability to approach zero asymptotically, i.e., $\lim_{K \to \infty} \lim_{N \to \infty} P(E) = 0$.

**Remark 3.1.** In certain models the worst-case mutual information $I(X_S^1; Y | X_S^2, \beta_{\min}, S)$ can be exactly equal to zero, such as the linear model with $\mathcal{B} \ni 0$ which would lead to a vacuous upper bound. However such cases can possibly be avoided for models where such $\beta_{\min}$ occur with zero probability (such as continuous $\beta_S$) by considering a typical set of $\beta_S$ where the worst-case mutual information is non-zero in the set and the atypical set has vanishing probability.

The sufficiency conditions in Theorem 3.1 are derived from an upper bound on the error probability $P(E_i)$ for each $i = 1, \ldots, K$. This upper bound is characterized by the error exponent $E_\omega(\rho)$, which is described by

$$E_\omega(\rho) = -\frac{1}{T} \log \sum_{Y^T} \sum_{X_{S_2}^T} \left[ \sum_{X_{S_1}^T} Q(X_{S_1}^T) p(Y^T, X_{S_2}^T | X_{S_1}^T)^{1+\rho} \right]^{1+\rho}, \quad 0 \leq \rho \leq 1,$$

and the following lemma provides the upper bound on $P(E_i)$, the probability of decoding error in $i$ variables.

---

1 Note that it is sufficient to compute $I(X_S^1; Y | X_S^2, \beta_{\min}, S = \hat{S})$ for one value of $\hat{S}$ (e.g. $S_1$) instead of averaging over all possible $S$, since the conditional mutual information expressions are identical due to our symmetry assumptions on the variable distribution and the observation model. Similarly, the bound need only be computed for one partitioning $(S^1, S^2)$ for each $i = |S^1| \in \{1, \ldots, K\}$, since our assumptions ensure that the mutual information is identical for all such partitions. Also similarly, since $\beta_k$ is IID for $k \in S$, $\beta_{\min}$ can be chosen as $b\hat{1}$ for some $b \in \mathcal{B}$.

2 "Sufficient condition" is defined formally in the problem setup, where in this case we have $g(T, K, N) = \text{LHS/RHS}.$
Lemma 3.1. The probability of the error event $E_i$ defined in (9) that a set which differs from the set $S_1$ in exactly $i$ variables is selected by the ML decoder (averaged over all data realizations and outcomes) is bounded from above by

$$P(E_i) \leq e^{-(TE_o(\rho)-\rho \log (N-K_i)-\log (K_i))}.$$  \hspace{1cm} (9)

The proof for Lemma 3.1 follows largely along the proof of Lemma III.1 of [6] for discrete variables and observations. We note certain differences in the proof and the result, and further extend it to continuous variables and observations in Section A.6.

The proof of Theorem 3.1 is provided in the Appendix. It follows from lower bounding the error exponent using a worst-case analysis for $\beta_S$ to reduce to a single-letter expression and performing a Taylor series analysis of the lower bound around $\rho = 0$, from which the worst-case mutual information condition is derived. This Taylor series analysis is similar to the analysis of the ML decoder in [27]. While our proof uses a similar methodology to the proof of Theorem III.1 in [6], there are very important conceptual and technical differences, including

- the generalization to discrete alphabets for both $X$ and $Y$ with arbitrary cardinality,
- the generalization to continuous alphabets,
- the handling of latent observation model parameters $\beta_S$, which complicates the error exponent and mutual information expressions and induces dependence between $(X^{(t)}, Y^{(t)})$ pairs across $t$,
- the second order analysis of the error exponent for scaling models.

All of the above are necessary for the analysis of general sparse signal processing problems and represent a significant technical contribution. In contrast, the group testing model considered in [6] can be viewed as a special case, which enabled the use of simpler analysis.

It is also important to highlight the main difference between the analysis of the error probability for the problem considered herein and the channel coding problem. In contrast to channel coding, the codewords of a candidate set and the true set are not independent since the two sets could be overlapping. To overcome this difficulty, we separate the error events $E_i$, $i = 1, \ldots, K$, of misidentifying the true set in $i$ items. Then, for every $i$ we fix the $K - i$ correctly identified elements of the true set and average over the set of possible codeword realizations for every candidate set with $i$ differing elements.

### 3.2 Necessity

In this section, we derive lower bounds on the required number of measurements using Fano’s inequality [3]. We state the following theorem:

**Theorem 3.2.** Let $(S^1, S^2)$ be any partition of true set $S$ to $i$ and $K - i$ indices respectively and define $I(X_{S^1}; Y | X_{S^2}, \beta_S, S)$ to be the conditional mutual information between $X_{S^1}$ and $Y$ conditioned on $X_{S^2}$, $\beta_S$ and true set $S$. For $N$ variables and a set $S$ of $K$ salient variables, a necessary condition \(^4\) on the number of samples required for the probability of error to be asymptotically bounded away from zero is given by

$$T \geq \max_{i=1, \ldots, K} \frac{\log \left( \frac{N-K+i}{i} \right)}{I(X_{S^1}; Y | X_{S^2}, \beta_S, S)}. \hspace{1cm} (10)$$

The proof follows along similar lines to the proof of Theorem IV.1 in [6] with some important differences regarding the latent variable $\beta_S$ in the observation model and explicit conditioning on $S_\omega$ and $S^2$. We detail the differences in the Appendix.

---

\(^3\) Note that this mutual information is averaged over $\beta_S$, instead of being defined for a fixed value of $\beta_S$.

\(^4\) “Necessary condition” is defined formally in the problem setup, where in this case we have $g(T; K, N) = \text{LHS/RHS}$. 
Remark 3.2. Given that the worst-case mutual information is equal to the average mutual information (or random $\beta_S$ does not exist), the upper bound in Theorem [3.1] is tight as it matches lower bound given in Theorem [3.2]. For non-scaling models, the bound is always order-wise tight if the worst-case mutual information is strictly positive.

Interpretation. Intuitively, the bounds in (7) and (10) can be explained as follows: For each $i$, the numerator is approximately the number of bits required to represent all sets $S_\omega$ that differ from $S$ in $i$ elements. The denominator represents the information given by the output variable $Y$ about the remaining $i$ indices $S^1_i$, given the subset $S^2$ of $K - i$ true indices. Hence, the ratio represents the number of samples needed to control $i$ support errors and the maximization accounts for all possible support errors.

Support recovery and support coefficients. In the sufficiency and necessity proofs above, we show that $\beta_S$ being unknown with prior $P(\beta_S)$ induces a penalty term in the denominator given by $I(\beta_S; X_{S^1}^T | X_{S^2}^T, Y^T, S)/T$, compared to the case where the support coefficients $\beta_S$ are fixed and known. We show that this term is always dominated by $I(X_{S^1}; Y|X_{S^2}, \beta_S, S)$, therefore does not affect the sample complexity asymptotically. This shows that recovering the support given the knowledge of the support coefficients is as hard as recovering the support with unknown coefficients, underlying the importance of recovering the support in sparse recovery problems.

Partial recovery. As we analyze the error probability separately for $i = 1, \ldots, K$ support errors in order to obtain the necessity and sufficiency results, it is trivial to determine necessary and sufficient conditions for partial support recovery instead of exact support recovery. By changing the maximization from over $i = 1, \ldots, K$ to $i = K - k + 1, \ldots, K$ in the two recovery bounds, the conditions to recover at least $k$ of the $K$ support indices can be determined.

3.3 Sufficiency for Models with Scaling

In this section we consider models with scaling distributions, i.e., models where the observation model and the variable distributions/densities and number of relevant variables $|S| = K$ may depend on scaling variables $N$ or $T$. While Theorem [3.1] characterizes precisely the constants (including constants related to $K$) in the sample complexity, it is also important to analyze models where $K$ is scaling with $N$ or where distributions depend on scaling variables. Group testing where $K$ scales with $N$ (e.g. $K = \Theta(\sqrt{N})$) is an example of such a model, as well as the normalized sparse linear regression model with SNR and matrix probabilities functions of $N$ and $T$ in Section [4.1.1]. Therefore in this section we consider the most general case where $Q(X_k)$ or $p(Y|X_S)$ can be functions of $K$, $N$ or $T$ and $K = O(N)$. Note that the necessity result in Theorem [3.2] holds also for scaling models thus does not need to be generalized.

For this case, in addition to the general assumptions (A1)-(A4), we require additional smoothness properties related to the second derivative of the error exponent $E_o(\rho)$ as defined in (8). We first present a sufficient condition involving multi-letter expressions. This theorem presents the most general results in this section with the least stringent second derivative conditions.

Theorem 3.3. (Multi-letter sufficiency condition). Let $(S^1, S^2)$ be any partition of true set $S$ to $i$ and $K - i$ indices respectively, and define $I(X_{S^1}^T; Y^T | X_{S^2}^T, S)$ to be the multi-letter conditional mutual information between $X_{S^1}^T$ and $Y^T$ conditioned on $X_{S^2}^T$ and true set $S$.

Let $\tau_N > 10$ be a sequence of numbers (which can be a function of $i, K, N$ etc.) such that $|E_o''(\rho)| \leq \frac{\tau}{2^4} I(X_{S^1}^T; Y^T | X_{S^2}^T, S)$ for all $0 \leq \rho \leq 1$. Then, if assumptions (A1)-(A4) are satisfied,

$$\min_{i=1, \ldots, K} \frac{I(X_{S^1}^T; Y^T | X_{S^2}^T, S)}{\tau_N \log \left( \frac{N-K}{N-i} \right)} > 1,$$

(11)
is a sufficient condition\footnote{The mutual information characterizations in Sections [3.1] and [3.2] were single-letter.} for the average error probability to asymptotically approach zero, i.e., $\lim_{K,N \to \infty} P(E) = 0$. 

As both the error exponent $E_o(\rho)$ and the mutual information expression in the above theorem are multi-letter expressions, the second derivative condition and the sufficiency bound may be difficult to analyze, in contrast to the single-letter characterization of Theorem 3.1. In the theorem below we present a single-letter simplification of Theorem 3.3, which has slightly stronger conditions and may have a looser sufficiency bound for certain cases.

**Theorem 3.4.** (Single-letter sufficiency condition). Let $\{S^1, S^2\}$ be any partition of true set $S$ to $i$ and $K-i$ indices respectively and let $I(X_{S^1}; Y | X_{S^2}, \beta_S = b, S)$ be the conditional mutual information conditioned on fixed $\beta_S = b$ and $I(X_{S^1}; Y | X_{S^2}, \beta_{\text{min}}, S)$ to be the worst-case (w.r.t. $\beta_S$) conditional mutual information conditioned on fixed $\beta_S = \beta_{\text{min}}$, as defined in (6).

Let $E_o(\rho, \beta_S)$ be the single-letter conditional error exponent as defined in (A.4), $p_{\text{min}} = \min_{b \in B} P(\beta_k)$ for any $k \in S$ and $\tau_N \to \infty$ be a sequence of numbers (which can be a function of $i$, $K$, $N$ etc.) such that

$$|E_o''(\rho, \beta_S = b)| \leq \frac{\tau_N}{5} I(X_{S^1}; Y | X_{S^2}, \beta_S = b, S),$$

for all $b \in B^K$, $0 \leq \rho \leq 1$ and $i = 1, \ldots, K$. Then, if asymptotically $I(X_{S^1}; Y | X_{S^2}, \beta_{\text{min}}, S) \geq \frac{K}{\tau_{p_{\text{min}}}}$, $T > \max_{i=1,\ldots,K} \log \left( \frac{(N-K)}{i} \right) I(X_{S^1}; Y | X_{S^2}, \beta_{\text{min}}, S) - \frac{K}{\tau_{p_{\text{min}}}} \cdot \tau_N,$

(13)
is a sufficient condition\(^6\) for the average error probability to asymptotically approach zero, i.e., $\lim_{K,N \to \infty} P(E) = 0$.

The single-letter theorem above is general and useful for all models including continuous and discrete models, and the second derivative condition can be checked easier indirectly using bounds that we present in the below lemma.

**Lemma 3.2.** Let

$$g_\rho \triangleq \left( \sum_{X_{S^1}} Q(X_{S^1}) p(Y, X_{S^2} | X_{S^1}, \beta_S) \right)^{1+\rho}, \quad u_\rho \triangleq \frac{p(Y | X_{S^1}, X_{S^2}, \beta_S) \left[ Q(X_{S^1}) p(Y | X_{S^1}, X_{S^2}, \beta_S) \right]^{1+\rho}}{\sum_{X_{S^1}'} Q(X_{S^1}') p(Y | X_{S^1}', X_{S^2}, \beta_S)^{1+\rho}}$$

and note that $E_o(\rho, \beta_S) = - \log \sum_{Y,X_{S^2}} g_\rho$. Then,

$$|E_o''(\rho, \beta_S)| \leq \frac{\sum_{Y,X_{S^2}} g_\rho E[u_\rho \log^2 u_\rho]}{\sum_{Y,X_{S^2}} g_\rho} \leq \sup_{Y,X_{S^2}} E[u_\rho \log^2 u_\rho],$$

(14)

for $0 \leq \rho \leq 1$ and $\beta_S \in B^K$.

**Remark 3.3.** The sufficiency bound \(^{[13]}\) in Theorems 3.3, 3.4 reduces to \(^{[7]}\) in Theorem 3.1 nearly exactly given that the model is sufficiently sparse ($K = O(\log N)$) and \(^{[12]}\) is satisfied such that $\tau_N$ can be chosen to scale arbitrarily slowly.

The proofs of Theorems 3.3 and 3.4 are provided in the Appendix. They follow using Lemma 3.1 similar to the proof of Theorem 3.1 but for Theorem 3.4 we further lower bound the error exponent using a worst-case analysis for $\beta_S$ to reduce to a single-letter expression and again perform a Taylor series analysis of the lower bound around $\rho = 0$, from which the worst-case mutual information condition is derived. The second derivative conditions such as \(^{[12]}\) are necessary to control the second order term in the Taylor series.

\(^6\)While we use notation for discrete variables and observations, continuous case follows by replacing sums with appropriate integrals.
4 Applications

In this section, we establish results for several problems for which our necessity and sufficiency results are applicable. In the first subsection, we look at linear observation models and derive results for sparse linear regression with measurement noise. Then we consider a multivariate regression model, where we deal with vector-valued variables and outcomes. In the second subsection, we analyze probit regression and group testing as examples of non-linear observation models. Finally, we look at a general framework where some of the variables are not observed, i.e., each variable is missing with some probability. Proofs are provided in the Appendix where necessary.

4.1 Linear Settings

4.1.1 Sparse Linear Regression

Using the bounds presented in this paper for general sparse models, we derive sufficient and necessary conditions for the sparse linear regression problem with measurement noise [4] and a Gaussian variable matrix with IID entries.

We consider the following normalized model [15],

\[ Y^T = X^T \beta + W^T, \]

where \( X^T \) is the \( T \times N \) variable matrix, \( \beta \) is a \( K \)-sparse vector of length \( N \) with support \( S \), \( W^T \) is the measurement noise of length \( T \) and \( Y^T \) is the observation vector of length \( T \). In particular, we assume \( X_n^{(t)} \) are Gaussian distributed random variables and the entries of the matrix are independent across rows \( t \) and columns \( n \). Each element \( X_n^{(t)} \) is zero mean and has variance \( \frac{1}{T} \). \( W^T \) denotes the observation noise of length \( T \). We assume each element is IID with \( W \sim \mathcal{N}(0, \frac{1}{SNR}) \). The coefficients of the support, \( \beta_S \), are IID zero-mean random variables with variance \( \sigma^2 \) and \( |\beta_k| \geq b_{\min} \) for \( k \in S \).

In order to analyze this problem using the proposed sparse signal processing framework, it is important to observe how the regression model as defined above relates to the general sparse model. The elements in a row of the matrix \( X^T \) correspond to variables \( X_1, \ldots, X_N \) as defined in Section 2. Each row of the matrix is a realization of \( X \), and the rows are generated independently and identically to form \( X^T \). It is easy to see that assumption (1) is satisfied in both models since each measurement \( Y^{(t)} \) depends only on the linear combination of the elements \( X_S^{(t)} \) that correspond to the support of \( \beta \). The coefficients of this combination are given by \( \beta_S \), the values of the non-zero elements of \( \beta \). \( \beta_S \) corresponds to the latent parameter of the observation model \( P(Y|X_S, \beta_S, S) \), which accounts for the noise \( W \).

Let \( i \) denote the ratio of misidentified elements of the support \( S \), where \( 1 \leq i \leq K \). For the conditions for recovery and SNR, we will first show that

\[ I(X_{S^1};Y|X_{S^2},\beta_S,S) \geq \frac{1}{2} \log \left( 1 + \frac{i\sigma^2 SNR}{T} \right), \quad I(X_{S^1};Y|X_{S^2},\beta_{\min},S) = \frac{1}{2} \log \left( 1 + \frac{i\sigma_{\min}^2 SNR}{T} \right). \]

We then consider all values of \( i \) and note that \( \log \left( \frac{N-K}{i} \right) = \Theta(i \log(N/i)) \), to state the following theorem.

**Theorem 4.1.** For sparse linear regression with IID Gaussian matrix, \( SNR = \Omega(\log N) \) is a necessary condition for recovery. Furthermore, for this SNR a necessary condition on the number of observations is \( T = \Omega \left( \frac{K \log(N/K)}{\log(1+\sigma^2)} \right) \), while a sufficient condition is \( T = \Omega \left( \frac{K \log(N/K)}{\log(1+b_{\min})} \right) \).

**Remark 4.1.** For the sparse linear regression problem, we showed that our relatively simple mutual information analysis gives us a bound that is asymptotically identical to the best-known bound \( T = \Omega(K \log(N/K)) \) [15] with an independent Gaussian variable matrix, in the sublinear sparsity regime, in addition to providing us with a necessary condition on SNR that matches the bound in [15].

Figure 5 illustrates the lower bound on the number of measurements and shows that a necessary condition on SNR has to be satisfied for recovery. Our necessity result holds for all scalings of \( K \) and \( N \), while the
Figure 5: $\frac{T}{LB}$ vs. $T$ for different SNR values, where $LB$ is the necessity bound given by (A.30) for $K = 16$, $D = 512$ and $\sigma^2 = b_{\text{min}} = 1$. For low levels of SNR the necessary condition ($T > LB$, above the dotted line) is not satisfied even for very large $T$, for fixed $K$ and $N$. This is due to $\log \left( 1 + \frac{c \text{SNR}}{T} \right)$ behaving linearly instead of logarithmically for low $\frac{\text{SNR}}{T}$ ratios.

Figure 6: Mapping the multiple linear regression problem to a vector-valued outcome and variable model. On the left is the representation for a single problem $r = 1$. On the right is the corresponding vector formulation, shown for sample index $t = 2$.

sufficiency result holds in the fixed sparsity regime considered in Section 3.1. Although we provided results for exact recovery with random $\beta_S$, it is easy to obtain results for partial recovery as we remark in Section 3.

Another interesting aspect of our analysis is that in addition to sample complexity bounds, an upper bound to the probability of error in recovery can be explicitly computed and obtained using Lemma 3.1 for any finite triplet $(T, K, N)$. Following this line of analysis, an upper bound is obtained for sparse linear regression and compared to the empirical performance of practical algorithms such as Lasso \cite{9,10} in \cite{2}. It is then seen that while certain practical recovery algorithms have provably optimal asymptotic sample complexity, there is still a gap between information-theoretically attainable recovery performance and empirical performance of such algorithms. We refer the reader to \cite{2} for details.

4.1.2 Multivariate Regression

In this problem, we consider the following linear model \cite{31}, where we have a total of $R$ linear regression problems,

$$ Y_{(r)}^T = X_{(r)}^T \beta_{(r)} + W_{(r)}^T, \quad r = 1, \ldots, R. $$

For each $r$, $\beta_{(r)} \in \mathbb{R}^N$ is a $K$-sparse vector, $X_{(r)}^T \in \mathbb{R}^{T \times N}$ and $Y_{(r)}^T \in \mathbb{R}^T$. The relation between different tasks is that $\beta_{(r)}$ have joint support $S$. This set-up is also called multiple linear regression or distributed compressive sensing \cite{32} and is useful in applications such as multi-task learning \cite{33}.
It is easy to see that this problem can be formulated in our sparse recovery framework, with vector-valued outcomes $Y$ and variables $X$. Namely, let $Y = (Y_1, \ldots, Y_{(R)}) \in \mathbb{R}^R$ be a vector-valued outcome, $X = (X_1, \ldots, X_{(R)})^T \in \mathbb{R}^{R \times N}$ be the collection of $N$ vector-valued variables and $\beta = (\beta_1, \ldots, \beta_{(R)}) \in \mathbb{R}^{N \times R}$ be the collection of $R$ sparse vectors sharing support $S$, making it block-sparse. This mapping is illustrated in Figure 6. Assuming independence between $X_{(r)}$ and support coefficients $\beta_{(r)} \in S$ across $r = 1, \ldots, R$, we have the following observation model:

$$P(Y|X,S) = \prod_{r=1}^R p(Y_{(r)}|X_{(r)},S) = \prod_{r=1}^R \int_{\mathbb{R}^K} p(Y_{(r)}|X_{(r)},S,\beta_{(r)},S)p(\beta_{(r)},S) d\beta_{(r)},S.$$ (16)

We state the following theorem for the specific linear model in Section 4.1.1 and IID variables, as a direct result of Theorem 4.1 and the fact that the joint mutual information decomposes to $R$ identical mutual information terms due to the above equality.

**Theorem 4.2.** The sample complexity $T$ per test for the linear multi-regression model above is $\frac{T_o}{R}$, where $T_o$ is the sample complexity in Theorem 4.1.

**Remark 4.2.** We showed that having $R$ problems with independent measurements and sparse vector coefficients decreases the number of measurements per problem by a factor of $1/R$. While having $R$ such problems increases the number of measurements $R$-fold, the inherent uncertainty in the problem is the same since the support is shared. It is then reasonable to expect such a decrease in the number of measurements.

### 4.2 Non-linear Settings

#### 4.2.1 Binary Regression

As an example of a non-linear observation model, we look at the following binary regression problem, also called 1-bit compressive sensing [34–36] or probit regression. Regression with 1-bit measurements is interesting as the extreme case of regression models with quantized measurements, which are of practical importance in many real world applications. The conditions on the number of measurements have been studied for both noiseless [35] and noisy [36] models and $T = \Omega(K \log N)$ has been established as a sufficient condition for Gaussian variable matrices.

Following the problem setup of [36], we have

$$Y^T = q(X^T \beta + W^T),$$ (16)

where $X^T$ is a $T \times N$ matrix with IID standard Gaussian elements, and $\beta$ is an $N \times 1$ vector that is $K$-sparse with support $S$. $W^T$ is a $T \times 1$ noise vector with IID standard Gaussian elements. $q(\cdot)$ is a 1-bit quantizer that outputs 1 if the input is non-negative and 0 otherwise, for each element in the input vector. This setup corresponds to the SNR = 1 regime in [36].

**Theorem 4.3.** For probit regression with IID Gaussian variable matrix and the above setup, $T = \Omega(K \log N)$ measurements are sufficient to recover $S$, the support of $\beta$, with an arbitrarily small average error probability.

**Remark 4.3.** Similar to linear regression, for probit regression with noise we provided a sufficiency bound that matches [36] for an IID Gaussian matrix, for the corresponding SNR regime.

#### 4.2.2 Group Testing - Boolean Model

In this section, we consider another non-linear model, namely, group testing. This problem has been covered comprehensively in [6] and the results derived therein can also be recovered using the generalized results we presented in this paper.

The problem of group testing can be summarized as follows. Among a population of $N$ items, $K$ unknown items are of interest. The collection of these $K$ items represents the defective set. The goal is to construct
pooling design, i.e., a collection of tests, to recover the defective set while reducing the number of required tests. In this case $X^T$ is a binary measurement matrix defining the assignment of items to tests. For the noise-free case, the outcome of the tests $Y^T$ is deterministic. It is the Boolean sum of the codewords corresponding to the defective set $S$, given by $Y^T = \bigvee_{i \in S} X_i^T$.

**Theorem 4.4.** For $N$ items and $K$ defectives, the number of tests $T = \Omega(K \log N)$ is sufficient to identify the defective set $S$ with an arbitrarily small average error probability.

The upper and lower bounds on the number of tests (given by Theorem 3.1 and 3.2 respectively) for the noiseless case are illustrated in Figure 7. The results in [6] also establish upper and lower bounds on the number of tests needed for testing with additive noise (leading to false alarms) and dilution effects (leading to potential misses), as well as worst-case errors. We refer the reader to [6] for the mutual information analysis that leads to the theorem and for further details.

### 4.3 Models with Missing Features

Consider the general sparse signal processing model as described in Section 2. However, assume that instead of fully observing outcomes $Y^T$ and features $X^T$, we observe a $T \times N$ matrix $Z^T$ instead of $X^T$, with the relation

$$Z_i^{(t)} = \begin{cases} X_i^{(t)}, & \text{w.p. } 1 - \rho \\ m_i, & \text{w.p. } \rho \end{cases} \quad \forall i, t$$

i.e., we observe a version of the feature matrix which may have entries missing with probability $\rho$, independently for each entry. We show how the sample complexity changes relative to the case where the features are fully observed.

First we present a universal lower bound on the number of samples for the missing data framework, by relating $I(Z_{S^1}; Y | Z_{S^2}, \beta_{S}, S)$ to $I(X_{S^1}; Y | X_{S^2}, \beta_{S}, S)$.

**Theorem 4.5.** Consider the missing data setup described above. Then we have the lower bound on the sample complexity $T_{\text{miss}} \geq \frac{T_o}{1 - \rho}$, where $T_o$ is the lower bound on the sample complexity for the fully observed variables case given in Theorem 3.2.

As a special case, we analyze the sparse linear regression model with missing data [5,37], where we obtain a model-specific upper bound on the sample complexity, in addition to the universal lower bound given by Theorem 4.5.
Theorem 4.6. For the sparse linear regression setting in Section 4.1.1 with variable matrix entries missing w.p. $\rho$, for $\text{SNR} = \Omega(\log N)$, $T = \Omega\left(\frac{K \log(N/K)}{\log \left(\frac{1+\rho_{\text{min}}}{1+\rho_{\text{min}}^2}\right)}\right)$ samples are sufficient to estimate $S$ with an arbitrarily small average error probability.

Remark 4.4. We observe that the number of sufficient samples increases by a factor of $\frac{1}{\log(1+\frac{1}{1-\rho})}$ for missing probability $\rho$ and constant $\sigma^2$. Compare this to the upper bound given by [37] with scaling $\frac{1}{(1-\rho)^T}$, where the authors propose and analyze an orthogonal matching pursuit algorithm to recover the support $S$ with noisy or missing data. In this example, we have shown an upper bound that improves upon the bounds in the literature, with an intuitive universal lower bound.

This example highlights the flexibility of our results in view of the mutual information characterization. This flexibility enables us to easily compute new bounds and establish new results for a very wide range of general models and their variants.

5 Conclusions

We have presented a unifying framework based on noisy channel coding for analyzing sparse recovery problems. This approach unifies linear and non-linear observation models and leads to an explicit, intuitive and universal mutual information formula for computing the sample complexity of sparse recovery problems. We explicitly focus on the inference of the combinatorial component corresponding to the support set, probably the main difficulty in sparse recovery. We unify sparse problems from an inference perspective based on a Markov conditional independence assumption. Our approach is not algorithmic and therefore must be used in conjunction with tractable algorithms. It is useful for identifying gaps between existing algorithms and the fundamental information limits of different sparse models. It also provides an understanding of the fundamental tradeoffs between different parameters of interest such as $K$, $N$, SNR and other model parameters.

6 Appendix

A.1 Proof of Theorem 3.1

We derive an upper bound on $P(E)$ by upper bounding the maximum probability of the $K$ error events $E_i$, $i = 1, \ldots, K$, where using the union bound we have

$$P(E) \leq \sum_{i=1}^{K} P(E_i) \leq K \max_i P(E_i) = \max_i K P(E_i). \tag{A.1}$$

For each error event $E_i$, we aim to derive a sufficient condition on $T$ such that $K P(E_i) \to 0$ as $N \to \infty$, with $P(E_i)$ given by [9]. Using Lemma 3.1, this is equivalent to finding a condition on $T$ such that

$$T \left( E_o(\rho) - \rho \frac{\log \binom{N-K}{i}}{T} - \frac{\log \binom{K}{i}}{T} - \frac{\log K}{T} \right) \to \infty, \tag{A.2}$$

where $E_o(\rho)$ is given by [9]. Note that since $\log \binom{K}{i} + \log K = \Theta(1)$ for fixed $K$ and $T \to \infty$, the following is a sufficient condition on $T$ for (A.2) to hold true:

$$T f(\rho) = T \left( E_o(\rho) - \rho \frac{\log \binom{N-K}{i}}{T} \right) \to \infty.$$

We now find a lower bound to the error exponent, which removes the dependence between samples $t$ by considering worst-case $\beta_S$ and reduces it to a single-letter expression.
Lemma A.1.

\[ E_o(\rho) \geq E_o(\rho) \triangleq -\frac{\rho K}{T_{p_{\min}}} + \min_{\beta_S \in B^K} E_o(\rho, \beta_S) \]  

where \( p_{\min} = \min_{\beta \in B} p(\beta) = \Theta(1) \) and we define

\[ E_o(\rho, \beta_S) = -\log \left( \sum_Y \sum_{X_{S^1}} \left[ \sum_{X_{S^2}} Q(X_{S^1})p(Y, X_{S^2}|X_{S^1}, b) \right] \right). \]

where for continuous models sums are replaced with the appropriate integrals.

The lemma is proven later in the Appendix. Note that \( E_o(\rho, \beta_S) \) is exactly equal to \( E_o(\rho) \) as defined in when \( \beta_S \) coefficients do not exist (or are fixed and known), as \( E_o(\rho) \) readily factorizes over \( t \) and becomes a single-letter expression in that case. Also note \( E_o(\rho, \beta_S) \) does not scale with \( N \) or \( T \).

To show that the condition \([7]\) is sufficient to ensure \([A.2]\), we define \( f(\rho) = E_o(\rho) - \rho \frac{\log(N-K)}{T} \) and analyze \( E_o(\rho) \) using its Taylor expansion around \( \rho = 0 \). Using the mean value theorem, we can write \( E_o(\rho, \beta_S) \) in the Lagrange form of the Taylor series expansion, i.e., in terms of its first derivative evaluated at zero and a remainder term,

\[ E_o(\rho, \beta_S) = E_o(0, \beta_S) + \rho E'_o(0, \beta_S) + \frac{\rho^2}{2} E''_o(\psi, \beta_S) \]

for some \( \psi \in [0, \rho] \). Note that \( E_o(0, b) = 0 \) and the derivative of \( E_o(\rho, b) \) for any \( b \in B^K \) evaluated at zero can be shown to be

\[ \frac{\partial E_o(\rho, b)}{\partial \rho} \bigg|_{\rho=0} = \sum_Y \sum_{X_{S^1}} \left[ \sum_{X_{S^2}} Q(X_{S^1})p(Y, X_{S^2}|X_{S^1}, b) \log p(Y, X_{S^2}|X_{S^1}, b) \right. \\
- \sum_{X_{S^1}} Q(X_{S^1})p(Y, X_{S^2}|X_{S^1}, b) \log \sum_{X_{S^1}} Q(X_{S^1})p(Y, X_{S^2}|X_{S^1}, b) \bigg] \\
= \sum_Y \sum_{X_{S^2}} \sum_{X_{S^1}} Q(X_{S^1})p(Y, X_{S^2}|X_{S^1}, b) \log \sum_{X_{S^1}} Q(X_{S^1})p(Y, X_{S^2}|X_{S^1}, b) \\
= I(X_{S^1}; X_{S^2}, Y|\beta_S = b, S) = I(X_{S^1}; Y|X_{S^2}, \beta_S = b, S). \]

Above we used notation for discrete variables and observations, i.e. sums, however the proof is valid for continuous models if the sums are replaced by the appropriate integrals.

For notational convenience, let \( I_b \triangleq I(X_{S^1}; Y|X_{S^2}, \beta_S = b, S) \) and \( I \triangleq I(X_{S^1}; Y|X_{S^2}, \beta_{min}, S) \). Then, with the Taylor expansion of \( E_o(\rho, b) \) above we have

\[ T f(\rho) \geq T \bar{f}(\rho) \geq T \left( \min_b \left[ \rho I_b + \frac{\rho^2}{2} E''_o(\psi, b) \right] - \rho \frac{K}{T_{p_{min}}} - \rho \frac{\log(N-K)}{T} \right) \]

and our aim is to show that the above quantity approaches infinity for some \( \rho \in [0, 1] \) as \( N \to \infty \).

Now assume that \( T \) satisfies

\[ T > (1 + \epsilon) \cdot \frac{\log(N-K)}{I(X_{S^1}; Y|X_{S^2}, \beta_{min}, S)} \]

\text{Footnote: The actual mutual information expressions we are computing are conditioned on } S = S_1 \text{ (e.g. } I(X_{S^1}; Y|X_{S^2}, \beta_{min}, S = S_1) \text{) since we assumed true set is } S_1 \text{ w.l.o.g., which are equal to the averaged mutual information expressions over } S \text{ (e.g. } I(X_{S^1}; Y|X_{S^2}, \beta_{min}, S) \text{) which we use throughout this section (see footnote^1).} \]
for all $i$, which is implied by condition \[7\]. Using above $T$ and \[A.6\], we can then write

$$Tf(\rho) \geq T\bar{f}(\rho) \geq T \left( \rho I + \frac{\rho^2}{2} \min_b E''_\beta(\psi, b) - \rho \frac{K}{T_{\min}} - \rho \frac{\log(N-K)}{T} \right)$$

$$\geq T \left( \rho I + \frac{\rho^2}{2} \min_b E''_\beta(\psi, b) - \rho o(1) - \rho \frac{I}{1+\epsilon} \right)$$

$$= T\rho \left( \epsilon'I + \frac{\rho^2}{2} \min_b E''_\beta(\psi, b) - o(1) \right),$$

for $\epsilon' = \frac{\epsilon}{1+\epsilon}$, where in the first inequality we lower bounded by separating the minimum of the sum to the sum of minimums and in the second inequality replaced $T$, noting that $K/T \to 0$.

We note that $E''_\beta(\psi, b)$ is independent of $N$ or $T$ and bounded from below (which can be trivially obtained using Lemma \[3.2\] for any $b \in \mathcal{B}^K$. Then, we pick $\rho$ small enough such that the second derivative term is dominated by the mutual information term; specifically we choose $\rho \leq \frac{\epsilon'I}{\min_b E''_\beta(\psi, b)}$ and note that it can be chosen such that $\rho \geq \delta > 0$ for a constant $\delta$, since $|E''_\beta(\psi, b)| = O(1)$. We then have

$$Tf(\rho) \geq T\rho \left( \epsilon'I - \epsilon'/2 - o(1) \right) = T\rho I \Theta(1) = \Theta(1) = \Omega(\log N) \to \infty,$$

showing that $KP(E_i)$ goes to zero for all $i$ given the conditions (A1)-(A4) are satisfied. From these two cases, it follows that $P(E) \leq \max_i KP(E_i)$ goes to zero for $N \to \infty$ for any $K$, therefore

$$\lim_{K \to \infty} \lim_{N \to \infty} P(E) = 0.$$

### A.2 Proof of Theorem 3.2

As we remark in the main body, the proof of the necessity condition follows along the lines of proof of Theorem IV.1 in \[6\]. However, we need to account for the present of latent variable $\beta_s$ and obtain a slightly different bound as a result, that matches the sufficiency bound in Theorem 3.1. We also note that the conditioning on $S^2$ in the expressions in \[6\] is implicit, which we make explicit below to ensure correctness and clarity.

We further explicitly condition on $S_\omega$ in the mutual information and entropy expressions similarly.

Let the true set $S = S_\omega$ for some $\omega \in \mathcal{I}$ and suppose $K - i$ elements of $S_\omega$ is revealed, denoted by $S^2$. We define the estimate of $\omega$ to be $\hat{\omega} = g(X^T, Y^T)$ and the probability of error in the estimation $P_e = P(E) = Pr[\hat{\omega} \neq \omega]$.

Then following along the proof of \[6\], in equation (29) we have the following Fano’s inequality:

$$H(\omega|Y^T, X^T, S^2) \leq 1 + P_e \log \left( \frac{N-K+i}{i} \right).$$

We now diverge from \[6\] and we have the following chain of inequalities for the left-hand term:

$$H(\omega|Y^T, X^T, S^2) = H(\omega|S^2) - I(\omega; Y^T, X^T|S^2)$$

$$= H(\omega|S^2) - I(\omega; X^T|S^2) - I(\omega; Y^T|X^T, S^2)$$

$$\overset{(a)}{=} H(\omega|S^2) - I(\omega; Y^T|X^T, S^2)$$

$$\overset{(b)}{=} H(\omega|S^2) - (H(Y^T|X^T, S^2) - H(Y^T|X^T, \omega))$$

$$\overset{(c)}{=} H(\omega|S^2) - (H(Y^T|X_{S_\omega}^T, S^2) - H(Y^T|X_{S_\omega}^T, \omega))$$

$$\overset{(d)}{=} H(\omega|S^2) - I(X_{S_\omega}^T; Y^T|X_{S_\omega}^T, S_\omega) = \log \left( \frac{N-K+i}{i} \right) - I(X_{S_\omega}^T; Y^T|X_{S_\omega}^T, S_\omega),$$

which completes the proof of Theorem 3.2.
where for continuous variables and/or observations we can replace the (conditional) entropy expressions with differential (conditional) entropy. (a) follows from the fact that $X^T$ is independent of $S^2$ and $\omega$; (b) follows from the fact that conditioning on $\omega$ includes conditioning on $S^2$; (c) follows from the fact that conditioning on less variables increases entropy and for the second term that $Y^T$ depends on $\omega$ only through $X^T_{S^1}$. (d) follows by noting that $S^1$ does not give any additional information on $Y^T$ when $X^T_{S^1}$ is marginalized, because of our assumption that the observation model is independent of indices themselves except through the variables and we have symmetrically distributed variables, and therefore $H(Y^T|X^T_{S^1}, S^2) = H(Y^T|X^T_{S^1}, S_\omega)$.

Note that we can further decompose $I(X^T_{S^1}; Y^T|X^T_{S^2}, S)$ using the following chain of equalities:

$$I(X^T_{S^1}; Y^T|X^T_{S^2}, S) + I(\beta_S; X^T_{S^1}|X^T_{S^2}, Y^T, S) = I(X^T_{S^1}; Y^T, \beta_S|X^T_{S^2}, S)$$

$$= I(X^T_{S^1}; \beta_S|X^T_{S^2}, S) + I(X^T_{S^1}; Y^T|X^T_{S^2}, \beta_S, S) = T I(X^T_{S^1}; Y|X^T_{S^2}, \beta_S, S),$$

where the last equality follows from the independence of $X$ and $\beta_S$, and the independence of the $(X^T, Y^T)$ pairs over $t$ given $\beta_S$. Therefore we have

$$\frac{I(X^T_{S^1}; Y^T|X^T_{S^2}, S)}{T} = I(X^T_{S^1}; Y|X^T_{S^2}, \beta_S, S) - \frac{I(\beta_S; X^T_{S^1}|X^T_{S^2}, Y^T, S)}{T}. \quad (A.9)$$

From (A.8) and (A.9), it then follows using standard Fano bound arguments (cf. [3,6]) that

$$T \geq \max_{i=1,...,K} \log \left( \frac{N-K+i}{i} \right) \frac{\log \left( N-K+i \right)}{I(X^T_{S^1}; Y|X^T_{S^2}, \beta_S, S)} - \frac{I(\beta_S; X^T_{S^1}|X^T_{S^2}, Y^T, S)}{T}$$

is a necessary condition for $P_e$ to not be strictly greater than zero. Finally, since $I(\beta_S; X^T_{S^1}|X^T_{S^2}, Y^T, S_\omega) \geq 0$, the expression in (10) is a lower bound to the expression above, proving that (10) is a necessary condition for recovery.

### A.3 Proof of Theorem 3.3

As in the proof of Theorem 3.1, we need to show that

$$T f(\rho) = T E_0(\rho) - \rho \log \left( \frac{N-K}{i} \right) - \log \left( \frac{N-K}{i} \right) - \log K \to \infty,$$

for some $0 \leq \rho \leq 1$ and for all $i = 1, \ldots, K$. Let $I_T \triangleq I(X^T_{S^1}; Y^T|X^T_{S^2}, S)$ and note that $E'_0(0) = I_T/T$, which can be derived similar to (A.5). So we have $E_0(\rho) \geq \rho I_T - \frac{\rho^2}{2} |E'_0(\psi)|$ for some $\psi \in [0, \rho]$.

Then, using above inequality and noting that $\log K + \log \left( \frac{N-K}{i} \right) \leq 2 \log \left( \frac{N-K}{i} \right)$ for $K < N/2$, we can write

$$T f(\rho) \geq \rho I_T - \frac{\rho^2}{2} |E'_0(\psi)| - \rho \log \left( \frac{N-K}{i} \right) - 2 \log \left( \frac{N-K}{i} \right),$$

and then using (11), asymptotically we have

$$T f(\rho) \geq \rho I_T - \frac{\rho^2}{2} T |E'_0(\psi)| - (\rho + 2) \frac{I_T}{\tau_N} = \rho(1 - 1/\tau_N) I_T - \frac{\rho^2}{2} T |E'_0(\psi)| - 2 \frac{I_T}{\tau_N}.$$

Using the condition on $|E'_0(\rho)|$, we upper bound it by $\frac{\tau_N I_T}{3T}$ to obtain

$$T f(\rho) \geq \rho(1 - 1/\tau_N - \rho \tau_N/10) I_T - 2 \frac{I_T}{\tau_N} = \left( 1 - \frac{1}{\tau_N} - \rho \frac{\tau_N}{10} - \frac{2}{\rho \tau_N} \right) I_T,$$

and by choosing $\rho = 4/\tau_N$, (which is $\leq 1$ for $\tau_N$ as chosen), we write

$$T f(\rho) \geq \rho(1 - 1/\tau_N - 2/5 - 1/2) I_T = \frac{4}{\tau_N} (1/10 - 1/\tau_N) I_T.$$
Finally, again using (11), for any \( \tau_N > 10 \) we have
\[
Tf(\rho) \geq 4(1/10 - 1/\tau_N) \log \left( \frac{N-K}{i} \right) \rightarrow \infty,
\]
and therefore \( \lim_{K,N \rightarrow \infty} P(E) = 0. \)

A.4 Proof of Theorem 3.4 and Lemma 3.2

Similar to the proof of Theorem 3.1 let \( I_b \triangleq I(X_{S1}; Y | X_{S2}, \beta_S = b, S) - \frac{K}{T_{\text{min}}} \) and \( I \triangleq I(X_{S1}; Y | X_{S2}, \beta_{\text{min}}, S) - \frac{K}{T_{\text{min}}} \). Then, with the Taylor expansion of \( \bar{E}_o(\rho) \) in Section A.1 we have
\[
Tf(\rho) \geq T \bar{f}(\rho) \geq T \left( \min_b \left[ \rho I_b + \frac{\rho^2}{2} E''_o(\psi, b) \right] - \rho \frac{\log (N-K)}{T} - \frac{\log (K)}{T} \right) \quad (A.10)
\]
and our aim is to show that the above quantity approaches infinity for some \( \rho \in [0, 1] \) as \( K, N \rightarrow \infty \). Note that we did not omit the \( \frac{\log (K) + \log K}{T} \) term as in (A.1) since \( K \) is also scaling with \( N \).

Now assume that \( T \) satisfies
\[
T > \tau_N \cdot \frac{\log (N-K)}{I}, \quad (A.11)
\]
for all \( i \), which is implied by condition (13).

Using above \( T \) and assuming the second derivative condition in (12) holds, we can then write
\[
T \bar{f}(\rho) \geq T \left( \min_b \left[ \rho I_b + \frac{\rho^2}{2} E''_o(\psi, b) \right] - \rho \frac{1}{\tau_N} I - \frac{2}{\tau_N} I \right)
\]
\[
\geq T \left( \min_b \left[ \rho I_b - \frac{\rho^2}{10} \tau_N I_b \right] - \rho \frac{1}{\tau_N} I - \frac{2}{\tau_N} I \right)
\]
\[
= T \left( \min_b \left[ \rho I_b - \frac{\rho^2}{10} \tau_N I_b \right] - \rho \frac{1}{\tau_N} I - \frac{2}{\tau_N} I \right)
\]
where in the second inequality we used (12). Then, choosing \( \rho = 4/\tau_N \), we have
\[
T \bar{f}(\rho) \geq T \left( \rho \left( \min_b \left[ I_b - \frac{2}{5} I_b \right] - \frac{1}{\tau_N} I - \frac{2}{\tau_N} I \right) \right)
\]
\[
\geq T \left( \rho \left( \frac{3}{5} I - \frac{1}{\tau_N} I \right) - \frac{2}{\tau_N} I \right)
\]
\[
= TI \left( \rho (3/5 - o(1)) - \frac{2}{\tau_N} \right) = T \frac{I}{\tau_N} (2/5 - o(1))
\]
\[
\geq \log \left( \frac{N-K}{I} \right) \Theta(1) \rightarrow \infty,
\]
showing that \( KP(E_i) \) goes to zero for all \( i \) and it follows that \( P(E) \leq \max_i KP(E_i) \) goes to zero for any \( K, N \rightarrow \infty \), therefore \( \lim_{K,N \rightarrow \infty} P(E) = 0. \)

While we have proven Theorem 3.4, below we state a lemma and note some inequalities that might be useful for bounding \( |E''_o(\rho, \beta_S)| \). To this end, we use a notation similar to the Appendix C of [6] and consider the following quantities defined in Lemma 3.2
\[
g_\rho \triangleq \left( \sum_{X_{S1}} Q(X_{S1}) p(Y, X_{S2} | X_{S1}, \beta_S) \right)^{1+\rho}, \quad u_\rho \triangleq \frac{p(Y | X_{S1}, X_{S2}, \beta_S) \cdots}{\sum_{X_{S1}} Q(X_{S1}) p(Y | X_{S1}, X_{S2}, \beta_S) \cdots}.
\]
As in the previous proofs, while we use notation for discrete variables and observations, continuous case follows by replacing sums with appropriate integrals. For the first and second derivatives of \( E_o(\rho, \beta_S) \) we have the following lemma.

**Lemma A.2.**

\[
E'_o(\rho, \beta_S) = -\sum_{Y,S} g'_\rho \sum_{Y,S} g_\rho E[u_\rho \log u_\rho], \quad E''_o(\rho, \beta_S) = \sum_{Y,S} g''_\rho \sum_{Y,S} g_\rho + \left( \sum_{Y,S} g'_\rho \right)^2,
\]

where the expectations are defined w.r.t. \( X_{S^1} \).

**Proof.** Note that \( g_\rho \geq 0 \), \( E[X_{S^1}|u_\rho] = 1 \) and \( E_o(\rho, \beta_S) = -\log \sum_{Y,S} g_\rho \). The first and second equalities follow directly from the above definitions and are established in the Appendix C of [6], along with the equality

\[
g''_\rho = \frac{g_\rho}{1+\rho} \left( E[u_\rho \log^2 u_\rho] + \rho E[u_\rho \log u_\rho]^2 \right).
\]

Since we know \( E_o(\rho, \beta_S) \) is a concave function [27], \( E''_o(\rho, \beta_S) \) is negative and as the second term in the equality is nonnegative, we can upper bound its magnitude with the magnitude of the first term,

\[
|E''_o(\rho, \beta_S)| \leq \left| \sum_{Y,S} g''_\rho \sum_{Y,S} g_\rho \right|.
\]

and the last equality follows from the nonnegativity of both numerator and denominator. We can further write

\[
(E[u_\rho \log u_\rho])^2 = \left( \sum_{X_{S^1}} Q(X_{S^1}) u_\rho \log u_\rho \right)^2 \leq \sum_{X_{S^1}} Q(X_{S^1}) u_\rho \log^2 u_\rho = E[u_\rho \log^2 u_\rho],
\]

where we used Jensen’s inequality by noting that \( \sum_{X_{S^1}} Q(X_{S^1}) u_\rho = E[u_\rho] = 1 \). Using the equation for \( E''_o(\rho, \beta_S) \) and the above bounds, we have

\[
|E''_o(\rho, \beta_S)| \leq \left| \sum_{Y,S} g''_\rho \sum_{Y,S} g_\rho E[u_\rho \log^2 u_\rho] \right| = \frac{\sum_{Y,S} g_\rho E[u_\rho \log^2 u_\rho]}{\sum_{Y,S} g_\rho}, \quad 0 \leq \rho \leq 1, \, \beta_S \in B^K.
\]

As a potentially easier to check alternative to above, we also note that we can trivially obtain

\[
|E''_o(\rho, \beta_S)| \leq \sup_{Y,S} E[u_\rho \log^2 u_\rho], \quad 0 \leq \rho \leq 1, \, \beta_S \in B^K.
\]

which proves Lemma 3.2

**A.5 Proof of Lemma A.1**

As in the previous proofs, while we use notation for discrete variables and observations, the proof below is valid for continuous models. For the error exponent as defined in [8], let

\[
g_\rho(Y^{T}, X_{S^2}^{T}) = \left( E_{\beta_S} \left[ p(Y^{T}, X_{S^2}^{T}|X_{S^1}^{T}, \beta_S) \right] \right)^{1+\rho},
\]

22
such that \( E_o(\rho) = -\frac{1}{\tau} \log \sum_{Y^T, X^T_{S_1}} g_p(Y^T, X^T_{S_2}) \). We then write the following chain of inequalities:

$$
\sum_{Y^T, X^T_{S_2}} g_p(Y^T, X^T_{S_2}) = \sum_{Y^T, X^T_{S_2}} \left( E_{X^T_{S_1}} \left[ \sum_{\beta_S} p(\beta_S) p(Y^T, X^T_{S_2} | X^T_{S_1}, \beta_S) \right] \right)^{1+\rho} \\
\leq \sum_{Y^T, X^T_{S_2}} \left( E_{X^T_{S_1}} \left[ \sum_{\beta_S} p(\beta_S) \frac{1}{1+\tau} p(Y^T, X^T_{S_2} | X^T_{S_1}, \beta_S) \right] \right)^{1+\rho} \\
= \sum_{Y^T, X^T_{S_2}} \left( \sum_{\beta_S} p(\beta_S) \frac{1}{1+\tau} E_{X^T_{S_1}} \left[ p(Y^T, X^T_{S_2} | X^T_{S_1}, \beta_S) \right] \right)^{1+\rho} \\
\leq R_\beta \sum_{Y^T, X^T_{S_2}} \left( \sum_{\beta_S} p(\beta_S) E_{X^T_{S_1}} \left[ p(Y^T, X^T_{S_2} | X^T_{S_1}, \beta_S) \right] \right)^{1+\rho} \\
\leq R_\beta \sum_{Y^T, X^T_{S_2}} \sum_{\beta_S} p(\beta_S) \left( E_{X^T_{S_1}} \left[ p(Y^T, X^T_{S_2} | X^T_{S_1}, \beta_S) \right] \right)^{1+\rho} \\
= R_\beta \sum_{\beta_S} p(\beta_S) \left( \sum_{Y^T, X^T_{S_2}} \left( E_{X^T_{S_1}} \left[ p(Y, X_{S_2} | X_{S_1}, \beta_S) \right] \right)^{1+\rho} \right)^T,
$$

where the first inequality follows from the subadditivity of exponentiating with \( \frac{1}{1+\rho} \) and the second follows by multiplying and dividing inside the sum by \( p(\beta_S)^{\rho/(1+\rho)} \) and upper bounding by defining \( R_\beta = \max_{\beta_S} p(\beta_S)^{-\rho} \). The third inequality follows using Jensen’s inequality. We obtain the final expression by noting that the expression in the square brackets factorizes over \( t = 1, \ldots, T \) and is IID over \( t \) when conditioned on \( \beta_S \).

Noting that \( R_\beta = \max_{\beta \in S} p(\beta)^{-K_\rho} \) since \( \beta_S \) is IID, we then have

$$
E_o(\rho) \geq \frac{1}{T} \log R_\beta - \frac{1}{T} \log \sum_{\beta_S} p(\beta_S) \left( \sum_{Y^T, X^T_{S_2}} \left( E_{X^T_{S_1}} \left[ p(Y, X_{S_2} | X_{S_1}, \beta_S) \right] \right)^{1+\rho} \right)^T \\
= -\frac{\rho K}{T \rho_{\min}} - \frac{1}{T} \log \sum_{\beta_S} p(\beta_S) \left( \sum_{Y^T, X^T_{S_2}} \left( E_{X^T_{S_1}} \left[ p(Y, X_{S_2} | X_{S_1}, \beta_S) \right] \right)^{1+\rho} \right)^T \\
\geq -\frac{\rho K}{T \rho_{\min}} - \frac{1}{T} \log \max_{\beta_S} \left( \sum_{Y^T, X^T_{S_2}} \left( E_{X^T_{S_1}} \left[ p(Y, X_{S_2} | X_{S_1}, \beta_S) \right] \right)^{1+\rho} \right)^T \\
= -\frac{\rho K}{T \rho_{\min}} + \min_{\beta_S} \left( -\log \left( \sum_{Y^T, X^T_{S_2}} \left( E_{X^T_{S_1}} \left[ p(Y, X_{S_2} | X_{S_1}, \beta_S) \right] \right)^{1+\rho} \right) \right) \\
= -\frac{\rho K}{T \rho_{\min}} + \min_{\beta_S} E_o(\rho, \beta_S),
$$

23
which follows from the definition of $E_o(\rho, \beta_S)$. $p_{\text{min}} = \Theta(1)$ follows from our assumption that $p(\beta)$ is lower bounded by a constant on its support independent of scaling variables $K$, $N$ and $T$.

A.6 Proof of Lemma 3.1 and extension to continuous models

Proof of Lemma 3.1

As we note in the main section, the proof of the lemma follows along the proof of Lemma III.1 in [6] which considers binary alphabets in [7], yet readily generalizes to discrete alphabets for $X$ and $Y$. However, because of the latent variables $\beta_S$, the final step in the bottom of p. 1888 of [6] does not hold true and the proof ends with the previous equation. As a result, our error exponent $E_o(\rho)$ is different than the error exponent in [6].

Furthermore, note that Lemma 3.1 is missing a $\rho$ multiplying the $\log(\epsilon)$ term compared to Lemma III.1 in [6]; this is due to the fact that we do not utilize the stronger proof argument in Appendix A of [6], but rather follow from the argument provided in the proof in the main body. Following that proof, Lemma 3.1 can be obtained by modifying inequality (c) in p. 1887 that upper bounds $\Pr[E_i|\omega_0 = 1, X_{S_1}, Y^T]$ such that $(N-K)^{i}$ is replaced with $(N-K)^{\rho}$.

Continuous models

Even though the results and proof ideas that were used in Sections 3.1 and 3.2 are fairly general, the proofs for the results in Section 3.1 specifically the proof of Lemma 3.1 were stated for discrete variables and outcomes. In this section we make the necessary generalizations to extend these proofs to continuous variable and observation models. We follow the methodology in [27] and [38].

To simplify the exposition, we consider the extension to continuous variables in the special case of fixed and known $\beta_S$. In that case, $E_o(\rho)$ as defined in [8] reduces to

$$E_o(\rho) = -\log \sum_{Y} \sum_{X_S^2} \sum_{X_{S_1}} Q(X_{S_1}) p(Y, X_{S_2}|X_{S_1}) \frac{1}{1+\rho}$$

subject to $0 \leq \rho \leq 1$ (A.15) with $\frac{\partial E_o(\rho)}{\partial \rho} \bigg|_{\rho=0} = I(X_{S_1}; X_{S_2}, Y|S) = I(X_{S_1}; Y|X_{S_2}, S)$, since $(X^{(t)}, Y^{(t)})$ pairs are independent across $t$ for fixed $\beta_S$.

We assume a continuous and bounded joint probability density function $Q(X)$ with joint cumulative distribution function $F$. The conditional probability density $p(Y = y|X_S = x)$ for the observation model is assumed to be a continuous and bounded function of both $x$ and $y$.

Let $X' \in \mathcal{X}'^N$ be the random vector and $Y' \in \mathcal{Y}'$ be the random variable generated by the quantization of $X \in \mathcal{X}^N = \mathbb{R}^N$ and $Y \in \mathcal{Y} = \mathbb{R}$, respectively, where each variable in $X$ is quantized to $L$ values and $Y$ quantized to $J$ values. Let $F'$ be the joint cumulative distribution function of $X'$. As before, let $S(X'T, Y'T)$ be the ML decoder with continuous inputs with probability of making $i$ errors in decoding denoted by $P'(E_i)$. Let $\tilde{S}(X'T, Y'T)$ be the ML decoder that quantizes inputs $X'T$ and $Y'T$ to $X'T$ and $Y'T$, and have the corresponding probability of error $P'(E_i)$. Define

$$E_o(\rho, X', Y') = -\log \sum_{y' \in \mathcal{Y}'} \sum_{x'_{S_2} \in \mathcal{X}'^{K-1}} \sum_{x'_{S_1} \in \mathcal{X}'^i} Q(x'_{S_1}) p(y', x'_{S_2}|x'_{S_1}) \frac{1}{1+\rho}$$

$$E_o(\rho, X, Y) = -\log \int_{\mathcal{Y}} \int_{X^{K-1}} \int_{X^i} Q(x_{S_1}) p(y, x_{S_2}|x_{S_1}) \frac{1}{1+\rho} dx_{S_1} d x_{S_2}$$

where the indexing denotes the random variables that the error exponents are computed with respect to.

Utilizing the results in Section 3.1 for discrete models, we will show that Lemma 3.1 holds for the continuous model, i.e.,

$$P(E_i) \leq e^{-(TE_o(\rho,X,Y) - \rho \log (N-K) - \log (K))}.$$  (A.16)
The rest of the proof of Theorem 3.1 will then follow as in the discrete case, by noting that
\[ \frac{\partial E_o(\rho, X, Y)}{\partial \rho} \bigg|_{\rho=0} = I(X; Y|X'_2, S), \]
with the mutual information definition for continuous variables.

Our approach can be described as follows. We will increase the number of quantization levels for \( Y' \) and \( X' \), respectively. Then, since the discrete result in (9) holds for any number of quantization levels, by taking limits we will be able to show that
\[ P'(E_i) \leq e^{-\left(TE_o(\rho, X, Y) - \rho \log \left( \frac{N-K}{K} \right) \right)}. \tag{A.17} \]

Since \( \hat{S}(X^T, Y^T) \) is the minimum probability of error decoder, any upper bound for \( P'(E_i) \) will also be an upper bound for \( P(E_i) \), thereby proving (A.16).

Assume \( Y \) is quantized with the quantization boundaries denoted by \( a_1, \ldots, a_{J-1} \), with \( Y' = a_j \) if \( a_{j-1} < Y \leq a_j \). For convenience denote \( a_0 = -\infty \) and \( a_J = \infty \). Furthermore, assume the quantization boundaries are equally spaced, i.e. \( a_j - a_{j-1} = \Delta_j \) for \( 2 \leq j \leq J - 1 \). Now, we have that
\[
E_o(\rho, X', Y') = -\log \sum_{j=1}^{J} \sum_{x'_s} \left[ \sum_{x'_{s1}} Q(x'_{s1}) \left( \int_{a_{j-1}}^{a_j} p(y, x'_{s2}|x'_{s1}) \, dy \right) \right]^{1+\rho} \\
= -\log \left\{ \sum_{j=2}^{J-1} \sum_{x'_{s2}} \left[ \sum_{x'_{s1}} Q(x'_{s1}) \left( \int_{a_{j-1}}^{a_j} \frac{p(y, x'_{s2}|x'_{s1}) \, dy}{\Delta_j} \right) \right]^{1+\rho} \\
+ \sum_{x'_{s2}} \left[ \sum_{x'_{s1}} Q(x'_{s1}) \left( \int_{-\infty}^{a_1} \frac{p(y, x'_{s2}|x'_{s1}) \, dy}{\Delta_j} \right) \right]^{1+\rho} \\
+ \sum_{x'_{s2}} \left[ \sum_{x'_{s1}} Q(x'_{s1}) \left( \int_{a_{j-1}}^{\infty} \frac{p(y, x'_{s2}|x'_{s1}) \, dy}{\Delta_j} \right) \right]^{1+\rho} \right\}. \tag{A.19} \tag{A.20} \tag{A.21}
\]

Let \( J \to \infty \) and for each \( J \) choose the sequence of quantization boundaries such that \( \lim \Delta_j = 0 \), \( \lim a_{J-1} = \infty \), \( \lim a_1 = -\infty \). Then the last two terms disappear and using the fundamental theorem of calculus, we obtain
\[
\lim_{J \to \infty} E_o(\rho, X', Y') = E_o(\rho, X', Y) = -\log \int_{y} \sum_{x'_{s2}} \left[ \sum_{x'_{s1}} Q(x'_{s1}) p(y, x'_{s2}|x'_{s1}) \right]^{1+\rho} \, dy. \tag{A.22}
\]

It can also be shown that \( E_o(\rho, X', Y') \) increases for finer quantizations of \( Y' \), therefore \( E_o(\rho, X', Y) \) gives the smallest upper bound over \( P'(E_i) \) over the quantizations of \( Y \), similar to (27). However, this is not necessary for the proof.

We repeat the same procedure for \( X \). Assume each variable \( X_n \) in \( X \) is quantized with the quantization boundaries denoted by \( b_1, \ldots, b_{L-1} \), with \( X'_n = b_l \) if \( b_{l-1} < X_n \leq b_l \). For convenience denote \( b_0 = -\infty \) and \( b_L = \infty \). Furthermore, assume that the quantization boundaries are equally spaced, i.e. \( b_l - b_{l-1} = \Delta_L \) for
2 \leq l \leq L - 1. \text{ Then we can write}

\begin{align*}
E_o(\rho, X', Y) &= -\log \int_y \left[ \sum_{l=1}^{L} \int_{x_{S^1}} Q(x_{S^1}) \left( \int_{b_{l-1}}^{b_l} p(y, x_{S^2} | x'_{S^1}) \, dx_{S^2} \right) \right]^{1+\rho} \, dy \\
&= -\log \int_y \left[ \sum_{l=1}^{L} \left( \int_{x_{S^1}} \left( \int_{b_{l-1}}^{b_l} p(y, x_{S^2} | x_{S^1}) \, dx_{S^2} \right)^{1+\rho} \, dF'(x_{S^1}) \right) \right] \, dy \\
&= -\log \left[ \sum_{l=2}^{L-1} \int_{x_{S^1}} \left( \int_{b_{l-1}}^{b_l} p(y, x_{S^2} | x_{S^1}) \, dx_{S^2} \right)^{1+\rho} \, dF'(x_{S^1}) \right] \, dy,
\end{align*}

where (A.24) follows with $F'(x_{S^1})$ being the step function that represents the cumulative distribution function of the quantized variables $X_{S^1}$.

Let $L \to \infty$, for each $L$ choose a set of quantization points such that $\lim \Delta_L = 0$, $\lim b_{L-1} = \infty$, $\lim b_1 = -\infty$. Again, the second and third terms disappear and the first sum converges to the integral over $X_{S^1}$. Note that $p(y, x_{S^2} | x_{S^1})$ is a bounded continuous function of all its variables since it was assumed that $Q(x)$ and $p(y|x)$ were bounded and continuous. Also note that $\lim_{L \to \infty} F' = F$, which implies the weak convergence of the probability measure of $X'$ to the probability measure of $X$. Given these facts, using the portmanteau theorem we obtain that $E_F [p(Y, X_{S^2} | X_{S^1})] \to E_F [p(Y, X_{S^2} | X_{S^1})]$, which leads to

$$
\lim_{L \to \infty} E_o(\rho, X', Y) = -\log \int_y \int_{x_{S^1}} \left( \int_{x_{S^1}} p(y, x_{S^2} | x_{S^1})^1 \, dF'(x_{S^1}) \right)^{1+\rho} \, dx_{S^2} \, dy = E_o(\rho, X, Y).
$$

This leads to the following result, completing the proof.

$$
P(E_i) \leq P'(E_i) \leq \lim_{J,L \to \infty} e^{-(T E_o(\rho, X', Y') - \rho \log (N, K) - \log (L))} = e^{-(T E_o(\rho, X, Y) - \rho \log (N, K) - \log (L))}.
$$

### A.7 Proof of Theorem 4.1

We first compute the mutual information $I(X_{S_1}; Y | X_{S^2}, \beta_S, S)$ for a fixed set $S$ (see footnote\(^1\)) and omit the explicit conditioning on $S$ on all expressions below.

\begin{align*}
I(X_{S_1}; Y | X_{S^2}, \beta_S) &= h(Y | X_{S^2}, \beta_S) - h(Y | X_{S}, \beta_S) \\
&= h(X_{S_1}^\top \beta_{S_1} + W | \beta_{S_1}) - h(W) \\
&= E_{\beta_{S_1}} \left[ \frac{1}{2} \log \left( 2\pi e \left( \text{var} \left( X_{S_1}^\top \beta_{S_1} | \beta_{S_1} \right) + \frac{1}{\text{SNR}} \right) \right) \right] - \frac{1}{2} \log \left( 2\pi e \frac{1}{\text{SNR}} \right) \\
&= E_{\beta_{S_1}} \left[ \frac{1}{2} \log \left( 1 + \frac{\beta_{S_1}^\top \beta_{S_1} \text{SNR}}{T} \right) \right],
\end{align*}

where the second equality follows from the independence of $X_{S_1}$ and $X_{S^2}$ and the last equality follows from the fact that $\text{var}(X_{S_1}^\top \beta_{S_1} | \beta_{S_1}) = \beta_{S_1}^\top E[X_{S_1} X_{S_1}^\top] \beta_{S_1} = \frac{\beta_{S_1}^\top \beta_{S_1}}{T}$. 

\[26\]
Using the mutual information expression above, a necessary condition on \( T \) for recovery is given by Theorem 3.2

\[
T \geq \max_{i=1,\ldots,K} \frac{\log \left( \frac{N-K+i}{i} \right)}{E_{\beta_{G}} \left[ \frac{1}{2} \log \left( 1 + \frac{\beta_{G}^T \beta_{G} \text{SNR}}{T} \right) \right]}.
\]  
(A.28)

We first show that SNR = \( \Omega(\log N) \) is necessary for recovery. For any \( N, K \) or SNR assume \( T \) scales much faster, e.g. \( T = \omega(K\sigma^2\text{SNR}) \), such that

\[
E_{\beta_{G}} \left[ \log \left( 1 + \frac{\beta_{G}^T \beta_{G} \text{SNR}}{T} \right) \right] \approx E_{\beta_{G}} \left[ \frac{\beta_{G}^T \beta_{G} \text{SNR}}{T} \right] = \frac{i\sigma^2\text{SNR}}{T},
\]

since \( \log(1+x) = \Theta(x) \) for \( x \to 0 \). Then, the necessary condition given by (A.28) is

\[
T > 2 \max_{i} \frac{\log \left( \frac{N-K+i}{i} \right)}{i\sigma^2\text{SNR}}
\]

which readily leads to the condition that

\[
\text{SNR} > 2 \max_{i} \frac{\log \left( \frac{N-K+i}{i} \right)}{i\sigma^2} \times \max \log(N/i) = \log N
\]  
(A.29)

for \( \sigma^2 \) constant. Note that if the above condition is necessary for any \( T = \omega(K\sigma^2\text{SNR}) \), it is also necessary for smaller scalings of \( T \).

To obtain a necessary condition on \( T \), note that

\[
E_{\beta_{G1}} \left[ \log \left( 1 + \frac{\beta_{G1}^T \beta_{G1} \text{SNR}}{T} \right) \right] \leq \log \left( E_{\beta_{G1}} \left[ 1 + \frac{\beta_{G1}^T \beta_{G1} \text{SNR}}{T} \right] \right) = \log \left( 1 + \frac{i\sigma^2\text{SNR}}{T} \right)
\]
due to Jensen’s inequality, therefore the following is also a necessary condition from Theorem 3.2

\[
T > \max_{i} \frac{\log \left( \frac{N-K+i}{i} \right)}{i\sigma^2},
\]  
(A.30)

and considering only the case \( i = K \), we have the necessary condition that

\[
T \geq \frac{\log \left( \frac{N}{K} \right)}{i\sigma^2} \approx \frac{K \log(N/K)}{\log \left( 1 + \frac{\sigma^2\text{SNR}}{T} \right)}.
\]  
(A.31)

Assume SNR = \( \Theta(\log(N/K)) \), which corresponds to the case \( i = K \) in (A.29). It is then clear that (A.31) does not hold for \( T = o(K \log(N/K)) \), since \( \frac{K \log(N/K)}{\log \left( 1 + \frac{\sigma^2\text{SNR}}{T} \right)} \geq \log \left( 1 + \frac{\sigma^2\text{SNR}}{T} \right) \) asymptotically, for \( \sigma^2 \) constant. However for \( T = \Omega(K \log(N/K)) \), the condition (A.31) is

\[
T = \Omega \left( \frac{K \log(N/K)}{\log \left( 1 + \sigma^2 \right)} \right),
\]

which proves the lower bound in Theorem 4.1. Note that we can assume \( \sigma^2 \) constant without loss of generality, since otherwise its scaling can be incorporated into SNR to obtain and analyze an equivalent model (where SNR\( \sigma^2 \) is equal in the two models).

We now show that \( T = \Omega \left( \frac{K \log(N/K)}{\log(1 + \beta_{\min}^2)} \right) \) is a sufficient condition for recovery. Note that for SNR = \( \Theta(\log N) \) and \( T = \Theta(K \log N) \) (which is equivalent to \( \Theta(K \log(N/K)) \) for \( K = o(N) \)), the normalized model we consider is equivalent to a non-scaling model with \( Q(X_k) \sim N(0, c/K) \) for a constant \( c \) and \( W_t \sim N(0, 1) \)
for all fixed $K$ and therefore we can use the results of Theorem 3.1. For the worst-case mutual information, we write

$$I(X_{S^1}; Y|X_{S^2}, \beta_{\min}, S) = \min_{\beta_S} \frac{1}{2} \log \left(1 + \frac{\beta_S^T \beta_S \text{SNR}}{T}\right) = \frac{1}{2} \log \left(1 + \frac{i b_{\min}^2 \text{SNR}}{T}\right),$$

where the minimum is achieved for any $|\beta_S| = b_{\min} 1_K$. Then, a sufficient condition given by Theorem 3.1 is

$$T > \Theta \left( \max_{i} \frac{i \log(N/i)}{\log(1 + \frac{1}{2} b_{\min}^2)} \right) = \Theta \left( \frac{K \log(N/K)}{\log(1 + b_{\min}^2)} \right)$$

which is satisfied for assumed $T$ and thus we have the upper bound for Theorem 4.1.

### A.8 Proof of Theorem 4.3

To simplify the analysis and exposition, we analyze the degenerate case of $\beta \in \{0, 1\}^N$, i.e. the latent variable $\beta_S$ is known and equal to the vector of 1's. However, the general case where $\beta_S$ are random with a known distribution can also be analyzed using the condition given by Theorem 3.1. In order to obtain the model-specific bounds, we analyze the mutual information term $I(X_{S^1}; Y|X_{S^2}, \beta_{\min}, S) = I(X_{S^1}; Y|X_{S^2}, \beta_S, S)$ for a fixed set $S$ and omit the explicit conditioning on $S$ (see footnote 1), which reduces to $I(X_{S^1}; Y|X_{S^2})$ for this case.

We write the mutual information term as

$$I(X_{S^1}; Y|X_{S^2}) = H(Y|X_{S^2}) - H(Y|X_S)$$

where we will analyze $H(Y|X_{S^2})$ and $H(Y|X_S)$ to obtain a lower bound for the mutual information expression.

Defining $Z_1 = \sum_{j \in S^1} X_j$, $Z_2 = \sum_{j \in S^2} X_j$ and $Z = Z_1 + Z_2$, we have $H(Y|X_{S^2}) = H(Y|Z_2)$ since the quantizer input $X\beta + W$ depends only on the sum of the elements of $X_S$. Note that $Z_2 \sim \mathcal{N}(0, D^2)$ with $D^2 = K - i$. Now we explicitly write the conditional entropy

$$H(Y|Z_2) = \int_{-\infty}^{\infty} P_{Z_2}(z) H(Y|Z_2 = z) \, dz = \int_{-\infty}^{\infty} P_{Z_2}(z) \left( p_1 \log \frac{1}{p_1} + p_0 \log \frac{1}{p_0} \right) \, dz \tag{A.32}$$

with $p_1 \triangleq \Pr[Y = 1|Z_2 = z]$ and $p_0 \triangleq 1 - p_1 = \Pr[Y = 0|Z_2 = z]$, which can be written as

$$p_1 = \Pr[Z_1 + Z_2 + W \geq 0|Z_2 = z] = \Pr[Z_1 + W \geq -z] = \Pr[\mathcal{N}(0, S^2) \geq -z] = Q \left( \frac{-z}{S} \right)$$

$$p_0 = \Pr[Z_1 + Z_2 + W < 0|Z_2 = z] = \Pr[Z_1 + W < -z] = \Pr[\mathcal{N}(0, S^2) < -z] = Q \left( \frac{z}{S} \right)$$

where $S^2 = i + 1$ and the $Q$ function defined as $Q(x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt$.

To lower bound $H(Y|Z)$, we make use of the following inequalities for $x > 0$ [39,40]:

$$\frac{1}{12} e^{-x^2} \leq Q(x) \leq \frac{1}{2} e^{-\frac{x^2}{2}} \tag{A.33}$$

$$\log 2 + \frac{x^2}{2} \leq \log(2e x^2) \leq \log \frac{1}{Q(x)} \leq \log 12 + x^2 \tag{A.34}$$
Then we write the following chain of inequalities:

\[ H(Y|Z_2) = 2 \int_0^\infty P_Z(z) \left( p_1 \log \frac{1}{p_1} + p_0 \log \frac{1}{p_0} \right) \, dz \]  

\[
\geq 2 \int_0^\infty P_Z(z) \cdot p_0 \log \frac{1}{p_0} \, dz
\]  

\[
\geq 2 \int_0^\infty \frac{1}{\sqrt{2\pi}D^2} \cdot e^{-z^2/2} \cdot \frac{1}{12} \cdot e^{-z^2} \cdot \left( \log 2 + \frac{z^2}{2S^2} \right) \, dz
\]  

\[
= \frac{1}{12\sqrt{2\pi}D} \int_{-\infty}^\infty e^{-A z^2} \cdot \left( \log 2 + \frac{z^2}{2S^2} \right) \, dz
\]  

\[
= \frac{1}{12\sqrt{2\pi}D} \left( \log 2 \frac{\sqrt{2\pi}}{\sqrt{A}} + \frac{\sqrt{\pi/2}}{A^{3/2}S^2} \right) = \frac{\log 2}{12\sqrt{AD}} + \frac{1}{24A^{3/2}DS^2}
\]  

Equality \(A.35\) follows from the evenness of the function inside the integral and we write \(A.36\) by noting that \(p_1 \log \frac{1}{p_1}\) and \(P_Z(z)\) are non-negative. \(P_Z(z)\) is expanded and the above bounds for the \(Q\) function are used to obtain \(A.37\) and \(A.38\) is a regrouping of terms by defining \(A = \frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}}\) and rewriting the limits of the integral by noting that the integrand is an even function. We obtain \(A.39\) by evaluating the integral. For \(A\), we have

\[ A = \frac{1}{K-i} + \frac{2}{i+1} = \frac{2K-i+1}{(i+1)(K-i)} \]

and replacing \(A\), \(D\) and \(S\), we can then write

\[ H(Y|X_S) = H(Y|Z_2) \geq c_1 \frac{\sqrt{i+1} \sqrt{K-i}}{\sqrt{2K-i} + \sqrt{K-i}} + c_2 \frac{(i+1)^{3/2}(K-i)^{3/2}}{(2K-i+1)^{3/2} \sqrt{K-i}(i+1)} \]

\[ \geq c \cdot \left( \sqrt{\alpha + \frac{1}{K}} + (1-\alpha) \sqrt{\alpha + \frac{1}{K}} \right) = \Omega \left( \sqrt{\alpha + \frac{1}{K}} \right) \]

for constants \(c, c_1, c_2 > 0\).

We now analyze the second term \(H(Y|X_S)\) to obtain an upper bound. Again, note that \(H(Y|X_S) = H(Y|Z)\), then

\[ H(Y|Z) = \int_0^\infty P_Z(z) H(Y|Z = z) \, dz = \int_0^\infty P_Z(z) \left( p_1 \log \frac{1}{p_1} + p_0 \log \frac{1}{p_0} \right) \, dz \]

where this time we define \(p_1 \triangleq \Pr[Y = 1|Z = z]\) and \(p_0 \triangleq \Pr[Y = 0|Z = z]\), which can be written as

\[ p_1 = \Pr[Z + W \geq z|Z = z] = \Pr[W \geq -z] = \Pr[V(0,1) \geq -z] = Q(-z) \]
\[ p_0 = \Pr[Z + W < z|Z = z] = \Pr[W < -z] = \Pr[V(0,1) < -z] = Q(z) \]

Then, write the following chain of inequalities:

\[ H(Y|Z) = 2 \int_0^\infty P_Z(z) \left( p_1 \log \frac{1}{p_1} + (1-p_1) \log \frac{1}{1-p_1} \right) \, dz \]  

\[
\leq 4 \int_0^\infty \frac{1}{\sqrt{2\pi}K} e^{-z^2/2} \frac{1}{2} e^{-z^2/2} \left( \log 12 + \frac{z^2}{2} \right) \, dz
\]  

\[
= \frac{1}{\sqrt{2\pi}K} \int_{-\infty}^{\infty} e^{-Bz^2} \left( \log 12 + \frac{z^2}{2} \right) \, dz
\]  

\[
= \frac{\log 12 \sqrt{2\pi}}{\sqrt{B} \sqrt{2\pi}K \sqrt{2B^{3/2}}} + \frac{1}{\sqrt{2\pi}K \sqrt{2B^{3/2}}} = \log 12 + \frac{1}{2\sqrt{B}} + \frac{1}{24A^{3/2}}
\]  

(41)
Equality \[ A.41 \] follows from the evenness of the function inside the integral and we write \[ A.42 \] by noting that \( p \log \frac{1}{p} \geq (1-p) \log \frac{1}{1-p} \) for \( 0 \leq p \leq \frac{1}{2} \). \( P_2(z) \) is expanded and the above bounds for the \( Q \) function are used to obtain \[ A.43 \] and \[ A.44 \] is a regrouping of terms by defining \( B = \frac{1}{\alpha} + 1 \) and rewriting the limits of the integral by noting that the integrand is an even function. We obtain \[ A.39 \] by evaluating the integral. Replacing \( B \), we then have

\[
H(Y|X_S) = H(Y|Z) \leq \frac{c_1}{\sqrt{K+1}} + \frac{c_2}{(\frac{1}{\sqrt{\alpha}} + 1)\sqrt{K+1}} = O\left(\frac{1}{\sqrt{K+1}}\right)
\]  

(A.46)

for constants \( c_1, c_2 > 0 \).

Looking at \[ A.40 \] and \[ A.46 \], we have the following:

\[
I(X_{S'; Y}|X_{S''}) = H(Y|X_{S''}) - H(Y|X_S) = \Omega(\sqrt{\alpha}).
\]  

(A.47)

Finally, since \( \log \binom{N-K}{i} = \Theta(i \log N) \), we can write

\[
\frac{\log \binom{N-K}{i}}{I(X_{S'; Y}|X_{S''})} = O\left(\frac{\alpha K \log N}{\sqrt{\alpha}}\right) = O(\sqrt{K \log N}) = O(K \log N)
\]

which is satisfied by \( T = \Omega(K \log N) \), proving Theorem 4.3.

### A.9 Proof of Theorem 4.5

We compute \( I(Z_{S'; Y}|Z_{S''}, \beta_S, S) \) in terms of \( I(X_{S'; Y}|X_{S''}, \beta_S, S) \). To do that, we compute \( H(Y|Z_S, \beta_S, S) \) for any set \( S \). To simplify the expressions, we omit the conditioning on \( \beta_S \) and \( S \) in all entropy and mutual information expressions below.

\[
H(Y|Z_S) = H(Y, Z_S) - H(Z_S)
\]

(A.48)

\[
= H(Y, Z_S, X_S) - H(X_S|Y, Z_S) - (H(Z_S, X_S) - H(X_S|Z_S))
\]

(A.49)

\[
= H(Y|Z_S, X_S) - H(X_S|Y, Z_S) + H(X_S|Z_S)
\]

(A.50)

\[
= H(Y|X_S) - H(X_S|Y, Z_S) + \sum_{k \in S} H(X_k|Z_k)
\]

(A.51)

\[
= H(Y|X_S) - H(X_S|Y, Z_S) + \sum_{k \in S} (\rho H(X_k|Z_k = m) + (1-\rho)H(X_k|Z_k = X_k))
\]

(A.52)

\[
= H(Y|X_S) - H(X_S|Y, Z_S) + \sum_{k \in S} \rho H(X_k)
\]

(A.53)

\[
= H(Y|X_S) - H(X_S|Y, Z_S) + \rho H(X_S)
\]

(A.54)

(A.48), (A.49) and (A.50) follow from the chain rule of entropy. (A.51) follows from the conditional independence of \( Y \) and \( Z_S \) given \( X_S \) and the independence of \( Z_S, X_S \) over \( k \in S \). In (A.52) we explicitly write the conditional entropies for two values of \( Z_i \). These expressions simplify to (A.53) and we group the terms over \( k \in S \) to obtain (A.54).
For any set $S$ with elements $1, \ldots, |S|$, we can further write

$$H(X_S|Y, Z_S) = \sum_{k=1}^{|S|} H(X_k|Y, Z_k, \ldots, |S|-1, X_1, \ldots, k-1)$$

(A.55)

$$= \rho \sum_{k=1}^{|S|} H(X_k|Y, Z_{k+1}, \ldots, |S|-1, X_1, \ldots, k-1)$$

(A.56)

$$= \rho \sum_{k=1}^{|S|} H(X_k|Y, X_1, \ldots, k-1) - I(X_k; Z_{k+1}, \ldots, |S|)|Y, X_1, \ldots, k-1)$$

(A.57)

$$= \rho H(X_S|Y) - \rho \sum_{k=1}^{|S|} I(X_k; Z_{k+1}, \ldots, |S|)|Y, X_1, \ldots, k-1),$$

(A.58)

where (A.55) follows from the chain rule and the independence of $X_j$ and $Z_k$ given $X_k$. (A.56) by expanding the conditioning on $Z_k$, (A.57) from the definition of mutual information and (A.58) from the chain rule.

W.l.o.g., assume $S = \{1, \ldots, K\}$ and $S^2 = \{1, \ldots, K - i\}$. Finally, using the above expressions we have

$$I(Z_{S^1}; Y|Z_{S^2}) = H(Y|Z_{S^2}) - H(Y|Z_S)$$

$$= H(Y|X_{S^2}) - H(Y|X_S) + \rho H(X_{S^2}) - H(X_S) - \rho H(X_{S^2}|Y) - H(X_S|Y))$$

$$+ \rho \left( \sum_{k=1}^{K-i} I(X_k; Z_{k+1}, \ldots, K-1|Y, X_1, \ldots, k-1) - \sum_{k=1}^{K} I(X_k; Z_{k+1}, \ldots, K|Y, X_1, \ldots, k-1) \right)$$

$$= I(X_{S^1}; Y|X_{S^2}) + \rho I(X_{S^2}; Y) - I(X_S; Y))$$

$$- \rho \left( \sum_{k=1}^{K-i} I(X_k; Z_{k+i+1}, \ldots, K|Y, X_1, \ldots, k-1, Z_{k+1}, \ldots, K-1) + \sum_{k=K-i+1}^{K} I(X_k; Z_{k+1}, \ldots, K|Y, X_1, \ldots, k-1) \right)$$

$$\leq I(X_{S^1}; Y|X_{S^2}) + \rho (H(Y|X_S) - H(Y|X_{S^2})) = (1 - \rho)I(X_{S^1}; Y|X_{S^2}).$$

The first two equalities follow from the expressions we found earlier and the third equality follows from the definition of the mutual information by rearranging the sums and using the chain rule of mutual information. The last inequality follows from the non-negativity of mutual information and expanding the mutual information expressions. The result then follows from Theorem 3.2.

A.10 Proof of Theorem 4.6

Define $Z_1 = Z_{S^1}$ and $Z_2 = Z_{S^2}$. For simplicity of exposition, we will assume $\beta_S = \pm \beta_{\text{min}}$, however the results can be generalized to random $\beta_S$ similar to the proof of Theorem 4.1. We also assume a fixed $S$ (see footnote) and omit the explicit conditioning on $S$ in the expressions below.

To prove the theorem, we will obtain a lower bound on $I(Z_1; Y|Z_2) = h(Y|Z_2) - h(Y|Z_1, Z_2)$. Let $M_1 = \{k \in S^1 : X_k = m\}$ and $M_2 = \{k \in S^2 : X_k = m\}$ denote the set of missing features in each set, then it simply follows that $Z_1 = (\{X_k\}_{k \in S^1 \cap M_1^c}, M_1)$ and $Z_2 = (\{X_k\}_{k \in S^2 \cap M_2^c}, M_2)$.
We will start by proving an upper bound on \( h(Y|Z_1, Z_2) \). We have,

\[
  h(Y|Z_1, Z_2) = E_{Z_1, Z_2} \left[ h(X_S^T \beta_S + W|Z_1, Z_2) \right]
\]

\[
  = E_{X_{S^1, M^1}, X_{S^2, M^2}, M_1, M_2} \left[ h(X_S^T \beta_S + W|X_{S^1, M^1}, X_{S^2, M^2}, M_1, M_2) \right]
\]

\[
  = E_{M_1, M_2} \left[ h(X_M^T \beta_{M_1} + X_M^T \beta_{M_2} + W|M_1, M_2) \right]
\]

\[
  = E_{M_1, M_2} \left[ \frac{1}{2} \log \left( 2\pi e b_{\text{min}}^2 \left[ \frac{|M_1|}{T} + \frac{|M_2|}{T} + \frac{1}{\text{SNR} b_{\text{min}}^2} \right] \right) \right]
\]

\[
  \leq E_{M_2} \left[ \frac{1}{2} \log \left( 2\pi e b_{\text{min}}^2 \left[ \frac{E_{M_1} |M_1|}{T} + \frac{|M_2|}{T} + \frac{1}{\text{SNR} b_{\text{min}}^2} \right] \right) \right]
\]

\[
  = E_{M_2} \left[ \frac{1}{2} \log \left( \frac{2\pi e b_{\text{min}}^2}{T} \left[ \frac{i \rho}{T} + \frac{|M_2|}{T} + \frac{1}{\text{SNR} b_{\text{min}}^2} \right] \right) \right].
\]

The first two equalities follow by expanding \( Y \) and \( Z_1, Z_2 \). The third equality follows by subtracting the known quantities related to \( X_{S^1}, X_{S^2}, X_M \) from the entropy expression. The fourth equality follows by noting that the variable inside the entropy conditioned on \( M_1 \) and \( M_2 \) is Gaussian and then computing its variance. We use the Jensen’s inequality over \( M_1 \) by noting that \( \log \) is a concave function to obtain the inequality. We then note that \( |M_1| \) is a binomially distributed random variable with expectation \( i \rho \).

Similar to what we did for \( h(Y|Z_1, Z_2) \), we can also write

\[
  h(Y|Z_2) = E_{M_2} \left[ h(X_S^T \beta_1 + X_M^T \beta_{M_2} + W|M_2) \right] = E_{M_2} \left[ \frac{1}{2} \log \left( \frac{2\pi e b_{\text{min}}^2}{T} \left[ \frac{i \rho}{T} + \frac{|M_2|}{T} + \frac{1}{\text{SNR} b_{\text{min}}^2} \right] \right) \right].
\]

Combining the two entropy expressions, we then have

\[
  I(Z_1; Y|Z_2) \geq E_{M_2} \left[ \frac{1}{2} \log \left( \frac{2\pi e b_{\text{min}}^2}{T} \left[ \frac{i}{T} + \frac{|M_2|}{T} + \frac{1}{\text{SNR} b_{\text{min}}^2} \right] \right) \right] - \frac{1}{2} \log \left( \frac{2\pi e b_{\text{min}}^2}{T} \left[ \frac{i \rho}{T} + \frac{|M_2|}{T} + \frac{1}{\text{SNR} b_{\text{min}}^2} \right] \right)
\]

\[
  = E_{M_2} \left[ \frac{1}{2} \log \left( \frac{i + |M_2| + \frac{T}{\text{SNR} b_{\text{min}}^2}}{\rho + |M_2| + \frac{T}{\text{SNR} b_{\text{min}}^2}} \right) \right]
\]

\[
  \geq \frac{1}{2} \log \left( \frac{1 + \frac{T}{\text{SNR} b_{\text{min}}^2}}{\rho + \frac{T}{\text{SNR} b_{\text{min}}^2}} \right) = \frac{1}{2} \log \left( 1 + \frac{(1 - \rho) i}{\rho K + \frac{T}{\text{SNR} b_{\text{min}}^2}} \right)
\]

\[
  = \frac{1}{2} \log \left( 1 + \frac{(1 - \rho) i}{\rho K + \frac{T}{\text{SNR} b_{\text{min}}^2}} \right)
\]

Note that the above expression reduces to the expression for fully observed case for \( \rho = 0 \).

Let SNR = \( c \log(N/K) \) (which is weaker than the assumed SNR = \( \Omega(\log N) \)) and \( T = \frac{K \log(N/K)}{\frac{1}{2} \log \left( 1 + \frac{b_{\text{min}}^2}{1 + \rho b_{\text{min}}^2} \right)} \), as assumed in the theorem. Then, for the mutual information we have,

\[
  I(Z_1; Y|Z_2) = \frac{1}{2} \log \left( 1 + \frac{(1 - \rho) i b_{\text{min}}^2}{\rho K + \frac{T}{\text{SNR} b_{\text{min}}^2}} \right)
\]

\[
  \geq \frac{1}{2} \log \left( 1 + \frac{(1 - \rho) i b_{\text{min}}^2}{1 + \rho b_{\text{min}}^2} \right)
\]

for a large enough constant \( c > 0 \).

Using the bound we have on the mutual information and Theorem 3.1, we have the sufficient condition

\[
  T > \max_i \frac{\Theta(i \log(N/i))}{\frac{1}{2} \log \left( 1 + \frac{(1 - \rho) i b_{\text{min}}^2}{1 + \rho b_{\text{min}}^2} \right)} = \frac{\Theta(K \log(N/K))}{\frac{1}{2} \log \left( 1 + \frac{1 - \rho}{1 + \rho b_{\text{min}}^2} \right)} = \Theta \left( \frac{K \log(N/K)}{\log \left( 1 + \frac{1 - \rho}{1 + \rho b_{\text{min}}^2} \right)} \right),
\]

which is satisfied by the \( T \) assumed in the theorem.
References

[1] C. Aksoylar, G. Atia, and V. Saligrama. Sample complexity of salient feature identification for sparse signal processing. In *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, pages 329–332, Aug. 2012.

[2] C. Aksoylar and V. Saligrama. Information-theoretic characterization of sparse recovery. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 38–46, Reykjavik, Iceland, April.

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

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

[5] 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.

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

[7] M. Cheraghchi, A. Karbasi, S. Mohajer, and V. Saligrama. Graph-constrained group testing. In *Proc. of the Int. Symp. on Information Theory (ISIT)*, pages 1913 –1917, June 2010.

[8] S. F. Cotter and B. D. Rao. Sparse channel estimation via matching pursuit with application to equalization. *IEEE Trans. Commun.*, 50(3):374 –377, March 2002.

[9] E. Candes and Y. Plan. Near-ideal model selection by ℓ₁ minimization. *The Annals of Statistics*, 2009.

[10] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ₁-constrained quadratic programs. In *Allerton Conf. on Communication, Control and Computing*, Monticello, IL, 2006.

[11] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri. Non-adaptive group testing: Explicit bounds and novel algorithms. *CoRR*, abs/1202.0206, 2012.

[12] Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *CoRR*, abs/1202.1212, 2012.

[13] 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.

[14] 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.

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

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

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

[18] Galen Reeves and Michael Gastpar. The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing. *IEEE Transactions on Information Theory*, May 2012.

[19] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(910):589–592, 2008.
[20] G. Tang and A. Nehorai. Performance analysis for sparse support recovery. *IEEE Trans. Inf. Theory*, 56(3):1383–1399, 2010.

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

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

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

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

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

[26] C. Aksoylar, G. Atia, and V. Saligrama. Compressive sensing bounds through a unifying framework for sparse models (submitted). In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

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

[28] G. Atia and V. Saligrama. Boolean compressed sensing and noisy group testing. *arXiv pre-print, arXiv:0907.1061v2*, Nov 2009.

[29] A. El Gamal and Y.-H. Kim. Elements of network information theory.

[30] X. Chen and D. Guo. Gaussian many-access channels: Definition and symmetric capacity. In *Information Theory Workshop (ITW), 2013 IEEE*, Seville, Spain, 2013.

[31] 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.

[32] M. B. Wakin, S. Sarvotham, M. F. Duarte, D. Baron, and R. G. Baraniuk. Recovery of jointly sparse signals from few random projections. In *Proceedings of the Workshop on Neural Information Processing Systems (NIPS)*, pages 1435–1442, Vancouver, Canada, Dec. 2005.

[33] A. Jalali, P. Ravikumar, S. Sanghavi, and C. Ruan. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 964–972. 2010.

[34] 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.

[35] L. Jacques, J.N. Laska, P.T. Boufounos, and R.G. Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *arXiv pre-print, arXiv:1104.3160*, 2011.

[36] A. Gupta, R. Nowak, and B. Recht. Sample complexity for 1-bit compressed sensing and sparse classification. In *Proc. of the Int. Symp. on Information Theory (ISIT)*, pages 1553–1557, June 2010.

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

[38] R. G. Gallager. Information theory. In *Mathematics of Physics and Chemistry, Vol. 2*. Van Nostrand, Princeton, NJ, USA, 1964.

[39] M. Chiani, D. Dardari, and M. K. Simon. New exponential bounds and approximations for the computation of error probability in fading channels. *Wireless Communications, IEEE Transactions on*, 2(4):840 – 845, July 2003.

34
[40] G. T. F. de Abreu. Supertight algebraic bounds on the Gaussian Q-function. In *Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on*, pages 948–951, Nov. 2009.