Limits on Support Recovery with Probabilistic Models: An Information-Theoretic Framework

Jonathan Scarlett and Volkan Cevher
Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
Email: {jonathan.scarlett,volkan.cevher}@epfl.ch

***Abstract*—The support recovery problem consists of determining a sparse subset of a set of variables that is relevant in generating a set of observations, and arises in a diverse range of settings such as group testing, compressive sensing, and subset selection in regression. In this paper, we provide a unified approach to support recovery problems, considering general probabilistic observation models relating a sparse data vector to an observation vector. We study the information-theoretic limits for both exact and partial support recovery, taking a novel approach motivated by thresholding techniques in channel coding. We provide general achievability and converse bounds characterizing the trade-off between the error probability and number of measurements, and we specialize these bounds the linear and 1-bit compressive sensing models. Our conditions not only provide scaling laws, but also explicit matching or near-matching constant factors. Moreover, our converse results not only provide conditions under which the error probability fails to vanish, but also conditions under which it tends to one.**

I. Introduction

The problem of support recovery (or model selection) consists of determining a sparse subset of variables that are relevant in producing a set of observations, and arises frequently in disciplines such as group testing [1], compressive sensing (CS) [2], and subset selection in regression [3]. In this paper, we study the information-theoretic limits for this problem, characterizing the number of measurements n required in terms of the sparsity level k and ambient dimension p regardless of the computational complexity.

Most of the previous works on the information-theoretic limits of support recovery have focused on the linear model, for which a variety of bounds are known for both exact support recovery [4]–[8] and partial support recovery [9], [10]. In contrast, we seek a unified approach for studying both linear and non-linear models, motivated by the fact that the latter are indispensable in several applications of interest.

We adopt a system model following those of a line of works seeking mutual information characterizations of sparsity problems [1], [11]–[13]. We consider an approach using thresholding techniques akin to those used in information-spectrum methods [14], thus providing a new alternative to previous approaches based on maximum-likelihood decoding and Fano's inequality. The advantages of our approach include the following: (i) Our achievability bounds provide precise characterizations with explicit constants under more general scalings of the sparsity level, signal-to-noise ratio, etc.; (ii) Our converse results provide necessary conditions for $\mathbb{P}[\text{error}] \not\to 1$, as opposed to simply $\mathbb{P}[\text{error}] \to 0$. This distinction was studied in [13] for discrete observation models, whereas we also allow for continuous models.

Notation: We write β_S to denote the subvector of β at the columns indexed by S, and we write $\mathbf{X}_S$ to denote the submatrix of $\mathbf{X}$ containing the columns indexed by S. The complement with respect to $\{1,\dots,p\}$ is denoted by $(\cdot)^c$. We define the function $[\cdot]^+ = \max\{0,\cdot\}$, and write the floor function as $\lfloor\cdot\rfloor$.

II. Problem Setup and Definitions

A. System Model

Let $\mathcal{S}$ be the set of subsets of $\{1,\dots,p\}$ having cardinality k. The key random variables in our setup are the support set $S \in \mathcal{S}$, the data vector $\beta \in \mathbb{R}^p$, the measurement matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, and the observation vector $\mathbf{Y} \in \mathbb{R}^n$.[1]

The support set S is assumed to be equiprobable on the $\binom{p}{k}$ subsets within $\mathcal{S}$. Given S, the entries of β_{S^c} are deterministically set to zero, and the remaining entries are generated according to some distribution $\beta_S \sim P_{\beta_S}$. We assume that these non-zero entries follow the same distribution for all of the $\binom{p}{k}$ possible realizations of S, and that this distribution is permutation-invariant.

The measurement matrix $\mathbf{X}$ is assumed to have entries that are i.i.d. on some distribution P_X. We write P_X^k, $P_X^{n\times p}$, etc. to denote the corresponding i.i.d. distributions for vectors and matrices. Given S, $\mathbf{X}$, and β, each entry of the observation vector $\mathbf{Y}$ is generated in a conditionally independent manner, with the i-th entry $Y^{(i)}$ distributed according to

$$(Y^{(i)}|S=s, X^{(i)}=x^{(i)}, \beta=b) \sim P_{Y|X_S\beta_S}(\,\cdot\,|x_s^{(i)}, b_s), \tag{1}$$

for some conditional distribution $P_{Y|X_S\beta_S}$. We again assume symmetry with respect to S, namely, that $P_{Y|X_S\beta_S}$ does not depend on the specific realization, and that the distribution is invariant when the columns of X_S and the entries of β_S undergo a common permutation. This covers a wide variety

This work was supported in part by the European Commission under Grant ERC Future Proof, SNF 200021-146750 and SNF CRSII2-147633.

[1] Here we consider entries on $\mathbb{R}$ for concreteness; extensions to the complex numbers $\mathbb{C}$ are immediate. The case of entries on $\mathbb{R}$ also covers problems on finite sets such as $\{0,1\}$, since we are free to choose the relevant probability distributions to be supported on these values.

of specific observation models of interest, including the linear, 1-bit, logistic, Gamma, Poisson, and group testing models.

Given $\mathbf{X}$ and $\mathbf{Y}$, a *decoder* forms an estimate $\hat{S}$ of S. Similarly to previous works, we assume that the decoder knows the system model (including k, P_{β_S} and $P_{Y|X_S\beta_S}$). Following [9], [10], the error probability is given as follows for some $d_{\max} \in \{0, \dots, k-1\}$:

$$P_{\mathrm{e}}(d_{\max}) := \mathbb{P}\big[|S \backslash \hat{S}| > d_{\max} \cup |\hat{S} \backslash S| > d_{\max}\big]. \quad (2)$$

where the probability is taken over the realizations of S, β, $\mathbf{X}$, and $\mathbf{Y}$ (the decoder is assumed to be deterministic). This reduces to exact support recovery when $d_{\max} = 0$.

B. Joint Distributions

It will prove convenient to work with random variables that are implicitly conditioned on a fixed value of S, say $s = \{1, \dots, k\}$. We write P_{β_s} and $P_{Y|X_s\beta_s}$ in place of P_{β_S} and $P_{Y|X_S\beta_S}$ to emphasize that $S = s$. Moreover, we define the corresponding joint distribution

$$P_{\beta_s X_s Y}(b_s, x_s, y) := P_{\beta_s}(b_s) P_X^k(x_s) P_{Y|X_s\beta_s}(y|x_s, b_s), \quad (3)$$

and its multiple-observation counterpart

$$P_{\beta_s \mathbf{X}_s \mathbf{Y}}(b_s, \mathbf{x}_s, \mathbf{y}) := P_{\beta_s}(b_s) P_X^{n\times k}(\mathbf{x}_s) P_{Y|X_s\beta_s}^n(\mathbf{y}|\mathbf{x}_s, b_s). \quad (4)$$

where $P_{Y|X_s\beta_s}^n(\cdot|\cdot, b_s)$ is the n-fold product of $P_{Y|X_s\beta_s}(\cdot|\cdot, b_s)$.

The random variables (β_s, X_s, Y) and $(\beta_s, \mathbf{X}_s, \mathbf{Y})$ appearing throughout this paper are distributed as

$$(\beta_s, X_s, Y) \sim P_{\beta_s X_s Y} \quad (5)$$

$$(\beta_s, \mathbf{X}_s, \mathbf{Y}) \sim P_{\beta_s \mathbf{X}_s \mathbf{Y}}, \quad (6)$$

with the remaining entries of the measurement matrix being distributed as $\mathbf{X}_{s^c} \sim P_X^{n\times(p-k)}$, and with $\beta_{s^c} = 0$ deterministically. Note that these distributions may be probability mass functions (discrete case), probability density functions (continuous case), or combinations of the two.

C. Information-Theoretic Definitions

Our framework treats the support recovery problem as a channel coding problem over a *mixed* channel [14, Sec. 3.3], where the input-output relation is conditionally i.i.d. given β_s.

As in [1], [12], we consider partitions of the support set $s \in \mathcal{S}$ into two sets $s_{\mathrm{dif}} \neq \emptyset$ and s_{eq}. We will see that s_{eq} corresponds to an overlap between s and some other set $\bar{s}$ (i.e. $s \cap \bar{s}$), whereas s_{dif} corresponds to the indices in one set but not the other (e.g. $s \backslash \bar{s}$). There are $2^k - 1$ ways of performing such a partition (the subtraction of one being due to the condition that s_{dif} is non-empty).

For fixed $s \in \mathcal{S}$ and a corresponding pair $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$, we introduce the notation

$$P_{Y|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{dif}}}, \mathbf{x}_{s_{\mathrm{eq}}}) := P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{y}|\mathbf{x}_s) \quad (7)$$

$$P_{Y|X_{s_{\mathrm{dif}}}X_{s_{\mathrm{eq}}}\beta_s}(y|x_{s_{\mathrm{dif}}}, x_{s_{\mathrm{eq}}}, b_s) := P_{Y|X_s\beta_s}(y|x_s, b_s), \quad (8)$$

where $P_{\mathbf{Y}|\mathbf{X}_s}$ is the marginal distribution of (4). While the left-hand sides of (7)–(8) represent the same quantity for any such $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$, it will prove convenient to work with these in place of the right-hand sides. In particular, this allows us to introduce the marginal distributions[2]

$$P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}}) := \sum_{\mathbf{x}_{s_{\mathrm{dif}}}} P_X^{n\times\ell}(\mathbf{x}_{s_{\mathrm{dif}}}) P_{Y|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{dif}}}, \mathbf{x}_{s_{\mathrm{eq}}}) \quad (9)$$

$$P_{Y|X_{s_{\mathrm{eq}}}\beta_s}(y|x_{s_{\mathrm{eq}}}, b_s) := \sum_{x_{s_{\mathrm{dif}}}} P_X^{\ell}(x_{s_{\mathrm{dif}}}) P_{Y|X_{s_{\mathrm{dif}}}X_{s_{\mathrm{eq}}}\beta_s}(y|x_{s_{\mathrm{dif}}}, x_{s_{\mathrm{eq}}}, b_s), \quad (10)$$

where $\ell := |s_{\mathrm{dif}}|$. Using the preceding definitions, we introduce two *information densities* (in the terminology of [15]). The first contains probabilities averaged over β_s,

$$\imath(\mathbf{x}_{s_{\mathrm{dif}}}; \mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}}) := \log \frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{dif}}}, \mathbf{x}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}})}, \quad (11)$$

and the second contains probabilities conditioned on $\beta_s = b_s$,

$$\imath^n(\mathbf{x}_{s_{\mathrm{dif}}}; \mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}}, b_s) := \sum_{i=1}^n \imath(x_{s_{\mathrm{dif}}}^{(i)}; y^{(i)}|x_{s_{\mathrm{eq}}}^{(i)}, b_s), \quad (12)$$

where the single-letter information density is

$$\imath(x_{s_{\mathrm{dif}}}; y|x_{s_{\mathrm{eq}}}, b_s) := \log \frac{P_{Y|X_{s_{\mathrm{dif}}}X_{s_{\mathrm{eq}}}\beta_s}(y|x_{s_{\mathrm{dif}}}, x_{s_{\mathrm{eq}}}, b_s)}{P_{Y|X_{s_{\mathrm{eq}}}\beta_s}(y|x_{s_{\mathrm{eq}}}, b_s)}. \quad (13)$$

Averaging (13) with respect to the random variables in (5) conditioned on $\beta_s = b_s$ yields a conditional mutual information, which we denote by

$$I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s) := I(X_{s_{\mathrm{dif}}}; Y|X_{s_{\mathrm{eq}}}, \beta_s = b_s). \quad (14)$$

III. General Achievability and Converse Bounds

A. Initial Non-Asymptotic Bounds

Here we provide our main non-asymptotic upper and lower bounds on the error probability, applying for arbitrary probabilistic models satisfying the assumptions given above.

Theorem 1. (Achievability) *For any constants γ and $\delta_1 > 0$, there exists a decoder such that*

$$P_{\mathrm{e}}(d_{\max}) \le \mathbb{P}\Bigg[\bigcup_{(s_{\mathrm{dif}}, s_{\mathrm{eq}}) \,:\, |s_{\mathrm{dif}}| > d_{\max}} \Bigg\{\imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}, \beta_s) \le \log\binom{p-k}{\ell} + \log\left(\frac{k^2}{\delta_1^2}\binom{k}{\ell}^2\right) + \gamma\Bigg\}\Bigg] + 2\delta_1 + P_0(\gamma), \quad (15)$$

where $\ell := |s_{\mathrm{dif}}|$, and

$$P_0(\gamma) := \mathbb{P}\left[\log \frac{P_{\mathbf{Y}|\mathbf{X}_s,\beta_s}(\mathbf{Y}|\mathbf{X}_s, \beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} > \gamma\right]. \quad (16)$$

[2] In the case that P_X is continuous, the summations should be replaced by integrals.

Theorem 2. (Converse) *Fix $\delta_1 > 0$, and let $(s_{\rm dif}(b_s), s_{\rm eq}(b_s))$ be an arbitrary partition of $s = \{1, \ldots, k\}$ with $|s_{\rm dif}| > d_{\max}$, depending on $b_s \in \mathbb{R}^k$. For any decoder, we have*

$$P_{\rm e} \geq \mathbb{P}\Bigg[\imath^n(\mathbf{X}_{s_{\rm dif}(\beta_s)}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}(\beta_s)}, \beta_s) \leq \log \binom{p-k+\ell}{\ell} - \log \sum_{d=0}^{d_{\max}} \binom{p-k}{d}\binom{\ell}{d} + \log \delta_1\Bigg] - \delta_1, \quad (17)$$

where $\ell := |s_{\rm dif}|$.

In Section V, we provide brief outlines of the proofs of these bounds. The details can be found in [16].

B. Techniques for Applying Theorems 1 and 2

The steps for applying the preceding theorems are similar, so we focus primarily on Theorem 1.

First, it is often useful to consider a "typical" set of sequences of non-zero entries $\mathcal{T}_\beta$ such that $\mathbb{P}[\beta_s \in \mathcal{T}_\beta] \to 1$, thus restricting the sequences for which the information density $\imath^n(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}, b_s)$ needs to be characterized.

The key idea is to bound the probabilities of the events appearing in (15) using a concentration inequality such as Chebyshev's inequality or Bernstein's inequality [17, Ch. 2]. Since $\imath^n(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}, b_s)$ is an i.i.d. summation, it concentrates sharply about its mean $nI_{s_{\rm dif}, s_{\rm eq}}(b_s)$, and hence the corresponding tail probability in (15) is small provided that

$$n \geq \frac{\log \binom{p-k}{\ell} + \log\left(\frac{k^2}{\delta_1^2}\binom{k}{\ell}^2\right) + \gamma}{I_{s_{\rm dif}, s_{\rm eq}}(b_s)(1-\delta_2)}, \quad (18)$$

where δ_2 is a suitably chosen "backoff constant".

The choice of γ and the bounding of P_0 in (16) can be done on a case-by-case basis. The examples in the present paper use simple bounds based on Markov's inequality and Chebyshev's inequality. In the case that P_{β_S} is discrete, the choice $\gamma = \log \frac{1}{\min_{b_s} P_{\beta_S}(b_s)}$ yields $P_0 = 0$.

For the converse, the analogous condition for ensuring a probability close to one in (17) conditioned on $\beta_s = b_s$ is

$$n \leq \frac{\log \binom{p-k+\ell}{\ell} - \log \sum_{d=0}^{d_{\max}} \binom{p-k}{d}\binom{\ell}{d} - \log \delta_1}{I_{s_{\rm dif}, s_{\rm eq}}(b_s)(1+\delta_2)}. \quad (19)$$

Assuming that (i) the remainder terms resulting from the concentration inequalities are small; (ii) the combinatorial terms in the numerators of (18)–(19) dominate the other terms; and (iii) both δ_1 and δ_2 are small, the preceding arguments lead to simplified bounds on the error probability of the form

$$P_{\rm e}(d_{\max}) \lesssim \mathbb{P}\left[n \leq \max_{(s_{\rm dif}, s_{\rm eq}) \,:\, |s_{\rm dif}| > d_{\max}} \frac{\log \binom{p-k}{\ell}}{I_{s_{\rm dif}, s_{\rm eq}}(\beta_s)}\right]. \quad (20)$$

$$P_{\rm e}(d_{\max}) \gtrsim \mathbb{P}\left[n \geq \max_{(s_{\rm dif}, s_{\rm eq}) \,:\, |s_{\rm dif}| > d_{\max}} \frac{\log \binom{p-k+\ell}{\ell} - \log \sum_{d=0}^{d_{\max}} \binom{p-k}{d}\binom{\ell}{d}}{I_{s_{\rm dif}, s_{\rm eq}}(\beta_s)}\right]. \quad (21)$$

The maximum in (20) (achievability) arises directly, whereas in (21) (converse) it arises by choosing the pair $(s_{\rm dif}(b_s), s_{\rm eq}(b_s))$ to achieve the maximum for each b_s.

IV. Applications to Specific Models

Here we present applications of our general theorems to the linear model and its 1-bit quantized counterpart.

A. Linear Model

The linear model is given by

$$Y = \langle X, \beta \rangle + Z, \quad (22)$$

where $Z \sim N(0, \sigma^2)$. We let β_s be i.i.d. on $N(0, \sigma_\beta^2)$, and we assume that $\sigma_\beta^2 = \frac{c_\beta}{k}$ for some $c_\beta > 0$ not depending on p. We consider the recovery condition (2) with

$$d_{\max} = \lfloor \alpha^* k \rfloor \quad (23)$$

for some $\alpha^* \in (0,1)$ (not varying with p). We consider Gaussian measurements, i.e. $P_X \sim N(0,1)$.

Due to space constraints, we only explain the high-level steps in obtaining the results; see [16] for details. As mentioned in the previous section, we restrict our attention to a typical set of β_s vectors. This is done using the following result, which follows in a straightforward fashion from the Glivenko-Cantelli theorem [18, Thm. 19.1] (stating that an empirical distribution converges uniformly to the true distribution). We define the random variable β_s' to be the permutation of β_s whose entries are listed in increasing order of magnitude.

Proposition 1. *For any $\alpha \in (0,1)$, we have*

$$\lim_{k\to\infty} \frac{1}{k\sigma_\beta^2} \sum_{i=1}^{\lfloor \alpha k \rfloor} (\beta_s')_i^2 = g(\alpha) \quad (24)$$

with probability one, where

$$g(\alpha) := \int_0^\infty \left[\alpha - F_{\chi^2}(u)\right]^+ du \quad (25)$$

and F_{χ^2} is the cumulative distribution function of a χ^2 random variable with one degree of freedom.

By a direct calculation, the mutual information in (14) is

$$I_{s_{\rm dif}, s_{\rm eq}}(b_s) = \frac{1}{2}\log\Big(1 + \frac{1}{\sigma^2}\sum_{i \in s_{\rm dif}} b_i^2\Big), \quad (26)$$

where $W \sim N(0,1)$. Proposition 1 implies that, within a high probability (typical) set, the minimum mutual information for a fixed value of $|s_{\rm dif}| = \lfloor \alpha k \rfloor$ behaves as follows:

$$I_{s_{\rm dif}, s_{\rm eq}}(b_s) \to \frac{1}{2}\log\Big(1 + \frac{c_\beta}{\sigma^2} g(\alpha)\Big), \quad (27)$$

where we recall that $c_\beta = k\sigma_\beta^2$ is a constant.

Since the random variables $(\beta_s, \mathbf{X}_s, \mathbf{Y})$ are jointly Gaussian, the information densities can be written explicitly as in [15]. Upon doing so, some standard bounding techniques for χ^2 random variables reveal that each information density $i^n(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}, b_s)$ (with $|s_{\rm dif}| > d_{\max}$) is within a multiplicative factor $1 \pm \delta_2$ of its mean $nI_{s_{\rm dif}, s_{\rm eq}}(b_s)$ with probability approaching one exponentially fast.

The bounding of P_0 in (16) is based on Chebyshev's inequality; the corresponding mean and variance can again

be evaluated explicitly, since the random variables are jointly Gaussian. We omit the details here.

Combining the above observations, we have the following formalized statement (and simplification) of (20)–(21).

Corollary 1. *Under the preceding setup for the linear model with $k \to \infty$, $k = o(p)$, $\sigma_\beta^2 = \frac{c_\beta}{k}$ for some $c_\beta > 0$, and $d_{\max} = \lfloor \alpha^* k \rfloor$ for some $\alpha^* \in (0,1)$, we have $P_{\mathrm{e}}(d_{\max}) \to 0$ as $p \to \infty$ provided that*

$$n \geq \max_{\alpha \in [\alpha^*, 1]} \frac{\alpha k \log \frac{p}{k}}{\frac{1}{2} \log \left(1 + \frac{c_\beta}{\sigma^2} g(\alpha)\right)} (1+\eta) \tag{28}$$

for some $\eta > 0$, where $g(\cdot)$ is defined in (25)*. Conversely, $P_{\mathrm{e}}(d_{\max}) \to 1$ as $p \to \infty$ whenever*

$$n \leq \max_{\alpha \in [\alpha^*, 1]} \frac{(\alpha - \alpha^*) k \log \frac{p}{k}}{\frac{1}{2} \log \left(1 + \frac{c_\beta}{\sigma^2} g(\alpha)\right)} (1-\eta), \tag{29}$$

for some $\eta > 0$.

The numerators in (28)–(29) arise by applying Stirling's approximation to the combinatorial terms in (18)–(19) (e.g. $\log \binom{p-k}{\lfloor \alpha k \rfloor} = \alpha (k \log \frac{p}{k})(1 + o(1))$). These conditions resemble those of [9], [10]; the focus therein was on linear sparsity $k = \Theta(p)$, whereas we have considered $k = o(p)$.

B. 1-bit Model

The 1-bit model is given by

$$Y = \operatorname{sign}\big(\langle X, \beta \rangle + Z\big). \tag{30}$$

The distributions of the random variables on the right-hand side are the same as those for the linear model. By a direct computation, the mutual information is

$$I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s) = \mathbb{E}\Bigg[H_2\Bigg(Q\Bigg(W \sqrt{\frac{\sum_{i \in s_{\mathrm{eq}}} b_i^2}{\sigma^2 + \sum_{i \in s_{\mathrm{dif}}} b_i^2}} \Bigg) \Bigg) - H_2\Bigg(Q\Bigg(W \sqrt{\frac{1}{\sigma^2} \sum_{i \in s} b_i^2} \Bigg) \Bigg) \Bigg], \tag{31}$$

where $W \sim N(0,1)$, H_2 is the binary entropy function, and Q is the Q-function. The analog of (27) is

$$I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s) \to \Psi(\alpha, c_\beta, \sigma), \tag{32}$$

where

$$\Psi(\alpha, c_\beta, \sigma) := \mathbb{E}\Bigg[H_2\Bigg(Q\Bigg(W \sqrt{\frac{c_\beta (1 - g(\alpha))}{\sigma^2 + c_\beta g(\alpha)}} \Bigg) \Bigg) - H_2\Bigg(Q\Bigg(W \sqrt{\frac{c_\beta}{\sigma^2}} \Bigg) \Bigg) \Bigg]. \tag{33}$$

The analysis is done in the same way as above, with two main differences. First, the information density tail probabilities are bounded using Bernstein's inequality along with the techniques of [14, Rmk 3.1.1] for bounding the corresponding moments. Second, P_0 in (16) is bounded differently. Here the variance of the logarithm therein is more difficult to characterize, so we instead surround the logarithm by $|\cdot|$ and apply Markov's inequality, using a relation between the averages of $\log \frac{P_{\mathbf{Y}|\mathbf{X}_s, \beta_s}}{P_{\mathbf{Y}|\mathbf{X}_s}}$ and $\big| \log \frac{P_{\mathbf{Y}|\mathbf{X}_s, \beta_s}}{P_{\mathbf{Y}|\mathbf{X}_s}} \big|$ from [19]. After some further manipulations, we obtain the following.

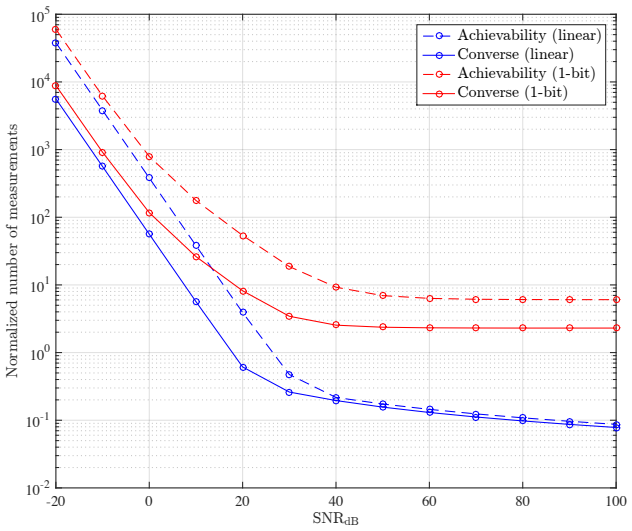

Figure 1: Asymptotic thresholds on the number of measurements required for partial support recovery for the linear and 1-bit models, with an allowable fraction of errors $\alpha^* = 0.1$.

Corollary 2. *Under the preceding setup for the 1-bit model with $k \to \infty$, $k = o(p)$, $\sigma^2 = \Theta(1)$, $\sigma_\beta^2 = \frac{c_\beta}{k}$ for some $c_\beta > 0$, and $d_{\max} = \lfloor \alpha^* k \rfloor$ for some $\alpha^* \in (0,1)$, we have $P_{\mathrm{e}}(d_{\max}) \to 0$ as $p \to \infty$ provided that*

$$n \geq \max_{\alpha \in [\alpha^*, 1]} \frac{\alpha k \log \frac{p}{k}}{\Psi(\alpha, c_\beta, \sigma)} (1+\eta) \tag{34}$$

for some $\eta > 0$, where Ψ is defined in (33)*. Conversely, $P_{\mathrm{e}}(d_{\max}) \to 1$ as $p \to \infty$ whenever*

$$n \leq \max_{\alpha \in [\alpha^*, 1]} \frac{(\alpha - \alpha^*) k \log \frac{p}{k}}{\Psi(\alpha, c_\beta, \sigma)} (1-\eta) \tag{35}$$

for some $\eta > 0$.

C. Numerical Evaluations

We set $\alpha^* = 0.1$ and $\sigma^2 = 1$, and define

$$\mathrm{SNR}_{\mathrm{dB}} := 10 \log \frac{k \sigma_\beta^2}{\sigma^2} = 10 \log c_\beta, \tag{36}$$

which represents the per-sample SNR in dB.

Figure 1 plots the asymptotic thresholds on the number of measurements from Corollaries 1 and 2 with $\alpha^* = 0.1$. More precisely, we replace the arbitrarily small constant η by zero, and we normalize the number of measurements by dividing by $k \log \frac{p}{k}$. Note that k plays no further rule after this normalization, and hence the plot applies to any sequence $k \to \infty$ with $k = o(p)$.

For both models, there is a close correspondence between the necessary and sufficient number of measurements. Interestingly, there is very little loss due to quantization in the low SNR regime, whereas the difference between the two models is

significant in the high SNR regime. Intuitively, this is because the number of measurements for the 1-bit model eventually becomes limited by the quantization, and not by the noise.

V. Outlines of Proofs

A. Achievability (Theorem 1)

We fix the constants $\gamma_{d_{\max}+1}, \dots, \gamma_k$ and consider a decoder that searches for a sparsity pattern $s \in \mathcal{S}$ such that

$$\imath(\mathbf{x}_{s_{\rm dif}}; \mathbf{y}|\mathbf{x}_{s_{\rm eq}}) > \gamma_{|s_{\rm dif}|} \tag{37}$$

for all partitions $(s_{\rm dif}, s_{\rm eq})$ of s with $|s_{\rm dif}| > d_{\max}$. If no such s exists, then an error is declared. If multiple exist, then one is chosen arbitrarily.

By the union bound, we have

$$P_{\rm e}(d_{\max}) \le \mathbb{P}\Bigg[\bigcup_{\substack{(s_{\rm dif}, s_{\rm eq}): \\ |s_{\rm dif}|>d_{\max}}} \Big\{ \imath(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}) \le \gamma_{|s_{\rm dif}|} \Big\} \Bigg] + \sum_{\substack{\bar{s} \in \mathcal{S}\setminus\{s\} \\ |\bar{s}\setminus s|>d_{\max}}} \mathbb{P}\Big[\imath(\mathbf{X}_{\bar{s}\setminus s}; \mathbf{Y}|\mathbf{X}_{\bar{s}\cap s}) > \gamma_{|s_{\rm dif}|} \Big]. \tag{38}$$

By standard bounding techniques [14, Sec. 3.3] and counting arguments [4], the second term is upper bounded by $\sum_{\ell=d_{\max}+1}^{k} \binom{p-k}{\ell}\binom{k}{\ell} e^{-\gamma_\ell}$. The remainder of the proof involves replacing the information density in (11) by that in (12). The denominator is handled as in the analysis of mixed channels [14, Sec. 3.3], whereas the handling of the numerator leads to the term P_0 in (16). The choice of γ_ℓ is

$$\gamma_\ell = \log\left(\frac{k}{\delta_1} \binom{p-k}{\ell} \binom{k}{\ell} \right), \tag{39}$$

though an additional factor of $\frac{k}{\delta_1}\binom{k}{\ell}$ also appears in (15) due to an additional threshold that needs to be chosen similarly.

B. Converse (Theorem 2)

As has been done in several previous proofs of information-theoretic converse bounds for sparsity pattern recovery [5], [10], [12], we consider an argument based on a genie. The genie reveals some of elements of the support set to the decoder, which is left to estimate the remaining entries. An important novelty in our arguments is that we also let the revealed indices depend on the random non-zero entries of β.

We first condition on fixed values of the revealed indices $s_{\rm eq}$ and the non-zero entries $(b_{\rm dif}, b_{\rm eq})$; here $b_{\rm eq}$ contains the entries corresponding to $s_{\rm eq}$, and $b_{\rm dif}$ contains the remaining entries. The decoder is left to estimate $S_{\rm dif}$, which is uniform on the $\binom{p-k+\ell}{\ell}$ subsets of $\{1,\dots,p\}\setminus s_{\rm eq}$ of size $\ell = k - |s_{\rm eq}|$. As noted by Reeves and Gastpar [10], for the criterion in (2) we can consider without loss of generality the case that the estimate $\hat{s}_{\rm dif}$ of $s_{\rm dif}$ also has cardinality ℓ. For any event $\mathcal{A}(s_{\rm eq}, b_{\rm dif}, b_{\rm eq})$, the genie-aided error probability satisfies

$$P_{\rm e}(d_{\max}, s_{\rm eq}, b_{\rm dif}, b_{\rm eq}) \ge \mathbb{P}[\mathcal{A}(s_{\rm eq}, b_{\rm dif}, b_{\rm eq})] - \mathbb{P}[\mathcal{A}(s_{\rm eq}, b_{\rm dif}, b_{\rm eq}) \cap \text{no error}] \tag{40}$$

by the simple identity $\mathbb{P}[\mathcal{A}] = \mathbb{P}[\mathcal{A}\cap\mathcal{E}] + \mathbb{P}[\mathcal{A}\cap\mathcal{E}^c]$. We fix the constant γ_ℓ and choose

$$\mathcal{A}(s_{\rm eq}, b_{\rm dif}, b_{\rm eq}) = \big\{ \imath^n(\mathbf{X}_{S_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}, b_s) \le \gamma_\ell \big\}, \tag{41}$$

where b_s is deterministically constructed from $(b_{\rm dif}, b_{\rm eq})$.

Again using standard bounding techniques and counting arguments, the second probability in (40) can be upper bounded by $\frac{\sum_{d=0}^{d_{\max}} \binom{p-k}{d}\binom{\ell}{d}}{\binom{p-k+\ell}{\ell}} e^{\gamma_\ell}$. Note that the summation in the numerator is the number of $\hat{s}_{\rm dif} \subseteq \{1,\dots,p\}\setminus s_{\rm eq}$ such that $|s_{\rm dif}\setminus\hat{s}_{\rm dif}| \le d_{\max}$ for some fixed $s_{\rm dif}$.

We obtain Theorem 2 by averaging the resulting bound over $(s_{\rm eq}, b_{\rm dif}, b_{\rm eq})$ and using the symmetry properties of P_{β_S} and $P_{Y|X_S\beta_S}$ from Section II; the functions $s_{\rm dif}(b_s)$ and $s_{\rm eq}(b_s)$ arise from the fact that we let the revealed indices depend on the random non-zero entries. We set $\gamma_\ell = \log \frac{\delta_1 \binom{p-k+\ell}{\ell}}{\sum_{d=0}^{d_{\max}} \binom{p-k}{d}\binom{\ell}{d}}$.

References

[1] G. Atia and V. Saligrama, "Boolean compressed sensing and noisy group testing," *IEEE Trans. Inf. Theory*, vol. 58, no. 3, pp. 1880–1901, March 2012.

[2] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer New York, 2013.

[3] A. Miller, *Subset Selection in Regression*. Chapman & Hall, 2002.

[4] M. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, Dec. 2009.

[5] W. Wang, M. Wainwright, and K. Ramchandran, "Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2967–2979, June 2010.

[6] K. Rahnama Rad, "Nearly sharp sufficient conditions on exact sparsity pattern recovery," *IEEE Trans. Inf. Theory*, vol. 57, no. 7, pp. 4672–4679, July 2011.

[7] A. Fletcher, S. Rangan, and V. Goyal, "Necessary and sufficient conditions for sparsity pattern recovery," *IEEE Trans. Inf. Theory*, vol. 55, no. 12, pp. 5758–5772, Dec. 2009.

[8] Y. Jin, Y.-H. Kim, and B. Rao, "Limits on support recovery of sparse signals via multiple-access communication techniques," *IEEE Trans. Inf. Theory*, vol. 57, no. 12, pp. 7877–7892, Dec 2011.

[9] G. Reeves and M. Gastpar, "The sampling rate-distortion tradeoff for sparsity pattern recovery in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 5, pp. 3065–3092, May 2012.

[10] ——, "Approximate sparsity pattern recovery: Information-theoretic lower bounds," *IEEE Trans. Inf. Theory*, vol. 59, no. 6, pp. 3451–3465, June 2013.

[11] G. Atia and V. Saligrama, "A mutual information characterization for sparse signal processing," in *Int. Colloq. Aut., Lang. and Prog. (ICALP)*, Zürich, 2011.

[12] C. Aksoylar, G. Atia, and V. Saligrama, "Sparse signal processing with linear and non-linear observations: A unified Shannon theoretic approach," April 2013, http://arxiv.org/abs/1304.0682.

[13] V. Tan and G. Atia, "Strong impossibility results for sparse signal processing," *IEEE Sig. Proc. Letters*, vol. 21, no. 3, pp. 260–264, March 2014.

[14] T. S. Han, *Information-Spectrum Methods in Information Theory*. Springer, 2003.

[15] Y. Polyanskiy, V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[16] J. Scarlett and V. Cevher, "Limits on support recovery with probabilistic models: An information-theoretic framework," 2015, http://infoscience.epfl.ch/record/204670.

[17] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

[18] A. van der Vaart, *Asymptotic Statistics*. Cambridge Univ. Press, 2000.

[19] A. R. Barron, "Limits of information, Markov chains, and projection," in *IEEE Int. Symp. Inf. Theory*, Serento, 2000.