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

Jonathan Scarlett and Volkan Cevher

***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 compressive sensing, and subset selection in regression, and group testing. In this paper, we take a unified approach to support recovery problems, considering general probabilistic models relating a sparse data vector to an observation vector. We study the information-theoretic limits of 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 to the linear, 1-bit, and group testing models. In several cases, our bounds not only provide matching scaling laws in the necessary and sufficient number of measurements, but also sharp thresholds with matching constant factors. Our approach has several advantages over previous approaches: For the achievability part, we obtain sharp thresholds under broader scalings of the sparsity level and other parameters (e.g., signal-to-noise ratio) compared to several previous works, and for the converse part, we not only provide conditions under which the error probability fails to vanish, but also conditions under which it tends to one.**

***Index Terms*—Support recovery, sparsity pattern recovery, information-theoretic limits, compressive sensing, non-linear models, 1-bit compressive sensing, group testing, phase transitions, strong converse**

I. Introduction

The support recovery problem consists of determining a sparse subset of a set of variables that is relevant in producing a set of observations, and arises frequently in disciplines such as group testing [1], [2], compressive sensing (CS) [3], and subset selection in regression [4]. The observation models can vary significantly among these disciplines, and it is of considerable interest to consider these in a unified fashion. This can be done via probabilistic models relating the sparse vector $\beta \in \mathbb{R}^p$ to a single observation $Y \in \mathbb{R}$ in the following manner:

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

where $S \subseteq \{1,\dots,p\}$ represents the set of relevant variables, $X \in \mathbb{R}^p$ is a measurement vector, X_S (respectively, β_S) is the subvector of X (respectively, β_S) containing the entries indexed by S, and $P_{Y|X_S\beta_S}$ is a given probability distribution. Given a collection of measurements $\mathbf{Y} \in \mathbb{R}^n$ and the corresponding measurement matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ (with each row containing a single measurement vector), the goal is to find the conditions under which the support S can be recovered either *perfectly* or *partially*. 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. Such studies are useful for assessing the performance of practical techniques and determining to what extent improvements are possible.

Before proceeding, we state some important examples of models that are captured by (1).

Linear Model: The linear model [5], [6] is ubiquitous in signal processing, statistics, and machine learning, and in itself covers an extensive range of applications. Each observation takes the form

$$Y = \langle X, \beta\rangle + Z, \qquad (2)$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product, and Z is additive noise. An important quantity in this setting is the signal-to-noise ratio (SNR) $\frac{\mathbb{E}[\langle X,\beta\rangle^2]}{\mathbb{E}[Z^2]}$, and in the context of support recovery, the smallest non-zero absolute value $\beta_{\min}$ in β has also been shown to play a key role [5], [7], [8].

Quantized Linear Models: Quantized variants of the linear model are of significant interest in applications with hardware limitations. An example that we will consider in this paper is the 1-bit model [9], given by

$$Y = \mathrm{sign}\big(\langle X, \beta\rangle + Z\big), \qquad (3)$$

where the sign function equals 1 if its argument is non-negative, and -1 if it is negative.

Group Testing: Studies of group testing problems began several decades ago [10], [11], and have recently regained significant attention [2], [12], with applications including medical testing, database systems, computational biology, and fault detection. The goal is to determine a small number of “defective” items within a larger subset of items. The items involved in a single test are indicated by $X \in \{0,1\}^p$, and each observation takes the form

$$Y = \mathbb{1}\bigg\{\bigcup_{i\in S}\{X_i = 1\}\bigg\} \oplus Z, \qquad (4)$$

The authors are with the Laboratory for Information and Inference Systems (LIONS), École Polytechnique Fédérale de Lausanne (EPFL) (e-mail: {jonathan.scarlett,volkan.cevher}@epfl.ch).

This work was supported in part by the European Commission under Grant ERC Future Proof, SNF 200021-146750 and SNF CRSII2-147633, and EPFL Fellows Horizon2020 grant 665667.

This work was presented in part at the IEEE International Symposium on Information Theory (2015), and at the ACM-SIAM Symposium on Discrete Algorithms (2016).

with S representing the defective items, Y indicating whether the test contains at least one defective item, and Z representing possible noise (here $\oplus$ denotes modulo-2 addition). In this setting, one can think of β as deterministically having entries equaling one on S, and zero on S^c.

The above examples highlight that (1) captures both discrete and continuous models. Beyond these examples, several other non-linear models are captured by (1), including the logistic, Poisson, and gamma models.

A. Previous Work and Contributions

Numerous previous works on the information-theoretic limits of support recovery have focused on the linear model [5], [7], [8], [13]–[22]. The main aim of these works, and of that the present paper, is to develop necessary and sufficient conditions for which an "error probability" vanishes as $p \to \infty$. However, there are several distinctions that can be made, including:

- Random measurement matrices [5], [7], [8], [13] vs. arbitrary measurement matrices [16], [18], [19];
- Exact support recovery [5], [7], [8], [13] vs. partial support recovery [15], [16], [20];
- Minimax characterizations for β in a given class [5], [7], [8], [13] vs. average performance bounds for random β [14], [16], [21].

Perhaps the most widely-studied combination of these is that of minimax characterizations for exact support recovery with random measurement matrices. In this setting, within the class of vectors β whose non-zero entries have an absolute value exceeding some threshold $\beta_{\min}$, necessary and sufficient conditions on n are available with matching scaling laws [7], [8]. See also [23], [24] for information-theoretic studies of the linear model with a mean square error criterion.

Compared to the linear model, research on the information-theoretic limits of support recovery for non-linear models is relatively scarce. The system model that we have adopted follows those of a line of works seeking mutual information characterizations of sparsity problems [2], [11], [14], [25], though we make use of significantly different analysis techniques. Similarly to these works, we focus on random measurement matrices and random non-zero entries of β. Other works considering non-linear models have used vastly different approaches such as regularized M-estimators [26], [27] and approximate message passing [28].

High-level Contributions: We consider an approach using thresholding techniques akin to those used in information-spectrum methods [29], thus providing a new alternative to previous approaches based on maximum-likelihood decoding and Fano's inequality. Our key contributions and the advantages of our framework are as follows:

1. Considering both exact and partial support recovery, we provide non-asymptotic performance bounds applying to general probabilistic models, along with a procedure for applying them to specific models (*cf.* Section III-B).
2. We explicitly provide the constant factors in our bounds, allowing for more precise characterizations of the performance compared to works focusing on scaling laws (e.g., see [5], [8], [20]). In several cases, the resulting necessary and sufficient conditions on the number of measurements coincide up to a multiplicative $1 + o(1)$ term, thus providing *exact* asymptotic thresholds (sometimes referred to as *phase transitions* [24], [30]) on the number of measurements.
3. As evidenced in our examples outlined below, our framework often leads to such exact or near-exact thresholds for significantly more general scalings of k, SNR, etc. compared to previous works.
4. The majority of previous works have developed converse results using Fano's inequality, leading to necessary conditions for $\mathbb{P}[\text{error}] \to 0$. In contrast, our converse results provide necessary conditions for $\mathbb{P}[\text{error}] \not\to 1$. The distinction between these two conditions is important from a practical perspective: One may not expect a condition such as $\mathbb{P}[\text{error}] \ge 10^{-10}$ to be significant, whereas the condition $\mathbb{P}[\text{error}] \to 1$ is inarguably so.

Contributions for Specific Models: An overview of our bounds for specific models is given in Table I, where we state the derived bounds with the asymptotically negligible terms omitted. All of the models and their parameters are defined precisely in Section IV; in particular, the functions $f_1, \dots, f_9$ and the remainder terms (Δ_1, Δ_9) are given explicitly, and are easy to evaluate. We proceed by discussing these contributions in more detail, and comparing them to various existing results in the literature:

1. *(Linear model)* In the case of exact recovery, we recover the exact thresholds on the required number of measurements given by Jin *et al.* [17], as well as handling a broader range of scalings of $\beta_{\min} := \min\{|\beta_i| : \beta_i \neq 0\}$ (see Section IV-A for details) and strengthening the converse by considering the more stringent condition $\mathbb{P}[\text{error}] \to 1$. Our results for partial recovery provide near-matching necessary and sufficient conditions under scalings with $k = o(p)$, thus complementing the extensive study of the scaling $k = \Theta(p)$ by Reeves and Gastpar [15], [16].
2. *(1-bit model)* We provide two surprising observations regarding the 1-bit model: Corollary 3 provides a low-SNR setting where the quantization only increases the asymptotic number of measurements by a factor of $\frac{\pi}{2}$, whereas Corollary 4 provides a high-SNR setting where the scaling law is strictly worse than the linear model. Similar behavior will be observed for partial recovery (Corollaries 2 and 5) by numerically comparing the bounds for various SNR values.
3. *(Group testing)* Asymptotic thresholds for group testing with $k = \Theta(1)$ were given previously by Malyutov [11] and Atia and Saligrama [2]. However, for the case that $k \to \infty$, the sufficient conditions of [2] that introduced additional logarithmic factors. In

Model	Result	Parameters	Distributions	Sufficient n for $\mathbb{P}[\text{error}] \to 0$	Necessary n for $\mathbb{P}[\text{error}] \not\to 1$
Linear	Cor. 1	$k = o(p)$	Discrete β_S Gaussian $\mathbf{X}$	$\max_{\ell=1,\dots k} \frac{(1+\Delta_1)\log\binom{p-k}{\ell}}{f_1(\ell)}$ ($\Delta_1 \to 0$ for various scalings)	$\max_{\ell=1,\dots k} \frac{\log\binom{p-k+\ell}{\ell}}{f_1(\ell)}$
	Cor. 2	$k \to \infty, k = o(p)$ Partial recovery of proportion $1-\alpha^*$	Gaussian β_S Gaussian $\mathbf{X}$	$\max_{\alpha\in[\alpha^*,1]} \frac{\alpha k \log \frac{p}{k}}{f_2(\alpha)}$	$\max_{\alpha\in[\alpha^*,1]} \frac{(\alpha-\alpha^*) k \log \frac{p}{k}}{f_2(\alpha)}$
1-bit	Cor. 3	$k = \Theta(1)$ Low SNR	Discrete β_S Gaussian $\mathbf{X}$	$\max_{\ell=1,\dots k} \frac{\ell \log p}{f_3(\ell)}$ (within a factor $\frac{\pi}{2}$ of linear model)	$\max_{\ell=1,\dots k} \frac{\ell \log p}{f_3(\ell)}$ (within a factor $\frac{\pi}{2}$ of linear model)
	Cor. 4	$k = \Theta(p)$ High SNR	Fixed β_S Gaussian $\mathbf{X}$	-	$\Omega(p\sqrt{\log p})$ (compared to $\Theta(p)$ for linear model)
	Cor. 5	$k \to \infty, k = o(p)$ Partial recovery of proportion $1-\alpha^*$	Gaussian β_S Gaussian $\mathbf{X}$	$\max_{\alpha\in[\alpha^*,1]} \frac{\alpha k \log \frac{p}{k}}{f_5(\alpha)}$	$\max_{\alpha\in[\alpha^*,1]} \frac{(\alpha-\alpha^*) k \log \frac{p}{k}}{f_5(\alpha)}$
Group testing	Cor. 6	$k = \Theta(p^\theta)$	Fixed β_S Bernoulli $\mathbf{X}$	$\frac{k \log \frac{p}{k}}{f_6(\theta)}$ ($f_6(\theta) = \log 2$ for $\theta \le \frac{1}{3}$)	$\frac{k \log \frac{p}{k}}{\log 2}$
	Cor. 7	$k = \Theta(p^\theta)$ Noisy (crossover probability ρ)	Fixed β_S Bernoulli $\mathbf{X}$	$\frac{k \log \frac{p}{k}}{f_7(\theta)}$ ($f_7(\theta) = \log 2 - H_2(\rho)$ for small θ)	$\frac{k \log \frac{p}{k}}{\log 2 - H_2(\rho)}$
	Cor. 8	$k \to \infty, k = o(p)$ Partial recovery of proportion $1-\alpha^*$	Fixed β_S Bernoulli $\mathbf{X}$	$\frac{k \log \frac{p}{k}}{\log 2 - H_2(\rho)}$	$\frac{(1-\alpha^*)\big(k \log \frac{p}{k}\big)}{\log 2 - H_2(\rho)}$
General discrete observations	Cor. 9	Arbitrary	Arbitrary	-	$\max_{\ell=1,\dots k} \frac{\log\binom{p-k+\ell}{\ell}}{f_9(\ell)+\Delta_9}$

Table I: Overview of main results for exact or partial support recovery under various observation models. In the necessary and sufficient number of measurements, asymptotically negligible terms have been omitted. All quantities are defined precisely in Section IV.

contrast, we obtain matching $\Theta\big(k \log \frac{p}{k}\big)$ scaling laws for any sublinear scaling of the form $k = O(p^\theta)$ ($\theta \in (0,1)$). Moreover, for sufficiently small θ we obtain exact thresholds. In particular, for the noiseless setting we show that $n \approx k \log_2 \frac{p}{k}$ measurements are both necessary and sufficient for $\theta \le \frac{1}{3}$. This is in fact the same threshold as that for adaptive group testing [31], thus proving that non-adaptive Bernoulli measurement matrices are *asymptotically optimal* even when adaptivity is allowed; this was previously known only in the limit as $\theta \to 0$ [32]. For the noisy case, we prove an analogous claim for sufficiently small θ. A shortened and simplified version of this paper focusing exclusively on group testing can be found in [33].

4. *(General discrete observations)* Our converse for the case of general discrete observations (Corollary 9) recovers that of Tan and Atia [25] for the case that β_S is fixed, strengthens it due to a smaller remainder term Δ_9, and provides a generalization to the case that β_S is random.

B. Structure of the Paper

In Section II, we introduce our system model. In Section III, we present our main non-asymptotic achievability and converse results for general observation models, and the procedure for applying them to specific problems. Several applications of our results to specific models are presented in Section IV. The proofs of the general bounds are given in Section V, and conclusions are drawn in Section VI.

C. Notation

We use upper-case letters for random variables, and lower-case variables for their realizations. A non-bold character may be a scalar or a vector, whereas a bold character refers to a collection of n scalars (e.g., $\mathbf{Y} \in \mathbb{R}^n$) or vectors (e.g., $\mathbf{X} \in \mathbb{R}^{n \times p}$). We write β_S to denote the subvector of β at the columns indexed by S, and $\mathbf{X}_S$ to denote the submatrix of $\mathbf{X}$ containing the columns indexed by S. The complement with respect to $\{1,\dots,p\}$ is denoted by $(\cdot)^c$.

The symbol $\sim$ means "distributed as". For a given joint distribution P_{XY}, the corresponding marginal distributions are denoted by P_X and P_Y, and similarly for conditional marginals (e.g., $P_{Y|X}$). We write $\mathbb{P}[\cdot]$ for probabilities, $\mathbb{E}[\cdot]$ for expectations, and $\mathrm{Var}[\cdot]$ for variances. We use usual notations for the entropy (e.g., $H(X)$) and mutual information (e.g., $I(X;Y)$), and their conditional counterparts (e.g., $H(X|Z)$, $I(X;Y|Z)$). Note that H may also denote the differential entropy for continuous random variables; the distinction will be clear from the context. We define the binary entropy function $H_2(\rho) := -\rho \log \rho - (1-\rho)\log(1-\rho)$, and the Q-function $Q(x) := \mathbb{P}[W \ge x]$ ($W \sim N(0,1)$).

We make use of the standard asymptotic notations $O(\cdot)$, $o(\cdot)$, $\Theta(\cdot)$, $\Omega(\cdot)$ and $\omega(\cdot)$. We define the function $[\cdot]^+ =$

$\max\{0,\cdot\}$, and write the floor function as $\lfloor\cdot\rfloor$. The function $\log$ has base e.

II. Problem Setup

A. Model and Assumptions

Recall that p denotes the ambient dimension, k denotes the sparsity level, and n denotes the number of measurements. We 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 i.i.d. values on some distribution P_X. We write $P_X^{n\times p}$, to denote the corresponding i.i.d. distributions for matrices, and we write P_X^k as a shorthand for $P_X^{k\times 1}$. 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

$$(Y^{(i)}|S=s, X^{(i)}=x^{(i)}, \beta=b) \sim P_{Y|X_S\beta_S}(\,\cdot\,|x_s^{(i)}, b_s), \tag{5}$$

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.

Given $\mathbf{X}$ and $\mathbf{Y}$, a decoder forms an estimate $\hat{S}$ of S. Similarly to previous works studying information-theoretic limits on support recovery, we assume that the decoder knows the system model. We consider two related performance measures. In the case of exact support recovery, the error probability is given by

$$P_{\mathrm{e}} := \mathbb{P}[\hat{S} \ne S], \tag{6}$$

and is taken with respect to the realizations of S, β, $\mathbf{X}$, and $\mathbf{Y}$; the decoder is assumed to be deterministic. We also consider a less stringent performance criterion requiring that only $k - d_{\max}$ entries of S are successfully recovered, for some $d_{\max} \in \{1,\dots,k-1\}$. Following [15], [16], the error probability is given by

$$P_{\mathrm{e}}(d_{\max}) := \mathbb{P}\big[|S\backslash\hat{S}| > d_{\max} \cup |\hat{S}\backslash S| > d_{\max}\big]. \tag{7}$$

Note that if both S and $\hat{S}$ have cardinality k with probability one, then the two events in the union are identical, and hence either of the two can be removed.

For clarity, we formally state our main assumptions as follows:

[A1] The support set S is uniform on the $\binom{p}{k}$ subsets of $\{1,\dots,p\}$ of size k, and the measurement matrix $\mathbf{X}$ is i.i.d. on some distribution P_X.

[A2] The non-zero entries β_S are distributed according to P_{β_S}, and this distribution is permutation-invariant and the same for all realizations of S.

[A3] The observation vector $\mathbf{Y}$ is conditionally i.i.d. according to $P_{Y|X_S\beta_S}$, and this distribution is the same for all realizations of S, and invariant to common permutations of the columns of X_S and entries of β_S.

[A4] The decoder is given $(\mathbf{X},\mathbf{Y})$, and also knows the system model including k, $P_{Y|X_S\beta_S}$, and P_{β_S}.

Our main goal is to derive necessary and sufficient conditions on n and k (as functions of p) such that P_{e} or $P_{\mathrm{e}}(d_{\max})$ vanishes as $p\to\infty$. Moreover, when considering converse results, we will not only be interested in conditions under which $P_{\mathrm{e}} \not\to 0$, but also conditions under which the stronger statement $P_{\mathrm{e}} \to 1$ holds.

In particular, we introduce the terminology that the *strong converse* holds if there exists a sequence of values n^*, indexed by p, such that for all $\eta > 0$, we have $P_{\mathrm{e}} \to 0$ when $n \ge n^*(1+\eta)$, and $P_{\mathrm{e}} \to 1$ when $n \le n^*(1-\eta)$. This is related to the notion of a *phase transition* [24], [30]. More generally, we will refer to conditions under which $P_{\mathrm{e}} \to 1$ as *strong impossibility results*, not necessarily requiring matching achievability bounds. That is, the strong converse conclusively gives a sharp threshold between failure and success, whereas a strong impossibility result may not.

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

$$P_{\beta_s X_s Y}(b_s, x_s, y) := P_{\beta_s}(b_s)P_X^k(x_s)P_{Y|X_s\beta_s}(y|x_s,b_s), \tag{8}$$

and its multiple-observation counterpart

$$P_{\beta_s \mathbf{X}_s \mathbf{Y}}(b_s, \mathbf{x}_s, \mathbf{y}) := P_{\beta_s}(b_s)P_X^{n\times k}(\mathbf{x}_s)P^n_{Y|X_s\beta_s}(\mathbf{y}|\mathbf{x}_s,b_s). \tag{9}$$

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

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

$$(\beta_s, X_s, Y) \sim P_{\beta_s X_s Y} \tag{10}$$

$$(\beta_s, \mathbf{X}_s, \mathbf{Y}) \sim P_{\beta_s \mathbf{X}_s \mathbf{Y}}, \tag{11}$$

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. That is, we condition on a fixed $S = s$ except where stated otherwise.

For notational convenience, the main parts of our analysis are presented with P_{β_s}, P_X and $P_{Y|X_s\beta_s}$ representing probability mass functions (PMFs), and with the corresponding averages written using summations. However, except where stated otherwise, our analysis is directly applicable to case

[1] Extensions to more general alphabets beyond $\mathbb{R}$ are straightforward.

that these distributions instead represent probability density functions (PDFs), with the summations replaced by integrals where necessary. The same applies to mixed discrete-continuous distributions.

B. Information-Theoretic Definitions

Before introducing the required definitions for support recovery, it is instructive to discuss thresholding techniques in channel coding studies. These commenced in early works such as [34], [35], and have recently been used extensively in information-spectrum methods [29], [36].

1) Channel Coding: We first recall the mutual information, which is ubiquitous in information theory:

$$I(X;Y) := \sum_{x,y} P_{XY}(x,y) \log \frac{P_{Y|X}(y|x)}{P_Y(y)}. \qquad (12)$$

In deriving asymptotic and non-asymptotic performance bounds, it is common to work directly with the logarithm,

$$\imath(x;y) := \log \frac{P_{Y|X}(y|x)}{P_Y(y)}, \qquad (13)$$

which is commonly known as the *information density*. The thresholding techniques work by manipulating probabilities of events of the form $\sum_{i=1}^n \imath(X_i;Y_i) \le \gamma$ and $\sum_{i=1}^n \imath(X_i;Y_i) > \gamma$. For the former, one can perform a *change of measure* from the conditional distribution $\mathbf{Y}$ given $\mathbf{X}$ to the unconditional distribution of $\mathbf{Y}$, with a multiplicative constant $e^{-\gamma}$. For the latter, one can similarly perform a change of measure from $\mathbf{Y}$ to $(\mathbf{Y}|\mathbf{X})$. Hence, in both cases, there is a simple relation between the conditional and unconditional probabilities of the output sequences.

Using these methods, one can get upper and lower bounds on the error probability such that the dominant term is

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \imath(X_i;Y_i) \le I(X;Y) + \zeta_n\right] \qquad (14)$$

for some $\zeta_n = o(1)$. Assuming that $\{(X_i,Y_i)\}_{i=1}^n$ has some form of i.i.d. structure, one can analyze this expression using tools from probability theory. The law of large numbers yields the channel capacity $C = \max_{P_X} I(X;Y)$, and refined characterizations can be obtained using variations of the central limit theorem [37].

Among the channel coding literature, our analysis is most similar to that of mixed channels [29, Sec. 3.3], where the relation between the input and output sequences is not i.i.d., but instead conditionally i.i.d. given another random variable. In our setting, β_s will play the role of this random variable. See Figure 1 for a depiction of this connection.

2) Support Recovery: As in [2], [14], we will consider partitions of the support set $s \in \mathcal{S}$ into two sets $s_{\text{dif}} \neq \emptyset$ and s_{eq}. As will be seen in the proofs, s_{eq} will typically correspond to an overlap between s and some other set $\bar{s}$ (i.e., $s \cap \bar{s}$), whereas s_{dif} will correspond 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 with $s_{\text{dif}} \neq \emptyset$.

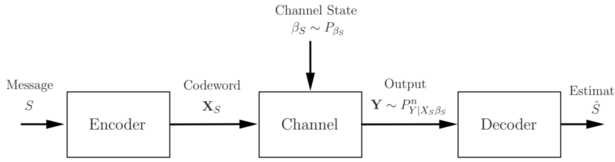

Figure 1: Connection between support recovery and coding over a mixed channel.

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

$$P_{Y|\mathbf{X}_{s_{\text{dif}}}\mathbf{X}_{s_{\text{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\text{dif}}},\mathbf{x}_{s_{\text{eq}}}) := P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{y}|\mathbf{x}_s) \qquad (15)$$

$$P_{Y|X_{s_{\text{dif}}}X_{s_{\text{eq}}}\beta_s}(y|x_{s_{\text{dif}}},x_{s_{\text{eq}}},b_s) := P_{Y|X_s\beta_s}(y|x_s,b_s), \qquad (16)$$

where $P_{\mathbf{Y}|\mathbf{X}_s}$ is the marginal distribution of (9). While the left-hand sides of (15)–(16) represent the same quantities for any such $(s_{\text{dif}}, s_{\text{eq}})$, it will still prove convenient to work with these in place of the right-hand sides. In particular, this allows us to introduce the marginal distributions

$$\begin{aligned} &P_{\mathbf{Y}|\mathbf{X}_{s_{\text{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\text{eq}}}) \\ &\quad := \sum_{\mathbf{x}_{s_{\text{dif}}}} P_X^{n\times\ell}(\mathbf{x}_{s_{\text{dif}}}) P_{Y|\mathbf{X}_{s_{\text{dif}}}\mathbf{X}_{s_{\text{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\text{dif}}},\mathbf{x}_{s_{\text{eq}}}) \end{aligned} \qquad (17)$$

$$\begin{aligned} &P_{Y|X_{s_{\text{eq}}}\beta_s}(y|x_{s_{\text{eq}}},b_s) \\ &\quad := \sum_{x_{s_{\text{dif}}}} P_X^{\ell}(x_{s_{\text{dif}}}) P_{Y|X_{s_{\text{dif}}}X_{s_{\text{eq}}}\beta_s}(y|x_{s_{\text{dif}}},x_{s_{\text{eq}}},b_s), \end{aligned} \qquad (18)$$

where $\ell := |s_{\text{dif}}|$. Using the preceding definitions, we introduce two information densities. The first contains probabilities averaged over β_s,

$$\imath(\mathbf{x}_{s_{\text{dif}}};\mathbf{y}|\mathbf{x}_{s_{\text{eq}}}) := \log \frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\text{dif}}}\mathbf{X}_{s_{\text{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\text{dif}}},\mathbf{x}_{s_{\text{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\text{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\text{eq}}})}, \qquad (19)$$

whereas the second conditions on $\beta_s = b_s$:

$$\imath^n(\mathbf{x}_{s_{\text{dif}}};\mathbf{y}|\mathbf{x}_{s_{\text{eq}}},b_s) := \sum_{i=1}^n \imath(x_{s_{\text{dif}}}^{(i)};y^{(i)}|x_{s_{\text{eq}}}^{(i)},b_s), \qquad (20)$$

where the single-letter information density is

$$\imath(x_{s_{\text{dif}}};y|x_{s_{\text{eq}}},b_s) := \log \frac{P_{Y|X_{s_{\text{dif}}}X_{s_{\text{eq}}}\beta_s}(y|x_{s_{\text{dif}}},x_{s_{\text{eq}}},b_s)}{P_{Y|X_{s_{\text{eq}}}\beta_s}(y|x_{s_{\text{eq}}},b_s)}. \qquad (21)$$

As mentioned above, we will generally work with discrete random variables for clarity of exposition, in which case the ratio is between two PMFs. In the case of continuous observations the ratio is instead between two PDFs, and more generally this can be replaced by the Radon-Nikodym derivative as in the channel coding setting [37].

Averaging (21) with respect to the random variables in (10) conditioned on $\beta_s = b_s$ yields a conditional mutual information, which we denote by

$$I_{s_{\text{dif}},s_{\text{eq}}}(b_s) := I(X_{s_{\text{dif}}};Y|X_{s_{\text{eq}}},\beta_s = b_s). \qquad (22)$$

This quantity will play a key role in our bounds, which will typically have the form

$$P_{\mathrm{e}} \approx \mathbb{P}\biggl[n \le \max_{(s_{\mathrm{dif}}, s_{\mathrm{eq}})} \frac{\binom{p}{|s_{\mathrm{dif}}|}}{I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(\beta_s)} \biggr], \tag{23}$$

as will be made more precise in the subsequent sections.

III. General Achievability and Converse Bounds

In this section, we provide general results holding for arbitrary models satisfying the assumptions given in Section II. Each of the results for exact recovery has a direct counterpart for partial recovery. For clarity, we focus on the former throughout Sections III-A and III-B, and then proceed with the latter in Section III-C.

A. Initial Non-Asymptotic Bounds

Here we provide our main non-asymptotic upper and lower bounds on the error probability. These bounds bear a strong resemblance to analogous bounds from the channel coding literature [29]; in each case, the dominant term involves tail probabilities of the information density given in (20). The mean of the information density is the mutual information in (22), which thus arises naturally in the subsequent necessary and sufficient conditions on n upon showing that the deviation from the mean is small with high probability. The procedure for doing this given a specific model will be given in Section III-B.

We start with our achievability result. Here and throughout this section, we make use of the random variables defined in (11).

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

$$P_{\mathrm{e}} \le \mathbb{P}\Biggl[\bigcup_{(s_{\mathrm{dif}}, s_{\mathrm{eq}}) \,:\, s_{\mathrm{dif}} \ne \emptyset} \biggl\{ \imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y} | \mathbf{X}_{s_{\mathrm{eq}}}, \beta_s) \\ \le \log \binom{p-k}{|s_{\mathrm{dif}}|} + \log \biggl(\frac{k^2}{\delta_1^2} \binom{k}{|s_{\mathrm{dif}}|}^2 \biggr) + \gamma \biggr\} \Biggr] + P_0(\gamma) + 2\delta_1, \tag{24}$$

where

$$P_0(\gamma) := \mathbb{P}\biggl[\log \frac{P_{\mathbf{Y}|\mathbf{X}_s, \beta_s}(\mathbf{Y}|\mathbf{X}_s, \beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} > \gamma \biggr]. \tag{25}$$

Proof. See Section V-A. □

Remark 1. The probability in the definition of $P_0(\gamma)$ is not an i.i.d. sum, and the techniques for ensuring that $P_0(\gamma) \to 0$ vary between different settings. The following approaches will suffice for all of the applications in this paper:

1. In the case that P_{β_S} is discrete, $P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{y}|\mathbf{x}_s) = \sum_{b_s} P_{\beta_s}(b_s) P_{\mathbf{Y}|\mathbf{X}_s, \beta_s}(\mathbf{y}|\mathbf{x}_s, b_s)$, and it follows that

$$\gamma = \log \frac{1}{\min_{b_s} P_{\beta_s}(b_s)} \implies P_0(\gamma) = 0. \tag{26}$$

Moreover, this can be strengthened by noting from the proof of Theorem 1 that γ may depend on β_s, and choosing $\gamma(b_s) = \log \frac{1}{P_{\beta_s}(b_s)}$ accordingly.

2. Defining

$$I_0 := I(\beta_s; \mathbf{Y} | \mathbf{X}_s) \tag{27}$$

$$V_0 := \mathrm{Var}\biggl[\log \frac{P_{\mathbf{Y}|\mathbf{X}_s, \beta_s}(\mathbf{Y}|\mathbf{X}_s, \beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} \biggr], \tag{28}$$

we have for any $\delta_0 > 0$ that

$$\gamma = I_0 + \sqrt{\frac{V_0}{\delta_0}} \implies P_0(\gamma) \le \delta_0. \tag{29}$$

This follows directly from Chebyshev's inequality.

3. Defining

$$I_{0,+} := \mathbb{E}\biggl[\biggl[\log \frac{P_{\mathbf{Y}|\mathbf{X}_s \beta_s}(\mathbf{Y}|\mathbf{X}_s, \beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} \biggr]^+ \biggr], \tag{30}$$

we have for any $\delta_0 > 0$ that

$$\gamma = \frac{I_{0,+}}{\delta_0} \implies P_0(\gamma) \le \delta_0. \tag{31}$$

This follows directly from Markov's inequality.

The proof of Theorem 1 is based on a decoder the searches for a unique support set s such that

$$\imath(\mathbf{x}_{s_{\mathrm{dif}}}; \mathbf{y} | \mathbf{x}_{s_{\mathrm{eq}}}) > \gamma_{|s_{\mathrm{dif}}|} \tag{32}$$

for some $\{\gamma_\ell\}_{\ell=1}^k$ and all $2^k - 1$ partitions $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$ of s with $s_{\mathrm{dif}} \ne \emptyset$. Since the numerator in (19) is the likelihood of $\mathbf{y}$ given $(\mathbf{x}_{s_{\mathrm{dif}}}, \mathbf{x}_{s_{\mathrm{eq}}})$, this decoder can be thought of a weakened version of the maximum-likelihood (ML) decoder. Like the ML decoder, computational considerations make its implementation intractable.

The following theorem provides a general non-asymptotic converse bound.

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

$$P_{\mathrm{e}} \ge \mathbb{P}\biggl[\imath^n(\mathbf{X}_{s_{\mathrm{dif}(\beta_s)}}; \mathbf{Y} | \mathbf{X}_{s_{\mathrm{eq}(\beta_s)}}, \beta_s) \\ \le \log \binom{p - k + |s_{\mathrm{dif}}(\beta_s)|}{|s_{\mathrm{dif}}(\beta_s)|} + \log \delta_1 \biggr] - \delta_1. \tag{33}$$

Proof. See Section V-B. □

The proof of Theorem 2 is based on Verdú-Han type bounding techniques [36].

B. Techniques for Applying Theorems 1 and 2

The bounds presented in the preceding theorems do not directly reveal the number of measurements required to achieving a vanishing error probability. In this subsection, we present the steps that can be used to obtain such conditions. We provide examples in Section IV.

The idea is to use a concentration inequality to bound the first term in (24) (or (33)), which is possible due to the fact that each summation $\imath^n$ is conditionally i.i.d. given β_s. We proceed by providing the details of these steps separately for the achievability and converse. We start with the former.

1. Observe that, conditioned on $\beta_s = b_s$, the mean of $\imath^n(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}, \beta_s)$ is $nI_{s_{\rm dif},s_{\rm eq}}(b_s)$, where $I_{s_{\rm dif},s_{\rm eq}}(b_s)$ is defined in (22).
2. Fix $\delta_2 \in (0,1)$, and suppose that for a fixed value b_s of β_s, we have for all $(s_{\rm dif}, s_{\rm eq})$ that

$$\log\binom{p-k}{|s_{\rm dif}|} + \log\bigg(\frac{k^2}{\delta_1^2}\binom{k}{|s_{\rm dif}|}^2\bigg) + \gamma \\ \le n(1-\delta_2)I_{s_{\rm dif},s_{\rm eq}}(b_s), \quad (34)$$

and

$$\mathbb{P}\Big[\imath^n(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}, b_s) \le n(1-\delta_2)I_{s_{\rm dif},s_{\rm eq}}(b_s) \,\big|\, \beta_s = b_s\Big] \\ \le \psi_{|s_{\rm dif}|}(n,\delta_2) \quad (35)$$

for some functions $\{\psi_\ell\}_{\ell=1}^k$ (e.g., these may arise from Chebyshev's inequality or Bernstein's inequality [38, Ch. 2]). Combining these conditions with the union bound, we obtain

$$\mathbb{P}\bigg[\bigcup_{(s_{\rm dif},s_{\rm eq})\,:\,s_{\rm dif}\neq\emptyset}\bigg\{\imath^n(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}, \beta_s) \\ \le \log\binom{p-k}{|s_{\rm dif}|} + \log\bigg(\frac{k^2}{\delta_1^2}\binom{k}{|s_{\rm dif}|}^2\bigg) + \gamma\bigg\} \,\bigg|\, \beta_s = b_s\bigg] \\ \le \sum_{\ell=1}^{k}\binom{k}{\ell}\psi_\ell(n,\delta_2). \quad (36)$$

3. Observe that the condition in (34) can be written as

$$n \ge \frac{\log\binom{p-k}{|s_{\rm dif}|} + \log\Big(\frac{k^2}{\delta_1^2}\binom{k}{|s_{\rm dif}|}^2\Big) + \gamma}{I_{s_{\rm dif},s_{\rm eq}}(b_s)(1-\delta_2)}. \quad (37)$$

We summarize the preceding findings in the following.

Theorem 3. *For any constants $\delta_1 > 0$, $\delta_2 \in (0,1)$ and γ, and functions $\{\psi_\ell\}_{\ell=1}^k$ ($\psi_\ell : \mathbb{Z}\times\mathbb{R}\to\mathbb{R}$), define the set*

$$\mathcal{B}(\delta_1,\delta_2,\gamma) := \big\{b_s \,:\, (35) \text{ and } (37) \text{ hold for all} \\ (s_{\rm dif}, s_{\rm eq}) \text{ with } s_{\rm dif} \neq \emptyset\big\}. \quad (38)$$

Then we have

$$P_{\rm e} \le \mathbb{P}\big[\beta_s \notin \mathcal{B}(\delta_1,\delta_2,\gamma)\big] + \sum_{\ell=1}^{k}\binom{k}{\ell}\psi_\ell(n,\delta_2) + P_0(\gamma) + 2\delta_1. \quad (39)$$

Remark 2. The preceding arguments remain unchanged when δ_2 also depends on $\ell = |s_{\rm dif}|$. We leave this possible dependence implicit throughout this section, since a fixed value will suffice for all but one of the models considered in Section IV.

In the case that (35) holds for all b_s (or more generally, within a set whose probability under P_{β_s} tends to one) and the final three terms in (39) vanish, the overall upper bound approaches the probability, with respect to P_{β_s}, that (37) fails to hold. In many cases, the second logarithm in the numerator therein is dominated by the first. It should be noted that the condition that the second term in (39) vanishes can also impose conditions on n. For most of the examples presented in Section IV, the condition in (37) will be the dominant one; however, this need not always be the case, and it depends on the concentration inequality used in (35).

The application of Theorem 2 is done using similar steps, so we provide less detail. Fix $\delta_2 > 0$, and suppose that, for a fixed value b_s of β_s, the pair $(s_{\rm dif}, s_{\rm eq}) = (s_{\rm dif}(b_s), s_{\rm eq}(b_s))$ is such that

$$\log\binom{p-k+|s_{\rm dif}|}{|s_{\rm dif}|} - \log\delta_1 \ge n(1+\delta_2)I_{s_{\rm dif},s_{\rm eq}}(b_s), \quad (40)$$

and

$$\mathbb{P}\Big[\imath^n(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}, b_s) \le n(1+\delta_2)I_{s_{\rm dif},s_{\rm eq}}(b_s) \,\big|\, \beta_s = b_s\Big] \\ \ge 1 - \psi'_{|s_{\rm dif}|}(n,\delta_2) \quad (41)$$

for some function $\psi'_{|s_{\rm dif}|}$. Combining these conditions, we see that the first probability in (33), with an added conditioning on $\beta_s = b_s$, is lower bounded by $1 - \psi'_{|s_{\rm dif}|}(n,\delta_2)$. In the case that ψ'_ℓ is defined for multiple ℓ values corresponding to different values of b_s, we can further lower bound this by $1 - \max_\ell \psi'_\ell(n,\delta_2)$.

Next, we observe that (40) holds if and only if

$$n \le \frac{\log\binom{p-k+|s_{\rm dif}|}{|s_{\rm dif}|} - \log\delta_1}{I_{s_{\rm dif},s_{\rm eq}}(b_s)(1+\delta_2)}. \quad (42)$$

Recalling that the partition $(s_{\rm dif}, s_{\rm eq})$ is an arbitrary function of β_s, we can ensure that this coincides with

$$n \le \max_{(s_{\rm dif},s_{\rm eq})\,:\,s_{\rm dif}\neq\emptyset} \frac{\log\binom{p-k+|s_{\rm dif}|}{|s_{\rm dif}|} - \log\delta_1}{I_{s_{\rm dif},s_{\rm eq}}(b_s)(1+\delta_2)} \quad (43)$$

by choosing each pair $(s_{\rm dif}, s_{\rm eq})$ as a function of b_s to achieve this maximum.

Finally, we note that the maximum over ℓ in the above-derived term $1 - \max_\ell \psi'_\ell(n,\delta_2)$ may be restricted to any set $\mathcal{L} \subseteq \{1,\dots,k\}$ provided that $|s_{\rm dif}|$ is constrained similarly in (43); one simply chooses the partition $(s_{\rm dif}(b_s), s_{\rm eq}(b_s))$ so that $\ell = |s_{\rm dif}|$ always lies in this set. Putting everything together, we have the following.

Theorem 4. *For any set $\mathcal{L} \subseteq \{1,\dots,k\}$, constants $\delta_1 > 0$ and $\delta_2 > 0$, and functions $\{\psi'_\ell\}_{\ell\in\mathcal{L}}$ ($\psi'_\ell : \mathbb{Z}\times\mathbb{R}\to\mathbb{R}$), define the set*

$$\mathcal{B}'(\delta_1,\delta_2) := \big\{b_s \,:\, (41) \text{ and } (42) \text{ hold for all} \\ (s_{\rm dif}, s_{\rm eq}) \text{ with } |s_{\rm dif}| \in \mathcal{L}\big\}. \quad (44)$$

Then we have

$$P_{\rm e} \ge \mathbb{P}\big[\beta_s \in \mathcal{B}'(\delta_1,\delta_2)\big]\Big(1 - \max_{\ell\in\mathcal{L}}\psi'_\ell(n,\delta_2)\Big) - \delta_1. \quad (45)$$

If the pair $(s_{\rm dif}, s_{\rm eq})$ had been fixed in Theorem 2, as opposed to being a function of β_s, then we would have only obtained a weaker result with the statement "for all $(s_{\rm dif}, s_{\rm eq})$"

Procedure 1: Steps for Obtaining Necessary and Sufficient Conditions on n from Theorems 3 and 4

1. *(Identify a Typical Set)* Construct a sequence of "typical" sets $\mathcal{T}_\beta \subseteq \mathbb{R}^k$ of non-zero entries, indexed by p, such that $\mathbb{P}\big[\beta_s \in \mathcal{T}_\beta\big] \to 1$, thus restricting the vectors b_s for which $\imath(X_{s_{\rm dif}}; Y | X_{s_{\rm eq}}, b_s)$ needs to be characterized.
2. *(Bound the Information Density Tail Probabilities)* Using a concentration inequality for i.i.d. summations (e.g., Chebyshev, Bernstein), bound the tail probabilities in (35) and (41) for each $(s_{\rm dif}, s_{\rm eq})$ and $b_s \in \mathcal{T}_\beta$, with a fixed constant δ_2. Upon making these dependent on $(s_{\rm dif}, s_{\rm eq}, b_s)$ only through $\ell := |s_{\rm dif}|$, the bounds are denoted by $\psi_\ell(n,\delta_2)$ and $\psi'_\ell(n,\delta_2)$.
3. *(Control the Remainder Terms)* By suitable rearrangements, find conditions on n under which the terms $\sum_\ell \binom{k}{\ell}\psi_\ell(n,\delta_2)$ and $\max_{\ell\in\mathcal{L}} \psi'_\ell(n,\delta_2)$ in (39) and (45) vanish, thus ensuring that their contribution is negligible. Similarly, choose δ_1 to vanish with p so that its contribution is negligible, and for the achievability part, choose γ such that the remainder term $P_0(\gamma)$ vanishes (*cf.* Remark 1).
4. *(Combine and Simplify)* Combine the previous steps as follows:
   a) Construct the set of non-zero entries $\mathcal{B}(\delta_1,\delta_2,\gamma) \subseteq \mathbb{R}^k$ (respectively, $\mathcal{B}'(\delta_1,\delta_2)$) in (38) (respectively, (44));
   b) Deduce from (39) (respectively, (45)) and Step 3 that $P_{\rm e} \le \mathbb{P}[\beta_s \notin \mathcal{B}(\delta_1,\delta_2,\gamma)] + o(1)$ (respectively, $P_{\rm e} \ge \mathbb{P}[\beta_s \in \mathcal{B}'(\delta_1,\delta_2)] + o(1)$);
   c) From the properties of the typical set $\mathcal{T}_\beta$ in Steps 1–2, deduce that $P_{\rm e} \to 0$ (respectively, $P_{\rm e} \to 1$) when n satisfies (37) (respectively, (42)) for all $b_s \in \mathcal{T}_\beta$;
   d) Augment this condition on n with Step 3.

in (44) replaced by a fixed pair. Assuming that the remainder terms in (45) are insignificant, this weaker result is of the form $P_{\rm e} \gtrsim \max_{(s_{\rm dif},s_{\rm eq})} \mathbb{P}\big[n \le f(s_{\rm dif}, s_{\rm eq}, \beta_s)\big]$ rather than $P_{\rm e} \gtrsim \mathbb{P}\big[n \le \max_{(s_{\rm dif},s_{\rm eq})} f(s_{\rm dif}, s_{\rm eq}, \beta_s)\big]$. This can lead to significantly different bounds on the sample complexity, and the distinction is crucial in our applications in Section IV. As described in the proof in Section V, the key to obtaining this difference is in applying a refined version of an argument based on a genie.

The general steps in applying Theorems 3 and 4 to specific problems are outlined in Procedure 1.

In our experience, the choice of $\mathcal{T}_\beta$ in the first step of Procedure 1 usually comes naturally given the specific model. On the other hand, it is often less straightforward to find a sufficiently powerful concentration inequality in Step 2. A simple choice is Chebyshev's inequality, which expresses ψ_ℓ and ψ'_ℓ in terms of $I_{s_{\rm dif},s_{\rm eq}}(b_s)$ (see (22)) and the corresponding variances of the information densities. This choice is usually effective for the converse, wheres the achievability part typically requires sharper concentration inequalities such as Bernstein's inequality, due to the combinatorial terms in (39).

C. Extensions to Partial Recovery

We now turn to the partial support recovery criterion in (7). The changes in the analysis required to generalize Theorems 1 and 2 to this setting are given in Section V-C; rather than repeating each of these, we focus our attention on the resulting analogues of Theorems 3 and 4.

Theorem 5. *For any constants $\delta_1 > 0$, $\delta_2 \in (0,1)$ and $\gamma > 0$, and functions $\{\psi_\ell\}_{\ell=d_{\max}+1}^k$ ($\psi_\ell : \mathbb{Z}\times\mathbb{R} \to \mathbb{R}$), define the set*

$$\mathcal{B}(\delta_1,\delta_2,\gamma) := \big\{b_s : \text{(35) and (37) hold for all } (s_{\rm dif}, s_{\rm eq}) \text{ with } |s_{\rm dif}| \in \{d_{\max}+1,\dots,k\}\big\}. \quad (46)$$

Then we have

$$P_{\rm e}(d_{\max}) \le \mathbb{P}\big[\beta_s \notin \mathcal{B}(\delta_1,\delta_2,\gamma)\big] + \sum_{\ell=d_{\max}+1}^{k} \binom{k}{\ell}\psi_\ell(n,\delta_2) + P_0(\gamma) + 2\delta_1, \quad (47)$$

where P_0 is defined in (25).

For the converse part, (42) is replaced by

$$n \ge \frac{\log\binom{p-k+|s_{\rm dif}|}{|s_{\rm dif}|} - \log\sum_{d=0}^{d_{\max}}\binom{p-k}{d}\binom{|s_{\rm dif}|}{d} - \log\delta_1}{I_{s_{\rm dif},s_{\rm eq}}(b_s)}, \quad (48)$$

and we have the following analog of Theorem 4.

Theorem 6. *For any set $\mathcal{L} \subseteq \{d_{\max}+1,\dots,k\}$, constants $\delta_1 > 0$ and $\delta_2 \in (0,1)$, and functions $\{\psi'_\ell\}_{\ell\in\mathcal{L}}$ ($\psi'_\ell : \mathbb{Z}\times\mathbb{R}\to\mathbb{R}$), define the set*

$$\mathcal{B}'(\delta_1,\delta_2) := \big\{b_s : \text{(41) and (48) hold for all } (s_{\rm dif}, s_{\rm eq}) \text{ with } |s_{\rm dif}| \in \mathcal{L}\big\}. \quad (49)$$

Then we have

$$P_{\rm e}(d_{\max}) \ge \mathbb{P}\big[\beta_s \in \mathcal{B}'(\delta_1,\delta_2)\big]\Big(1 - \max_{\ell\in\mathcal{L}}\psi'_\ell(n,\delta_2)\Big) - \delta_1. \quad (50)$$

The applications of Theorems 5 and 6 follow identical steps to Procedure 1. However, it will be seen that the restriction $|s_{\rm dif}| > d_{\max}$ can in fact considerably simplify these steps, since it removes the need to obtain concentration inequalities for smaller values of $|s_{\rm dif}|$.

D. Comparison to Fano's Inequality

Most previous works on the information-theoretic limits of sparsity recovery have made use of Fano's inequality [39, Sec. 2.11]. For this reason, we provide here a discussion on the relative merits of this approach and our approach. To this end, we consider the following bound, which can be obtained by combining the analysis of [2], [14] with our refined genie argument:

$$P_{\mathrm{e}} \geq \sum_{b_s} P_{\beta_S}(b_s) \max\left\{0, 1 - \frac{n I_{s_{\mathrm{dif}}(b_s), s_{\mathrm{eq}}(b_s)}(b_s) + 1}{\log \binom{p-k+|s_{\mathrm{dif}}(b_s)|}{|s_{\mathrm{dif}}(b_s)|}}\right\} \tag{51}$$

in the notation of Theorem 2. By analyzing this bound similarly to Section III-B, we obtain for any $\delta_2 > 0$ that

$$P_{\mathrm{e}} \geq \delta_2 \, \mathbb{P}\big[\beta_s \in \mathcal{B}'_{\mathrm{Fano}}(\delta_2)\big] - \frac{1}{\log(p-k+1)}, \tag{52}$$

where

$$\mathcal{B}'_{\mathrm{Fano}}(\delta_2) := \Bigg\{ b_s \,:\, n \leq \frac{\log \binom{p-k+|s_{\mathrm{dif}}|}{|s_{\mathrm{dif}}|}}{I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s)} (1 - \delta_2) \quad \text{for all } (s_{\mathrm{dif}}, s_{\mathrm{eq}}) \text{ with } s_{\mathrm{dif}} \neq \emptyset \Bigg\}. \tag{53}$$

A similar result for partial recovery can also be derived by incorporating the arguments from [16] and the present paper.

As discussed in the introduction, the key advantage of Theorem 4 is that it provides a more precise characterization of how far the error probability is from zero, and in particular, the conditions under which $P_{\mathrm{e}} \to 1$ (strong impossibility results). On the other hand, the bound on P_{e} in (52) is always bounded away from one for fixed δ_2, and becomes increasingly weak for small δ_2.

The advantage of Fano's inequality is that it only requires the mutual information to be computed, whereas our approach also requires the application of a concentration inequality. This, in turn, typically requires the variance of the information density to be characterized, which is not always straightforward. However, as discussed following Procedure 1, the main difficulty associated with these concentration inequalities is typically in finding one which is sufficiently powerful for the *achievability* part. Thus, the added difficulty in the converse may not add to the overall difficulty in deriving matching achievability and converse bounds.

IV. Applications to Specific Models

In this section, we present applications of Theorems 3–6 to the linear [5], 1-bit [9], and group testing [2] models, and to more general models with discrete observations [25]. Throughout the section, we make use of general concentration inequalities given in Appendix A. We also make use of the following variance quantity:

$$V_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s) := \mathrm{Var}\big[\imath(X_{s_{\mathrm{dif}}}; Y | X_{s_{\mathrm{eq}}}, b_s) \,\big|\, \beta_s = b_s\big]. \tag{54}$$

A. Linear Model with Discrete β_s

Here we consider the linear model, where each observation takes the form

$$Y = \langle X, \beta \rangle + Z, \tag{55}$$

where $Z \sim N(0, \sigma^2)$ for some $\sigma > 0$.

Without loss of generality, we consider the fixed support set $s = \{1, \dots, k\}$. Following the setup of [17], we let β_s be a uniformly random permutation of a fixed vector $(b_1, \dots, b_k)$, and we choose $P_X \sim N(0, 1)$. Since both the measurement matrices and the noise are Gaussian, the mutual information in (22) is given by [39, Ch. 10]

$$I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s) = \frac{1}{2} \log\Big(1 + \frac{1}{\sigma^2} \sum_{i \in s_{\mathrm{dif}}} b_i^2\Big). \tag{56}$$

Throughout this subsection, we denote $b_{\min} := \min_i |b_i|$ and $b_{\max} := \max_i |b_i|$. We assume that $\sigma^2 = \Theta(1)$, and that $b_{\min} = \Theta(b_{\max})$ and $0 < b_{\min} = O(1)$; note that $b_{\min} = o(1)$ is allowed. The steps of Procedure 1 are as follows.

Step 1: We trivially choose the typical set $\mathcal{T}_\beta$ to contain all vectors on the support of P_{β_S}.

Step 2: We make use of the following concentration inequality based on Bernstein's inequality.

Proposition 1. *Under the preceding setup for the linear model, we have for all $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$ and b_s that*

$$\mathbb{P}\bigg[\Big|\imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y} | \mathbf{X}_{s_{\mathrm{eq}}}, b_s) - n I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s)\Big| \geq n\delta \,\Big|\, \beta_s = b_s\bigg] \leq 2 \exp\bigg(-\frac{\delta^2 n}{2(4\alpha_{s_{\mathrm{dif}}}^2 + \delta \alpha_{s_{\mathrm{dif}}})}\bigg), \tag{57}$$

where

$$\alpha_{s_{\mathrm{dif}}} := \frac{2\sigma_{s_{\mathrm{dif}}}(\sigma + \sigma_{s_{\mathrm{dif}}})}{\sigma^2 + \sigma_{s_{\mathrm{dif}}}^2} \tag{58}$$

with $\sigma_{s_{\mathrm{dif}}}^2 := \sum_{i \in s_{\mathrm{dif}}} b_i^2$.

Proof. See Appendix B. □

Setting $\delta = \delta_2 I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s)$, it follows that in (35) and (41) we can set

$$\psi_\ell(n, \delta_2) = \psi'_\ell(n, \delta_2) = 2 \max_{(s_{\mathrm{dif}}, s_{\mathrm{eq}}, b_s) \,:\, |s_{\mathrm{dif}}| = \ell} \exp\bigg(-\frac{(\delta_2 I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s))^2 n}{2(4\alpha_{s_{\mathrm{dif}}} + \delta_2 I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s))\alpha_{s_{\mathrm{dif}}}}\bigg). \tag{59}$$

Step 3: In accordance with the first item of Remark 1, we set γ as in (26) so that $P_0(\gamma) = 0$.

We focus on the conditions on n under which the term $\sum_{\ell=1}^k \binom{k}{\ell} \psi_\ell(n, \delta_2)$ in (39) vanishes; the term containing ψ'_ℓ in (45) can be handled in a similar yet simpler fashion. By the assumptions $\sigma^2 = \Theta(1)$ and $b_{\max} = \Theta(b_{\min})$, we readily obtain $I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s) = \Theta(\log(1 + \ell b_{\min}^2))$ and $\alpha_{s_{\mathrm{dif}}}^2 = \Theta(\min\{1, \ell b_{\min}^2\})$ using (56) and (58), where $\ell = |s_{\mathrm{dif}}|$. Using these growth rates and upper bounding the summation in (39) by k times the corresponding maximum, we see that $\sum_{\ell=1}^k \binom{k}{\ell} \psi_\ell(n, \delta_2) \to 0$ provided that the following holds for some sufficiently small constant ζ (depending on δ_2):

$$\frac{n \log^2(1 + \ell b_{\min}^2)}{\min\{1, \ell b_{\min}^2\} + \log(1 + \ell b_{\min}^2)\sqrt{\min\{1, \ell b_{\min}^2\}}} \zeta - \ell \log \frac{k}{\ell} - \log k \to \infty \tag{60}$$

for all ℓ. We now treat two cases separately:

- If $\ell b_{\min}^2 = o(1)$, the first term in (60) behaves as $\Theta(n \ell b_{\min}^2)$; by rearranging, we conclude that it suffices that $n \to \infty$ and $n = \Omega\big(\frac{\log k}{b_{\min}^2}\big)$ with a sufficiently large implied constant.

- If $\ell b_{\min}^2 = \Omega(1)$, the first term in (60) behaves as $\Omega(n)$, and it thus suffices that $n \to \infty$ and $n = \Omega(k)$ with a sufficiently large implied constant.

Thus, the overall condition that we require is $n \to \infty$ and

$$n = \Omega\Big(\frac{\log k}{b_{\min}^2}\Big) \quad \text{and} \quad n = \Omega(k), \tag{61}$$

with sufficiently large implied constants. For the converse, the analogous condition to (60) contains only the first term on the left-hand side (the difference being due to the fact that the combinatorial term in (39) is not present in (45)), and a similar argument reveals that it suffices that $n = \omega\big(\frac{1}{b_{\min}^2}\big)$.

Step 4: Combining the preceding steps and applying asymptotic simplifications, we obtain the following.

Corollary 1. *Under the preceding setup for the linear model with $\sigma^2 = \Theta(1)$, $b_{\min} = \Theta(b_{\max})$, $b_{\min}^2 = O(1)$, $k = o(p)$, and m_β distinct elements in $(b_1, \dots, b_k)$, we have $P_{\mathrm{e}} \to 0$ as $p \to \infty$ provided that*

$$n \ge \max_{s_{\mathrm{dif}} \ne \emptyset} \frac{\log \binom{p-k}{|s_{\mathrm{dif}}|}}{\frac{1}{2}\log\Big(1 + \frac{1}{\sigma^2}\sum_{i \in s_{\mathrm{dif}}} b_i^2\Big)}(1+\eta), \tag{62}$$

under any one of the following additional conditions: (i) $k = \Theta(1)$; (ii) $k = o(\log p)$ and $m_\beta = \Theta(1)$; (iii) $k = O((\log p)^\theta)$ for some $\theta > 0$, and $m_\beta = 1$; (iv) $k = \Theta(p^\theta)$ for some $\theta \in (0,1)$, $b_{\min}^2 = \Theta\big(\frac{\log k}{k}\big)$, and $m_\beta = 1$.

Conversely, without any additional conditions, we have $P_{\mathrm{e}} \to 1$ as $p \to \infty$ whenever

$$n \le \max_{s_{\mathrm{dif}} \ne \emptyset} \frac{\log \binom{p-k+|s_{\mathrm{dif}}|}{|s_{\mathrm{dif}}|}}{\frac{1}{2}\log\Big(1 + \frac{1}{\sigma^2}\sum_{i \in s_{\mathrm{dif}}} b_i^2\Big)}(1-\eta) \tag{63}$$

for some $\eta > 0$.

Proof. The converse part follows from (42) with $\delta_1 \to 0$ sufficiently slowly. To check the condition $n = \omega\big(\frac{1}{b_{\min}^2}\big)$ stated following (61), we may assume without loss of generality that (63) holds with equality, since the decoder can always choose to ignore additional measurements. When equality holds, we observe that for the worst-case s_{dif} with $\ell = 1$, the denominator therein behaves as $O(b_{\min}^2)$ (since $b_{\min}^2 = O(1)$) and the numerator behaves as $\Theta(\log p)$, and hence, the condition $n = \omega\big(\frac{1}{b_{\min}^2}\big)$ is satisfied.

For the achievability part, we first use (37) to obtain

$$n \ge \max_{s_{\mathrm{dif}} \ne \emptyset} \frac{\log \binom{p-k}{|s_{\mathrm{dif}}|} + 2\log\Big(k\binom{k}{|s_{\mathrm{dif}}|}\Big) + k\log m_\beta}{\frac{1}{2}\log\Big(1 + \frac{1}{\sigma^2}\sum_{i \in s_{\mathrm{dif}}} b_i^2\Big)}(1+\eta), \tag{64}$$

where the final term in the numerator arises from (26) since $P_{\beta_s}(b_s)$ is the same for all permutations of $(b_1, \dots, b_k)$, and is lower bounded by m_β^{-k}. Observe that the first term in the numerator behaves as $\Theta(|s_{\mathrm{dif}}|\log p)$ for each of the cases in the corollary statement, and the second term behaves as $\Theta\big(\log k + |s_{\mathrm{dif}}|\log\frac{k}{|s_{\mathrm{dif}}|}\big)$.

In cases (i)–(iii), we have $\log k = o(\log p)$, and it immediately follows that the numerator in (64) is dominated by the first term, and hence, the others can be factored into η in (62). Moreover, in case (i), both conditions in (61) are dominated by the objective in (64) with $\ell := |s_{\mathrm{dif}}| = 1$, which behaves as $\Theta\big(\frac{\log p}{b_{\min}^2}\big)$. In cases (ii)–(iii), the first condition in (61) is again dominated by the term in (64) with $\ell = 1$. The second condition is dominated by the term with $\ell = k$, which behaves as $\Theta\big(\frac{k\log p}{\log(1+kb_{\min}^2)}\big) = \Omega\big(k\frac{\log p}{\log k}\big)$.

In case (iv), the first term in the numerator of (64) may not be dominant for small $\ell := |s_{\mathrm{dif}}|$, since $\log k = \Theta(\log p)$. However, by observing that the objective scales as $\Theta\big(\frac{\ell \log p}{\log(1+\ell b_{\min}^2)}\big)$ and using the assumed scaling of $b_{\min}^2$, it is readily verified that the maximum can only be achieved with $\ell = \Theta(k)$. For any such maximizer, we have $\log\binom{p-k}{\ell} = \Theta(k\log p)$, and hence, the second term in the numerator of (64) can be factored into η, as it behaves as $O(k)$. The two conditions in (61) are identical under the given scaling of $b_{\min}^2$, and are dominated by the objective in (62) with $\ell = k$, which behaves as $\Theta\big(\frac{k\log k}{\log\log k}\big)$. □

In the case that $b_{\min} = \Theta(1)$, the thresholds given in Corollary 1 coincide with those given in the main results of [17]. Our framework has the advantage of handling the case that $b_{\min} = o(1)$, as well as providing the strong converse ($P_{\mathrm{e}} \to 1$) instead of the weak converse ($P_{\mathrm{e}} \not\to 0$). However, it should be noted that the achievability parts of [17] have the notable advantage of using a decoder that does not depend on the distribution of β_s.

On first glance, the bounds in (62)–(63) may appear to be difficult to evaluate, since the maximizations are over $2^k - 1$ non-empty subsets s_{dif}. However, it is in fact only k of them that need to be computed, since for any given $\ell = |s_{\mathrm{dif}}|$ the maximizing s_{dif} is the one with the smallest corresponding value of $\sum_{i \in s_{\mathrm{dif}}} b_i^2$.

Comparison to the LASSO: Conditions for the support recovery of the computationally tractable LASSO algorithm were given by Wainwright [6]. Several comparisons to the information-theoretic limits were given in [5], [6] in terms of scaling laws; here we complement these comparisons by briefly discussing the corresponding constant factors. For simplicity, we focus on the case that the non-zero entries are all equal to a common value $b_0 = \frac{c_\beta}{k}$ (for some constant c_β representing the per-sample SNR) and k is poly-logarithmic in p, corresponding to case (iii) of Corollary 1.

The results of [6] state that LASSO requires at least $(2k\log p)(1+o(1))$ measurements regardless of c_β, and that this bound is also achievable in the limit as $c_\beta \to \infty$. On the other hand, Corollary 1 reveals that for the optimal decoder, the coefficient to $k\log p$ can be arbitrarily small provided that c_β is large enough. More precisely, applying some simple manipulations to (62), we find that the coefficient to $k\log p$ is $\sup_{\alpha \in (0,1]} \frac{\alpha}{\frac{1}{2}\log(1+c_\beta\alpha)}$, where α represents the ratio $\frac{|s_{\mathrm{dif}}|}{k}$. It is easy to verify that the maximum is achieved at $\alpha = 1$, yielding the constant $\frac{2}{\log(1+c_\beta)}$. We conclude that the LASSO provably yields a suboptimal constant when $c_\beta > 1$, and fails to achieve the optimal logarithmic decay. However, it should be noted that our decoder requires knowledge of k and c_β, whereas the LASSO does not (except possibly via their role in determining the regularization parameter).

B. Linear Model with Gaussian β_S and Partial Recovery

In this subsection, we consider the setup of Section IV-A with two changes: We let the distribution of β_s be continuous rather than discrete, and we consider partial recovery instead of exact recovery. More specifically, we let β_s be i.i.d. on $N(0,\sigma_\beta^2)$ for some variance σ_β^2, and we consider the recovery condition in (7) with

$$d_{\max} = \lfloor \alpha^* k \rfloor \tag{65}$$

for some $\alpha^* \in (0,1)$ not varying with p. We again choose $P_X \sim N(0,1)$. We assume $\sigma_\beta^2 = \frac{c_\beta}{k}$ for some $c_\beta > 0$ not depending on p, corresponding to a fixed per-sample SNR.

We begin with the following auxiliary result.

Proposition 2. *Under the preceding setup for the linear model, the quantities I_0 and V_0 defined in* (27)–(28) *satisfy*

$$I_0 \le \frac{k}{2}\log\Big(1+\frac{n\sigma_\beta^2}{\sigma^2}\Big) \tag{66}$$

$$V_0 \le 2n. \tag{67}$$

Proof. See Appendix B. □

We now proceed with the steps of Procedure 1 (with the suitable changes from exact recovery to partial recovery, *cf.* Section III-C).

Step 1: Our choice of the typical set $\mathcal{T}_\beta$ is based on the following proposition characterizing the behavior of the $\lfloor \alpha k \rfloor$ entries of β_s having the smallest magnitude for fixed α. We define the random variable β_s' to be the permutation of β_s whose entries are listed in increasing order of magnitude.

Proposition 3. *For any $\alpha \in (0,1]$, we have*

$$\lim_{k\to\infty} \frac{1}{k\sigma_\beta^2}\sum_{i=1}^{\lfloor \alpha k \rfloor} (\beta_s')_i^2 = g(\alpha) \tag{68}$$

with probability one, where

$$g(\alpha) := \int_0^\infty \big[\alpha - F_{\chi^2}(u)\big]^+ \, du, \tag{69}$$

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

Proof. Letting $\hat{F}_k$ be the empirical distribution of the values $\big\{\frac{1}{\sigma_\beta^2}\beta_i^2\big\}_{i=1}^k$, we have from the Glivenko-Cantelli theorem [40, Thm. 19.1] that $\sup_u |\hat{F}_k(u) - F_{\chi^2}(u)| \to 0$ almost surely. This immediately implies that the sum of the $\lfloor \alpha k \rfloor$ smallest values in $\big\{\frac{1}{\sigma_\beta^2}\beta_i^2\big\}$, normalized by the number of values k, converges almost surely to the integral of $F_{\chi^2}^{-1}(u)$ from 0 to α. It is easily verified graphically that this integral can equivalently be written as (69). □

Based on this result and its proof, we set $\mathcal{T}_\beta$ to be the set of vectors b_s such that $\sup_u |\hat{F}_k(u) - F_{\chi^2}(u)| \le \epsilon$, where ϵ is chosen to decay sufficiently slowly so that $\mathbb{P}[\beta_s \in \mathcal{T}_\beta] \to 1$. Thus, within the typical set, the empirical distribution of the non-zero entries closely follows a χ^2 random variable.

An important consequence of this choice of typical set regards the behavior of the mutual information in (56). For a fixed set size $|s_{\rm dif}|$, the partition $(s_{\rm dif}, s_{\rm eq})$ minimizing this mutual information is the one with the smallest value of $\sum_{i\in s_{\rm dif}} b_i^2$. Within the typical set, we immediately obtain from Proposition 3 that the corresponding mutual information behaves as follows when $|s_{\rm dif}| = \lfloor \alpha k \rfloor$:

$$I_{s_{\rm dif},s_{\rm eq}}(b_s) \to \frac{1}{2}\log\Big(1+\frac{c_\beta}{\sigma^2}g(\alpha)\Big), \tag{70}$$

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

Step 2: We again make use of Proposition 1 and its subsequent expression for ψ_ℓ and ψ_ℓ' in (59).

Step 3: We choose $\gamma = I_0 + \sqrt{\frac{V_0}{\delta_0}}$ as in (29) for some $\delta_0 > 0$, thus ensuring that $P_0(\gamma) \le \delta_0$.

For the terms in Theorems 5–6 containing ψ_ℓ and ψ_ℓ', we first note that since we are considering partial recovery, we may focus on values of $\ell = |s_{\rm dif}|$ greater than $\alpha^* k$. By our choice of $\mathcal{T}_\beta$, we may also focus on realizations b_s of β_s satisfying (68). For such realizations, we have for all $s_{\rm dif}$ with $|s_{\rm dif}| = \ell = \Theta(k)$ that $\sum_{i\in s_{\rm dif}} b_i^2 = \Omega(1)$, which implies that $\alpha_{s_{\rm dif}}^2 = \Theta(1)$ in (58) and $I_{s_{\rm dif},s_{\rm eq}}(b_s) = \Omega(1)$ in (56). The analogous condition to (60) thus simplifies to $nI' \gg k$ for some $I' = \Omega(1)$, giving the following condition under which the second term in (39) vanishes:

$$n = \Omega(k), \tag{71}$$

with a sufficiently large implied constant. For the converse part, it suffices to have the weaker condition $n = \omega(1)$.

Step 4: Combining the above steps, we get the following.

Corollary 2. *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 $P_{\rm e}(d_{\max}) \to 0$ as $p \to \infty$ provided that*

$$n \ge \max_{\alpha\in[\alpha^*,1]} \frac{\alpha k \log\frac{p}{k}}{\frac{1}{2}\log\Big(1+\frac{c_\beta}{\sigma^2}g(\alpha)\Big)}(1+\eta) \tag{72}$$

for some $\eta > 0$, where $g(\cdot)$ is defined in (69)*. Conversely, $P_{\rm e}(d_{\max}) \to 1$ as $p \to \infty$ whenever*

$$n \le \max_{\alpha\in[\alpha^*,1]} \frac{(\alpha-\alpha^*) k \log\frac{p}{k}}{\frac{1}{2}\log\Big(1+\frac{c_\beta}{\sigma^2}g(\alpha)\Big)}(1-\eta) \tag{73}$$

for some $\eta > 0$.

Proof. The condition in (72) is obtained using (37) and (70). By the assumption $k = o(p)$, the numerator in (72) coincides with $\log\binom{p-k}{\lfloor \alpha k \rfloor}$ up to remainder terms in Stirling's approximation that can be factored into η. The factor $\log\Big(\frac{k^2}{\delta_1^2}\binom{k}{|s_{\rm dif}|}^2\Big)$ in (37) has been factored into η; this is valid when $\delta_1 \to 0$ sufficiently slowly due to the fact that $\log\big(k\binom{k}{|s_{\mathrm{dif}}|}\big) = O(k)$, whereas (again using the assumption $k = o(p)$) the numerator in (72) behaves as $\omega(k)$. We claim that the factor $\gamma = I_0 + \sqrt{\frac{V_0}{\delta_0}}$ resulting from (29) can also be factored into η for some vanishing sequence of parameters δ_0 indexed by p. To see this, we consider without loss of generality the "worst-case" setting in which (72) holds with equality. We readily obtain $n = \Theta(k \log \frac{p}{k})$, which in turn implies from Proposition 2 that $I_0 = O\big(k\log(1 + k\sigma_\beta^2 \log\frac{p}{k})\big) = O(k \log\log\frac{p}{k})$ and $\sqrt{V_0} = O(\sqrt{k\log\frac{p}{k}})$. Thus, $I_0 + \sqrt{\frac{V_0}{\delta_0}}$ is dominated by the numerator of (72) if δ_0 is chosen to decay as (for example) $\Theta\big(\frac{1}{\log k}\big)$. The fact that $n = \Theta(k\log\frac{p}{k})$ also implies (71).

The converse bound in (72) is obtained similarly using (37), except for the term α^* in the numerator. To see how this arises, we consider an arbitrary value of $\alpha \in (\alpha^*, 1]$ and set $\ell = \lfloor \alpha k \rfloor$; the case $\alpha = \alpha^*$ follows by continuity. The term $\log\binom{p-k+\ell}{\ell}$ is handled in the same way as the term $\log\binom{p-k}{\ell}$ above, so we focus on the term $\log\sum_{d=0}^{d_{\max}}\binom{p-k}{d}\binom{\ell}{d}$. This is upper bounded by $\max_{d=0,\dotsc,d_{\max}} \log\big((1+d_{\max})\binom{p-k}{d}\binom{k}{d}\big)$. Similarly to the achievability part, we can factor $\log\big((1+d_{\max})\log\binom{k}{d}\big)$ into η, so we are left with $\log\binom{p-k}{d_{\max}}$. Approximating this using Stirling's approximation as before, and recalling that $d_{\max} = \lfloor \alpha^* k \rfloor$, we obtain the desired term $\alpha^* k \log\frac{p}{k}$. □

While the achievability and converse bounds in Corollary 2 do not have the same constants, the two are similar, and always have the same scaling laws. In the limit as $c_\beta \to \infty$, we have $\frac{1}{2}\log\big(1 + \frac{c_\beta}{\sigma^2}g(\alpha)\big) = \frac{1}{2}(\log c_\beta)(1+o(1))$; in this case, the maxima in (72)–(73) are both achieved with $\alpha \to 1$, and hence, the two bounds coincide to within a multiplicative factor of $\frac{1}{1-\alpha^*}$.

Corollary 2 is related to the setting studied by Reeves and Gastpar [15], [16], but considers $k = o(p)$ instead of $k = \Theta(p)$. Despite this difference, it is instructive to compare the bounds upon letting the implied constant in the $\Theta(p)$ scaling tend to zero. A careful comparison reveals that the converse bounds coincide in this limit, whereas our achievability bound is slightly better, in that the analogous bound in [15] multiplies $\frac{c_\beta}{\sigma^2}g(\alpha)$ by $(\sqrt{2}-1)^2 \approx 0.17$; see [15, Eq. (21)] and [16, Eq. (25)].

In Section IV-E, we present some numerical results for this setting.

C. 1-bit Model with Discrete β_S

We now turn to the quantized counterpart of (55):

$$Y = \mathrm{sign}\big(\langle X_S, \beta_S\rangle + Z\big). \tag{74}$$

As in Section IV-A, we fix $s = \{1,\dotsc,k\}$ and let β_s be a uniformly random permutation of a fixed vector $(b_1,\dotsc,b_k)$, and we set $P_X \sim N(0,1)$. We again write the minimum and maximum absolute values of $\{b_i\}_{i=1}^k$ as $b_{\min}$ and $b_{\max}$.

The following proposition gives the required characterizations on the mutual information terms and the corresponding variance terms. Recall the binary entropy function $H_2(\cdot)$ and the Q-function $Q(\cdot)$ defined in Section I-C.

Proposition 4. *Under the preceding setup for the 1-bit model, we have the following:*

(i) The mutual information $I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)$ is given by

$$\begin{aligned} I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s) = \mathbb{E}\Bigg[& H_2\Bigg(Q\Bigg(W\sqrt{\frac{\sum_{i\in s_{\mathrm{eq}}} b_i^2}{\sigma^2 + \sum_{i\in s_{\mathrm{dif}}} b_i^2}}\Bigg)\Bigg) \\ & - H_2\Bigg(Q\Bigg(W\sqrt{\frac{1}{\sigma^2}\sum_{i\in s} b_i^2}\Bigg)\Bigg)\Bigg], \end{aligned} \tag{75}$$

where $W \sim N(0,1)$.

(ii) If $k = \Theta(1)$, $\sigma^2 = \Theta(1)$, $b_{\min} = \Theta(b_{\max})$, and $b_{\min}^2 = o(1)$, then

$$I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s) = \Bigg(\frac{1}{\pi\sigma^2}\sum_{i\in s_{\mathrm{dif}}} b_i^2\Bigg)\big(1+o(1)\big). \tag{76}$$

(iii) If $k = \Theta(p)$, $\sigma^2 = \Theta(1)$, and the entries of b_s all equal a common value b_0 such that $b_0^2 = \Theta\big(\frac{\log p}{p}\big)$, then the mutual information quantities $I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)$ with $|s_{\mathrm{dif}}| = 1$ all equal a common value I_1 satisfying

$$\begin{aligned} I_1 &= \frac{1}{2}\frac{\frac{b_0^2}{\sigma^2}}{\sqrt{2\pi k \frac{b_0^2}{\sigma^2}}}\mathbb{E}\Big[W\log\frac{1-Q(W)}{Q(W)}\Big](1+o(1)) \qquad (77)\\ &= \Theta\Bigg(\frac{\sqrt{\log p}}{p}\Bigg), \qquad (78) \end{aligned}$$

where $W \sim N(0,1)$.

(iv) The variance $V_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)$ defined in (54) *satisfies*

$$\begin{aligned} V_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s) \le c_0\Bigg(&\frac{1}{\sigma^2}\sum_{i\in s_{\mathrm{dif}}} b_i^2 + \Bigg(\frac{1}{\sigma^2}\sum_{i\in s_{\mathrm{dif}}} b_i^2\Bigg)^2 \\ &+ \min\Bigg\{1, \Bigg(\frac{1}{\sigma^2}\sum_{i\in s_{\mathrm{dif}}} b_i^2\Bigg)^2\Bigg\}\frac{1}{\sigma^2}\sum_{i\in s_{\mathrm{eq}}} b_i^2\Bigg) \end{aligned} \tag{79}$$

for some universal constant c_0.

Proof. See Appendix C. □

Below we present two corollaries corresponding to different scalings of k and the SNR, namely, those given in parts (ii) and (iii) of Proposition 4. We proceed by simultaneously presenting the steps of Procedure 1 for both settings.

Step 1: As in Section IV-A, we choose the trivial typical set $\mathcal{T}_\beta$ containing all vectors on the support of P_{β_S}.

Step 2: We make use of Chebyshev's inequality in Proposition 9 in Appendix A. Choosing $\delta = \delta_2 I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)$ in (143), it follows that we may set

$$\begin{aligned} \psi_\ell(n,\delta_2) &= \psi'_\ell(n,\delta_2) \\ &= \max_{(s_{\mathrm{dif}},s_{\mathrm{eq}},b_s)\,:\,|s_{\mathrm{dif}}|=\ell} \frac{V_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)}{n\delta_2^2 I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)^2}. \end{aligned} \tag{80}$$

Step 3: We again choose γ as in (26) so that $P_0(\gamma) = 0$. Consider the setting described in part (ii) of Proposition 4. Under the scalings therein, (76) and (79) both behave as $\Theta(b_{\min}^2)$. Hence, and using (80) and the fact that $k = \Theta(1)$, the second term in (39) vanishes provided that

$$n = \omega\Big(\frac{1}{b_{\min}^2}\Big). \tag{81}$$

The setting described in part (iii) of Proposition 4 is handled similarly. We set $\mathcal{L} = \{1\}$ in Theorem 4, thus focusing only on $\ell := |s_{\mathrm{dif}}| = 1$. Denoting the corresponding variance $V_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)$ by V_1, it follows by substituting the scalings of k, σ^2 and b_0^2 into (79) that $V_1 = O\big(\frac{\log p}{p}\big)$. It thus follows from (78) and (80) that $\psi_1'(n, \delta_2)$ vanishes provided that

$$n = \omega(p). \tag{82}$$

Step 4: Combining the above steps and applying asymptotic simplifications, we obtain the following corollaries.

Corollary 3. *Under the preceding setup for the 1-bit model with $k = \Theta(1)$, $\sigma^2 = \Theta(1)$, $b_{\min} = \Theta(b_{\max})$, and $b_{\min} = o(1)$, we have $P_{\mathrm{e}} \to 0$ as $p \to \infty$ provided that*

$$n \ge \max_{s_{\mathrm{dif}} \ne \emptyset} \frac{|s_{\mathrm{dif}}| \log p}{\frac{1}{\pi\sigma^2} \sum_{i \in s_{\mathrm{dif}}} b_i^2} (1 + \eta) \tag{83}$$

for some $\eta > 0$. Conversely, $P_{\mathrm{e}} \to 1$ as $p \to \infty$ whenever

$$n \le \max_{s_{\mathrm{dif}} \ne \emptyset} \frac{|s_{\mathrm{dif}}| \log p}{\frac{1}{\pi\sigma^2} \sum_{i \in s_{\mathrm{dif}}} b_i^2} (1 - \eta) \tag{84}$$

for some $\eta > 0$.

Proof. We obtain (83) and (84) from (37) and (42) respectively. The denominators are obtained directly from part (ii) of Proposition 4, and the numerators follow from the identity $\log\binom{p}{|s_{\mathrm{dif}}|} = (|s_{\mathrm{dif}}| \log p)(1 + o(1))$, which holds whenever $k = \Theta(1)$ and hence $|s_{\mathrm{dif}}| = \Theta(1)$. By the assumption $k = \Theta(1)$, the remaining terms in (37) (including the choice of γ in (26)) can be factored into η. The condition in (81) is implied by (83) (or by (84) when equality holds) by the same argument as Corollary 1. □

Corollary 4. *Under the preceding setup for the 1-bit model with $k = \Theta(p)$, $\sigma^2 = \Theta(1)$, and the entries of β_s deterministically equaling a common value b_0 such that $b_0^2 = \Theta\big(\frac{\log p}{p}\big)$, we have $P_{\mathrm{e}} \to 1$ provided that*

$$\begin{aligned} n &\le \frac{\log p}{\frac{1}{2} \frac{\frac{b_0^2}{\sigma^2}}{\sqrt{2\pi k \frac{b_0^2}{\sigma^2}}} \mathbb{E}\Big[W \log \frac{1 - Q(W)}{Q(W)}\Big]} (1 - \eta) && (85) \\ &= \Theta\big(p\sqrt{\log p}\big) && (86) \end{aligned}$$

for some $\eta \in (0,1)$, where $W \sim N(0,1)$.

Proof. The condition in (85) follows using (42) with $|s_{\mathrm{dif}}| = 1$; the numerator behaves as $(\log p)(1 + o(1))$, and the denominator behaves according to (77). The additional condition in (82) is satisfied when (85) holds with equality. □

In the same way as (62)–(63), one can compute (83)–(84) without evaluating all $2^k - 1$ objective values; for a given value of $|s_{\mathrm{dif}}|$, the maximum is achieved by the set s_{dif} with the smallest value of $\sum_{i \in s_{\mathrm{dif}}} b_i^2$.

The asymptotic identities used in the proof of Corollary 3 can directly be applied to (62)–(63) with $k = \Theta(1)$ and $b_{\min} = o(1)$, and the resulting expressions are precisely those in (83)–(84) with $\frac{1}{\pi}$ replaced by $\frac{1}{2}$. Thus, this is a case where there is only a minor loss in the performance due to the quantization; the corresponding asymptotic number of measurements only increases by a factor of $\frac{\pi}{2} \approx 1.57$.

In contrast, Corollary 4 describes a setting where the linear model and its 1-bit counterpart lead to significantly different requirements on the number of measurements. Under the scaling described therein, the necessary and sufficient number of measurements for the linear model behaves as $\Theta(p)$ [8, Table I]. Thus, the 1-bit quantization increases the required number of measurements from linear to super-linear in the ambient dimension.

D. 1-bit Model with Gaussian β_S and Partial Recovery

We now consider the 1-bit counterpart of the setting studied in Section IV-B, where β_s is i.i.d. on $N(0, \sigma_\beta^2)$ for some $\sigma_\beta^2 = \frac{c_\beta}{k}$, and we seek partial recovery as in (7) with $d_{\max} = \lfloor \alpha^* k \rfloor$. We make use of the following.

Proposition 5. *Under the preceding setup for the 1-bit model, the quantity $I_{0,+}$ in (30) satisfies*

$$I_{0,+} \le \frac{k}{2} \log\Big(1 + \frac{n\sigma_\beta^2}{\sigma^2}\Big) + \sqrt{k \log\Big(1 + \frac{n\sigma_\beta^2}{\sigma^2}\Big)}. \tag{87}$$

for some universal constant c_0'.

Proof. By the data processing inequality, I_0 must satisfy (66) even in the 1-bit setting. We immediately obtain (87) from the identity $I_{0,+} \le I_0 + \sqrt{2I_0}$ given in [41]. □

We now turn to the steps for providing a counterpart to Corollary 2. We define the function

$$\begin{aligned} \Psi(\alpha, c_\beta, \sigma) := \mathbb{E}\Bigg[& H_2\Bigg(Q\Bigg(W\sqrt{\frac{c_\beta(1 - g(\alpha))}{\sigma^2 + c_\beta g(\alpha)}}\Bigg)\Bigg) \\ & - H_2\Bigg(Q\Bigg(W\sqrt{\frac{c_\beta}{\sigma^2}}\Bigg)\Bigg)\Bigg], \end{aligned} \tag{88}$$

where $W \sim N(0,1)$, and $g(\alpha)$ is defined in (69).

Step 1: We choose the same typical set $\mathcal{T}_\beta$ as that in Section IV-B, thus ensuring that (68) holds for all sequences of typical vectors. It follows that $\sum_{i \in s_{\mathrm{dif}}} b_i^2 \to c_\beta g(\alpha)$ and $\sum_{i \in s_{\mathrm{eq}}} b_i^2 \to c_\beta(1 - g(\alpha))$ for the pair $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$ with corresponding sizes $(\ell, k - \ell)$ ($\ell = \lfloor \alpha k \rfloor$) such that $\sum_{i \in s_{\mathrm{dif}}} b_i^2$ is minimized. We observe from (75) that minimizing $\sum_{i \in s_{\mathrm{dif}}} b_i^2$ also amounts to minimizing $I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)$ for a fixed value of $|s_{\mathrm{dif}}|$, as was the case for the linear model. If $\frac{|s_{\mathrm{dif}}|}{k}$ converges to a given constant α, then the corresponding mutual information converges as follows, in accordance with (75) and (88):

$$I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s) \to \Psi(\alpha, c_\beta, \sigma). \tag{89}$$

Step 2: We make use of the general concentration inequality given in Proposition 10 in Appendix A; setting $\delta = \delta_2 I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)$ in (145) in Appendix A gives

$$\psi_\ell(n,\delta_2) - \psi'_\ell(n,\delta_2)$$
$$= \max_{(s_{\mathrm{dif}},s_{\mathrm{eq}},b_s)\,:\,|s_{\mathrm{dif}}|=\ell} 2\exp\left(- \frac{(\delta_2 I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s))^2 n}{2(8|\mathcal{Y}| + 2\delta_2 I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s))} \right). \tag{90}$$

Step 3: We choose $\gamma = \frac{I_{0,+}}{\delta_0}$ as in (31), ensuring that $P_0(\gamma) \le \delta_0$. The other remainder terms are controlled in the same way as Section IV-B. We again use the fact that the typical realizations b_s of β_s satisfy (68), and yield $\sum_{i\in s_{\mathrm{dif}}} b_i^2 = \Omega(1)$, and hence $\alpha^2_{s_{\mathrm{dif}}} = \Theta(1)$ in (58). We also have $I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s) = \Theta(1)$ in (75); this is seen by noting that the smallest mutual information for a fixed $|s_{\mathrm{dif}}| = \lfloor \alpha k \rfloor$ satisfies (89), and the mutual information upper bounded by $\log 2$ since the observations are binary. It follows that the exponent in (90) behaves as $\Theta(n)$; hence, following the arguments in Section IV-B, we conclude that the second term in (39) vanishes provided that

$$n = \Omega(k) \tag{91}$$

with a sufficiently large implied constant. Once again, for the converse part, one can analogously show that the weaker condition $n = \omega(1)$ suffices.

Step 4: Combining the above steps, we get the following.

Corollary 5. *Under the preceding setup for the 1-bit model with $k \to \infty$ and $k = o(p)$, $\sigma^2 = \Theta(1)$, $\sigma^2_\beta = \frac{c_\beta}{k}$ 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*

$$n \ge \max_{\alpha \in [\alpha^*,1]} \frac{\alpha k \log \frac{p}{k}}{\Psi(\alpha, c_\beta, \sigma)}(1+\eta) \tag{92}$$

for some $\eta > 0$, where Ψ is defined in (88). Conversely, $P_{\mathrm{e}}(d_{\max}) \to 1$ as $p \to \infty$ whenever

$$n \le \max_{\alpha \in [\alpha^*,1]} \frac{(\alpha - \alpha^*) k \log \frac{p}{k}}{\Psi(\alpha, c_\beta, \sigma)}(1-\eta) \tag{93}$$

for some $\eta > 0$.

Proof. As usual, we begin with the conditions in (37) and (42). The denominators in (92)–(93) follow directly by applying (89). Moreover, the terms $\alpha k \log \frac{p}{k}$ and $(\alpha - \alpha^*) k \log \frac{p}{k}$ in the numerators are obtained in an identical fashion to Corollary 2 once we show that there exists a vanishing sequence of constants δ_0, indexed by p, such that the remainder term $\gamma = \frac{I_{0,+}}{\delta_0}$ resulting from (31) can be factored into η. To see this, we note that the right-hand side of (92) behaves as $\Theta(k \log p)$, whereas from Proposition 5 (with the scalings $n = \Theta(k \log p)$ and $\sigma^2_\beta = \Theta\big(\frac{1}{k}\big)$), $I_{0,+}$ behaves as $O\big(k \log\log p\big)$. We may thus set δ_0 to be (for example) $\frac{\log\log p}{\sqrt{\log p}}$. Finally, we observe that (91) holds whenever (92) holds, and similarly for the converse part. □

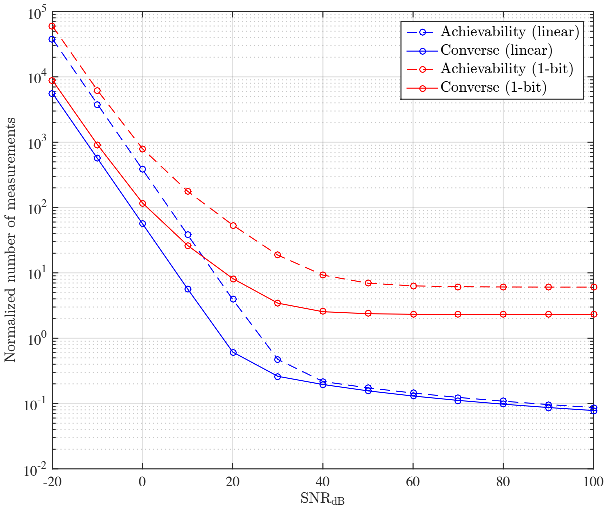

Figure 2: Asymptotic thresholds on the number of measurements required for partial support recovery for the linear and 1-bit models, with $\alpha^* = 0.1$. The number of measurements is normalized by $k \log \frac{p}{k}$, and $\mathrm{SNR}_{\mathrm{dB}}$ is defined in (94).

The main difference in (92)–(93) compared to the linear counterparts in (72)–(73) is the behavior in the limit as $c_\beta := k\sigma^2_\beta \to \infty$. As stated following Corollary 2, the denominator in the linear setting behaves as $(\log c_\beta)(1 + o(1))$, thus tending towards infinity. In contrast, for the 1-bit setting, we have $\Psi \le \log 2$ due to the fact that $H_2(\cdot) \in [0, \log 2]$, and thus the denominator cannot grow unbounded. These observations are consistent with Corollary 4, which shows that 1-bit CS can require significantly more measurements compared to the linear setting when the signal-to-noise ratio (SNR) is sufficiently high.

E. Numerical Evaluations for Partial Recovery

In this subsection, we present numerical calculations for the settings considered in Sections IV-B and IV-D. We set $\alpha^* = 0.1$, $\sigma^2 = 1$, and $k = o(p)$. We consider values of σ^2_β of the form $\sigma^2_\beta = \frac{c_\beta}{k}$ for fixed c_β. Similarly to [15], [16], we present our results in terms of

$$\mathrm{SNR}_{\mathrm{dB}} := 10 \log \frac{k\sigma^2_\beta}{\sigma^2} = 10 \log c_\beta, \tag{94}$$

which represents the per-sample SNR in dB.

Figure 2 plots the asymptotic thresholds on the number of measurements from Corollaries 2 and 5. For both the linear and 1-bit settings, there is a close correspondence between the necessary and sufficient number of measurements. The bounds for the two models nearly coincide at low SNR, which is consistent with Corollary 3.

The behavior of the bounds at high SNRs is also consistent with our previous discussions. In the linear setting, the ratio between the bounds narrows to approximately

1.11 as the SNR grows large, which coincides with the value $\frac{1}{1-\alpha^*}$ given in the discussion following Corollary 2. Moreover, as discussed following Corollary 5, the number of measurements steadily decreases for increasing SNRs for the linear model, while saturating at an asymptotic limit for the 1-bit model.

F. Group Testing

1) Noiseless Case with Exact Recovery: Here we consider the noiseless group testing problem, where each observation is deterministically generated according to

$$Y = \mathbb{1}\Big\{ \bigcup_{i \in S} \{X_i = 1\} \Big\}. \tag{95}$$

We consider Bernoulli measurement matrices with $P_X(1) = 1 - P_X(0) = \frac{\nu}{k}$, where ν is a constant not depending on p. Here there is no latent variable β_s, which can equivalently be thought of as corresponding to β_s equaling the vector of ones deterministically. This implies that $I_{s_{\mathrm{dif}}, s_{\mathrm{eq}}}(b_s)$ depends only on $\ell = |s_{\mathrm{dif}}|$, and we emphasize this by writing it as I_ℓ. Our setting readily handles both fixed and growing k; since the former is already well-understood [2], [11], [42], we focus our attention on the case that $k \to \infty$, and in particular on the case that $k = \Theta(p^\theta)$ for some $\theta \in (0,1)$.

Proposition 6. *Under the noiseless group testing setup, consider arbitrary sequences of sparsity levels $k \to \infty$ and $\ell \in \{1, \dots, k\}$, indexed by p. If $\frac{\ell}{k} = o(1)$, then*

$$I_\ell = \left(e^{-\nu} \nu \frac{\ell}{k} \log \frac{k}{\ell} \right)(1 + o(1)). \tag{96}$$

Moreover, if $\frac{\ell}{k} \to \alpha \in (0,1]$, then

$$I_\ell = e^{-(1-\alpha)\nu} H_2\big(e^{-\alpha\nu}\big)(1 + o(1)). \tag{97}$$

Proof. See Appendix D. □

We proceed with the steps of Procedure 1.

Step 1: The first step is trivial; β_s is deterministic, and thus the typical set $\mathcal{T}_\beta$ is a singleton.

Step 2: In contrast to the previous examples, we use different concentration inequalities to handle different values of ℓ. Moreover, in accordance with Remark 2, we let δ_2 depend on ℓ, writing it as $\delta_{2,\ell}$. For "large" values of ℓ (to be made precise below), we will apply the general bound in Proposition 10 in Appendix A. For "small" values of ℓ, we use the following to obtain an improved bound.

Proposition 7. *For the noiseless group testing problem, consider sequences $k \to \infty$ and ℓ, indexed by p, such that $\frac{\ell}{k} \to 0$. For any $\epsilon > 0$ and $\delta_{2,\ell} \in (0,1)$ bounded away from zero and one, the following holds for sufficiently large p:*

$$\begin{aligned} &\mathbb{P}\Big[\imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y} | \mathbf{X}_{s_{\mathrm{eq}}}, b_s) \le n I_\ell (1 - \delta_{2,\ell}) \Big] \\ &\le \exp\left(-n \frac{\ell}{k} e^{-\nu} \nu \Big((1-\delta_{2,\ell}) \log(1-\delta_{2,\ell}) + \delta_{2,\ell} \Big) (1-\epsilon) \right) \end{aligned} \tag{98}$$

for all $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$ with $|s_{\mathrm{dif}}| = \ell$.

Proof. See Appendix D. □

From the bounds in (98) and (145) in Appendix A, we may fix $\epsilon > 0$ and choose the following when p is sufficiently large:

- For $\ell \le \ell \le \lfloor \frac{k}{\log k} \rfloor$:

$$\begin{aligned} &\psi_\ell(n, \delta_{2,\ell}) = \\ &\exp\left(-n \frac{\ell}{k} e^{-\nu} \nu \Big((1-\delta_{2,\ell}) \log(1-\delta_{2,\ell}) + \delta_{2,\ell} \Big) (1-\epsilon) \right). \end{aligned} \tag{99}$$

- For $\ell > \lfloor \frac{k}{\log k} \rfloor$:

$$\psi_\ell(n, \delta_{2,\ell}) = 2 \exp\left(- \frac{(\delta_{2,\ell} I_\ell)^2 n}{2(16 + 2\delta_{2,\ell} I_\ell)} \right). \tag{100}$$

For the converse, we only use the latter of these two cases, setting $\psi'_\ell(n, \delta_{2,\ell}) = 2 \exp\big(- \frac{(\delta_{2,\ell} I_\ell)^2 n}{2(16 + 2\delta_{2,\ell} I_\ell)} \big)$.

Step 3: Since β_s is deterministic, we may trivially set $\gamma = 0$ to obtain $P_0(\gamma) = 0$ in (25).

For the converse, we set $\mathcal{L} = \{k\}$ in Theorem 4. From the above choice of ψ'_ℓ and the growth of I_k in (97), we immediately obtain that $\psi'_k \to 0$ whenever $n = \omega(1)$. The achievability part requires more effort; we summarize the findings in the following proposition.

Proposition 8. *Let $k = \Theta(p^\theta)$ for some $\theta \in (0,1)$.*

(i) For any $\eta > 0$, there exists $\delta_2^{(1)} \in (0,1)$ and a choice of $\epsilon > 0$ in (99) such that $\sum_{\ell=1}^{\lfloor \frac{k}{\log k} \rfloor} \binom{k}{\ell} \psi_\ell(n, \delta_2^{(1)}) \to 0$ provided that

$$n \ge \frac{\frac{\theta}{1-\theta} k \log \frac{p}{k}}{e^{-\nu} \nu} (1 + \eta). \tag{101}$$

(ii) For any $\delta_2^{(2)} \in (0,1)$, we have $\sum_{\lfloor \frac{k}{\log k} \rfloor + 1}^{k} \binom{k}{\ell} \psi_\ell(n, \delta_2^{(2)}) \to 0$ provided $n = \Omega\big(k \log \frac{p}{k}\big)$.

Proof. See Appendix D. □

The idea here is that for the smaller values of ℓ, it is the concentration inequality that dominates the final bound, so we let $\delta_{2,\ell} = \delta_2^{(1)}$ be closer to one to provide better concentration behavior. For large values of ℓ, the opposite is true, so we let $\delta_{2,\ell} = \delta_2^{(2)}$ be close to zero.

Step 4: We obtain the following corollary by combining the previous steps and applying asymptotic simplifications.

Corollary 6. *For the noiseless group testing problem with $k = \Theta(p^\theta)$ ($\theta \in (0,1)$) and an optimized parameter ν, we have $P_{\mathrm{e}} \to 0$ as $p \to \infty$ provided that*

$$n \ge \inf_{\nu > 0} \max\left\{ \frac{\theta}{e^{-\nu}\nu(1-\theta)}, \frac{1}{H_2(e^{-\nu})} \right\} \Big(k \log \frac{p}{k} \Big) (1 + \eta) \tag{102}$$

for some $\eta > 0$. Conversely, we have $P_{\mathrm{e}} \to 1$ as $p \to \infty$ whenever

$$n \le \frac{k \log \frac{p}{k}}{\log 2} (1 - \eta) \tag{103}$$

for some $\eta > 0$.

Proof. We first consider the achievability part. We immediately obtain the first term in the maximum in (102) from (101), so it remains to derive the second term. We start with (37); by substituting $\gamma = 0$ and taking $\delta_1 \to 0$ sufficiently slowly, we obtain

$$n \geq \max_{\ell=1,\dots,k} \frac{\log\binom{p-k}{\ell} + 2\log\big(k\binom{k}{\ell}\big)}{I_\ell(1-\delta_{2,\ell})}\big(1+o(1)\big). \tag{104}$$

Using (96)–(97) and the asymptotic identity $\log\binom{p-k}{\ell} = \Theta\big(\ell\log\frac{p}{\ell}\big)$ we see that the objective in (104) behaves as

$$\Theta\left(\frac{k\log\frac{p}{\ell}}{1+\log\frac{k}{\ell}}\right) \tag{105}$$

whenever the constants $\{\delta_{2,\ell}\}$ are bounded away from one. This behaves as $\Theta\big(k\log\frac{p}{k}\big)$ when $\frac{\ell}{k} = \Theta(1)$, and as $\Theta\big(\frac{k\log\frac{p}{k}}{\log\frac{k}{\ell}} + k\big)$ when $\frac{\ell}{k} = o(1)$ (the latter of these is seen by writing $\log\frac{p}{\ell} = \log\frac{p}{k} + \log\frac{k}{\ell}$). Thus, the maximum in (104) can only be achieved by a sequence such that $\frac{\ell}{k} = \Theta(1)$. Moreover, with $\frac{\ell}{k} = \Theta(1)$, we see from the assumption $k = o(p)$ that the term $2\log\big(k\binom{k}{\ell}\big) = O(k)$ is dominated by $\log\binom{p-k}{\ell} = \Theta\big(k\log\frac{p}{k}\big)$, and can thus be factored into the $o(1)$ remainder term in (104). This yields the condition

$$n \geq \max_{\ell=1,\dots,k} \frac{\ell\log\frac{p}{\ell}}{I_\ell(1-\delta_{2,\ell})}\big(1+o(1)\big). \tag{106}$$

Since the maximum can only be achieved asymptotically if $\frac{\ell}{k} = \Theta(1)$, we proceed by considering $\frac{\ell}{k} \to \alpha$ for some arbitrary $\alpha \in (0,1]$. Under this scaling, $\ell\log\frac{p}{\ell}$ behaves as $\big(\alpha k\log\frac{p}{k}\big)(1+o(1))$. Moreover, according to Proposition 8, we can choose $\delta_{2,\ell}$ to be arbitrarily small for all ℓ values except those below $\lfloor\frac{k}{\log k}\rfloor$. Such values behave as $o(k)$, and thus do achieve the maximum in (106). Combining these observations with (97), the right-hand side of (106) yields the condition

$$n \geq \max_{\alpha\in(0,1]} \frac{\alpha k\log\frac{p}{k}}{e^{-(1-\alpha)\nu}H_2\big(e^{-\alpha\nu}\big)}\big(1+\eta\big), \tag{107}$$

where η may be arbitrarily small. By a change of variable $\lambda = e^{-\alpha\nu}$, the coefficient to $k\log\frac{p}{k}$ can be written as $\frac{1}{\nu}e^{\nu}\frac{\lambda\log\frac{1}{\lambda}}{H_2(\lambda)}$. This is easily verified to be decreasing in $\lambda \in [0,1]$, which implies that the maximizing value of α is one, and yields the second term in (102).

The converse part is similar but considerably simpler; by setting $\mathcal{L} = \{k\}$ in Theorem 3, we obtain $\alpha = 1$ immediately. The denominator $\log 2$ in (103) is obtained by maximizing $H_2(e^{-\nu})$ over ν, and the condition $n \to \infty$ stated before Proposition 8 is clearly satisfied when (103) holds with equality. □

By setting $\nu = \log 2$ in (102), it is readily verified that the necessary and sufficient conditions coincide for $\theta \leq \frac{1}{3}$, and in fact yield the same threshold as *adaptive* group testing [31]. To our knowledge, this was only known previously in the limit as $\theta \to 0$ [32]. Further comparisons to previous works are provided at the end of this subsection.

2) Noisy Case with Exact Recovery: We now turn to the noisy counterpart of (95):

$$Y = \mathbb{1}\bigg\{\bigcup_{i\in S}\{X_i = 1\}\bigg\} \oplus Z, \tag{108}$$

where $Z \in \{0,1\}$ is additive noise, and $\oplus$ denotes modulo-2 addition. For concreteness, we focus on the case that $Z \sim \text{Bernoulli}(\rho)$ for some $\rho \in (0,\frac{1}{2})$ not varying with p, though other noise models also fall into our framework (e.g., see [2]). As discussed below, we do not attempt to provide results with constants that are optimized to the same extent as the noiseless case, and we thus set $\nu = \log 2$, i.e., $P_X \sim \text{Bernoulli}\big(\frac{\log 2}{k}\big)$.

We follow Procedure 1 in a similar fashion to the noiseless case, altering the statements of Proposition 6–8 accordingly. To avoid repetition, we give the modified propositions and their proofs in Appendix E, and state the resulting corollary here. The main difference is that in the analog of Proposition 8, we let $\delta_2^{(1)}$ remain arbitrary, thus leading to the optimization parameter δ_2 in the following.

Corollary 7. *Under the preceding setup for the noisy group testing problem with $\rho \in (0, 0.5)$, $\nu = \log 2$, and $k = \Theta(p^\theta)$ ($\theta \in (0,1)$), we have $P_\mathrm{e} \to 0$ as $p \to \infty$ provided that*

$$n \geq \inf_{\delta_2\in(0,1)} \max\left\{\zeta(\rho,\delta_2,\theta), \frac{1}{\log 2 - H_2(\rho)}\right\}\left(k\log\frac{p}{k}\right) \times (1+\eta) \tag{109}$$

for some $\eta > 0$, where

$$\zeta(\rho,\delta_2,\theta) := \frac{2}{\log 2}\max\left\{\frac{2(1+\frac{1}{3}\delta_2(1-2\rho))\frac{\theta}{1-\theta}}{\delta_2^2(1-2\rho)^2}, \frac{\frac{1+4\theta}{1-\theta}}{(1-2\rho)\log\frac{1-\rho}{\rho}(1-\delta_2)}\right\}. \tag{110}$$

Conversely, we have $P_\mathrm{e} \to 1$ as $p \to \infty$ whenever

$$n \leq \frac{k\log\frac{p}{k}}{\log 2 - H_2(\rho)}(1-\eta). \tag{111}$$

for some $\eta > 0$.

Proof. See Appendix E. □

As we will see in the numerical examples below, Corollary 7 provides an exact asymptotic threshold for a narrower range of θ values compared to the noiseless case. This is due to the difficulty in precisely characterizing the concentration behavior of the information density tail probabilities. Nevertheless, the second term in the maximum in (109) is always dominant for sufficiently small θ, thus matching the converse. To see this, we first note that the first term in the maximum in (110) tends to zero as $\theta \to 0$, and cannot be dominant in this limit. This implies that δ_2 may be arbitrarily close to zero provided that θ is sufficiently small. Assuming then that δ_2 and θ are small and the maximum in (110) is achieved by the second term, we can write $\zeta(\rho,\delta_2,\theta) \approx \frac{2}{\log 2}\frac{1}{(1-2\rho)\log\frac{1-\rho}{\rho}}$. This is strictly smaller than $\frac{1}{\log 2 - H_2(\rho)}$; see Proposition 14 in Appendix E.

3) Partial Recovery: The consideration of partial recovery (*cf.* (7)) in fact leads to *simpler* expressions and proofs, as seen in the following.

Corollary 8. *Under the preceding setup for the group testing problem with $\rho \in [0, 0.5)$ (i.e., possibly noiseless), $\nu = \log 2$, $k \to \infty$, $k = o(p)$, and $d_{\max} = \lfloor \alpha^* k \rfloor$ for some $\alpha^* \in (0,1)$, we have $P_{\rm e}(d_{\max}) \to 0$ as $p \to \infty$ provided*

$$n \geq \frac{k \log \frac{p}{k}}{\log 2 - H_2(\rho)}(1+\eta) \tag{112}$$

for some $\eta > 0$. Conversely, $P_{\rm e}(d_{\max}) \to 1$ as $p \to \infty$ whenever

$$n \leq \frac{(1-\alpha^*)\big(k \log \frac{p}{k}\big)}{\log 2 - H_2(\rho)}(1-\eta) \tag{113}$$

for some $\eta > 0$.

Proof. The achievability part follows the proofs of Corollaries 6 and 7, except that the "small" values of ℓ need not be handled. That is, we only make use of the general concentration inequality in (145) in Appendix A, and we end up with the single condition in (112). For the converse part, we again choose $\mathcal{L} = \{k\}$ in Theorem 6, and the steps are again similar, with the multiplicative factor $1-\alpha^*$ arising via identical reasoning to Corollary 2. □

Corollary 8 shows that at least for sufficiently small θ (e.g., $k = O\big(p^{\frac{1}{3}}\big)$ in the noiseless case), there is not much to be saved by moving from exact recovery to partial recovery: Allowing for a fraction α^* of errors leads to at most a reduction in the number of measurements of a multiplicative factor $1-\alpha^*$.

4) Numerical Evaluations: In Figure 3, we compare the bounds in Corollary 6 with existing asymptotic bounds in the literature. For convenience, we switch to base-2 logarithms and plot the asymptotic limit of the ratio $\frac{k \log_2 \frac{p}{k}}{n}$, so that a higher value corresponds to fewer measurements. We see that our achievability bound improves on all of the existing bounds; however, we note that the Combinatorial Optimal Matching Pursuit (COMP) [12] and Definite Defective (DD) [43] algorithms are computationally tractable and do not require knowledge of k.

The converse bound shown is known to hold even for adaptive measurement matrices [31]. Thus, a key implication of our results is that adaptivity provides no asymptotic gain over non-adaptive Bernoulli measurements when $k = O(p^{\frac{1}{3}})$. It remains an important open problem to derive *practical* decoding schemes for achieving the bound in the non-adaptive setting.

Figure 4 provides an analogous plot for the noisy case, with three different noise levels (i.e., values of ρ). In each case, we obtain an exact threshold for sufficiently small θ, albeit over a narrower range than the noiseless case. Once again, the converse is known to hold even in the adaptive setting [12], and we have thus provided cases where non-adaptive Bernoulli measurements yield the same asymptotics as optimal adaptive measurements. To our knowledge, this has not been shown previously even in the limit as $\theta \to 0$.

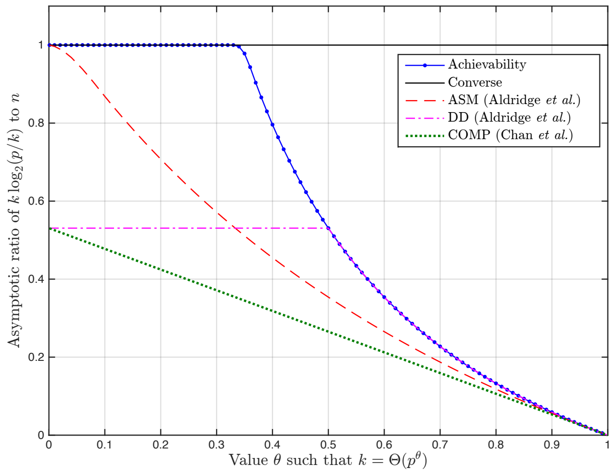

Figure 3: Asymptotic thresholds on the number of measurements required for noiseless group testing.

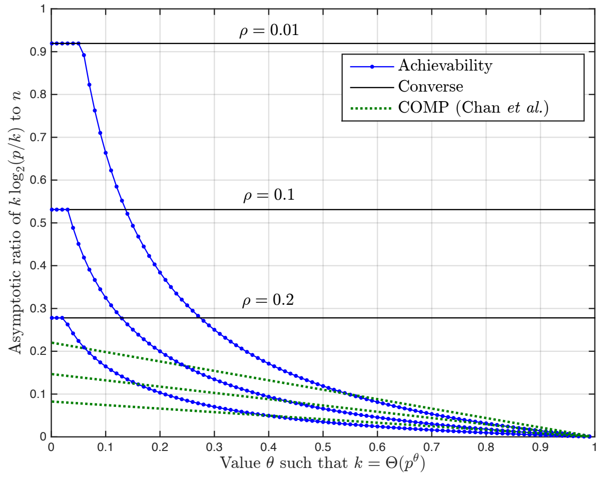

Figure 4: Asymptotic thresholds on the number of measurements required for noisy group testing.

G. General Strong Impossibility result for Discrete Observation Models

Equation (144) in Appendix A bounds the variance of the information density uniformly in terms of the output alphabet size for models with discrete observations. Notable examples include group testing, the 1-bit model (or more generally, quantizations with more than two levels), and logistic regression. We obtain the following general strong impossibility result (i.e., conditions under which $P_{\rm e} \to 1$) by combining Proposition 9 in Appendix A with a variant of Theorem 4.

Corollary 9. *If the observations lie in a finite set $\mathcal{Y} \subset \mathbb{R}$ with probability one, then $P_{\rm e} \to 1$ whenever there exist vanishing sequences $\delta_{1,p} \to 0$ and $\epsilon_p \to 0$ such that*

$$n \geq \max_{(s_{\rm dif}, s_{\rm eq})\,:\, s_{\rm dif} \neq \emptyset} \frac{\log \binom{p-k+|s_{\rm dif}|}{|s_{\rm dif}|} - \log \delta_{1,p}}{I_{s_{\rm dif}, s_{\rm eq}}(b_s) + \sqrt{\frac{|\mathcal{Y}|}{n\epsilon_p}}} \tag{114}$$

for all $b_s \in \mathbb{R}^k$ within a set whose probability under P_{β_s} approaches one.

Proof. In this application, we do not use Theorem 4 directly, but instead follow the arguments leading up to it with (40)–(41) replaced by

$$\log\binom{p-k+|s_{\mathrm{dif}}|}{|s_{\mathrm{dif}}|} - \log\delta_1 \ge n(I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)+\delta); \tag{115}$$

and

$$\mathbb{P}\Big[\imath^n(\mathbf{X}_{s_{\mathrm{dif}}};\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},b_s) \le n(I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)+\delta)\,\Big|\,\beta_s=b_s\Big] \ge 1-\frac{|\mathcal{Y}|(\frac{4}{e})^2}{\delta^2 n}. \tag{116}$$

By Proposition 9 in Appendix A, we have for all $(s_{\mathrm{dif}},s_{\mathrm{eq}},b_s)$ that (116) holds, so the analogous probability to that on the right-hand side of (45) is dictated only by (115). Moreover, the right-hand side of (116) tends to one upon setting $\delta=\sqrt{\frac{|\mathcal{Y}|}{n\epsilon_p}}$ for some $\epsilon_p \to 0$. By also setting $\delta_1=\delta_{1,p}\to 0$ (so that the analogous additive term to that of δ_1 in (45) vanishes), we see that (115) coincides with (114). □

When β_s deterministic, this theorem recovers a recent result by Tan and Atia [25], which was proved using combinatorial techniques. Our result is in fact slightly stronger in the sense that the additive term in the denominator only behaves as $\omega\big(\frac{1}{\sqrt{n}}\big)$, whereas the corresponding term in [25] behaves as $\omega\big(\frac{1}{n^{1/4}}\big)$. Thus, in our result, the mutual information term remains the dominant one in a wider range of settings.

V. Proofs of General Bounds

Here we provide the proofs of Theorems 1 and 2, and then give the changes required to obtain the results for partial recovery in Section III-C. As mentioned previously, the proofs bear some resemblance to those of mixed channels in channel coding [29, Sec. 3.3]. However, the analysis here is more involved, primarily due to the fact that the "codewords" $\mathbf{X}_s$ are not independent for different values of $s\in\mathcal{S}$, but instead share common columns corresponding to the overlapping parts of the support set. See [5], [7], [17] for further discussions on the differences between support recovery and channel coding.

A. Proof of Theorem 1

1) Initial Non-Asymptotic Bound: Recall the definitions of the random variables in (10)–(11), and the information densities in (19)–(21). We fix the constants $\gamma_1,\dots,\gamma_k$ arbitrarily, and consider a decoder that searches for the unique set $s\in\mathcal{S}$ such that

$$\imath(\mathbf{x}_{s_{\mathrm{dif}}};\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}}) > \gamma_{|s_{\mathrm{dif}}|} \tag{117}$$

for all 2^k-1 partitions $(s_{\mathrm{dif}},s_{\mathrm{eq}})$ of s with $s_{\mathrm{dif}}\ne\emptyset$. An error occurs if no such s exists, if multiple exist, or if such a set differs from the true value.

Since the joint distribution of $(\beta_s,\mathbf{X}_s,\mathbf{Y}_s\,|\,S=s)$ is the same for all s in our setup (*cf.* Section II), and the decoder that we have chosen exhibits a similar symmetry, we can condition on a fixed and arbitrary value of S, say $s=\{1,\dots,k\}$. By the union bound, the error probability is upper bounded by

$$P_{\mathrm{e}} \le \mathbb{P}\bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\Big\{\imath(\mathbf{X}_{s_{\mathrm{dif}}};\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})\le\gamma_{|s_{\mathrm{dif}}|}\Big\}\bigg] + \sum_{\bar{s}\in\mathcal{S}\setminus\{s\}}\mathbb{P}\Big[\imath(\mathbf{X}_{\bar{s}\setminus s};\mathbf{Y}|\mathbf{X}_{\bar{s}\cap s})>\gamma_{|s_{\mathrm{dif}}|}\Big], \tag{118}$$

where here and subsequently we let the condition $s_{\mathrm{dif}}\ne\emptyset$ remain implicit. The first term in (118) corresponds to the true set failing the threshold test, and the second term corresponds to some incorrect set $\bar{s}$ passing the threshold test. In the summand of the second term, we have upper bounded the probability of an intersection of 2^k-1 events by just one such event, namely, the one corresponding to $s_{\mathrm{dif}}=\bar{s}\setminus s$ and $s_{\mathrm{eq}}=s\cap\bar{s}$.

Using the shorthand $\ell:=|\bar{s}\setminus s|$, we can weaken the second probability in (118) as follows:

$$\begin{aligned}
&\mathbb{P}\Big[\imath(\mathbf{X}_{\bar{s}\setminus s};\mathbf{Y}|\mathbf{X}_{\bar{s}\cap s})>\gamma_\ell\Big] \\
&= \sum_{\mathbf{x}_{\bar{s}\cap s},\mathbf{x}_{\bar{s}\setminus s},\mathbf{y}} P_X^{n\times(k-\ell)}(\mathbf{x}_{\bar{s}\cap s})P_X^{n\times\ell}(\mathbf{x}_{\bar{s}\setminus s})P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\cap s}) \\
&\qquad \times \mathbb{1}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\setminus s},\mathbf{x}_{\bar{s}\cap s})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\cap s})}>\gamma_\ell\bigg\} && (119)\\
&\le \sum_{\mathbf{x}_{\bar{s}\cap s},\mathbf{x}_{\bar{s}\setminus s},\mathbf{y}} P_X^{n\times(k-\ell)}(\mathbf{x}_{\bar{s}\cap s})P_X^{n\times\ell}(\mathbf{x}_{\bar{s}\setminus s}) \\
&\qquad \times P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{\bar{s}\setminus s},\mathbf{x}_{\bar{s}\cap s})e^{-\gamma_\ell} && (120)\\
&= e^{-\gamma_\ell}, && (121)
\end{aligned}$$

where in (119) we used the fact that the output vector depends only on the columns of $\mathbf{x}_{\bar{s}}$ corresponding to entries of $\bar{s}$ that are also in s, and (120) follows by bounding $P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}$ using the event within the indicator function, and then upper bounding the indicator function by one. Substituting (121) into (118) gives

$$P_{\mathrm{e}} \le \mathbb{P}\bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\Big\{\imath(\mathbf{X}_{s_{\mathrm{dif}}};\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})\le\gamma_\ell\Big\}\bigg] + \sum_{\ell=1}^{k}\binom{p-k}{\ell}\binom{k}{\ell}e^{-\gamma_\ell}, \tag{122}$$

where the combinatorial terms arise from a standard counting argument [5].

Note that while the bound in (122) appears to be simpler than that in the theorem statement, it is difficult to directly apply it to specific problems, since $\imath(\mathbf{X}_{s_{\mathrm{dif}}};\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})$ is not an i.i.d. summation in general.

2) Completion of the Proof: We fix the constants $\gamma_1',\dots,\gamma_\ell'$ arbitrarily, and apply the following elementary steps with $\ell = |s_{\mathrm{dif}}|$:

$$\begin{aligned}
&\mathbb{P}\Bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\Big\{\imath(\mathbf{X}_{s_{\mathrm{dif}}};\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}) \le \gamma_\ell\Big\}\Bigg] \\
&= \mathbb{P}\Bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}},\mathbf{X}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})} \le \gamma_\ell\bigg\}\Bigg]
\end{aligned} \tag{123}$$

$$\begin{aligned}
&\le \mathbb{P}\Bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}},\mathbf{X}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})} \le \gamma_\ell \\
&\qquad\qquad \cap \log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},\beta_s)} \le \gamma'_\ell\bigg\}\Bigg] \\
&\quad + \mathbb{P}\Bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},\beta_s)} > \gamma'_\ell\bigg\}\Bigg]
\end{aligned} \tag{124}$$

$$\begin{aligned}
&\le \mathbb{P}\Bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}},\mathbf{X}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},\beta_s)} \le \tilde{\gamma}_\ell\bigg\}\Bigg] \\
&\quad + \mathbb{P}\Bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},\beta_s)} > \gamma'_\ell\bigg\}\Bigg],
\end{aligned} \tag{125}$$

where $\tilde{\gamma}_\ell = \gamma_\ell + \gamma'_\ell$. The second term in (125) is upper bounded as

$$\begin{aligned}
&\mathbb{P}\Bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},\beta_s)} > \gamma'_\ell\bigg\}\Bigg] \\
&\le \sum_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\mathbb{P}\Bigg[\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},\beta_s)} > \gamma'_\ell\Bigg] && (126) \\
&= \sum_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\sum_{b_s,\mathbf{x}_{s_{\mathrm{eq}}},\mathbf{y}} P_{\beta_s}(b_s)P_X^{n\times(k-\ell)}(\mathbf{x}_{s_{\mathrm{eq}}}) \\
&\qquad \times P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}},b_s) \\
&\qquad \times \mathbb{1}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}})}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}},b_s)} > \gamma'_\ell\bigg\} && (127) \\
&\le \sum_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\sum_{b_s,\mathbf{x}_{s_{\mathrm{eq}}},\mathbf{y}} P_{\beta_s}(b_s)P_X^{n\times(k-\ell)}(\mathbf{x}_{s_{\mathrm{eq}}}) \\
&\qquad\qquad \times P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}}(\mathbf{y}|\mathbf{x}_{s_{\mathrm{eq}}})e^{-\gamma'_\ell} && (128) \\
&= \sum_{\ell=1}^{k}\binom{k}{\ell}e^{-\gamma'_\ell}, && (129)
\end{aligned}$$

where (126) follows from the union bound, and the remaining steps follow the arguments used in (119)–(121).

We now upper bound the first term in (125). The numerator in (125) equals $P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)$ for all $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$ (*cf.*, (15)), and we can thus write the overall term as

$$\begin{aligned}
&\mathbb{P}\Bigg[\log P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s) \\
&\quad \le \max_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\Big\{\log P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},\beta_s) + \gamma_\ell + \gamma'_\ell\Big\}\Bigg].
\end{aligned} \tag{130}$$

Using the same steps as those used in (123)–(125), we can upper bound this by

$$\begin{aligned}
&\mathbb{P}\Bigg[\log P_{\mathbf{Y}|\mathbf{X}_s\beta_s}(\mathbf{Y}|\mathbf{X}_s,\beta_s) \\
&\quad \le \max_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\Big\{\log P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},\beta_s) + \gamma_\ell + \gamma'_\ell + \gamma\Big\}\Bigg] \\
&\qquad\qquad + \mathbb{P}\Bigg[\log\frac{P_{\mathbf{Y}|\mathbf{X}_s,\beta_s}(\mathbf{Y}|\mathbf{X}_s,\beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} > \gamma\Bigg]
\end{aligned} \tag{131}$$

for any constant γ. Reversing the step in (130), this can equivalently be written as

$$\begin{aligned}
&\mathbb{P}\Bigg[\bigcup_{(s_{\mathrm{dif}},s_{\mathrm{eq}})}\bigg\{\log\frac{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}}\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{dif}}},\mathbf{X}_{s_{\mathrm{eq}}},\beta_s)}{P_{\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}\beta_s}(\mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}},\beta_s)} \\
&\quad \le \gamma_\ell + \gamma'_\ell + \gamma\bigg\}\Bigg] + \mathbb{P}\Bigg[\log\frac{P_{\mathbf{Y}|\mathbf{X}_s,\beta_s}(\mathbf{Y}|\mathbf{X}_s,\beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)} > \gamma\Bigg].
\end{aligned} \tag{132}$$

Observe that the first logarithm appearing here is precisely the information density in (20). Moreover, the choices

$$\gamma_\ell = \log\left(\frac{k}{\delta_1}\binom{p-k}{\ell}\binom{k}{\ell}\right) \tag{133}$$

$$\gamma'_\ell = \log\left(\frac{k}{\delta_1}\binom{k}{\ell}\right) \tag{134}$$

make (129) and the second term in (122) be upper bounded by δ_1 each. Hence, and combining (125) with (129) and (132), and recalling that $\ell = |s_{\mathrm{dif}}|$, we obtain (24).

B. Proof of Theorem 2

As has been done in several previous proofs of information-theoretic converse bounds for sparsity pattern recovery [2], [7], [16], we consider an argument based on a genie. As explained formally below, 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; this leads to the improvement stated following Theorem 4.

It will prove convenient to present the proof under the following assumption of symmetry.

Assumption 1. The pair $(s_{\mathrm{dif}}(b_s), s_{\mathrm{eq}}(b_s))$ in Theorem 2 satisfies the following property: If b'_s is a permutation of b_s, then the entries of b'_s indexed by $s_{\mathrm{dif}}(b'_s)$ (respectively, $s_{\mathrm{eq}}(b'_s)$) are a permutation of the entries of b_s indexed by $s_{\mathrm{dif}}(b_s)$ (respectively, $s_{\mathrm{eq}}(b_s)$).

We claim that the theorem statement under this assumption also implies the more general case. To see this, we use the symmetry of $P_{Y|X_S\beta_S}$ with respect to S from in Section II, and the fact that $\mathbf{X}$ has an i.i.d. distribution. Among all the possible choices of functions $(s_{\text{dif}}(\cdot), s_{\text{eq}}(\cdot))$, there always exists a pair that maximizes the lower bound in (33) and satisfies Assumption 1. More precisely, for any realization of β_s, the probability in (33) is determined by entries appearing in the partition $(\beta_{s_{\text{dif}}}, \beta_{s_{\text{eq}}})$ but not by their order, so one can always maximize (33) by forming this partition in a manner which is symmetric with respect to permutations of β_s.

We now formally define the genie-aided setup as follows:

1. Generate a random k-dimensional vector $\widetilde{\beta}' \sim P_{\beta_S}$.
2. Given $\widetilde{\beta}'$, let $\widetilde{\beta}'_{\text{dif}}$ and $\widetilde{\beta}'_{\text{eq}}$ be the subvectors indexed by $s_{\text{dif}}(\widetilde{\beta}')$ and $s_{\text{eq}}(\widetilde{\beta}')$ respectively.
3. Let $\widetilde{\beta}_{\text{dif}}$ (respectively, $\widetilde{\beta}_{\text{eq}}$) be a uniformly random permutation of $\widetilde{\beta}'_{\text{dif}}$ (respectively, $\widetilde{\beta}'_{\text{eq}}$).
4. Generate S_{eq} uniformly on $\mathcal{S}_{\text{eq}}(\ell)$, defined to contain the $\binom{p}{k-\ell}$ subsets of $\{1,\dots,p\}$ having cardinality $k-\ell$, where $\ell = |\widetilde{\beta}_{\text{dif}}|$. Set $\beta_{S_{\text{eq}}} = \widetilde{\beta}_{\text{eq}}$.
5. Generate S_{dif} uniformly on $\mathcal{S}_{\text{dif}}(S_{\text{eq}})$, defined to contain the $\binom{p-k+\ell}{\ell}$ subsets of $\{1,\dots,p\}\backslash S_{\text{eq}}$ having cardinality ℓ. Set $\beta_{S_{\text{dif}}} = \widetilde{\beta}_{\text{dif}}$.
6. Set $S = S_{\text{dif}} \cup S_{\text{eq}}$ and $\beta_{S^c} = 0$. The measurement matrix $\mathbf{X}$ is i.i.d. on P_X, and the observation vector $\mathbf{Y}$ is generated from S, $\mathbf{X}$, and β according to (5), as in the original problem setup.
7. Reveal the indices S_{eq} and the vectors $\widetilde{\beta}_{\text{dif}}$ and $\widetilde{\beta}_{\text{eq}}$ to the decoder. The decoder forms an estimate $\hat{S}_{\text{dif}}$ of S_{dif}, and an error occurs if $\hat{S}_{\text{dif}} \neq S_{\text{dif}}$.

The joint distribution of S and β is the same here as in the original setup: The support set is uniform on the $\binom{p}{k}$ elements of $\mathcal{S}$, and the distribution of the non-zero entries β_S is that of a uniformly random permutation of $\widetilde{\beta}' \sim P_{\beta_S}$. Since P_{β_S} is permutation-invariant by assumption, this yields $\beta_S \sim P_{\beta_S}$ as required. Thus, the only difference in this modified setup is that the decoder has further information, and it follows that any converse for this setup implies the same converse for the original setup.

Throughout the proof, we make use of the random variables defined in the preceding steps, departing from the notation implicitly conditioned on S equaling a fixed value s (see (11)) until the final step in obtaining (33).

We first study the error probability for the genie-aided setting conditioned on $(S_{\text{eq}}, \widetilde{\beta}_{\text{dif}}, \widetilde{\beta}_{\text{eq}}) = (s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}})$, denoted by $P_{\text{e}}(s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}})$. By the identity $\mathbb{P}[\mathcal{A}] = \mathbb{P}[\mathcal{A} \cap \mathcal{E}] + \mathbb{P}[\mathcal{A} \cap \mathcal{E}^c]$, we have for any event $\mathcal{A}(s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}})$ that

$$\begin{aligned} P_{\text{e}}(s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}}) &\geq \mathbb{P}[\mathcal{A}(s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}})] \\ &\quad - \mathbb{P}[\mathcal{A}(s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}}) \cap \text{no error}]. \quad (135) \end{aligned}$$

We fix the constant γ_ℓ and choose

$$\mathcal{A}(s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}}) = \big\{ \imath^n(\mathbf{X}_{S_{\text{dif}}}; \mathbf{Y}|\mathbf{X}_{s_{\text{eq}}}, \widetilde{b}) \leq \gamma_\ell \big\}, \quad (136)$$

where $\ell = k - |s_{\text{eq}}|$, and $\widetilde{b} := \widetilde{b}(\widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}}, s_{\text{dif}}, s_{\text{eq}})$ equals $\widetilde{b}_{\text{dif}}$ (respectively, $\widetilde{b}_{\text{eq}}$) on the entries indexed by s_{dif} (respectively, s_{eq}). Using the definitions in (20)–(21), and defining $\mathcal{D}(s_{\text{dif}}|s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}})$ to be the set of pairs $(\mathbf{x}, \mathbf{y})$ such that the decoder outputs s_{dif} given $(s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}}, \mathbf{x}, \mathbf{y})$, we obtain

$$\begin{aligned} &\mathbb{P}[\mathcal{A}(s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}}) \cap \text{no error}] \\ &\quad = \sum_{s_{\text{dif}} \in \mathcal{S}_{\text{dif}}(s_{\text{eq}})} \frac{1}{\binom{p-k+\ell}{\ell}} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}(s_{\text{dif}}|s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}})} P_X^{n\times p}(\mathbf{x}) \\ &\qquad \times P^n_{Y|X_{s_{\text{dif}}}X_{s_{\text{eq}}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\text{dif}}}, \mathbf{x}_{s_{\text{eq}}}, \widetilde{b}) \\ &\qquad \times \mathbb{1}\left\{ \log \frac{P^n_{Y|X_{s_{\text{dif}}}X_{s_{\text{eq}}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\text{dif}}}, \mathbf{x}_{s_{\text{eq}}}, \widetilde{b})}{P^n_{Y|X_{s_{\text{eq}}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\text{eq}}}, \widetilde{b})} \leq \gamma_\ell \right\} \quad (137) \\ &\quad \leq \frac{1}{\binom{p-k+\ell}{\ell}} \sum_{s_{\text{dif}} \in \mathcal{S}_{\text{dif}}(s_{\text{eq}})} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}(s_{\text{dif}}|s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}})} P_X^{n\times p}(\mathbf{x}) \\ &\qquad \times P^n_{Y|X_{s_{\text{eq}}}\beta_s}(\mathbf{y}|\mathbf{x}_{s_{\text{eq}}}, \widetilde{b}) e^{\gamma_\ell} \quad (138) \\ &\quad = \frac{e^{\gamma_\ell}}{\binom{p-k+\ell}{\ell}}, \quad (139) \end{aligned}$$

where (137) follows since an error occurs if and only if $(\mathbf{x}, \mathbf{y}) \notin \mathcal{D}(s_{\text{dif}}|s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}})$, (138) follows by upper bounding $P^n_{Y|X_{s_{\text{eq}}}\beta_s}$ using the event in the indicator function, and (139) follows since the sets $\mathcal{D}(s_{\text{dif}}|s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}})$ are disjoint, and their union over s_{dif} is the entire space of $(\mathbf{x}, \mathbf{y})$ pairs.

Averaging (135) over $(S_{\text{eq}}, \widetilde{\beta}', \widetilde{\beta}_{\text{dif}}, \widetilde{\beta}_{\text{eq}})$ and applying (139), we obtain

$$\begin{aligned} P_{\text{e}} &\geq \sum_{\widetilde{b}'} P_{\beta_S}(\widetilde{b}') \sum_{\widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}}} \mathbb{P}\big[(\widetilde{\beta}_{\text{dif}}, \widetilde{\beta}_{\text{eq}}) = (\widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}}) \,|\, \widetilde{b}'\big] \\ &\quad \times \sum_{s_{\text{eq}} \in S_{\text{eq}}(\ell)} \sum_{s_{\text{dif}} \in S_{\text{dif}}(s_{\text{eq}})} \frac{1}{\binom{p}{k-\ell}} \frac{1}{\binom{p-k+\ell}{\ell}} \\ &\quad \times \Bigg(\mathbb{P}\Big[\imath^n(\mathbf{X}_{s_{\text{dif}}}; \mathbf{Y}|\mathbf{X}_{s_{\text{eq}}}, \widetilde{b}) \leq \gamma_\ell \,\Big|\, s_{\text{dif}}, s_{\text{eq}}, \widetilde{b}_{\text{dif}}, \widetilde{b}_{\text{eq}} \Big] \\ &\qquad - \frac{e^{\gamma_\ell}}{\binom{p-k+\ell}{\ell}} \Bigg), \quad (140) \end{aligned}$$

where $\ell = |\widetilde{b}_{\text{dif}}|$, and the conditioning on $\widetilde{b}'$ is a shorthand for $\widetilde{\beta}' = \widetilde{b}'$, and similarly for the second probability. Finally, we claim that this recovers (33) upon setting

$$\gamma_\ell = \log \binom{p-k+\ell}{\ell} + \log \delta_1. \quad (141)$$

To see this, we first note that all of the terms in the summations over s_{dif} and s_{eq} in (140) are equal, since in the probability appearing in the summand, the entries $\widetilde{b}_{s_{\text{dif}}}$ and $\widetilde{b}_{s_{\text{eq}}}$ are the same for any such pair, namely, $\widetilde{b}_{s_{\text{dif}}} = \widetilde{b}_{\text{dif}}$ and $\widetilde{b}_{s_{\text{eq}}} = \widetilde{b}_{\text{eq}}$ (recall also the symmetry of $P_{Y|X_S\beta_S}$ with respect to S assumed in Section II). Due to Assumption 1, this probability also coincides with that in (33) with $b_s := \widetilde{b}'$, regardless of the realization of $(\widetilde{\beta}_{\text{dif}}, \widetilde{\beta}_{\text{eq}})$ given $\widetilde{\beta}'$; the only randomness in the corresponding distribution is that of the two random permutations of the subvectors.

C. Extensions to Partial Recovery

The achievability analysis in Section V-A extends immediately to handle the partial recovery criterion in (7), since we have already split the error events according to the amount of overlap between the true support and the incorrect support. The only difference is that the decoder searches for a set s such that (117) holds whenever $|s_{\mathrm{dif}}| > d_{\max}$ (as opposed to $s_{\mathrm{dif}} \ne \emptyset$), and chooses one arbitrarily if multiple such s exist. It follows that Theorem 1 remains true when the union in (24) is restricted to $|s_{\mathrm{dif}}| \in \{d_{\max}+1, \dots, k\}$.

The extension of the converse analysis in Section V-B is less immediate, but still straightforward. We first recall the observation from [16] that the performance metric in (7) allows us to focus without loss of generality on decoders such that the estimated support $\hat{S}$ (or $\hat{S}_{\mathrm{dif}} \cup S_{\mathrm{eq}}$ in the genie-aided setting) has cardinality k almost surely. For any such decoder, the definition in (7) is unchanged when the second term in the union is removed.

We restrict the partitions $(s_{\mathrm{dif}}(b_s), s_{\mathrm{eq}}(b_s))$ of s to satisfy $|s_{\mathrm{dif}}(b_s)| > d_{\max}$. In (137)–(138), we change the definition of $\mathcal{D}(s_{\mathrm{dif}}|s_{\mathrm{eq}}, \widetilde{b}_{\mathrm{dif}}, \widetilde{b}_{\mathrm{eq}})$ to be the set of pairs $(\mathbf{x}, \mathbf{y})$ such that the decoder outputs a sequence $\hat{s}_{\mathrm{dif}}$ such that $|s_{\mathrm{dif}} \backslash \hat{s}_{\mathrm{dif}}| \le d_{\max}$. This means that the sets $\mathcal{D}(\cdot|s_{\mathrm{eq}}, \widetilde{b}_{\mathrm{dif}}, \widetilde{b}_{\mathrm{eq}})$ are no longer disjoint. However, we can easily count the number of such sets that each $(\mathbf{x}, \mathbf{y})$ pair falls into. For fixed $(s_{\mathrm{eq}}, s_{\mathrm{dif}})$ and $d \in \{0, \dots, d_{\max}\}$, the number of sets $\hat{s}_{\mathrm{dif}} \subseteq \{1, \dots, p\} \backslash s_{\mathrm{eq}}$ such that $|s_{\mathrm{dif}} \backslash \hat{s}_{\mathrm{dif}}| = d$ is $\binom{p-k}{d}\binom{|s_{\mathrm{dif}}|}{|s_{\mathrm{dif}}|-d} = \binom{p-k}{d}\binom{|s_{\mathrm{dif}}|}{d}$. Thus, each $(\mathbf{x}, \mathbf{y})$ pair is included in $\sum_{d=0}^{d_{\max}} \binom{p-k}{d}\binom{|s_{\mathrm{dif}}|}{d}$ of the sets $\mathcal{D}(\cdot|s_{\mathrm{eq}}, \widetilde{b}_{\mathrm{dif}}, \widetilde{b}_{\mathrm{eq}})$, and (139) is replaced by

$$\mathbb{P}[\mathcal{A}(s_{\mathrm{eq}}, \widetilde{b}_{\mathrm{dif}}, \widetilde{b}_{\mathrm{eq}}) \cap \text{no error}] \le \frac{\sum_{d=0}^{d_{\max}} \binom{p-k}{d}\binom{\ell}{d}}{\binom{p-k+\ell}{\ell}} e^{-\gamma_\ell}. \tag{142}$$

Thus, Theorem 2 remains true when the pair $(s_{\mathrm{dif}}(\cdot), s_{\mathrm{eq}}(\cdot))$ is constrained to satisfy $|s_{\mathrm{dif}}| \in \{d_{\max}+1, \dots, k\}$, and $\binom{p-k+|s_{\mathrm{dif}}|}{|s_{\mathrm{dif}}|}$ is replaced by $\binom{p-k+|s_{\mathrm{dif}}|}{|s_{\mathrm{dif}}|} - \sum_{d=0}^{d_{\max}} \binom{p-k}{d}\binom{|s_{\mathrm{dif}}|}{d}$.

VI. Conclusion

Taking an approach motivated by thresholding techniques in channel coding, we have presented a framework for developing necessary and sufficient conditions on the number of measurements for exact and partial support recovery with probabilistic models. We have provided several new results for the linear, 1-bit, and group testing models, as well as general discrete observation models. In several cases, we have provided exact asymptotic thresholds on the number of measurements with strong converse results.

There are several possible directions for future research. While we have focused on i.i.d. measurement matrices, it would be of significant interest to consider other types of random matrices, and to present converse results that hold for arbitrary measurement matrices, subject to suitable constraints such as power constraints. We provided some work in these directions for specific models in [44], [45].

One could also attempt to move from standard sparsity models to structured sparsity models [46], and from probabilistic guarantees with random β to minimax guarantees. There are several additional non-linear models that our general results could be applied to, such as the Poisson and gamma models. Finally, it may be interesting to apply similar analysis techniques to other statistical problems beyond support recovery.

Appendix A
Concentration Inequalities

In order to apply our general bounds to specific models, we use concentration inequalities to obtain expressions for ψ_ℓ and ψ'_ℓ in (35) and (41), seeking to make the corresponding terms in (39) and (45) vanish. Here we present two general inequalities that will be used throughout Section IV.

Proposition 9. *For general observation models, we have for all $(s_{\mathrm{dif}}, s_{\mathrm{eq}}, b_s)$ and $\delta > 0$ that*

$$\mathbb{P}\Big[\big|\imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}, b_s) - nI_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)\big| \ge n\delta \,\Big|\, \beta_s = b_s\Big] \le \frac{V_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)}{\delta^2 n}, \tag{143}$$

where $V_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)$ is defined in (54)*. Moreover, if the observations lie in a finite set $\mathcal{Y} \subset \mathbb{R}$ with probability one, then the following holds for all $(s_{\mathrm{dif}}, s_{\mathrm{eq}}, b_s)$ and $\delta > 0$:*

$$V_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s) \le |\mathcal{Y}|\Big(\frac{4}{e}\Big)^2. \tag{144}$$

Before providing the proof, we state the following generalization of (144) to higher-order moments.

Proposition 10. *If the observations lie in a finite set $\mathcal{Y} \subset \mathbb{R}$ with probability one, then the following holds for all $(s_{\mathrm{dif}}, s_{\mathrm{eq}}, b_s)$ and $\delta > 0$:*

$$\mathbb{P}\Big[\big|\imath^n(\mathbf{X}_{s_{\mathrm{dif}}}; \mathbf{Y}|\mathbf{X}_{s_{\mathrm{eq}}}, b_s) - nI_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s)\big| \ge n\delta \,\Big|\, \beta_s = b_s\Big] \le 2\exp\bigg(-\frac{\delta^2 n}{2(8|\mathcal{Y}| + 2\delta)}\bigg). \tag{145}$$

In the remainder of this appendix, we prove these propositions. Equation (143) follows from Chebyshev's inequality, so we focus our attention on (144)–(145). We make use of the following form of Bernstein's inequality [38, Sec. 2.8].

Lemma 1. *Let $W_1, \dots, W_n$ be independent real-valued random variables such that*

$$\sum_{i=1}^n \mathbb{E}[W_i^2] \le \tau \tag{146}$$

$$\sum_{i=1}^n \mathbb{E}[|W_i|^q] \le \frac{q!}{2}\tau c^{q-2} \qquad (q \ge 3) \tag{147}$$

for some $\tau, c > 0$. Then

$$\mathbb{P}\bigg[\sum_{i=1}^{n}\big(W_i - \mathbb{E}[W_i]\big) \geq t\bigg] \leq \exp\bigg(\frac{t^2}{2(\tau + ct)}\bigg) \tag{148}$$

for all $t > 0$.

To bound the moments of $\imath$, we follow the arguments of [29, Rmk. 3.1.1] and [47, App. D]. Recall the definition of the information density in (21). For any $q \geq 2$, we have from Minkowski's inequality that

$$\begin{aligned}
&\mathbb{E}\big[|\imath(X_{s_{\rm dif}}; Y|X_{s_{\rm eq}}, b_s)|^q\big]^{1/q} \\
&\quad\leq \mathbb{E}\bigg[\bigg(\log\frac{1}{P_{Y|X_{s_{\rm dif}}X_{s_{\rm eq}}\beta_s}(Y|X_{s_{\rm dif}}, X_{s_{\rm eq}}, b_s)}\bigg)^q\bigg]^{1/q} \\
&\qquad\qquad + \mathbb{E}\bigg[\bigg(\log\frac{1}{P_{Y|X_{s_{\rm eq}}\beta_s}(Y|X_{s_{\rm eq}}, b_s)}\bigg)^q\bigg]^{1/q}, \quad (149)
\end{aligned}$$

where here and subsequently we implicitly condition on $\beta_s = b_s$. For any given $(x_{s_{\rm dif}}, x_{s_{\rm eq}})$, the remaining averaging over Y in the first term has the form

$$\begin{aligned}
&\sum_y P_{Y|X_{s_{\rm dif}}X_{s_{\rm eq}}\beta_s}(y|x_{s_{\rm dif}}, x_{s_{\rm eq}}, b_s) \\
&\qquad\times\bigg(\log\frac{1}{P_{Y|X_{s_{\rm dif}}X_{s_{\rm eq}}\beta_s}(y|x_{s_{\rm dif}}, x_{s_{\rm eq}}, b_s)}\bigg)^q, \quad (150)
\end{aligned}$$

and is thus upper bounded by $|\mathcal{Y}|\big(\frac{q}{e}\big)^{1/q}$, since the function $f(z) = z\log^q\frac{1}{z}$ has a maximum value of $\big(\frac{q}{e}\big)^{1/q}$ for $z \in [0, 1]$. Handling the second term in (149) similarly, we obtain

$$\mathbb{E}\big[|\imath(X_{s_{\rm dif}}; Y|X_{s_{\rm eq}}, b_s)|^q\big]^{1/q} \leq 2\bigg(|\mathcal{Y}|\bigg(\frac{q}{e}\bigg)^q\bigg)^{1/q}, \tag{151}$$

or equivalently

$$\begin{aligned}
\mathbb{E}\big[|\imath(X_{s_{\rm dif}}; Y|X_{s_{\rm eq}}, b_s)|^q\big] &\leq \bigg(\frac{q}{e}\bigg)^q 4|\mathcal{Y}|2^{q-2} &(152)\\
&\leq \frac{q!}{2}8|\mathcal{Y}|2^{q-2}, &(153)
\end{aligned}$$

where (153) follows since $\big(\frac{q}{e}\big)^q \leq q!$.

We obtain (144) by setting $q = 2$ in (152). Furthermore, we obtain Proposition 10 using Lemma 1 with $c = 2$, $\tau = n \cdot 8|\mathcal{Y}|$, and $t = \delta n$.

Appendix B
Proofs of Auxiliary Results for the Linear Model

A. *Proof of Proposition 1*

We again use Lemma 1, and we thus seek suitable values for τ and c. Throughout the proof, we consider the random variables $(X_{s_{\rm dif}}, X_{s_{\rm eq}}, Y)$ distributed according to (8), implicitly conditioning on $\beta_s = b_s$. From (55), we have $Z = Y - \sum_{i\in s} X_i b_i$, and a direct calculation gives

$$P_{Y|X_{s_{\rm dif}}X_{s_{\rm eq}}\beta_s}(Y|X_{s_{\rm dif}}, X_{s_{\rm eq}}, b_s) = \phi(Z; 0, \sigma^2) \tag{154}$$

$$\begin{aligned}
&P_{Y|X_{s_{\rm eq}}\beta_s}(Y|X_{s_{\rm eq}}, b_s) \\
&\qquad = \phi\Big(\sum_{i\in s_{\rm dif}} X_i b_i + Z; 0, \sigma^2 + \sum_{i\in s_{\rm dif}} b_i^2\Big), \quad (155)
\end{aligned}$$

where $\phi(\cdot; \mu, \sigma^2)$ is the $N(\mu, \sigma^2)$ density function. Substituting these into (21) gives

$$\begin{aligned}
&\imath(X_{s_{\rm dif}}; Y|X_{s_{\rm eq}}, b_s) = I_{s_{\rm dif},s_{\rm eq}}(b_s) - \frac{Z^2}{2\sigma^2} \\
&\qquad + \frac{1}{2\big(\sigma^2 + \sum_{i\in s_{\rm dif}} b_i^2\big)}\bigg(\sum_{i\in s_{\rm dif}} X_i b_i + Z\bigg)^2, \quad (156)
\end{aligned}$$

where $I_{s_{\rm dif},s_{\rm eq}}(b_s)$ is given in (56).

The mean of (156) is $I_{s_{\rm dif},s_{\rm eq}}(b_s)$, and we will apply Lemma 1 with W_i corresponding to the sum of the second and third terms on the right-hand side. We can write these in terms of independent $N(0, 1)$ random variables (denoted by $\hat{Z}_1$ and $\hat{Z}_2$) as follows:

$$\begin{aligned}
W &= -\frac{\hat{Z}_1^2}{2} + \frac{1}{2(\sigma^2 + \sigma_{s_{\rm dif}}^2)}\big(\sigma\hat{Z}_1 + \sigma_{s_{\rm dif}}\hat{Z}_2\big)^2 &(157)\\
&= \frac{\sigma_{s_{\rm dif}}^2}{2(\sigma^2 + \sigma_{s_{\rm dif}}^2)}\big(\hat{Z}_2^2 - \hat{Z}_1^2\big) + \frac{\sigma\sigma_{s_{\rm dif}}}{\sigma^2 + \sigma_{s_{\rm dif}}^2}\hat{Z}_1\hat{Z}_2, &(158)
\end{aligned}$$

where we have used the definitions in the proposition statement, and (158) follows from simple manipulations. Defining $\hat{Z}_{\max} = \max\{|\hat{Z}_1|, |\hat{Z}_2|\}$, we have the following with probability one:

$$\begin{aligned}
|W| &\leq \frac{\sigma_{s_{\rm dif}}^2}{2(\sigma^2 + \sigma_{s_{\rm dif}}^2)}2\hat{Z}_{\max}^2 + \frac{\sigma\sigma_{s_{\rm dif}}}{\sigma^2 + \sigma_{s_{\rm dif}}^2}\hat{Z}_{\max}^2 &(159)\\
&= \frac{\sigma_{s_{\rm dif}}(\sigma + \sigma_{s_{\rm dif}})}{\sigma^2 + \sigma_{s_{\rm dif}}^2}\hat{Z}_{\max}^2. &(160)
\end{aligned}$$

Since $\mathbb{E}[\hat{Z}_{\max}^4] \leq \mathbb{E}[\hat{Z}_1^4 + \hat{Z}_2^4] = 6$, we obtain

$$\mathbb{E}[W^2] \leq 6\bigg(\frac{\sigma_{s_{\rm dif}}(\sigma + \sigma_{s_{\rm dif}})}{\sigma^2 + \sigma_{s_{\rm dif}}^2}\bigg)^2. \tag{161}$$

Similarly, we can bound the higher moments as follows:

$$\begin{aligned}
\mathbb{E}[|W|^q] &\leq \bigg(\frac{\sigma_{s_{\rm dif}}(\sigma + \sigma_{s_{\rm dif}})}{\sigma^2 + \sigma_{s_{\rm dif}}^2}\bigg)^q\mathbb{E}[\hat{Z}_1^{2q} + \hat{Z}_2^{2q}] &(162)\\
&\leq \bigg(\frac{2\sigma_{s_{\rm dif}}(\sigma + \sigma_{s_{\rm dif}})}{\sigma^2 + \sigma_{s_{\rm dif}}^2}\bigg)^q\frac{2}{\sqrt{\pi}}\Gamma\Big(q + \frac{1}{2}\Big) &(163)\\
&\leq 2\cdot\bigg(\frac{2\sigma_{s_{\rm dif}}(\sigma + \sigma_{s_{\rm dif}})}{\sigma^2 + \sigma_{s_{\rm dif}}^2}\bigg)^q\cdot q!, &(164)
\end{aligned}$$

where (163) follows by the same argument as (160) and the fact that the $2q$-th moment of an $N(0, 1)$ random variable is $\frac{2^q}{\sqrt{\pi}}\Gamma\big(q + \frac{1}{2}\big)$, and (164) follows since $\Gamma\big(q + \frac{1}{2}\big) \leq \sqrt{\pi}q!$.

Combining (161) and (164), we see that the random variables $W_i = \imath(X_{s_{\rm dif}}^{(i)}; Y^{(i)}|X_{s_{\rm eq}}^{(i)}, b_s) - I_{s_{\rm dif},s_{\rm eq}}(b_s)$ satisfy the conditions of Lemma 1 with $\tau = n\cdot 4\alpha_{s_{\rm dif}}^2$ and $c = \alpha_{s_{\rm dif}}$ (see (58)). We thus obtain the desired result from (148) by identifying $t = \delta n$.

B. Proof of Proposition 2

Since $\mathbf{Y} = \mathbf{X}_s\beta_s + \mathbf{Z}$, we have

$$I_0 = I(\beta_s; \mathbf{Y}|\mathbf{X}_s) = H(\mathbf{Y}|\mathbf{X}_s) - H(\mathbf{Y}|\mathbf{X}_s, \beta_s) \tag{165}$$

$$= H(\mathbf{X}_s\beta_s + \mathbf{Z}|\mathbf{X}_s) - H(\mathbf{Z}). \tag{166}$$

From [39, Ch. 9], we have $H(\mathbf{Z}) = \frac{n}{2}\log(2\pi e\sigma^2)$ and $H(\mathbf{X}_s\beta_s + \mathbf{Z}|\mathbf{X}_s = \mathbf{x}_s) = \frac{1}{2}\log\big((2\pi e)^n \det(\sigma^2\mathbf{I}_n + \sigma_\beta^2\mathbf{x}_s\mathbf{x}_s^T)\big)$, where $\mathbf{I}_n$ is the $n\times n$ identity matrix. Averaging the latter over $\mathbf{X}_s$ and substituting these into (166) gives

$$I_0 = \frac{1}{2}\mathbb{E}\Big[\log\det\Big(\mathbf{I}_n + \frac{\sigma_\beta^2}{\sigma^2}\mathbf{X}_s\mathbf{X}_s^T\Big)\Big] \tag{167}$$

$$= \frac{1}{2}\mathbb{E}\Big[\log\det\Big(\mathbf{I}_k + \frac{\sigma_\beta^2}{\sigma^2}\mathbf{X}_s^T\mathbf{X}_s\Big)\Big] \tag{168}$$

$$= \frac{1}{2}\sum_{i=1}^k \mathbb{E}\Big[\log\Big(1 + \frac{\sigma_\beta^2}{\sigma^2}\lambda_i(\mathbf{X}_s^T\mathbf{X}_s)\Big)\Big] \tag{169}$$

$$\leq \frac{k}{2}\log\Big(1 + \frac{n\sigma_\beta^2}{\sigma^2}\Big), \tag{170}$$

where (168) follows from the identity $\det(\mathbf{I} + \mathbf{AB}) = \det(\mathbf{I} + \mathbf{BA})$, (169) follows by writing the determinant as a product of eigenvalues (denoted by $\lambda_i(\cdot)$), and (170) follows from Jensen's inequality and the following calculation:

$$\frac{1}{k}\mathbb{E}\Big[\sum_{i=1}^k \lambda_i(\mathbf{X}_s^T\mathbf{X}_s)\Big] = \frac{1}{k}\mathbb{E}[\mathrm{Tr}(\mathbf{X}_s^T\mathbf{X}_s)] = \mathbb{E}[\mathbf{X}_1^T\mathbf{X}_1] = n. \tag{171}$$

This concludes the proof of (66).

We now turn to the bounding of the variance. Again using the fact that $\mathbf{Y} = \mathbf{X}_s\beta_s + \mathbf{Z}$, we have

$$\log\frac{P_{\mathbf{Y}|\mathbf{X}_s,\beta_s}(\mathbf{Y}|\mathbf{X}_s,\beta_s)}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{Y}|\mathbf{X}_s)}$$

$$= \log\frac{P_{\mathbf{Z}}(\mathbf{Z})}{P_{\mathbf{Y}|\mathbf{X}_s}(\mathbf{X}_s\beta_s + \mathbf{Z}|\mathbf{X}_s)} \tag{172}$$

$$= I_0 - \frac{1}{2\sigma^2}\mathbf{Z}^T\mathbf{Z} + \frac{1}{2}(\mathbf{X}_s\beta_s + \mathbf{Z})^T\big(\sigma^2\mathbf{I} + \sigma_\beta^2\mathbf{X}_s\mathbf{X}_s^T\big)^{-1}(\mathbf{X}_s\beta_s + \mathbf{Z}), \tag{173}$$

where $P_{\mathbf{Z}}$ is the density of $\mathbf{Z}$, and (173) follows by a direct substitution of the densities $P_{\mathbf{Z}} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ and $P_{\mathbf{Y}|\mathbf{X}_S}(\cdot|\mathbf{x}_s) \sim N(\mathbf{0}, \sigma^2\mathbf{I} + \sigma_\beta^2\mathbf{x}_s\mathbf{x}_s^T)$, where $\mathbf{0}$ is the zero vector. Observe now that $\frac{1}{\sigma^2}\mathbf{Z}^T\mathbf{Z}$ is a sum of n independent χ^2 random variables with one degree of freedom (each having a variance of 2), and hence, the second term in (173) has a variance of $\frac{n}{2}$. Moreover, by writing $\mathbf{M}^{-1} = (\mathbf{M}^{-\frac{1}{2}})^T\mathbf{M}^{-\frac{1}{2}}$ for the symmetric positive definite matrix $\mathbf{M} = \sigma^2\mathbf{I} + \sigma_\beta^2\mathbf{X}_s\mathbf{X}_s^T$, where $(\cdot)^{-\frac{1}{2}}$ denotes the positive definite matrix square root of the inverse, we find that the final term in (173) is distributed as a sum of χ^2 variables when conditioned on any value of $\mathbf{X}_s$, and hence, the same is true unconditionally. We therefore again obtain a variance of $\frac{n}{2}$, and (67) follows using the identity $\mathrm{Var}[A + B] \leq \mathrm{Var}[A] + \mathrm{Var}[B] + 2\max\{\mathrm{Var}[A], \mathrm{Var}[B]\}$.

Appendix C
Proofs of Auxiliary Results for the 1-bit Model

We first write down the relevant probability distributions and information densities conditioned on a fixed value b_s of β_s. Under the model $Y = \mathrm{sign}\big(\sum_{i\in s} X_i b_i + Z\big)$ with $X_i \sim N(0,1)$ and $Z \sim N(0,\sigma^2)$, we have

$$P_{Y|X_s\beta_s}(1|x_s, b_s) = \mathbb{P}\Big[Z \geq -\sum_{i\in s} x_i b_i\Big] \tag{174}$$

$$= Q\Big(-\frac{1}{\sigma}\sum_{i\in s} x_i b_i\Big). \tag{175}$$

Similarly, for any partition of s into $(s_{\mathrm{dif}}, s_{\mathrm{eq}})$, we can write $Y = \mathrm{sign}\big(\sum_{i\in s_{\mathrm{eq}}} X_i b_i + \sum_{i\in s_{\mathrm{dif}}} X_i b_i + Z\big)$ and use the same steps to conclude that

$$P_{Y|X_{s_{\mathrm{eq}}}\beta_s}(1|x_{s_{\mathrm{eq}}}, b_s) = Q\bigg(\frac{-\sum_{i\in s_{\mathrm{eq}}} x_i b_i}{\sqrt{\sigma^2 + \sum_{i\in s_{\mathrm{dif}}} b_i^2}}\bigg). \tag{176}$$

The corresponding probabilities for $y = 0$ are one minus these expressions, which amounts to multiplying the argument to the Q-function by -1. Substitution into (21) gives

$$\imath(x_{s_{\mathrm{dif}}}; y|x_{s_{\mathrm{eq}}}, b_s) = \log\frac{Q\Big(-y\frac{1}{\sigma}\sum_{i\in s} x_i b_i\Big)}{Q\Big(\frac{-y\sum_{i\in s_{\mathrm{eq}}} x_i b_i}{\sqrt{\sigma^2 + \sum_{i\in s_{\mathrm{dif}}} b_i^2}}\Big)} \tag{177}$$

for $y \in \{-1, 1\}$.

Throughout this appendix, we will use the fact that the first two derivatives of the function

$$f(x) := H_2(Q(x)) \tag{178}$$

are given by

$$f'(x) = \log\frac{1 - Q(x)}{Q(x)}\frac{-1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{179}$$

$$f''(x) = -\frac{1}{2\pi}e^{-x^2}\frac{1}{Q(x)(1 - Q(x))} + \log\frac{1 - Q(x)}{Q(x)}\frac{x}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}. \tag{180}$$

A. Proof of Proposition 4 Part (i)

Recalling that the coefficients X_i ($i \in s$) are i.i.d. on $N(0,1)$, we directly obtain from (175) that

$$H(Y|X_s, \beta_s = b_s) = \mathbb{E}\Big[H_2\Big(Q\Big(\frac{1}{\sigma}\sum_{i\in s} X_i b_i\Big)\Big)\Big] \tag{181}$$

$$= \mathbb{E}\Big[H_2\Big(Q\Big(W\sqrt{\frac{1}{\sigma^2}\sum_{i\in s} b_i^2}\Big)\Big)\Big], \tag{182}$$

where $W \sim N(0,1)$. By evaluating $H(Y|X_{s_{\mathrm{eq}}}, \beta_s = b_s)$ similarly using (176) and taking the difference between the two, we obtain (75).

B. Proof of Proposition 4 Part (ii)

We obtain from (179)–(180) that $f'(0) = 0$ and $f''(0) = -\frac{2}{\pi}$. By performing further differentiations, one can also verify that $f^{(3)}(0) = 0$, and that $|f^{(4)}(x)|$ is uniformly upper bounded by $f^{(4)}(0) = \frac{8(\pi-1)}{\pi^2}$. We thus obtain via a fourth-order Taylor expansion that

$$\log 2 - \frac{1}{\pi}x^2 - \frac{4(\pi-1)}{3\pi^2}x^4 \le H_2(Q(x)) \le \log 2 - \frac{1}{\pi}x^2 + \frac{4(\pi-1)}{3\pi^2}x^4 \quad (183)$$

for all $x \in \mathbb{R}$. Substituting (183) into (75) and noting that the fourth moments of the arguments to $H_2(Q(\cdot))$ therein decay to zero strictly faster than the second moments (by the assumptions on k, $b_{\min}$ and $b_{\max}$), we obtain

$$I_{s_{\mathrm{dif}},s_{\mathrm{eq}}}(b_s) = \frac{1}{\pi}\bigg(\frac{1}{\sigma^2}\sum_{i\in s} b_i^2 - \frac{\sum_{i\in s_{\mathrm{eq}}} b_i^2}{\sigma^2 + \sum_{i\in s_{\mathrm{dif}}} b_i^2}\bigg)(1+o(1)). \quad (184)$$

Again using the assumptions on k, $b_{\min}$ and $b_{\max}$, we observe that the denominator is dominated by the term σ^2, thus yielding (76).

C. Proof of Proposition 4 Part (iii)

In this part, we have assumed that the values $\{b_i\}$ take a common value b_0. Since $\sigma^2 = \Theta(1)$, we may set $\sigma^2 = 1$ without loss of generality; the implied constant can be factored into b_0. In this case, (75) with $\ell = 1$ simplifies to

$$I_1 = \mathbb{E}\bigg[H_2\bigg(Q\bigg(W\sqrt{\frac{(k-1)b_0^2}{1+b_0^2}}\bigg)\bigg) - H_2\bigg(Q\Big(W\sqrt{kb_0^2}\Big)\bigg)\bigg]. \quad (185)$$

By the assumptions $k = \Theta(p)$ and $b_0^2 = \Theta\big(\frac{\log p}{p}\big)$, it is easily verified by a Taylor expansion of the function $f(z) = \frac{1}{\sqrt{1+z}}$ as $z \to 0$ that $\sqrt{\frac{(k-1)b_0^2}{1+b_0^2}} = \sqrt{kb_0^2}\big(1 - \frac{b_0^2}{2} + o(b_0^2)\big)$. For convenience, we write this identity as

$$\sqrt{\frac{(k-1)b_0^2}{1+b_0^2}} = \sqrt{kb_0^2}\big(1-\zeta b_0^2\big), \quad (186)$$

where ζ is a constant depending on p such that $\zeta \to \frac{1}{2}$. Substituting (186) into (185), we obtain

$$I_1 = \mathbb{E}\bigg[H_2\bigg(Q\Big(W\sqrt{kb_0^2}\big(1-\zeta b_0^2\big)\Big)\bigg) - H_2\bigg(Q\Big(W\sqrt{kb_0^2}\Big)\bigg)\bigg]. \quad (187)$$

The next step is to Taylor expand the function $f(x) = H_2(Q(x))$. For any x and $\delta > 0$, we have

$$f(x-\delta) = f(x) + \frac{\delta}{\sqrt{2\pi}}\log\frac{1-Q(x)}{Q(x)}e^{-\frac{x^2}{2}} + \frac{\delta^2}{2}f''(x-\delta_0) \quad (188)$$

for some $\delta_0 \in [0,\delta]$, where the middle term follows from (179). Next, we claim that f'' in (180) is bounded as follows:

$$|f''(x)| \le \frac{2}{\sqrt{2\pi}}(1+|x|)e^{\frac{-x^2}{2}} + \frac{|x|^3}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}. \quad (189)$$

In the case that $x \ge 0$, this is seen by applying $Q(x) \ge \frac{1}{\sqrt{2\pi}(1+x)}e^{-\frac{x^2}{2}}$ and $1 - Q(x) \ge \frac{1}{2}$ to obtain the first term, and applying $Q(x) \le e^{-x^2}$ (and hence $\log\frac{1-Q(x)}{Q(x)} = \log\big(\frac{1}{Q(x)} - 1\big) \le x^2$) to obtain the second term (e.g., see [48] for bounds on the Q-function). The case $x < 0$ follows since (180) is symmetric about zero.

Substituting (188) into (187) with the identifications $x = W\sqrt{kb_0^2}$ and $\delta = W\sqrt{kb_0^2}\zeta b_0^2$, we can write

$$T_1 - T_2 - T_3 \le I_1 \le T_1 + T_2 + T_3, \quad (190)$$

where

$$T_1 := \zeta b_0^2\mathbb{E}\bigg[\frac{W\sqrt{kb_0^2}}{\sqrt{2\pi}}\log\frac{1-Q(W\sqrt{kb_0})}{Q(W\sqrt{kb_0})}e^{-\frac{W^2kb_0^2}{2}}\bigg] \quad (191)$$

$$T_2 := (\zeta b_0^2)^2\mathbb{E}\bigg[\frac{W^2kb_0^2}{\sqrt{2\pi}}\big(1+|W|\sqrt{kb_0^2}\big)e^{-\frac{W^2kb_0^2}{2}\left(1-\zeta b_0^2\right)^2}\bigg] \quad (192)$$

$$T_3 := (\zeta b_0^2)^2\mathbb{E}\bigg[\frac{W^2kb_0^2}{2\sqrt{2\pi}}|W|^3(kb_0^2)^{3/2}e^{-\frac{W^2kb_0^2}{2}\left(1-\zeta b_0^2\right)^2}\bigg], \quad (193)$$

and where for T_2 and T_3 we used the fact that $\delta_0 \in [0,\delta]$ in (188) to upper bound the corresponding terms by the value at $\delta_0 = 0$ or $\delta_0 = \delta$.

We will complete the proof by showing that T_1 behaves as (77) (with $\sigma^2 = 1$), and that T_2 and T_3 behave as $o\big(\frac{\sqrt{\log p}}{p}\big)$. Letting $\phi(\cdot)$ denote the standard normal PDF, we have

$$T_1 = \frac{\zeta b_0^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi(w)w\sqrt{kb_0^2}\log\frac{1-Q(w\sqrt{kb_0^2})}{Q(w\sqrt{kb_0^2}} \times e^{-\frac{w^2kb_0^2}{2}}dw \quad (194)$$

$$= \frac{\zeta b_0^2}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\phi\Big(\frac{t}{\sqrt{kb_0^2}}\Big)t\log\frac{1-Q(t)}{Q(t)}e^{-\frac{t^2}{2}}\frac{1}{\sqrt{kb_0^2}}dt \quad (195)$$

$$= \frac{\zeta b_0^2}{\sqrt{2\pi kb_0^2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}(1+\frac{1}{kb_0^2})}t\log\frac{1-Q(t)}{Q(t)}dt \quad (196)$$

$$= \frac{\zeta b_0^2}{\sqrt{2\pi kb_0^2}}\frac{1}{\sqrt{1+\frac{1}{kb_0^2}}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(1+\frac{1}{kb_0^2})^{-1}}} \times e^{-\frac{t^2}{2}(1+\frac{1}{kb_0^2})}t\log\frac{1-Q(t)}{Q(t)}dt \quad (197)$$

$$= \frac{1}{2}\frac{b_0^2}{\sqrt{2\pi kb_0^2}}\mathbb{E}\Big[W\log\frac{1-Q(W)}{Q(W)}\Big](1+o(1)), \quad (198)$$

where (195) follows by a change of variable of the form $t = w\sqrt{kb_0^2}$, (196) follows from the definition of ϕ, and (198) follows since $\zeta \to \frac{1}{2}$, and since the integral in (197) is the average of $t\log\frac{1-Q(t)}{Q(t)}\mathbb{1}\{t \ge 0\}$ over an $N(0,(1+$

$\frac{1}{kb_0^2})^{-1})$ random variable; since $kb_0^2 \to \infty$, this converges to the corresponding average over $W \sim N(0,1)$, which is easily verified to be finite.

The terms T_2 and T_3 are handled similarly to T_1, so we only briefly comment on the analysis of T_3. By the same arguments as those leading to (196), we obtain

$$T_3 = \frac{(\zeta b_0^2)^2}{2\sqrt{2\pi k b_0^2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}\left((1-\zeta b_0^2)^2 + \frac{1}{kb_0^2}\right)} |t|^5 dt. \tag{199}$$

The integral is once again $\Theta(1)$, and thus $T_3 = \Theta\Big(\frac{b_0^4}{\sqrt{kb_0^2}}\Big)$, which decays to zero strictly faster than (198).

D. Proof of Proposition 4 Part (iv)

We again assume without loss of generality that $\sigma^2 = 1$. Defining $W_{\rm eq} := \sum_{i \in s_{\rm eq}} X_i b_i$ and $W_{\rm dif} := \sum_{i \in s_{\rm dif}} X_i b_i$, it follows from (177) that

$$\imath(X_{s_{\rm dif}}; Y | X_{s_{\rm eq}}, b_s) = \log \frac{Q\big(-Y(W_{\rm dif} + W_{\rm eq})\big)}{Q\big(-Y\tau W_{\rm eq}\big)}, \tag{200}$$

where $\tau := \frac{1}{1 + \sum_{i \in s_{\rm dif}} b_i^2}$, and we implicitly condition on $\beta_s = b_s$. Using (175) and the fact that the variance is upper bounded by the second moment, we have

$$\begin{aligned} &V_{s_{\rm dif}, s_{\rm eq}}(b_s) \\ &\le \mathbb{E}\bigg[Q\big(-(W_{\rm dif} + W_{\rm eq})\big)\bigg(\log \frac{Q\big(-(W_{\rm dif} + W_{\rm eq})\big)}{Q\big(-\tau W_{\rm eq}\big)}\bigg)^2 \\ &\qquad + Q\big(W_{\rm dif} + W_{\rm eq}\big)\bigg(\log \frac{Q\big(W_{\rm dif} + W_{\rm eq}\big)}{Q\big(\tau W_{\rm eq}\big)}\bigg)^2\bigg] \qquad (201) \\ &= 2\mathbb{E}\bigg[Q\big(W_{\rm dif} + W_{\rm eq}\big)\bigg(\log \frac{Q\big(W_{\rm dif} + W_{\rm eq}\big)}{Q\big(\tau W_{\rm eq}\big)}\bigg)^2\bigg], \qquad (202) \end{aligned}$$

where (202) follows since the distributions of $W_{\rm dif}$ and $W_{\rm eq}$ are symmetric about zero, and the two are independent.

The function $g(x) := -\log Q(x)$ is convex, and hence it lies above any given tangent vector. This implies that

$$\begin{aligned} |g(x_1) - g(x_2)| &\le \max\big\{|g'(x_1)|, |g'(x_2)|\big\}|x_1 - x_2| \qquad (203) \\ &\le \big(|g'(x_1)| + |g'(x_2)|\big)|x_1 - x_2|, \qquad (204) \end{aligned}$$

where $g'(x) = -\frac{\phi(x)}{Q(x)}$ is the derivative of g. Writing the logarithm of the ratio in (202) as a difference of logarithms and applying (204), we obtain

$$V_{s_{\rm dif}, s_{\rm eq}}(b_s) \le 2(T_1 + T_2), \tag{205}$$

where, overloading the notation from part (iii), we define

$$\begin{aligned} T_1 := &\iint f_{\rm dif}(w_{\rm dif}) f_{\rm eq}(w_{\rm eq}) Q(w_{\rm dif} + w_{\rm eq}) \\ &\times \bigg(\frac{\phi(w_{\rm dif} + w_{\rm eq})}{Q(w_{\rm dif} + w_{\rm eq})}\bigg)^2 \big(|w_{\rm dif}| + (1-\tau)|w_{\rm eq}|\big)^2 dw_{\rm dif} dw_{\rm eq} \end{aligned} \tag{206}$$

$$\begin{aligned} T_2 := &\iint f_{\rm dif}(w_{\rm dif}) f_{\rm eq}(w_{\rm eq}) Q(w_{\rm dif} + w_{\rm eq}) \\ &\times \bigg(\frac{\phi(\tau w_{\rm eq})}{Q(\tau w_{\rm eq})}\bigg)^2 \big(|w_{\rm dif}| + (1-\tau)|w_{\rm eq}|\big)^2 dw_{\rm dif} dw_{\rm eq} \end{aligned} \tag{207}$$

with $f_{\rm dif}$ and $f_{\rm eq}$ denoting the densities of $W_{\rm dif}$ and $W_{\rm eq}$. The function $Q(x)\big(\frac{\phi(x)}{Q(x)}\big)^2$ lies between 0 and $\frac{1}{2}$, and hence $T_1 \le \frac{1}{2}\mathbb{E}\big[\big(|W_{\rm dif}| + (1-\tau)|W_{\rm eq}|\big)^2\big]$, yielding

$$T_1 = O\big(\mathbb{E}[W_{\rm dif}]^2 + (1-\tau)^2\mathbb{E}[W_{\rm eq}]^2\big). \tag{208}$$

We will further simplify this expression below, but we first bound T_2, which requires more effort.

We split the integral over $\mathbb{R}^2$ in (207) according to whether $|w_{\rm dif}| \le \frac{1}{2}|w_{\rm eq}|$ or $|w_{\rm dif}| > \frac{1}{2}|w_{\rm eq}|$; the resulting expressions are denoted by $T_{1,1}$ and $T_{1,2}$ respectively. In each case, we use the following standard bounds on the Q-function (e.g., see [48]):

$$\frac{\phi(\tau w_{\rm eq})}{Q(\tau w_{\rm eq})} \le \begin{cases} 1 + \tau w_{\rm eq} & w_{\rm eq} \ge 0 \\ 1 & w_{\rm eq} < 0 \end{cases} \tag{209}$$

$$Q(w_{\rm dif} + w_{\rm eq}) \le \begin{cases} \frac{1}{2} e^{-\frac{(w_{\rm dif} + w_{\rm eq})^2}{2}} & w_{\rm dif} + w_{\rm eq} \ge 0 \\ 1 & w_{\rm dif} + w_{\rm eq} < 0. \end{cases} \tag{210}$$

To bound $T_{1,1}$, we note that the condition $|w_{\rm dif}| \le \frac{1}{2}|w_{\rm eq}|$ implies that $\mathrm{sign}(w_{\rm dif} + w_{\rm eq}) = \mathrm{sign}(w_{\rm eq})$, and hence only two of the four combinations of the cases in (209)–(210) can occur. When $w_{\rm eq} < 0$, we can use the second of each of these cases to upper bound the integrand in (207) by $f_{\rm dif}(w_{\rm dif}) f_{\rm eq}(w_{\rm eq})\big(|w_{\rm dif}| + (1-\tau)|w_{\rm eq}|\big)^2$. On the other hand, when $w_{\rm eq} \ge 0$ we can use the first of each of the cases to upper bound the integrand by

$$\begin{aligned} &f_{\rm dif}(w_{\rm dif}) f_{\rm eq}(w_{\rm eq}) \frac{1}{2} e^{-\frac{(w_{\rm dif} + w_{\rm eq})^2}{2}} \\ &\qquad \times \big(1 + \tau|w_{\rm eq}|\big)^2 \big(|w_{\rm dif}| + (1-\tau)|w_{\rm eq}|\big)^2. \end{aligned} \tag{211}$$

Again using the condition $|w_{\rm dif}| \le \frac{1}{2}|w_{\rm eq}|$, we find that $e^{-\frac{(w_{\rm dif} + w_{\rm eq})^2}{2}} \le e^{-\frac{1}{8}w_{\rm eq}^2}$. Since $\tau \le 1$ by its definition following (200), it follows that $e^{-\frac{(w_{\rm dif} + w_{\rm eq})^2}{2}}\big(1 + \tau w_{\rm eq}\big)^2$ is upper bounded by a universal constant, and we are again left only with $f_{\rm dif}(w_{\rm dif}) f_{\rm eq}(w_{\rm eq})\big(|w_{\rm dif}| + (1-\tau)|w_{\rm eq}|\big)^2$. Combining the two cases, we conclude that

$$T_{2,1} = O\big(\mathbb{E}[W_{\rm dif}^2] + (1-\tau)^2\mathbb{E}[W_{\rm eq}^2]\big). \tag{212}$$

To upper bound $T_{2,2}$, we upper bound the integrand in (207) by

$$f_{\rm dif}(w_{\rm dif})f_{\rm eq}(w_{\rm eq})\big(1+\tau|w_{\rm eq}|\big)^2\big(|w_{\rm dif}|+(1-\tau)|w_{\rm eq}|\big)^2 \tag{213}$$

$$\leq f_{\rm dif}(w_{\rm dif})f_{\rm eq}(w_{\rm eq})\big(1+2|w_{\rm dif}|\big)^2\big(3|w_{\rm dif}|\big)^2, \tag{214}$$

where (213) follows by taking the higher of the two cases in both (209) and (210), and (214) follows since $|w_{\rm dif}| > \frac{1}{2}|w_{\rm eq}|$ and $\tau \in [0,1]$. It follows that

$$T_{2,2} = O\big(\mathbb{E}[W_{\rm dif}^2] + \mathbb{E}[W_{\rm dif}^4]\big). \tag{215}$$

We now observe that the first two terms in (79) account for all of the terms in (208), (212) and (215) except for $(1-\tau)^2\mathbb{E}[W_{\rm eq}^2]$. Recalling that $\tau = \frac{1}{1+\sum_{i\in s_{\rm dif}} b_i^2}$, we see that $(1-\tau)^2 = \Theta(1)$ whenever $\sum_{i\in s_{\rm dif}} b_i^2 = \Omega(1)$, whereas a Taylor expansion yields $(1-\tau)^2 = \Theta\big(\big(\sum_{i\in s_{\rm dif}} b_i^2\big)^2\big)$ whenever $\sum_{i\in s_{\rm dif}} b_i^2 = o(1)$. Combining these cases, we obtain the third term in (79); recall that $\sigma^2 = 1$ throughout this proof.

Appendix D
Proofs of Auxiliary Results for Noiseless Group Testing

A. Proof of Proposition 6

As stated in [2, Eq. (36)], we have $I_\ell = \big(1-\frac{\nu}{k}\big)^{k-\ell}H_2\big(\big(1-\frac{\nu}{k}\big)^\ell\big)$, where $H_2(p)$ is the binary entropy function. For $k\to\infty$ and $\frac{\ell}{k}\to\alpha$, we immediately obtain (97) using the limits $\big(1-\frac{\nu}{k}\big)^{k-\ell} \to e^{-(1-\alpha)\nu}$ and $\big(1-\frac{\nu}{k}\big)^\ell \to e^{-\alpha\nu}$, along with the continuity of the binary entropy function. In the case that $\frac{\ell}{k}\to 0$, the analogous limits are $\big(1-\frac{\nu}{k}\big)^{k-\ell}\to e^{-\nu}$ and $\big(1-\frac{\nu}{k}\big)^\ell = 1-\frac{\nu\ell}{k}(1+o(1))$, and we obtain (96) using the fact that $H_2(1-\epsilon) = (-\epsilon\log\epsilon)(1+o(1))$ as $\epsilon\to 0$. Note also that $\log\frac{k}{\nu\ell} = \big(\log\frac{k}{\ell}\big)(1+o(1))$ since $\frac{k}{\ell}\to\infty$.

B. Proof of Proposition 7

We begin by evaluating the information density in (21); for brevity, we write $\imath_\ell := \imath(X_{s_{\rm dif}};Y|X_{s_{\rm eq}},b_s)$ and $\imath_\ell^n := \imath(\mathbf{X}_{s_{\rm dif}};\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},b_s)$. Recalling that $P_X \sim \mathrm{Bernoulli}\big(\frac{\nu}{k}\big)$, $\ell = o(k)$, and we are considering the noiseless case, we obtain the following:

1. We have $X_{s_{\rm eq}} \neq \mathbf{0}$ with probability $1-\big(1-\frac{\nu}{k}\big)^{k-\ell} = (1-e^{-\nu})(1+o(1))$, and in this case we have $\imath_\ell = 0$.
2. Given $X_{s_{\rm eq}} = \mathbf{0}$, we have $X_{s_{\rm dif}} \neq \mathbf{0}$ with probability $1-\big(1-\frac{\nu}{k}\big)^\ell = \frac{\nu\ell}{k}(1+o(1))$, and in this case we have $\imath_\ell = \log\frac{1}{1-(1-\frac{\nu}{k})^\ell} = \big(\log\frac{k}{\ell}\big)(1+o(1))$.
3. Given $X_{s_{\rm eq}} = \mathbf{0}$, we have $X_{s_{\rm dif}} = \mathbf{0}$ with probability $\big(1-\frac{\nu}{k}\big)^\ell = 1+o(1)$, and in this case we have $\imath_\ell = \log\frac{1}{(1-\frac{\nu}{k})^\ell} = \frac{\nu\ell}{k}(1+o(1))$.

The asymptotic identities given here follow from the assumption $\ell = o(k)$, along with standard Taylor expansions.

Let N_0 (respectively, N_1) be the random number of measurements such that $X_{s_{\rm eq}} = \mathbf{0}$ and $X_{s_{\rm dif}} = \mathbf{0}$ (respectively, $X_{s_{\rm eq}} = \mathbf{0}$ and $X_{s_{\rm dif}} \neq \mathbf{0}$). For any $\epsilon_1 \in (0,1)$, the above observations imply the following with probability one when p is sufficiently large:

$$\imath_\ell^n \geq N_1\Big(\log\frac{k}{\ell}\Big)(1-\epsilon_1) + N_0\nu\frac{\ell}{k}(1-\epsilon_1) \tag{216}$$

$$\geq N_1\Big(\log\frac{k}{\ell}\Big)(1-\epsilon_1). \tag{217}$$

We also have from (96) that $I_\ell \leq \big(e^{-\nu}\nu\frac{\ell}{k}\log\frac{k}{\ell}\big)(1+\epsilon_1)$ for sufficiently large p. Combining these, we conclude that

$$N_1 > n\frac{1+\epsilon_1}{1-\epsilon_1}e^{-\nu}\nu\frac{\ell}{k}(1-\delta_2) \implies \imath_\ell^n > nI_\ell(1-\delta_2). \tag{218}$$

By considering the contrapositive statement, we have for any $\epsilon_2 > 0$ and sufficiently large p that

$$\mathbb{P}\Big[\imath^n(\mathbf{X}_{s_{\rm dif}};\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},b_s) \leq nI_\ell(1-\delta_2)\Big]$$

$$\leq \mathbb{P}\Big[N_1 \leq ne^{-\nu}\nu\frac{\ell}{k}(1-\delta_2)(1+\epsilon_2)\Big]. \tag{219}$$

By the observations at the start of this subsection, we have $N_1 \sim \mathrm{Binomial}(n,q)$ with $q = e^{-\nu}\nu\frac{\ell}{k}(1+o(1))$. We can thus further upper bound the right-hand of (219) by

$$\mathbb{P}\big[N_1 \leq nq(1-\delta_2(1-\epsilon_3))\big] \tag{220}$$

for any $\epsilon_3 \in (0,1)$ and sufficiently large p; here we have used the fact that $(1-\delta_2)(1+o(1)) = (1-\delta_2(1+o(1))$, since δ_2 is fixed. It follows from a standard Chernoff-based tail bound for Binomial random variables (e.g., see [49, Sec. 4.1]) that

$$\mathbb{P}\Big[\imath^n(\mathbf{X}_{s_{\rm dif}};\mathbf{Y}|\mathbf{X}_{s_{\rm eq}},b_s) \leq nI_\ell(1-\delta_2)\Big]$$

$$\leq e^{-nq\big((1-\delta_2(1-\epsilon_3))\log(1-\delta_2(1-\epsilon_3))+\delta_2(1-\epsilon_3)\big)}. \tag{221}$$

The proof is concluded by substituting $q = e^{-\nu}\nu\frac{\ell}{k}(1+o(1))$ and noting that ϵ_3 may be arbitrarily small.

C. Proof of Proposition 8

For the first part, we write $\sum_{\ell=1}^{\lfloor\frac{k}{\log k}\rfloor}\binom{k}{\ell}\psi_\ell(n,\delta_2^{(1)}) =: T_1 + T_2$, where T_1 sums the terms from 1 to $\lfloor\log k\rfloor$, and T_2 sums the terms from $\lfloor\log k\rfloor+1$ to $\lfloor\frac{k}{\log k}\rfloor$. For each of these, we upper bound the summation by the number of terms times the maximum term.

For T_1, there are at most $\log k$ terms, and we apply (99), with $\delta_{2,\ell} = \delta_2^{(1)}$. The term $(1-\delta_2^{(1)})\log(1-\delta_2^{(1)})+\delta_2^{(1)}$ can be made arbitrarily close to one by choosing $\delta_2^{(1)}$ to be sufficiently close to one. Writing $\log\binom{k}{\ell} = \big(\ell\log\frac{k}{\ell}\big)(1+o(1))$ and performing some simple rearrangements, we obtain the following condition for $T_1\to 0$:

$$n \geq \max_\ell \frac{k\log\frac{k}{\ell}+\frac{k}{\ell}\log\log k}{e^{-\nu}\nu}(1+\eta_1), \tag{222}$$

where η_1 may be arbitrarily small. Note that $\log\log k$ arises as the logarithm of the number of terms in the summation.

We obtain (101) by noting that this bound is minimized at $\ell = 1$ and writing $k \log k = \big(\frac{\theta}{1-\theta} k \log \frac{p}{k}\big)(1+o(1))$, which follows from $k = \Theta(p^{\theta})$.

For T_2, a similar argument yields (222) with $\frac{1}{\ell}\log k$ in place of $\frac{1}{\ell}\log\log k$; this follows by upper bounding the number of terms in the summation by k. Since $\ell \ge \log k$, we have $\frac{1}{\ell}\log k = O(1)$, and we conclude that $T_2 \to 0$ provided that (101) holds.

Finally, for the second part of the proposition, we substitute (100). By an analogous argument to that leading to (222), along with the scaling laws of I_ℓ in (96)–(97), it is readily verified that it suffices that $n = \Omega\big(\max_\ell \frac{\ell \log \frac{k}{\ell}}{1+(\frac{\ell}{k}\log\frac{k}{\ell})^2}\big)$ with a sufficiently large implied constant. Using the fact that $\ell > \frac{k}{\log k}$ for this part, this reduces to $\Omega\big(\frac{k \log k}{\log\log k}\big)$. Thus, any $\Omega(k \log k)$ scaling suffices, and the proof is concluded by noting that $\log k = \Theta\big(\log\frac{p}{k}\big)$.

Appendix E
Noisy Group Testing

Here we provide the relevant details for noisy group testing, leading to Corollary 7. We focus our attention on the parts that differ from the noiseless case. Throughout the appendix, we use the notation $q_1 \star q_2 := q_1 q_2 + (1-q_1)(1-q_2)$. We work with an arbitrary Bernoulli distribution $P_X \sim \mathrm{Bernoulli}\big(\frac{\nu}{k}\big)$ to begin, and later substitute $\nu = \log 2$.

Before proceeding, we analyze the values taken by the information density $\imath_\ell := \imath(X_{s_{\rm dif}}; Y | X_{s_{\rm eq}}, b_s)$ (with $\ell := |s_{\rm dif}|$) given in (21), under the model in (108):

1. We have $X_{s_{\rm eq}} \ne \mathbf{0}$ with probability $1 - \big(1-\frac{\nu}{k}\big)^{k-\ell}$, and in this case we have $\imath_\ell = 0$.
2. Given $X_{s_{\rm eq}} = \mathbf{0}$, we have the following, where we define $\xi := \big(1-\frac{\nu}{k}\big)^{\ell}$:
   - $X_{s_{\rm dif}} = \mathbf{0} \cap Y = 0$ with probability $(1-\rho)\xi$, yielding $\imath_\ell = \log\frac{1-\rho}{(1-\rho)\xi+\rho(1-\xi)}$;
   - $X_{s_{\rm dif}} = \mathbf{0} \cap Y = 1$ with probability $\rho\xi$, yielding $\imath_\ell = \log\frac{\rho}{\rho\xi+(1-\rho)(1-\xi)}$;
   - $X_{s_{\rm dif}} \ne \mathbf{0} \cap Y = 0$ with probability $\rho(1-\xi)$, yielding $\imath_\ell = \log\frac{\rho}{(1-\rho)\xi+\rho(1-\xi)}$;
   - $X_{s_{\rm dif}} \ne \mathbf{0} \cap Y = 1$ with probability $(1-\rho)(1-\xi)$, yielding $\imath_\ell = \log\frac{1-\rho}{\rho\xi+(1-\rho)(1-\xi)}$.

In the case that $\ell = o(k)$, we can write $\xi = 1 - \frac{\nu\ell}{k}(1+o(1))$, yielding the following simplifications:

1. The preceding four probabilities behave as $(1-\rho)\big(1-\frac{\nu\ell}{k}(1+o(1))\big)$, $\rho\big(1-\frac{\nu\ell}{k}(1+o(1))\big)$, $\rho\frac{\nu\ell}{k}(1+o(1))$, and $(1-\rho)\frac{\nu\ell}{k}(1+o(1))$.
2. The corresponding information densities behave as $\frac{1-2\rho}{1-\rho}\frac{\nu\ell}{k}(1+o(1))$, $-\frac{1-2\rho}{\rho}\frac{\nu\ell}{k}(1+o(1))$, $-\log\frac{1-\rho}{\rho}(1+o(1))$ and $\log\frac{1-\rho}{\rho}(1+o(1))$. For example, the first of these follows by writing $\log\frac{1-\rho}{(1-\rho)(1-\frac{\nu\ell}{k})+\rho\frac{\nu\ell}{k}} = \log\frac{1-\rho}{1-\rho-(1-2\rho)\frac{\nu\ell}{k}}$, dividing the numerator and denominator by $1-\rho$, and Taylor expanding the logarithm.

A. Analogs of Propositions 6–8

The analog of Proposition 6 is as follows.

Proposition 11. *Under the noisy group testing setup in Section IV-F, consider arbitrary sequences of sparsity levels $k \to \infty$ and $\ell \in \{1,\dotsc,k\}$ (both indexed by p). If $\frac{\ell}{k} = o(1)$, then*

$$I_\ell = \left(e^{-\nu}\nu\frac{\ell}{k}(1-2\rho)\log\frac{1-\rho}{\rho}\right)(1+o(1)). \tag{223}$$

Moreover, if $\frac{\ell}{k} \to \alpha \in (0,1]$, then

$$I_\ell = e^{-(1-\alpha)\nu}\big(H_2\big(e^{-\alpha\nu}\star\rho\big) - H_2(\rho)\big)(1+o(1)). \tag{224}$$

Proof. We obtain (223) by recalling that the mutual information is the average of the information density, and applying the above-given asymptotic expansions, along with $1-\big(1-\frac{\nu}{k}\big)^{k-\ell} \to e^{-\nu}$.

To prove (224), we write $I(X_{s_{\rm dif}}; Y | X_{s_{\rm eq}}) = H(Y|X_{s_{\rm eq}}) - H(Y|X_{s_{\rm eq}}, X_{s_{\rm dif}})$. The system model (108) immediately gives $H(Y|X_{s_{\rm eq}}, X_{s_{\rm dif}}) = H_2(\rho)$. Moreover, a direct calculation reveals that $H(Y|X_{s_{\rm eq}} = x_{s_{\rm eq}})$ equals $H_2(\rho)$ if $x_{s_{\rm eq}}$ has an entry equal to one, and $H_2\big(\xi \star \rho\big)$ otherwise, where we again write $\xi := \big(1-\frac{\nu}{k}\big)^{\ell}$. The proof is concluded by noting that $\xi \to e^{-\alpha\nu}$ when $\frac{\ell}{k} \to \alpha$, and by similarly noting that $\mathbb{P}[X_{s_{\rm eq}} = \mathbf{0}] = \big(1-\frac{\nu}{k}\big)^{k-\ell} \to e^{-(1-\alpha)\nu}$. □

As in the noiseless case, we use Proposition 10 to characterize ψ_ℓ for $\ell > \lfloor\frac{k}{\log k}\rfloor$, and ψ'_ℓ for $\ell = k$. For $\ell \le \lfloor\frac{k}{\log k}\rfloor$, we instead use the following.

Proposition 12. *Under the noisy group testing setup in Section IV-F, consider sequences $k \to \infty$ and ℓ, indexed by p, such that $\frac{\ell}{k} \to 0$. For any $\epsilon > 0$ and $\delta_2 > 0$ not depending on p, the following holds for sufficiently large p:*

$$\mathbb{P}\Big[\imath^n(\mathbf{X}_{s_{\rm dif}}; \mathbf{Y}|\mathbf{X}_{s_{\rm eq}}, b_s) \le nI_\ell(1-\delta_2)\Big] \le \exp\left(-n\frac{\ell}{k}e^{-\nu}\nu\Big(\frac{\delta_2^2(1-2\rho)^2}{2(1+\frac{1}{3}\delta_2(1-2\rho))}\Big)(1-\epsilon)\right). \tag{225}$$

for all $(s_{\rm dif}, s_{\rm eq})$ with $|s_{\rm dif}| = \ell$.

Proof. We make use of the asymptotic identities for $\imath_\ell$ at the start of this appendix. We first note that by simple averaging analogous to that used to obtain (223), we have $v := \mathbb{E}[\imath_\ell^2] = e^{-\nu}\nu\frac{\ell}{k}\big(\log^2\frac{1-\rho}{\rho}\big)(1+o(1))$. Moreover, we have $\imath_\ell \le \big(\log\frac{1-\rho}{\rho}\big)(1+o(1))$ with probability one. Using the form of Bernstein's inequality based on Bennet's inequality [38, Sec. 2.7], we have $\mathbb{P}[\imath^n \le n(I_\ell - \delta)] \exp\big(-n\frac{\delta^2}{2(v+\frac{1}{3}\delta M)}\big)$, where M is any almost-sure upper bound on $\imath_\ell$. Setting $\delta = \delta_2 I_\ell$, substituting (223) and the preceding expressions for v and M, and canceling the common terms in the numerator and denominator, we obtain (225). □

Letting ψ_ℓ equal the right-hand side of (225) for $\ell \le \lfloor\frac{k}{\log k}\rfloor$, while being the same as in (100) for $\ell > \lfloor\frac{k}{\log k}\rfloor$, we obtain the following.

Proposition 13. *Let $k = \Theta(p^{\theta})$ for some $\theta \in (0,1)$.*

(i) For any $\eta > 0$ and $\delta_2 \in (0,1)$, there exists a choice of $\epsilon > 0$ in (99) such that $\sum_{\ell=1}^{\lfloor \frac{k}{\log k} \rfloor} \binom{k}{\ell} \psi_\ell(n, \delta_2) \to 0$ provided

$$n \geq \frac{2(1 + \frac{1}{3}\delta_2(1-2\rho))\frac{\theta}{1-\theta}}{e^{-\nu}\nu\delta_2^2(1-2\rho)^2}\Big(k \log \frac{p}{k}\Big)(1+\eta). \tag{226}$$

(ii) For any $\delta_2 \in (0,1)$, we have $\sum_{\lfloor \frac{k}{\log k} \rfloor + 1}^{k} \binom{k}{\ell} \psi_\ell(n, \delta_2) \to 0$ provided that $n = \Omega\big(k \log \frac{p}{k}\big)$.

Proof. The proof is nearly identical to that of Proposition 8, except that (225) is used in place of (98), and δ_2 is kept arbitrary in the first part. □

Note that the choices of δ_2 in the two cases above need not coincide; see Remark 2.

B. Remaining Details in the Proof of Corollary 7

Recall that we have set $\nu = \log 2$. This yields $e^{-\nu}\nu = \frac{\log 2}{2}$, and thus the first term in (110) follows from (226).

Next, we consider the condition in (37) with $\ell = |s_{\mathrm{dif}}| \leq \lfloor \frac{k}{\log k} \rfloor$. Setting $\gamma = 0$, letting $\delta_1 \to 0$ sufficiently slowly, applying Stirling's approximation, and substituting (223), we obtain the condition

$$n \geq \max_{\ell} \frac{k \log \frac{p}{\ell} + 2k \log k + 2\frac{k}{\ell}\log k}{e^{-\nu}\nu(1-2\rho)\log\frac{1-\rho}{\rho}(1-\delta_2)}(1+o(1)). \tag{227}$$

This is maximized for $\ell = 1$, thus yielding the second term in (110) upon writing $k \log k = \frac{\theta}{1-\theta}\big(k \log \frac{p}{k}\big)(1+o(1))$ and $k \log p = \frac{1}{1-\theta}\big(k \log \frac{p}{k}\big)(1+o(1))$ (since $k = \Theta(p^\theta)$).

Finally, we consider (37) with $\ell > \lfloor \frac{k}{\log k} \rfloor$. In this case, the numerator is dominated by the first term, and for the case that $\frac{\ell}{k} \to \alpha \in (0,1]$, we obtain the condition

$$n \geq \frac{\alpha k \log \frac{p}{k}}{e^{-(1-\alpha)\nu}\big(H_2(e^{-\alpha\nu} \star \rho) - H_2(\rho)\big)(1-\delta_2)}(1+o(1)), \tag{228}$$

where we have used (224). For the case that $\frac{\ell}{k} \to 0$ with $\ell > \lfloor \frac{k}{\log k} \rfloor$, we obtain a condition of the form (227) where only the first term of the numerator is kept. Such a condition is clearly dominated by (227).

Using the result in [11, Thm. 3a] in the limiting case that the number of defective items grows large, we have for the worst-case choice of $\alpha \in [0,1]$ and an optimized choice of $\nu > 0$ that the minimax threshold resulting from (228) is obtained with $\alpha = 1$ and $\nu = \log 2$. Substituting these values yields the second term in (109).

C. An Auxiliary Result for Comparing the Terms

The following result allows us to compare the terms appearing in the achievability part of Corollary 7.

Proposition 14. *For all $\rho \in (0, 0.5)$, we have*

$$(1-2\rho)\log\frac{1-\rho}{\rho} \geq 4\big(\log 2 - H_2(\rho)\big). \tag{229}$$

Proof. By some simple manipulations, the left-hand side can be written as $\log\frac{1}{\rho(1-\rho)} - 2H_2(\rho)$, and we may thus equivalently prove that $\log\frac{1}{\rho(1-\rho)} + 2H_2(\rho) \geq 4\log 2$. This, in turn, can be verified by showing that the minimum of the function $\log\frac{1}{\rho(1-\rho)} + 2H_2(\rho)$ occurs at $\rho = 0.5$, i.e., the point about which it is symmetric. □

Acknowledgment

We gratefully acknowledge Ya-Ping Hsieh for helpful comments and suggestions.

References

[1] M. B. Malyutov, "Search for sparse active inputs: A review," in *Inf. Theory, Comb. and Search Theory*, 2013, pp. 609–647.

[2] 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.

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

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

[5] 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.

[6] ——, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[7] W. Wang, M. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *IEEE Int. Symp. Inf. Theory*, 2010.

[8] 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.

[9] P. Boufounos and R. Baraniuk, "1-bit compressive sensing," in *Conf. on Inf. Sci. and Sys.*, March 2008.

[10] R. Dorfman, "The detection of defective members of large populations," *Ann. Math. Stats.*, vol. 14, no. 4, pp. 436–440, 1943.

[11] M. Malyutov, "The separating property of random matrices," *Math. notes Acad. Sci. USSR*, vol. 23, no. 1, pp. 84–91, 1978.

[12] C. L. Chan, P. H. Che, S. Jaggi, and V. Saligrama, "Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms," in *Allerton Conf. Comm., Ctrl., Comp.*, Sep. 2011, pp. 1832–1839.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] S. Aeron, V. Saligrama, and M. Zhao, "Information theoretic bounds for compressed sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 5111–5130, Oct. 2010.

[19] G. Tang and A. Nehorai, "Performance analysis for sparse support recovery," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1383–1399, March 2010.

[20] M. Akcakaya and V. Tarokh, "Shannon-theoretic limits on noisy compressive sampling," *IEEE Trans. Inf. Theory*, vol. 56, no. 1, pp. 492–504, Jan. 2010.

[21] A. Tulino, G. Caire, S. Verdú, and S. Shamai, "Support recovery with sparsely sampled free random matrices," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4243–4271, July 2013.

[22] J. Scarlett, J. Evans, and S. Dey, "Compressed sensing with prior information: Information-theoretic limits and practical decoders," *IEEE Trans. Sig. Proc.*, vol. 61, no. 2, pp. 427–439, Jan. 2013.

[23] Y. Wu and S. Verdú, "Rényi information dimension: Fundamental limits of almost lossless analog compression," *IEEE Trans. Inf. Theory*, no. 8, pp. 3721–3748, Aug. 2010.

[24] ——, "Optimal phase transitions in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6241–6263, Oct. 2012.

[25] 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.

[26] J. D. Lee, Y. Sun, J. E. Taylor *et al.*, "On model selection consistency of regularized m-estimators," *Elec. J. Stats.*, vol. 9, no. 1, pp. 608–642, 2015.

[27] Y.-H. Li, J. Scarlett, P. Ravikumar, and V. Cevher, "Sparsistency of ℓ_1-regularized m-estimators," in *Int. Conf. Art. Intel. Stats. (AISTATS)*, 2015, pp. 644–652.

[28] J. Tan, D. Carmon, and D. Baron, "Signal estimation with additive error metrics in compressed sensing," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 150–158, 2014.

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

[30] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," *Information and Inference*, vol. 3, no. 3, pp. 224–294, 2014.

[31] L. Baldassini, O. Johnson, and M. Aldridge, "The capacity of adaptive group testing," in *IEEE Int. Symp. Inf. Theory*, July 2013, pp. 2676–2680.

[32] M. Aldridge, L. Baldassini, and K. Gunderson, "Almost separable matrices," *J. Comb. Opt.*, pp. 1–22, 2015.

[33] J. Scarlett and V. Cevher, "Phase transitions in group testing," in *Proc. ACM-SIAM Symp. Disc. Alg. (SODA)*, 2016.

[34] A. Feinstein, "A new basic theorem of information theory," *IRE Prof. Group. on Inf. Theory*, vol. 4, no. 4, pp. 2–22, Sept. 1954.

[35] C. E. Shannon, "Certain results in coding theory for noisy channels," *Information and Control*, vol. 1, no. 1, pp. 6–25, 1957.

[36] S. Verdú and T. S. Han, "A general formula for channel capacity," *IEEE Trans. Inf. Theory*, vol. 40, no. 4, pp. 1147–1157, July 1994.

[37] 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.

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

[39] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

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

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

[42] T. Laarhoven, "Asymptotics of fingerprinting and group testing: Tight bounds from channel capacities," *IEEE Trans. Inf. Forens. Sec.*, vol. 10, no. 9, pp. 1967–1980, 2015.

[43] M. Aldridge, L. Baldassini, and O. Johnson, "Group testing algorithms: Bounds and simulations," *IEEE Trans. Inf. Theory*, vol. 60, no. 6, pp. 3671–3687, June 2014.

[44] J. Scarlett and V. Cevher, "Converse bounds for noisy group testing with arbitrary measurement matrices," in *IEEE Int. Symp. Inf. Theory*, Barcelona, 2016.

[45] ——, "Limits on sparse support recovery via linear sketching with random expander matrices," in *Int. Conf. Art. Intel. Stats. (AISTATS)*, Cadiz, Spain, 2016, pp. 149–158.

[46] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, April 2010.

[47] V. Tan and O. Kosut, "On the dispersions of three network information theory problems," *IEEE Trans. Inf. Theory*, vol. 60, no. 2, pp. 881–903, Feb. 2014.

[48] P. Fan, "New inequalities of Mill's ratio and its application to the inverse Q-function approximation," http://arxiv.org/abs/1212.4899.

[49] R. Motwani and P. Raghavan, *Randomized Algorithms*. Chapman & Hall/CRC, 2010.

Jonathan Scarlett (S'14 – M'15) received the B.Eng. degree in electrical engineering and the B.Sci. degree in computer science from the University of Melbourne, Australia. In 2011, he was a research assistant at the Department of Electrical & Electronic Engineering, University of Melbourne. From October 2011 to August 2014, he was a Ph.D. student in the Signal Processing and Communications Group at the University of Cambridge, United Kingdom. He is now a post-doctoral researcher with the Laboratory for Information and Inference Systems at the École Polytechnique Fédérale de Lausanne, Switzerland. His research interests are in the areas of information theory, signal processing, machine learning, and high-dimensional statistics. He received the Cambridge Australia Poynton International Scholarship, and the EPFL Fellows postdoctoral fellowship co-funded by Marie Skłodowska-Curie.

Volkan Cevher (SM'10) received the B.Sc. (valedictorian) in electrical engineering from Bilkent University in Ankara, Turkey, in 1999 and the Ph.D. in electrical and computer engineering from the Georgia Institute of Technology in Atlanta, GA in 2005. He was a Research Scientist with the University of Maryland, College Park from 2006-2007 and also with Rice University in Houston, TX, from 2008-2009. Currently, he is an Associate Professor at the Swiss Federal Institute of Technology Lausanne and a Faculty Fellow in the Electrical and Computer Engineering Department at Rice University. His research interests include signal processing theory, machine learning, convex optimization, and information theory. Dr. Cevher was the recipient of a Best Paper Award at SPARS in 2009, a Best Paper Award at CAMSAP in 2015, and an ERC StG in 2011.